Las ciencias de la administración y el análisis multivariante

Proyectos de investigación, análisis y discusión de resultados

Tomo I

Las técnicas dependientes

Juan Mejía Trejo

Las ciencias de la administración y el análisis multivariante

Proyectos de investigación, análisis y discusión de los resultados

Tomo I

Las técnicas dependientes

Las ciencias de la administración y el análisis multivariante

Proyectos de investigación, análisis y discusión de los resultados

Tomo I

Las técnicas dependientes

Juan Mejía Trejo

Universidad de Guadalajara
Centro Universitario de Ciencias Económico Administrativas

Este libro fue sometido a un proceso de dictamen por pares de acuerdo con las normas establecidas por el Comité Editorial del Centro Universitario de Ciencias Económico Administrativas de la Universidad de Guadalajara.

Primera edición 2017

Centro Universitario de Ciencias Económico Administrativas
Periférico Norte No. 799, Núcleo Universitario Los Belenes
45100, Zapopan, Jalisco, México

ISBN Tomo I: 978-607-742-773-5
ISBN Obra completa: 978-607-742-772-8
Impreso y hecho en México
Printed and made in Mexico

Contenido

Introducción

Al tener mayores capacidades y disponibilidad de los recursos de cómputo, hoy en día, hace que el análisis multivariante se presente en diversas aplicaciones de software por lo que se incrementan las posibilidades de ser usado en diversas disciplinas como las Ciencias de la Administración, siendo el Statistical Package for the Social Sciences (SPSS, de IBM), el Analytics, Business Intelligence and Data Management (SAS, de SAS Institute y/o de World Programming), Statistica (de STATISTICA), el lenguaje R (software libre) sólo por mencionar algunos como de los más utilizados en los campos académico y profesional a nivel mundial. Así que, no es de extrañar que las Ciencias de la Administración apoyen el desarrollo académico presentándose en diversos posgrados, así como en el mundo laboral que corresponde a las Ciencias Sociales y por lo tanto, se observa de manera creciente un repunte en la presentación de reportes, artículos, capítulos de libro o libros que discutan diversos aspectos teórico empíricos y su interpretación basados en dichas aplicaciones de software. En nuestro caso, adoptamos SPSS 20 de IBM, para el desarrollo de los temas de este libro.

Basados en lo anterior, presentamos la obra: *Las Ciencias de la Administración y el Análisis Multivariante bajo el enfoque de las Técnicas Dependientes. Proyectos de Investigación, Análisis y Discusión de Resultados.Tomo I*, con un propósito triple:

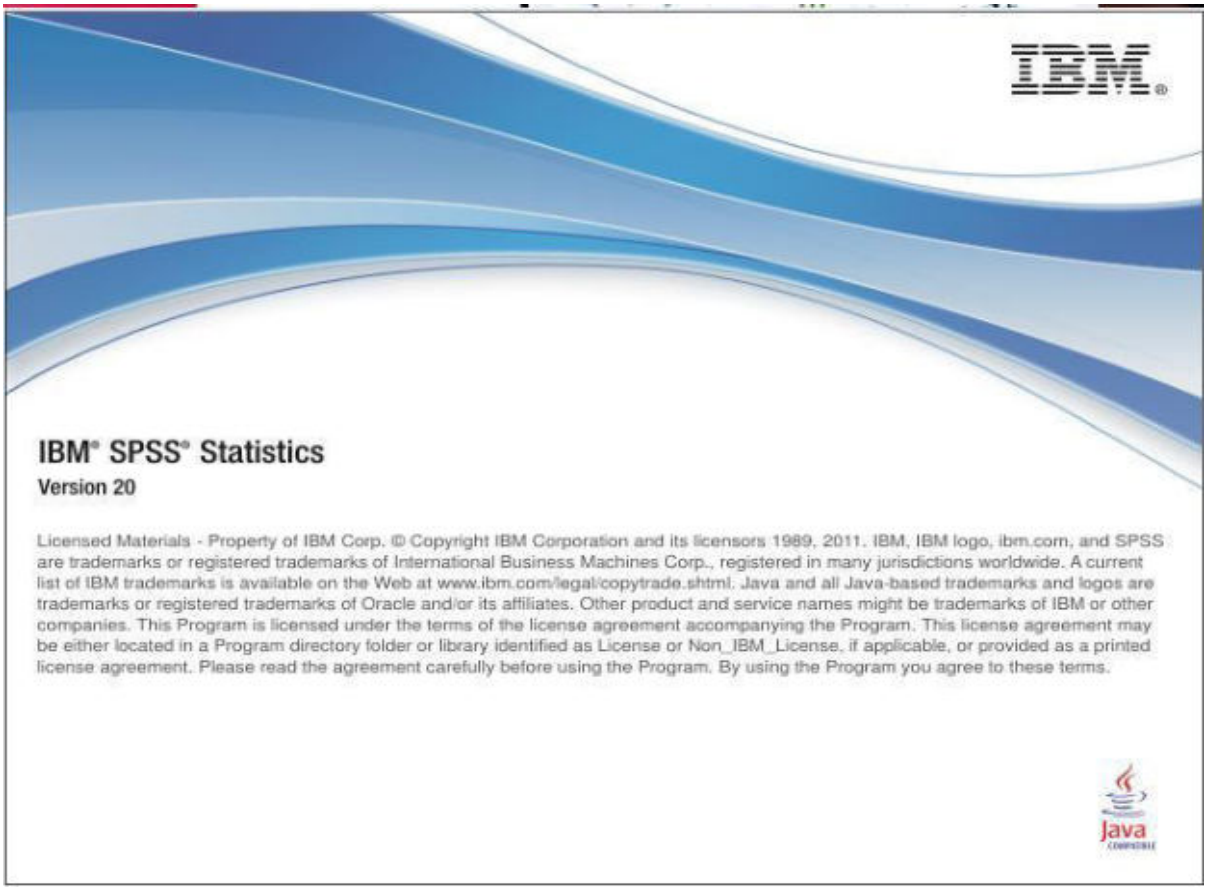
1.-Presentar un documento que sirva a propios y extraños al tema, que tengan la necesidad de conocer tanto los conceptos tratados en este tomo, como el de manipular los diversos comandos que ofrece SPSS 20 de IBM al respecto de los casos problema, presentados como ejemplo.

2.-Para una mayor comprensión del tratamiento de los casos, se expone la secuencia propuesta por Hair et al. (1999) de los 6 pasos: objetivos, diseño, supuestos, ejecución, interpretación y validación, como el eje de presentación y resolución de dichos caso.

3.-Como Coordinador del Doctorado en Ciencias de la Administración del Centro Universitario de Ciencias Económico Administrativas (CUCEA), de la Universidad de Guadalajara (UdG), presentar el libro base para la asignatura de Métodos Cuantitativos I y II.

Es deseo del autor, contribuir en el lector en la adquisición de conocimiento que se aplique en el mundo práctico y que ayude a su interpretación teórica. Si no fuere el caso, se espera que al menos sirva como otro peldaño útil a escalar en el logro de su formación académica y/o profesional.

Capítulo 1. Software SPSS



Fuente: SPSS 20 IBM

1.1 SPSS. ¿Qué es?

SPSS es un programa estadístico informático muy usado en las ciencias exactas, sociales y aplicadas, además de las empresas de investigación de mercado. Originalmente SPSS fue creado como el acrónimo de **Statistical Package for the Social Sciences** aunque también se ha referido como **Statistical Product and Service Solutions**. Sin embargo, en la actualidad la parte SPSS del nombre completo del software IBM SPSS se constituye como una marca. Es uno de los programas estadísticos más conocidos teniendo en cuenta su capacidad para trabajar con grandes bases de datos y un sencillo interface para la mayoría de los análisis. En la versión 12 de SPSS se podían realizar análisis con **2 millones de registros y 250.000 variables**. El programa consiste en un módulo base y módulos anexos que se han ido actualizando constantemente con nuevos procedimientos estadísticos. Cada uno de estos módulos se compra por separado. Actualmente, compite no sólo con software de licencia como lo son: **SAS, MATLAB, Statistica, Stata,** sino también con software de código abierto y libre, de los cuales el más destacado es el **Lenguaje R**. Recientemente ha sido desarrollado un paquete libre llamado **PSPP,** con una interfaz llamada **PSPPire** que ha sido compilada para diversos sistemas operativos como **Linux**, además de versiones

para **Windows** y **OSX.** Este último paquete pretende ser un clon de código abierto que emule todas las posibilidades del SPSS.

1.2 SPSS. Historia

Fue creado en 1968 por Norman H. Nie, C. Hadlai (Tex) Hull y Dale H. Bent. Entre 1969 y 1975 la Universidad de Chicago por medio de su *National Opinion Research Center* estuvo a cargo del desarrollo, distribución y venta del programa. A partir de 1975 corresponde a SPSS Inc. Para saber más, consulte: (IBM 2011a, IBM 2011b, IBM, 2011c. Originalmente el programa fue creado para grandes computadores. En 1970 se publica el primer manual de usuario del SPSS por Nie y Hall. Este manual populariza el programa entre las instituciones de educación superior en EUA. En 1984 sale la primera versión para computadores personales.

Desde la versión 14, pero más específicamente desde la versión 15 se ha implantado la posibilidad de hacer uso de las librerías de objetos del SPSS desde diversos lenguajes de programación. Aunque principalmente se ha implementado para Python, también existe la posibilidad de trabajar desde Visual Basic, C++ y otros lenguajes.

El 28 de junio de 2009 se anuncia que IBM, meses después de ver frustrado su intento de compra de Sun Microsystems, adquiere SPSS, por 1.200 millones de dólares.

1.3 SPSS. Versiones

SPSS, tiene un historial muy largo de versiones, las cuales se podrán apreciar en la **Figura 1.1.**

Figura 1.1SPSS Versiones

- SPSS 1 - 1968
- SPSSx release 2 - 1983 (para grandes servidores tipo UNIX)
- SPSS 5.0 - diciembre 1993
- SPSS 6.1 - febrero 1995
- SPSS 7.5 - enero 1997
- SPSS 8.0 - 1998
- SPSS 9.0 - marzo 1999
- SPSS 10.0.5 - diciembre 1999
- SPSS 10.0.7 - julio 2000
- SPSS 10.1.4 - enero 2002
- SPSS 11.0.1 - abril 2002
- SPSS 11.5.1 - abril 2003
- SPSS 12.0.1 - julio 2004
- SPSS 13.0.1 - marzo 2005 (Permite por primera vez trabajar con múltiples bases de datos al mismo tiempo.)
- SPSS 14.0.1 - enero 2006
- SPSS 15.0.1 - noviembre 2006
- SPSS 16.0.1 - noviembre 2007 SPSS 16.0.2 - abril 2008
- SPSS Statistics 17.0.1 SPSS Statistics 17.0.2 - marzo 2009
- PASW Statistics 17.0.3 - septiembre 2009 (IBM adquiere los derechos y cambia su denominación de SPSS por PASW 18)
- PASW Statistics 18.0 - agosto 2009

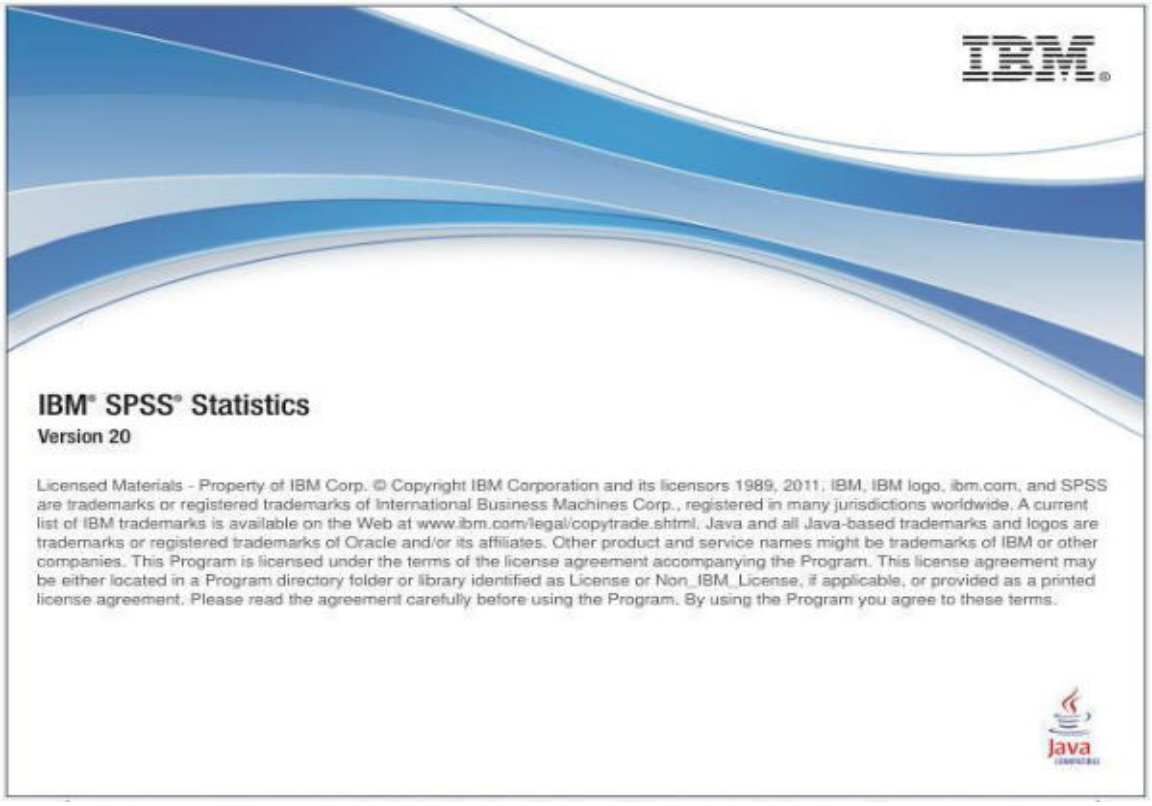
- PASW Statistics 18.0.1 - diciembre 2009
- PASW Statistics 18.0.2 - abril 2010
- PASW Statistics 18.0.3 - septiembre 2010
- IBM SPSS Statistics 19.0 - agosto 2010 (Pasa a denominarse IBM SPSS)
- IBM SPSS Statistics 19.0.1 - diciembre 2010
- IBM SPSS Statistics 20.0 - agosto 2011
- IBM SPSS Statistics 20.0.1 - marzo 2012
- IBM SPSS Statistics 21.0 - agosto 2012
- IBM SPSS Statistics 22.0 - agosto 2013
- IBM SPSS Statistics 23.0 - agosto 2014
- IBM SPSS Statistics 24.0 - junio 2016

Fuente: recopilación propia

1.4 SPSS. Presentación

SPSS es un conjunto de potentes herramientas de tratamiento de datos y análisis estadístico; funciona mediante **menús desplegables** y **cuadros de diálogo.** La pantalla de versión se presenta como la **Figura 1.2**

Figura 1.2. SPSS Pantalla de Presentación con Versión

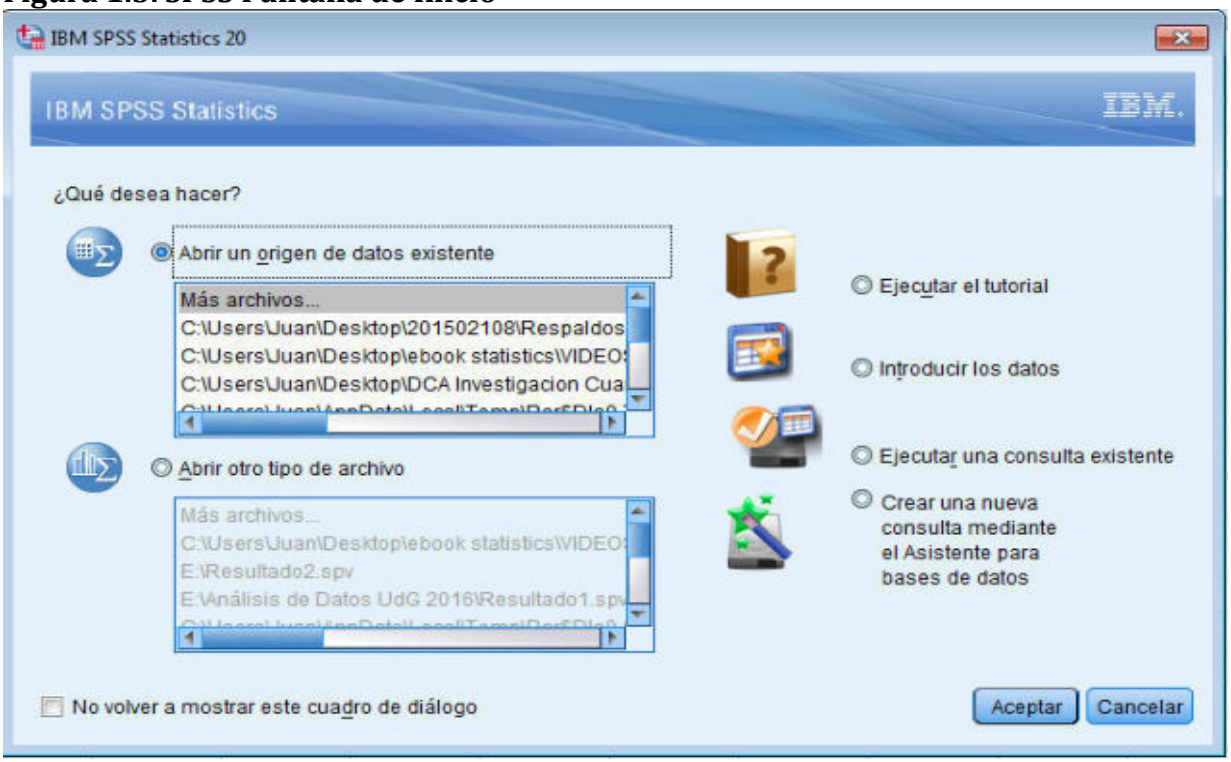


Fuente: SPSS 20 IBM

1.5 SPSS. Opciones de trabajo

Enseguida de la carga de la pantalla de presentación con versión, se despliega la pantalla de inicio. (IBM 2011a).Ver **Figura 1.3**

Figura 1.3. SPSS Pantalla de Inicio



Fuente: SPSS 20 IBM

La ventana anterior nos presenta las posibilidades de (IBM, 2011a) :

- Abrir un origen de datos existentes (Si ya tenemos un archivo SPSS).
- Abrir otro tipo de archivo (Lo seleccionamos si deseamos abrir un archivo que no sea SPSS, puede ser una hoja de cálculo, etc.).
- Ejecutar el tutorial (Nos presentara en otra ventana el tutorial, con diferentes Estudios de caso).
- Introducir datos (Si se desea introducir los datos directamente en el SPSS)
- Ejecutar una consulta existente.
- Crear una nueva consulta mediante al Asistente para bases de datos

Se puede obviar la anterior ventana presionando:

Cancelar.

Si se desea que no vuelva aparecer otra vea esta ventana al abrir el SPSS, se puede marcar la opción:

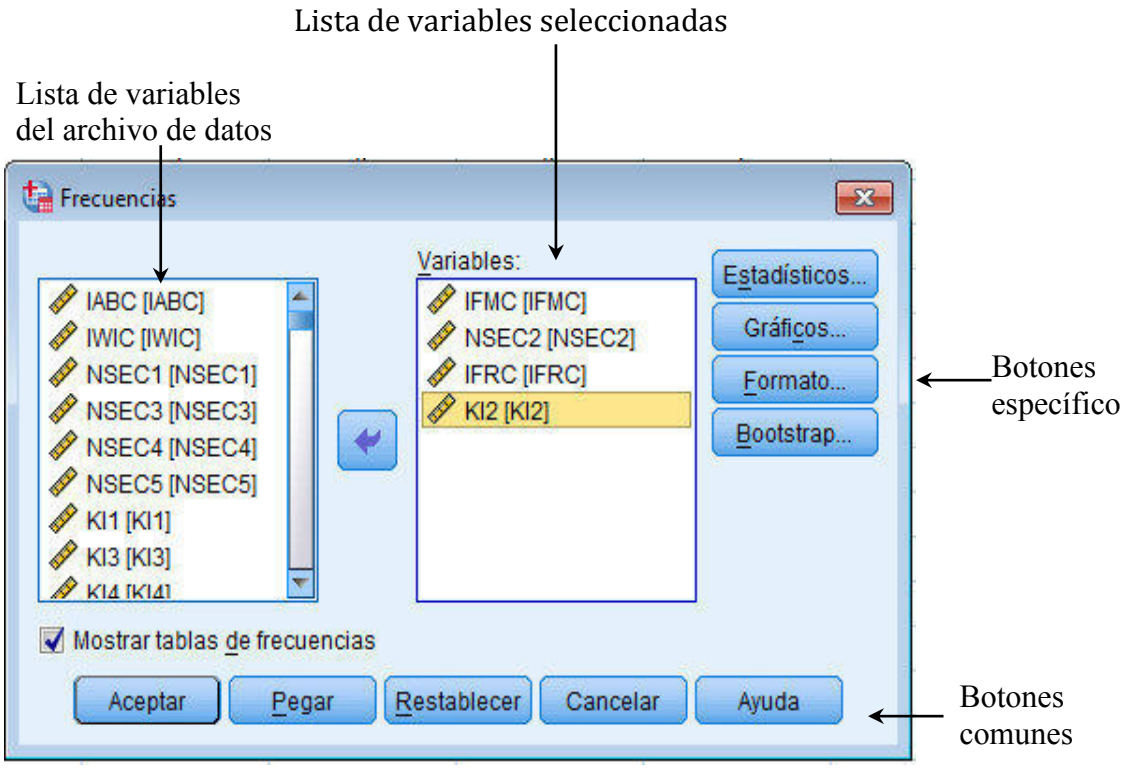
 No volver a mostrar este cuadro de dialogo

 Aceptar

1.6 SPSS. Cuadros de diálogo

Los **cuadros de diálogo** se representan con ventanas que apoyan al usuario a realizar cualquier actividad de la forma más sencilla, (IBM,2011a). Por ejemplo, cuando se intenta abrir un archivo se accede al **cuadro de diálogo Abrir archivo** (ver **Capítulo 3**). Los cuadros de diálogo permiten utilizar la mayoría de las funciones del SPSS simplemente señalando y pinchando con el puntero del ratón. Al intentar, por ejemplo, ejecutar el procedimiento **Frecuencias (Analizar-Estadísticos-Descriptivos)**, se mostrará el cuadro de diálogo Frecuencias de la **Figura 1.4**

Figura 1.4. Contenido del cuadro de diálogo Frecuencias



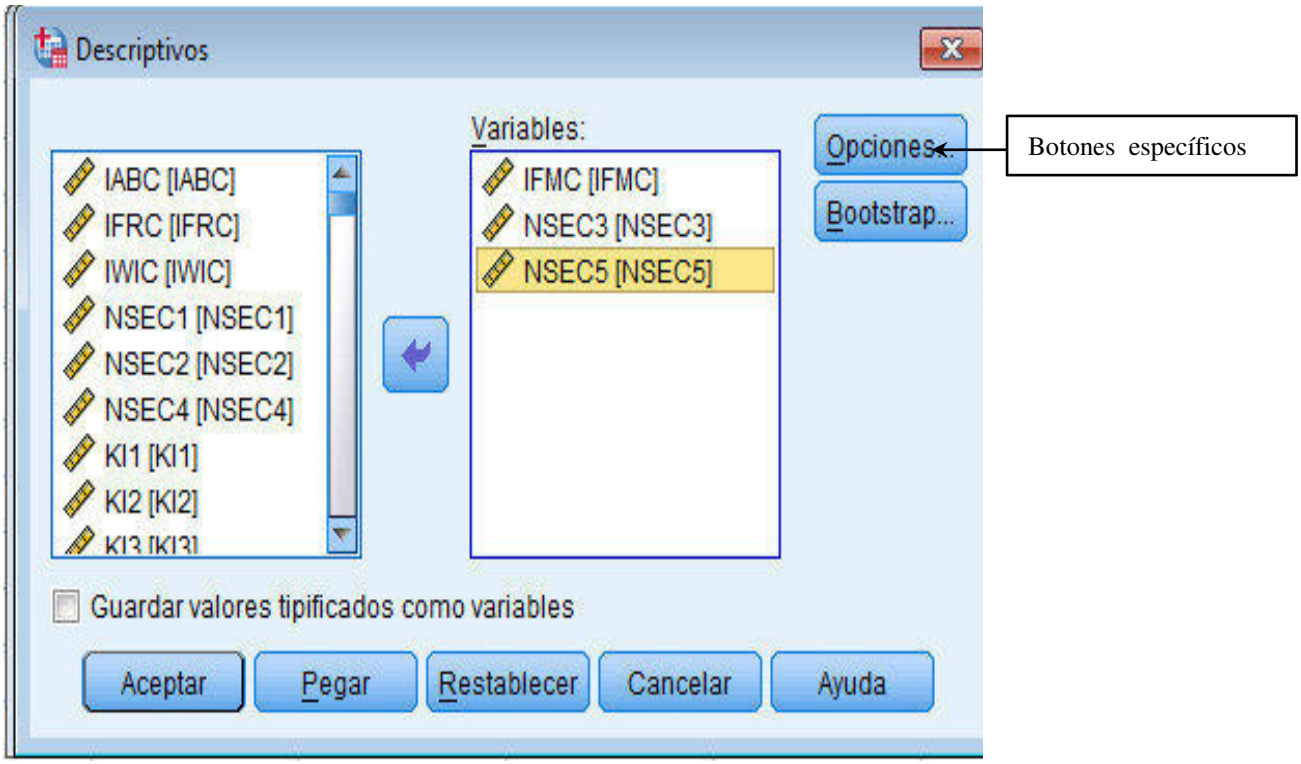
Fuente: SPSS 20 IBM

• **Lista de variables del archivo de datos**. El primer recuadro ofrece un listado de todas las variables del archivo de datos. Las variables nume ricas van precedidas del s mbolo "**#**"; las variables de cadena corta, del símbolo "A<"; y las de cadena larga, del símbolo "***A>***" (IBM,2011a).

Este listado muestra el **nombre de las variables** o su **etiqueta**; pueden aparecer en orden alfabético o en el orden en el que se encuentran en el **Editor de datos.** Ambos detalles pueden controlarse desde el menú **Edición** > **Opciones...**, en la pestaña **General**, dentro del recuadro **Listas de variables**. Oprimiendo el botón derecho del ratón sobre el nombre o la etiqueta de cualquiera de las variables del listado, se obtiene información adicional sobre esa variable: nivel de medida y etiquetas de los valores, si existen.

• **Lista de variables seleccionadas.** Lista (a veces más de una) a la que deben trasladarse las variables con las que se desea trabajar.
• Para trasladar variables desde el listado de **variables del archivo** hasta el listado de **variables seleccionadas**, marcar con el puntero de ratón la variable que se desea trasladar y pulsar el botón flecha situado entre ambos listados. 
• Para devolver al listado de **variables del archivo** una variable previamente seleccionada, Marcar esa variable en el listado de **variables seleccionadas** y pulsar el botón flecha, el cual apunta ahora en la dirección contraria.  .
• Cuando existe un único listado de **variables seleccionadas** (como ocurre en la figura anterior), es posible desplazar variables de un listado a otro pulsando dos veces con el puntero del ratón sobre la variable deseada.
• **Botones comunes**. Son botones que se encuentran en la mayoría de los cuadros de diálogo y siempre con el mismo significado:
• **Aceptar**. Cierra el cuadro de diálogo y ejecuta el procedimiento seleccionado teniendo en cuenta las opciones marcadas y las variables seleccionadas.
• **Pegar**. Genera la sintaxis SPSS correspondiente a las selecciones efectuadas en
El cuadro de diálogo y las pegas en la ventana de sintaxis designada (si no existe ninguna ventana de sintaxis abierta, el SPSS abre una y le asigna el nombre *Sintaxis#*). Cierra el cuadro de diálogo pero no ejecuta el procedimiento.
• **Restablecer**. Limpia el listado de variables seleccionadas y cualquier otra opción
marcada y devuelve a sus valores originales (los valores por defecto) todas las opciones del cuadro de diálogo. No cierra el cuadro de diálogo.
• **Cancelar**. Cancela todos los cambios introducidos en el cuadro de diálogo desde
la última vez que fue abierto y lo cierra.
• **Ayuda**. Ofrece ayuda específica sobre los contenidos del cuadro de diálogo
• **Botones específicos.** Estos van cambiando de un cuadro de diálogo a otro. Así, por ejemplo, en el cuadro de diálogo Frecuencias de la **Figura 1.3**, los botones específicos son **Estadísticos..., Gráficos... y Formato...** Pero si abrimos otro cuadro de diálogo como, por ejemplo, Descriptivos (**Analizar-Estadísticos Descriptivos**) como la **Figura 4**, encontraremos que los botones específicos se limitan a **2**: **Opciones y Bootstrap...** Los botones comunes son ahora exactamente los mismos que antes, pero los botones específicos han cambiado. **Ver Figura 1.5**.

Figura 1.5. Cuadro de diálogo Descriptivos

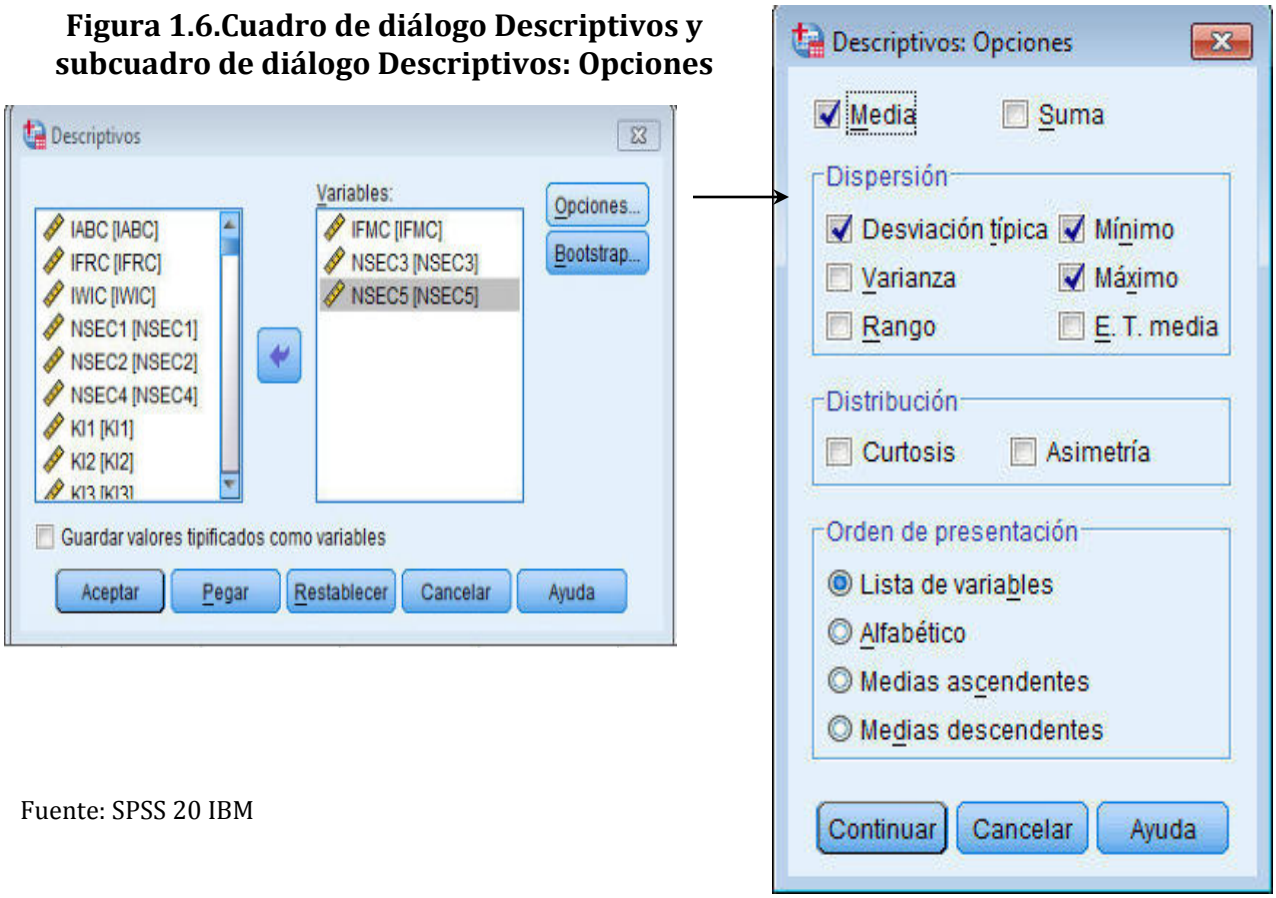


Fuente: SPSS 20 IBM

1.7 SPSS. Subcuadros de diálogo

Los botones específicos poseen la peculiaridad de ir acompañados de **puntos suspensivos**: **Estadísticos...**, **Opciones...** (IBM,2011a).Sirve para recordarnos que se trata de botones que conducen a **subcuadros de diálogo** que están colgando del **cuadro de diálogo principal**. Ver la **Figura 1.6**

Figura 1.6.Cuadro de diálogo Descriptivos y subcuadro de diálogo Descriptivos: Opciones



Fuente: SPSS 20 IBM

1.8 SPSS. Ventanas

Existen **5 tipos de ventanas SPSS (IBM, 2011a).**

1.- El Editor de datos. Contiene el archivo de datos donde se realizan la mayoría de los análisis del SPSS; se abre automáticamente cuando se entra al software, mostrando **2** vistas diferentes: la **Vista de datos** y la **Vista de variables** del archivo con el conjunto de características que las definen. Todas las ventanas SPSS, como el Editor de datos, contienen una **barra de menús** (menús desplegables), una **barra de herramientas** (una serie de botones- iconos para acceso rápido a funciones SPSS) y una **barra de estado** (con información precisa del estado del programa). Es posible trabajar con varios archivos de datos simultáneamente, al abrir más de un Editor de datos; los datos que interese analizar juntos deberán estar en el mismo archivo.

2.- Ventana de resultados y Navegador de resultados. Este recoge toda la información de: estadísticos, tablas, gráficos, etc. que el SPSS genera, permitiendo la edición de los resultados y su guardado para su uso posterior. Es posible tener abiertas varias ventanas del Visor por cada Editor de datos. El Visor de resultados se presentan en **3** formatos distintos: **tablas, gráficos y texto**, y se asocian a un editor con ventana distinta, para cada uno de estos **3** formatos básicos:

-El Editor de tablas. Permite múltiples posibilidades de edición de los resultados presentados en formato de **tabla pivotante**

-El Editor de gráficos. Permite modificar los colores, posición de los ejes, los tipos de letra, las etiquetas, la y muchos otros detalles de los gráficos del Visor.

-El Editor de texto. Permite modificar diferentes atributos (tipo, tamaño, color, etc., de las fuentes) de los resultados tipo texto: **títulos, subtítulos y notas**

3.- El Borrador de la Ventana de resultados. Ofrece la misma información que el Visor en modo normal, pero en formato **texto**, es decir, con un aspecto menos depurado y sin las posibilidades de edición del Visor en modo normal (no es posible, por ejemplo, pivotar tablas o editar gráficos).

4.-El Editor de sintaxis. Permite utilizar las posibilidades de programación del SPSS. Las acciones que el SPSS lleva a cabo como resultado de las selecciones hechas en los menús y cuadros de diálogo se basan en un conjunto de instrucciones construidas con una sintaxis propia del **SPSS**. Estas instrucciones pueden **pegarse** en una ventana de sintaxis desde cualquier cuadro de diálogo. El **botón Pegar** disponible en la mayor parte de los cuadros de diálogo siempre tiene el mismo efecto: convierte en sintaxis SPSS las selecciones hechas, las cuales, al pegarse pueden editarse para, por ejemplo, ejecutar algunas acciones no disponibles desde los cuadros de diálogo, o para salvarla en un archivo y volver a utilizarla en una sesión diferente. Es posible tener abiertas simultáneamente varias ventanas de sintaxis. Aunque el Editor de sintaxis no es imprescindible para trabajar con el SPSS, su capacidad para, entre otras cosas, automatizar trabajos repetitivos, lo convierte en una ventana de especial utilidad

5.-Editor de procesos. Personaliza y automatiza algunas de las tareas que el SPSS lleva a cabo, especialmente en lo relacionado con el contenido y el aspecto de las tablas de resultados.

1.8.1 SPSS. Ventana Editor de datos

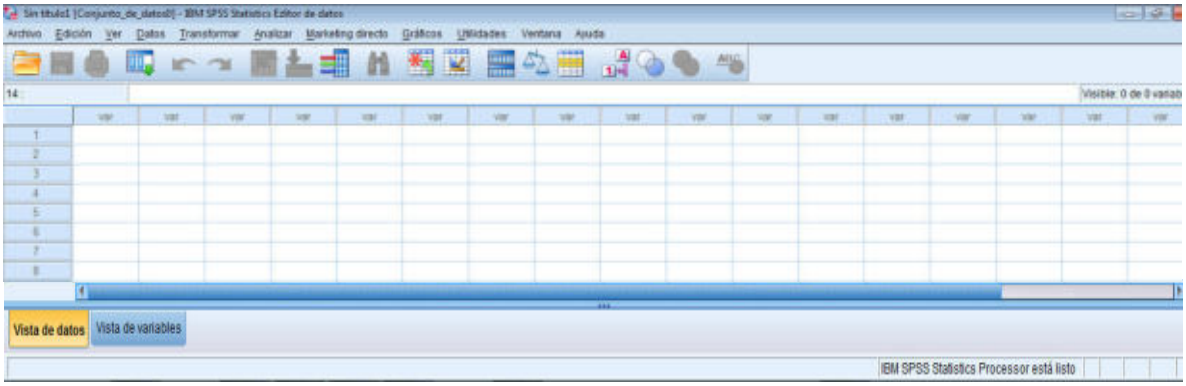
El **Editor de datos** (IBM,2011a) es la primera ventana que nos presenta y se abre automáticamente cuando se inicia la sesión, similar a las hojas de cálculo para la creación y edición de archivos de datos. El **Editor de datos** proporciona dos vistas: **Vista de datos** y **Vista de variables.**

Ingresamos o nos movemos entre ambas ventanas, seleccionando en las pestañas inferiores:

.Vista de datos y Vista de variables. Ver **Figura 1.7 y Figura 1.8.**

Figura 1.7. SPSS Editor de datos: Vista de Editor de datos

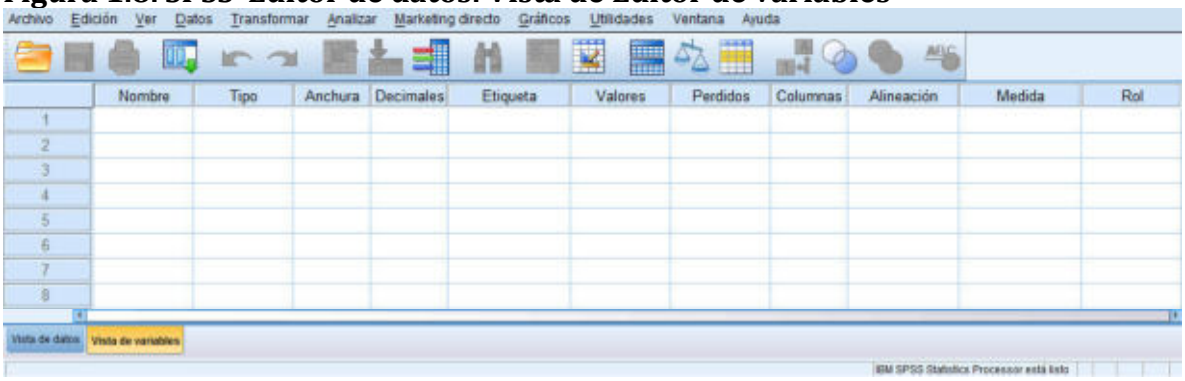
1.7. SPSS. Ventana Editor de datos: vista de datos



Fuente: SPSS 20 IBM

- **Las filas.** Cada fila representa un **caso** o una observación. Por ejemplo, las respuestas de un **cuestionario** corresponde a una fila.
- **Las columnas.** Cada columna representa una **variable** o una característica que se mide o una pregunta formulada. En un **cuestionario** la columna corresponderá a cada **pregunta** del cuestionario.
- **Las celdas.** Contienen valores de las **variables**, siendo este un valor único de una variable. La celda se encuentra en la intersección del **caso** y la **variable**. La diferencia con las hojas de cálculo, es que **no pueden contener fórmulas o realizarse operaciones entre celdas.**
- **Vista de variables.** Aquí se muestran la información de definición de las variables, que incluye: las etiquetas de la variable, tipo de datos (**cadena, fecha o numérico**), nivel de medida (**nominal, ordinal o de escala**) y los valores perdidos definidos por el usuario. Para ubicarnos en **vista de variables** seleccionamos, en las pestañas inferiores, **a describir 11 celdas**. **Ver Figura 1.8.**

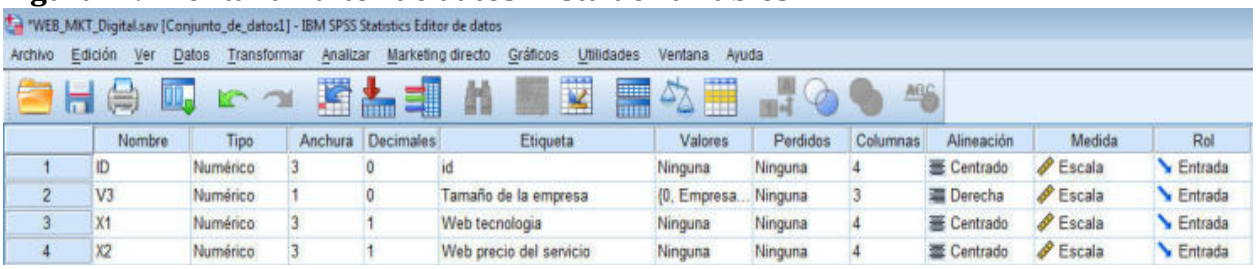
Figura 1.8. SPSS Editor de datos: Vista de Editor de variables



Fuente: SPSS 20 IBM

I.-Nombre.-El nombre asignado a cada variable debe ser único, no se puede tener nombres duplicados, el primer carácter del nombre de la variable debe ser una letra o uno de estos caracteres: @, # o $; los caracteres posteriores puede cualquier combinación de letras, números, que no sean signos de puntuación, punto **(.)**, lo recomendable es que este nombre sea una abreviación o simbología del nombre real de la variable (esto se realiza dándole una codificación, el código o nombre no debe ser mayor de **8 caracteres**). Por ejemplo, si tenemos las siguientes variables. Ver **Figura 1.9**.

Figura 1.9. Ventana Editor de datos: vista de variables



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	ID	Numérico	3	0	id	Ninguna	Ninguna	4	Centrado	Escala	Entrada
2	V3	Numérico	1	0	Tamaño de la empresa	{0, Empresa...	Ninguna	3	Derecha	Escala	Entrada
3	X1	Numérico	3	1	Web tecnologia	Ninguna	Ninguna	4	Centrado	Escala	Entrada
4	X2	Numérico	3	1	Web precio del servicio	Ninguna	Ninguna	4	Centrado	Escala	Entrada

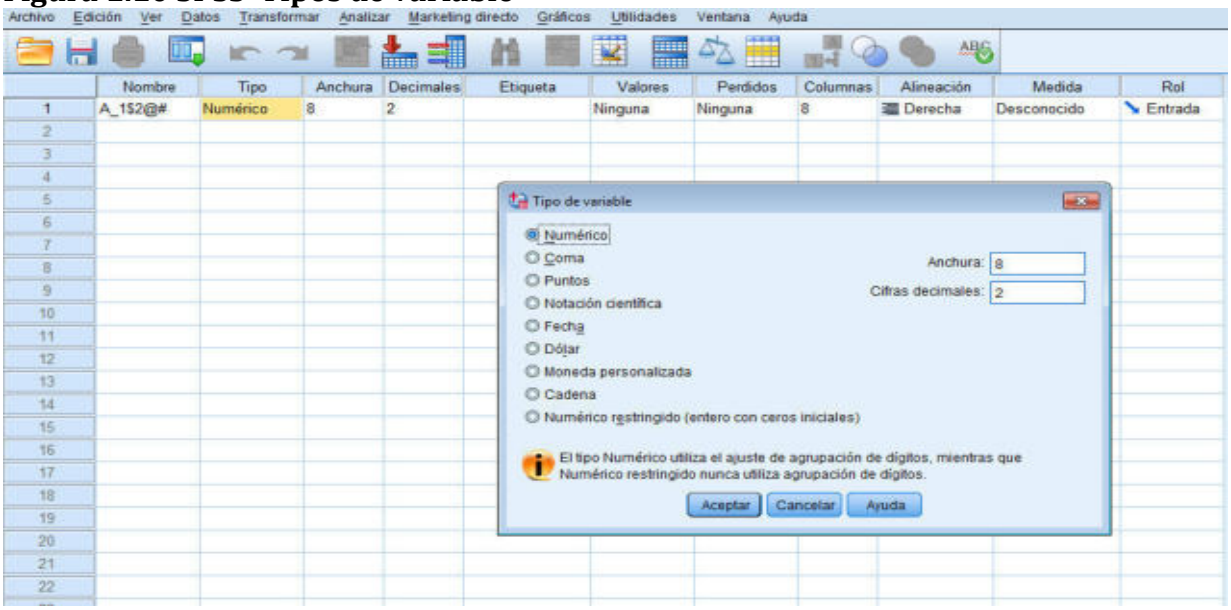
Fuente: SPSS 20 IBM

Para los **nombres de las variables** se debe considerar (IBM,2011a) :

- Es posible emplear los caracteres **$, # y @** dentro de los nombres de variable. Por ejemplo: **A_1$2@# es un nombre de variable válido**.
- Evitar que los nombres de variable que terminen en **punto (.).**
- Evitar que los nombres de variable que terminan con un carácter de subrayado.
- No se pueden utilizar como nombres de variable: **ALL, AND, BY, EQ, GE, GT, LE, LT, NE, NOT, OR, TO** y **WITH**; por ser parte del sistema.
- En los nombres de variable se pueden colocar con **mayúsculas y/o minúsculas**, siendo que el programa realiza una distinción entre estas.

II.-Tipo. Definir el tipo de datos de cada variable, de inicio se asume que todas las variables nuevas son **numéricas**. Para definir el **Tipo**, debemos hacer clic en la casilla de la variable de interés, de manera que aparezca en el costado derecho de la casilla un botón cuadrado con **puntos suspensivos (…)**. Al seleccionar el botón (hacer clic), aparece el **cuadro de diálogo** Tipo de variable en donde se apreciara las diferentes opciones. Ver **Figura 1.10**.

Figura 1.10 SPSS Tipos de variable



Fuente: SPSS 20 IBM

Existen **9 diferentes tipos de variables (IBM,2011a)** a elegir:

1.-Numérico. Se emplea en una variable numérica cuyos valores representan magnitudes o cantidades; este es el tipo de variable más usado, está relacionado con el formato estándar que se maneja en Windows, donde el separador decimal es coma (,) y no se tiene separación de miles. Por ejemplo: 1000,00. Así mismo se puede definir el número de decimales.
2.-Coma. Se emplea para variables numéricas, en el caso de que la separación de miles sea coma (,) y el punto como separador decimal (.). Por ejemplo: 1,000.00. Así mismo se puede definir el número de decimales.
3-Puntos. Se emplea cuando para variables numéricas, donde el separador de miles es punto (.) y el separador decimal es coma (,). Por ejemplo: 1.000,00. Así mismo se puede definir el número de decimales.
4.-Notación científica. Se utiliza cuando se emplea un exponente con signo que representa una potencia en base diez. 1'000.000.00 = 1.0E+6 o 0.000001 = 1.0E (-6). SPSS nos permite representarlo de varias formas como 1000000, 1.0E6, 1.0D6, 1.0E+6, 1.0+6. La notación es útil cuando manejamos cifras extremas de lo contrario es mejor manejarlo de forma numérica.
5.-Fecha. Este tipo de variable se emplea cuando los valores de la variable representan fechas de calendario u horas de reloj; al seleccionarla aparece en el cuadro de diálogo una casilla con el listado de los diferentes formatos que el programa.
6.-Dólar. Se emplea en una variable numérica cuyos valores representan dinero en dólares. La diferencia con el tipo numérico es que al seleccionar este tipo de variable se adicionara el símbolo del dólar ($) en el valor.
7.-Moneda personalizada. Es una variable numérica se emplea cuando los valores de una variable representan sumas de dinero diferentes al dólar (Pesos, Euros, etc.); al seleccionar no representa una moneda específica, si no que por el contrario el programa asume que la moneda es de origen distinto al dólar. La diferencia con el tipo dólar es que nos permite trabajar con cinco (5) diferentes tipos de moneda.
8.-Cadena. Se emplea cuando la variable no es numérica, es decir puede contener textos. Las mayúsculas y las minúsculas se consideran diferentes. Este tipo también se conoce como variable alfanumérica porque puede contener texto con número. Las variables de cadena pueden contener cualquier tipo de caracteres siempre que no exceda la longitud máxima de 255; las mayúsculas y las minúsculas se consideran diferentes ya que el programa trabaja bajo el código ASCII.
9.-Númerico restringido (entero con ceros iniciales). Se lo emplea si la variable es numérica entera y se desea apreciar ceros a la izquierda.

Para definir alguno de los tipos de variable, basta con hacer clic sobre cualquiera de las opciones y definirla (IBM,2011a):

III.-Anchura.-Por medio de esta propiedad podemos definir el máximo de dígitos que contienen los registros de una variable; para el cálculo del ancho se incluyen los dígitos enteros y los decimales. Por ejemplo; Anchura 5 = XXX.XX o X,XXX.X o XX,XXX donde X representa un número aleatorio.

IV.-Decimales.-A través de esta opción definimos el número de dígitos decimales que pueden contener los registros de una variable numérica (**Tipo Numérico, Coma o Puntos**).

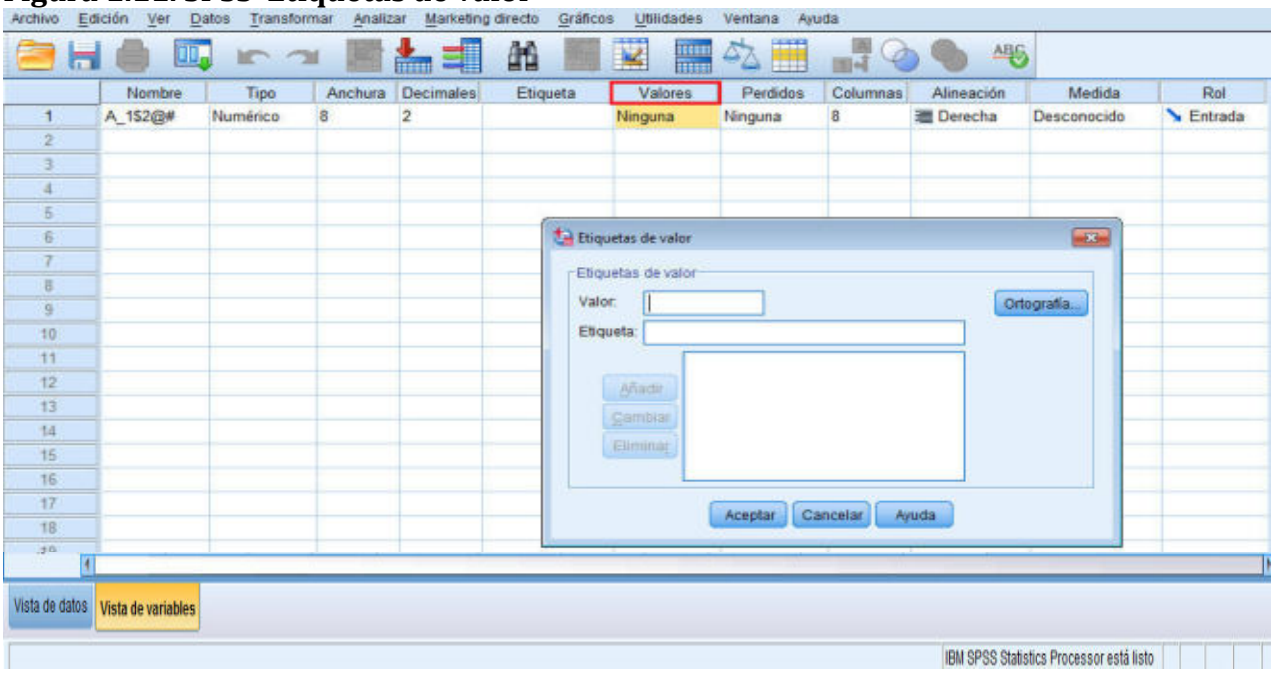
Las propiedades Anchura y Decimales pueden ser editadas directamente desde la ventana de Tipo de variable, ya que al seleccionar estas opciones se habilita en el cuadro de diálogo las casillas Anchura y Decimales.

V.-Etiquetas.-Nos sirve para colocar el nombre largo de variable, se puede asignar etiquetas de variable descriptivas de hasta **256 caracteres de longitud**. Las etiquetas de variable pueden contener espacios y caracteres reservados que no se admiten en los nombres de variable. El uso de la etiqueta es bastante útil para facilitar la interpretación de los resultados (**Tablas, Gráficos o estadísticos**), para las personas que no han participado en la generación de los procedimientos y desconocen el significado del nombre de la variable. El uso de la etiqueta es opcional, el programa en caso de no existir una etiqueta utiliza el nombre de la variable para generar los resultados.

VI.-Valores.- En el caso de que se utilice **códigos numéricos o literales** para representar categorías variables numéricas o de cadena; por ejemplo:
-Variable: Sexo, donde: 1 = hombre; 2 = mujer
-Variable: Tamaño, donde: P = pequeño; M = mediano; G = grande
-Variable: Altura de planta (si esta tiene intervalos), donde: 1 = 49 – 54; 2 = 55 -60; 3 = 61 – 66; 4 = 67 – 72; 5 = 73 – 78; 6 = 79 – 84; 7 = 85 – 90

Por ejemplo: Para **especificar etiquetas de valor de una variable**, ej. **Teclee celda Valores –>Ninguna (…)** de la que observará la ventana emergente de la **Figura 1.11.**

Figura 1.11. SPSS Etiquetas de valor



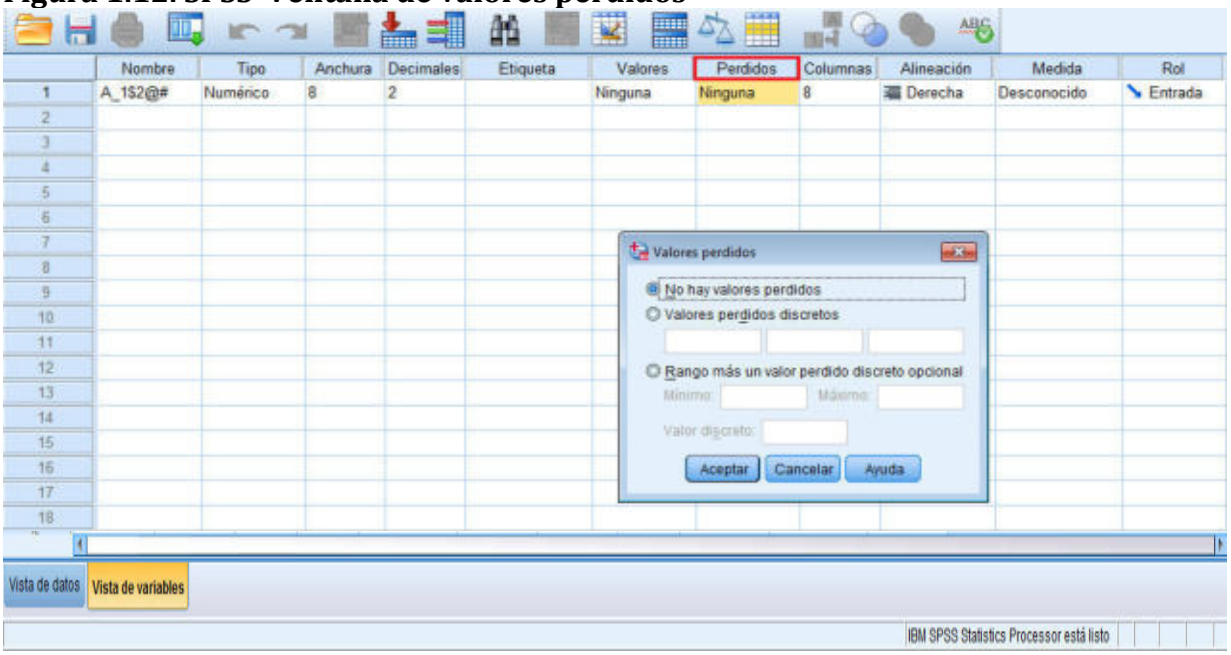
Fuente: SPSS 20 IBM

Una vez que estamos en la ventana de etiquetas de valor, en el caso del ejemplo de la **variable Sexo**, en la **celda Valor** se escribe el código (número o letra) y en la **celda**

Etiqueta escribimos el significado del código, una vez introducidos todos los códigos y etiquetas, presionamos aceptar:
Valor: 1; **Etiqueta:** Hombre
Añadir
Valor: 2; **Etiqueta:** Mujer
Añadir
Aceptar

VII.-Perdidos.-Con esta opción, se indica los valores de los datos definidos como perdidos por el usuario. SPSS maneja dos tipos de valores perdidos; el primero es perdido por el sistema, el cual se identifica por la ausencia total de datos; es decir, casillas vacías y el segundo corresponde a los datos perdidos definidos por el usuario:
-No sabe
-No responde o se niega a responder
-No aplica o sencillamente la pregunta no lo afecta, por ejemplo: preguntarle a una persona soltera la edad a la que se casó por primera vez, si no se ha casado nunca esta pregunta no lo afecta.
El programa detecta automáticamente los valores perdidos por el sistema y los omite, adicionando un punto en la celda correspondiente **(.)** como valor perdido, mientras que los valores perdidos por el usuario deben ser definidos al programa o de lo contrario los cálculos se realizarán contando con estos valores, lo cual puede afectar severamente los resultados.
Por ejemplo: para **definir un valor perdido** por el usuario para una variable, ej. **Teclee celda Perdidos ->Ninguna (...)**, de lo que observará la ventana emergente de la **Figura 1.12.**

Figura 1.12. SPSS Ventana de valores perdidos



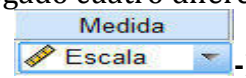
Fuente: SPSS 20 IBM

En este cuadro encontramos **3** diferentes posibilidades.
-No hay valores perdidos. Los cálculos se realizan con la totalidad de los registros.
-Valores perdidos discretos. Nos permite un máximo de tres valores perdidos que se pueden definir para una variable; se puede emplear los valores (números) que se deseen. Para este tipo de valores se recomienda que exista una distancia considerable entre los valores representativos y los perdidos con el fin de facilitar su identificación. Por ejemplo 999,9999. Para definir como perdidos los valores nulos o vacíos de una variable de cadena, escriba un espacio en blanco en uno de los campos debajo de la selección Valores perdidos discretos.
-Rango más un valor discreto opcional. Se utiliza cuando tenemos varios valores perdidos, los cuales se encuentran dentro de un rango. Esta opción solo es para variables numéricas.

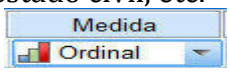
En el caso de **variables del tipo cadena (IBM, 2011a):**
-Se considera como válidos todos los valores de cadena, incluidos los valores vacíos o nulos, a no ser que se definan explícitamente como perdidos.
-Los valores perdidos de las variables de cadena no pueden tener más de ocho bytes.

VIII.-Columnas. Se puede especificar un número de caracteres para el ancho de la columna. Los anchos de columna también se pueden cambiar en la **Vista de datos** pulsando y arrastrando los bordes de las columnas.
IX.-Alineación.-La Alineación determina la alineación de los datos dentro de la casilla (Izquierda, derecha y centro). Por defecto es a la derecha para las variables numéricas y a la izquierda para las variables de cadena.

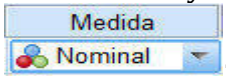
X.-Medidas.-Este es el parámetro más importante de las variables, de su definición depende el tipo de análisis que podemos realizar con el programa. Dentro de la estadística se han catalogado cuatro diferentes escalas de medida, pero el SPSS la resume en tres:



-Nominal. Son variables numéricas cuyos valores (Números) indican una categoría de pertenencia. Para este tipo de medida, las categorías no cuentan con un orden lógico que nos permita establecer una comparación de superioridad u ordenación entre ellas. Por ejemplo:
El género, el estado civil, etc.



-Ordinal. Son variables numéricas cuyos valores indican una categoría de pertenencia y a su vez las categorías poseen un orden lógico que nos indica una superioridad u ordenación. Por ejemplo: nivel de ingresos, nivel educativo, etc. Entre las variables ordinales se incluyen escalas de Likert.



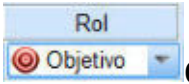
-Escala. Son variables numéricas sean estas discretas o continuas cuyos valores representan una magnitud o cantidad y no una categoría; los valores de este tipo de medida pueden ser empleados en operaciones aritméticas como la

suma, la resta, la multiplicación y la división. Como por ejemplo: Edad, altura, peso, rendimiento, etc.

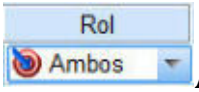
Para **variables de cadena** si estos son **ordinales**, se debe tener en cuenta que el SPSS asume el **orden alfabético** de los valores de cadena, por ejemplo si tenemos **chico, mediano y grande**, el SPSS nos presentara **chico, grande y mediano**, por lo que es mejor emplear números en la codificación de datos.

XI.-Rol. Se usa cuando se quiere predefinir el rol que cumplirá una determinada variable, siendo:

Entrada. La variable se utilizará como una entrada (por ejemplo, predictor, variable independiente).

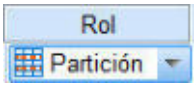


Objetivo. La variable se utilizará como una salida u objetivo (por ejemplo, variable dependiente).



Ambos. La variable se utilizará como entrada y salida.

Ninguna. La variable no tiene asignación de función



Partición. La variable se utilizará para dividir los datos en muestras diferentes para entrenamiento, prueba y validación.

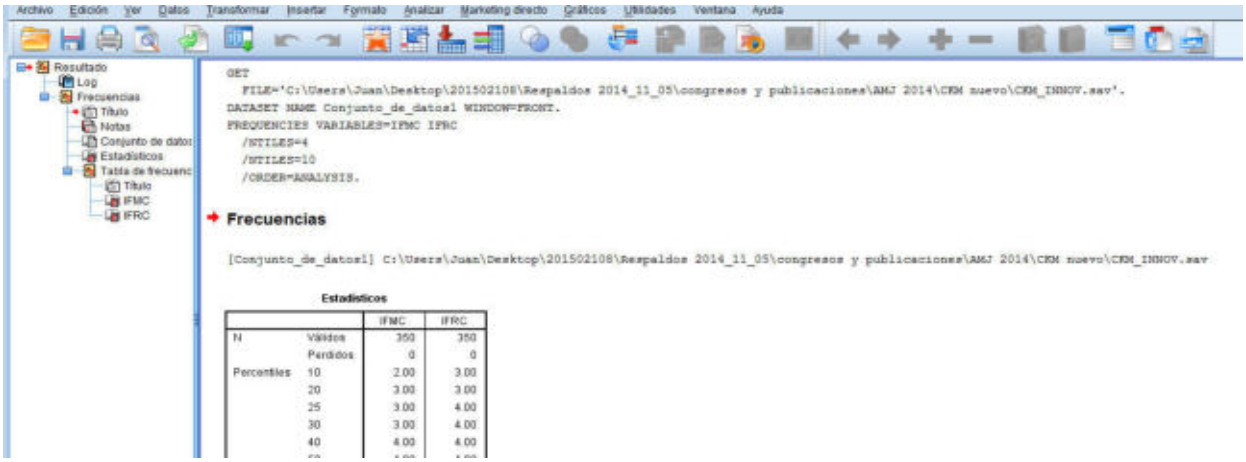
Dividir. Las variables no se utilizan como variables de archivos divididos en IBM® SPSS® Statistics

La asignación de roles sólo afecta a los cuadros de diálogo que admiten asignaciones de roles.

1.8.4. SPSS. Ventana de resultados y Navegador de resultados

Es la ventana donde aparecen los resultados de los análisis realizados con el programa (ej. **teclee Analizar->Estadísticos descriptivos->Frecuencias,** con una base de datos previamente cargada, eligiendo las variables), se presentará una ventana emergente como la **Figura 1.13.**

Figura 1.13. SPSS Ventana de resultados



Fuente: SPSS 20 IBM

Se puede archivar estos resultados para su utilización posterior. En el lado izquierdo el **navegador de resultados**, muestra donde explorar todos los resultados mediante distintos procedimientos del paquete.

Nota: Se debe tener cuidado que cada vez que se ejecutan los resultados, estos se van acumulando, por lo cual será necesario en muchos casos **borrar** su contenido, para identificar el de interés.

1.8.5 SPSS. Ventana de gráficos

Se la activa cuando realizamos gráficos y, nos permite modificar y archivar Gráficos. (ej. **teclee Gráficos->Generador de gráficos**, con una base de datos previamente cargada, eligiendo las variables) Ver **Figura 1.14.**

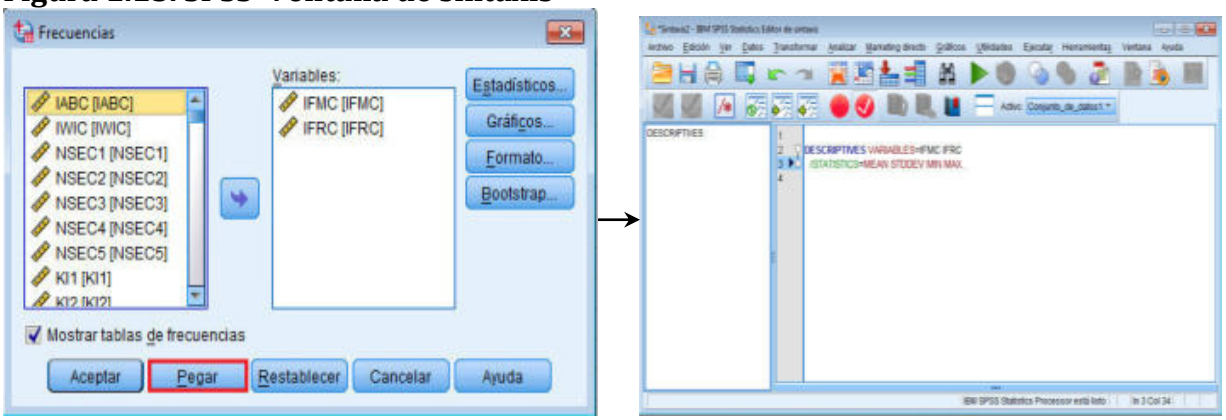
Figura 1.14. SPSS Ventana de gráficos

Fuente: SPSS 20 IBM

1.8.6 SPSS. Ventana de sintaxis

No solo se puede trabajar con los menús de las barras de herramientas, también se puede trabajar con la **Sintaxis** (IBM,2011a), esta forma es recomendable cuando se tiene análisis repetitivos, se puede pegar en esta ventana la sintaxis de los comandos seleccionados desde la ventana de dialogo de cualquier opción. La apertura de la ventana de **Sintaxis**, se la realiza cuando se está realizando algún tipo de análisis, mediante la opción **Pegar**. Por ej. **Teclee Analizar->Estadísticos descriptivos->Frecuencias** con una base de datos previamente cargada, eligiendo las variables Lo que nos proporcionar la ventana respectiva de análisis seleccionado, en esta se selecciona el botón: Pegar, lo que nos mostrara la ventana de **Sintaxis:** Ver **Figura 1.15.**

Figura 1.15. SPSS Ventana de sintaxis

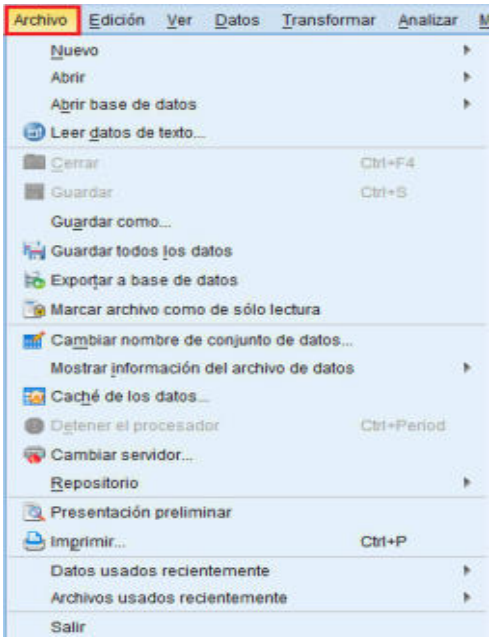


Fuente: SPSS 20 IBM

Esta ventana permite editar la sintaxis de los comandos y ampliarla con aquellas opciones que tiene el lenguaje SPSS, pero que no están disponibles a través de menús. Estos comandos pueden archivarse (**en archivos de texto con extensión .sps**.) y recuperarlos en sesiones posteriores con SPSS.

1.9 SPSS. Barras de menú

Son **11 menús desplegables** (IBM,2011a) que permiten controlar la mayor parte de las acciones que el SPSS puede llevar a cabo, siendo:



1.-Menú Archivo.- En el caso de Editor de datos, si nos situamos en el menú Archivo, de la barra de menús, entre las **principales opciones** podemos rescatar:

- **Nuevo.** Crea un nuevo Archivo de: Datos, Sintaxis, Resultado y Procesos.
- **Abrir.** Abre un archivo de: Datos, Sintaxis, Resultado y Proceso.
- **Abrir base de datos.** Nos permite: Abrir una nueva consulta, Editar consulta y Ejecutar consulta.
- **Abrir datos de SPSS Data Collection.** Abrirá los datos que se encuentran en la colección de SPSS

Fuente: SPSS 20 IBM

- **Leer datos de texto.** Nos permite leer datos de tipo texto que tengan las terminaciones en: *.txt, *.dat, *.csv.
- **Cerrar.** Cerrar el archivo de datos una vez que este se haya guardado.
- **Guardar.** Nos permite guardar el archivo de datos en formato SPSS, con la terminación *.sav.
- **Guardar como.** No permite guardar en una dirección y sitio personalizado en formato SPSS, con la terminación *.sav.
- **Guardar todos los datos.** Al igual que los anteriores nos permite guardar los datos en formato SPSS, con la terminación *.sav.
- **Exportar a base de datos.** Nos permite exportar los datos.
- **Cambiar el nombre de conjunto de datos.** Con esta opción podemos cambiar el nombre al conjunto de datos.
- **Presentación preliminar.** Nos presenta una vista previa del conjunto de datos.
- **Imprimir.** Imprimirá el conjunto de datos.
- **Datos usados recientemente.** Nos presenta un listado de los datos usados recientemente.
- **Archivos usados recientemente.** Nos presenta un listado de archivos usados recientemente.
- **Salir.** Saldrá del programa.
- Para el caso de la **Ventana de resultados**, en el menú Archivo, se activa la opción:
- **Exportar.** El cual nos permite exportar los resultados, en este caso se puede definir el tipo de archivo a ser exportado (*.doc, *.htm, *.pdf, *.xls, etc.).

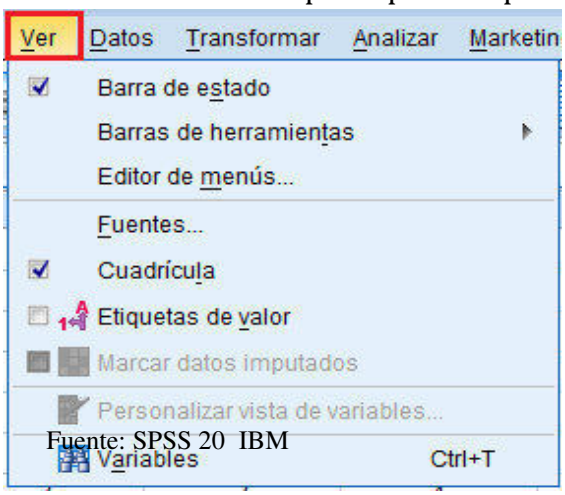
2.-Menú Edición.- Las opciones de **Edición** son las habituales opciones de Windows (Deshacer, Rehacer, Cortar, Copiar, Pegar, Buscar, Buscar siguiente, Reemplazar).

Fuente: SPSS 20 IBM

A esta se suman:

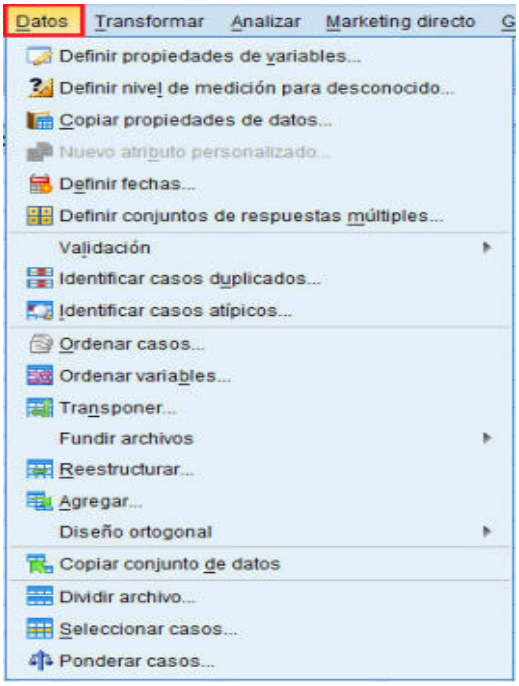
- **Borrar.** Nos permite borrar las ya sea las columnas o las filas.
- **Insertar variables.** Con esta opción se puede insertar una columna que corresponderá a una variable.
- **Insertar caso.** Inserta una fila en el conjunto de datos, donde se ubique el cursor.
- **Ir al caso.** Nos desplaza a un caso determinado.
- **Ir a la variable.** Nos permite desplazarnos a una variable determinada.
- **Opciones.** Nos permite personalizar y definir diferentes opciones como son: General, Visor, Datos, Moneda, Etiquetas de los resultados, Gráficos, Tablas de pivote, Ubicación de archivos, Procesos, Imputación múltiples, Editor de sintaxis.

3.-Menú Ver.- Esta opción permite presentarnos diferentes barras:



Fuente: SPSS 20 IBM

- **Barra de estado.**
- **Barras de herramientas.**
- **Editor de menús.**
- **Fuentes.** Permite seleccionar la fuente.
- **Líneas de cuadricula.** Nos muestra las cuadriculas o las esconde.
- **Etiquetas de valor.** En el caso de que los números o letras signifiquen categorías, nos presentara las etiquetas de los valores.
- **Variables.** Nos desplazamos a la ventana de Variables.

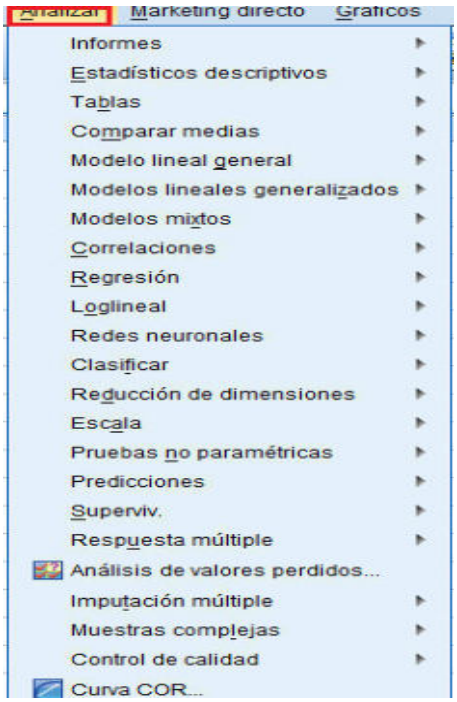


Fuente: SPSS 20 IBM

4.-Menú Datos.-Contiene opciones para hacer cambios que afectan a todo el archivo de datos: unir archivos, trasponer variables y casos, crear subconjunto de casos, etc. Estos cambios son temporales mientras no se guarde explícitamente el archivo.

Fuente: SPSS 20 IBM

5.-Menú Transformar.-Podemos realizar cambios sobre variables seleccionadas, creación de nuevas variables. Estos cambios son temporales mientras no se guarde explícitamente el archivo.



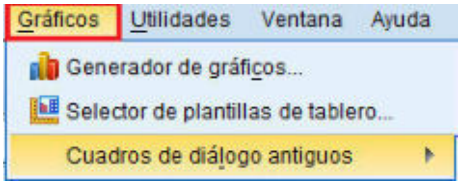
Fuente: SPSS 20 IBM

6. Menú Analizar.-Desde esta opción se ejecutan todos los procedimientos estadísticos

7.- Menú Marketing directo.-La opción ofrece un conjunto de herramientas diseñadas para mejorar el resultado de campañas de marketing directo identificando y adquiriendo características y otras características que definen a diferentes grupos de consumidores y dirigiéndose a grupos concretos para aumentar al máximo los índices de respuesta positivos.

Fuente: SPSS 20 IBM

8.-Menú Gráficos.-Con la opción Gráficos, se puede crear gráficos a partir de los gráficos predefinidos de la galería o a partir de los elementos individuales (por ejemplo, ejes y barras). Se puede crear un gráfico arrastrando y colocando los gráficos de la galería o los elementos básicos en el lienzo, que es la zona grande situada a la derecha de la lista Variables del cuadro de diálogo Generador de gráficos.

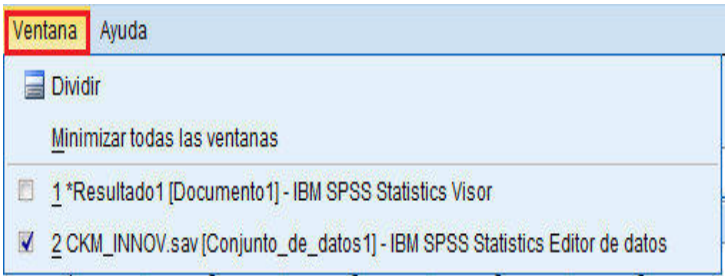


Fuente: SPSS 20 IBM

9.-Menú Utilidades.-Permite cambiar fuentes, obtener información completa del archivo de datos, acceder a un índice de comandos SPSS, etc.

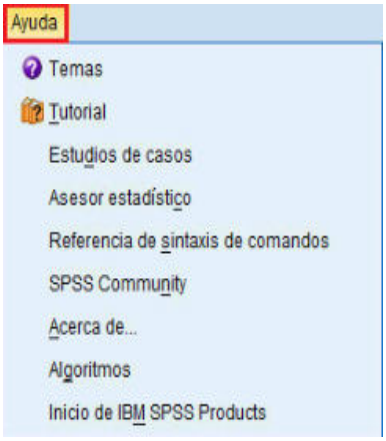
Fuente: SPSS 20 IBM

10.-Menú Ventana.- Permite ordenar, seleccionar, controlar atributos de las ventanas abiertas.



Fuente: SPSS 20 IBM

11.- Menú Ayuda.- Abre un archivo estándar de ayuda, como ser: Temas, Tutorial, Estudios de casos, Asesor estadístico, Referencia de sintaxis de comandos, etc.



Fuente: SPSS 20 IBM

1.10 SPSS. Barras de herramientas

Situada debajo de la barra de menús, permite un acceso rápido a funciones habituales del SPSS (IBM, 2011a, IBM 2011c) . La barra de herramientas del Editor de datos es:

Las cuales describiéndolas son:

Insertar caso

Deshacer una acción del usuario

Guardar este documento

Seleccionar casos

Insertar variable

Volver a hacer una acción del usuario

Imprimir

Etiquetas de valor

Dividir archivo

Ir a caso

Recuperar el cuadro de diálogo

Utilizar conjunto de variables Recientes

Fuente: SPSS 20 IBM

En el caso de la ventana de resultados, la barra de herramientas contiene (IBM, 2011a):

Abrir resultados

Ir a caso

Buscar

Guardar

Seleccionar últimos resultados

Imprimir

Asociar auto proceso

Ver presentación preliminar de este conjunto de datos

Crear/editar auto proceso

Exportar

Ejecutar proceso

Ir a datos

Designar ventana

Ir a la variable

Variables

Mostrar todas las variables

Corregir ortografía

Fuente: SPSS 20 IBM

Barra de titulares del visor. Esta se presenta en la ventana de resultados:

Describiendo alguna de las opciones, tenemos:

Ascender

Ocultar elementos seleccionados

 Degradar

 Insertar encabezado

 Expandir elementos de titulares seleccionados

 Nuevo título

 Contraer elementos de titulares seleccionados

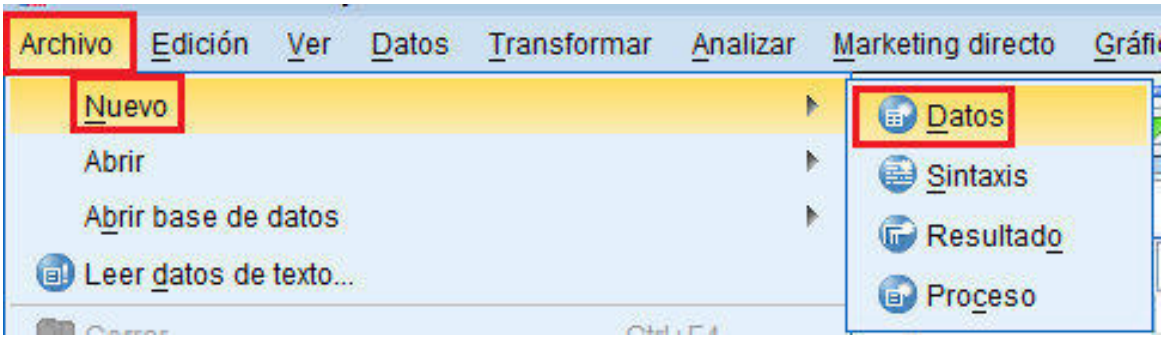
 Nuevo texto

 Mostrar elementos seleccionados

Fuente: SPSS 20 IBM

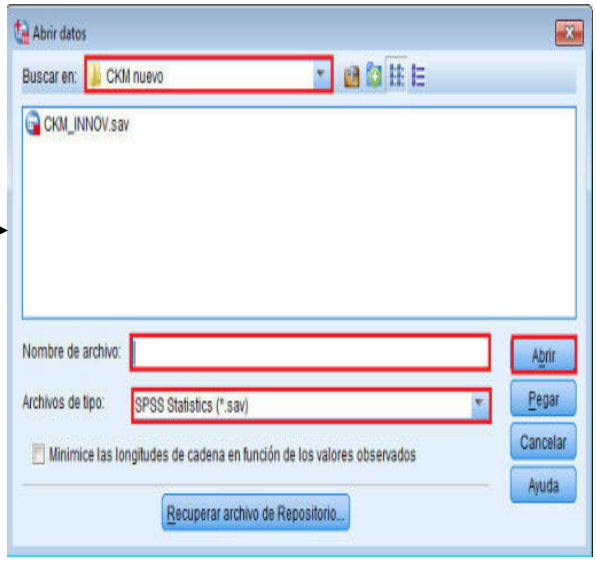
1.11 SPSS. Archivos de trabajo

Crear un archivo.-Podemos utilizar el Editor de datos de SPSS para introducir los datos y crear un archivo de datos, tecleando: **Archivo->Nuevo->Datos...** Una vez abierta la ventana del **Editor de datos,** se definen las variables con la **Vista de variables**, y posteriormente en la **Vista de datos** procedemos a introducir los datos. (IBM, 2011a)



Fuente: SPSS 20 IBM

Abrir un archivo.-Podemos abrir un archivo de datos SPSS, que este previamente almacenado, tecleando: **Archivo->Abrir ->Datos.**



Fuente: SPSS 20 IBM

Así en el **cuadro de dialogo de Abrir datos**, aparecen las opciones:
-Buscar en: (Localizamos el lugar donde se guardó el archivo.)
-Nombre de archivo: (Seleccionamos de la lista el archivo, SPSS nos dará una relación de los archivos con extensión ***.sav**)
-Archivos tipo: (Permite seleccionar entre distintos tipos de archivos de datos. Por defecto tendremos la opción SPSS -***.sav**- seleccionada.)

El SPSS reconoce los siguientes **tipos de archivos:**
SPSS Statistics (*.sav). Es el tipo por defecto, son archivos creados y/o grabados en SPSS para Windows.
SPSS/PC+ (*.sys). Archivos creados y/o grabados en SPSS/PC+. Solo está disponible en los sistemas operativos **Windows.**
Systat (*.syd, *.sys). Abre archivo de datos de SYSTAT.
Portable. Abre archivos de datos guardados con formato portátil. El almacenamiento de archivos en este formato lleva mucho más tiempo que guardarlos en formato SPSS Statistics.
Excel (*.xls, *xlsx, *.xlsm). Abre archivos de Excel.
Lotus (*.w*). Abre archivos de datos guardados en formato de Lotus.
Sylk (*.slk). Abre archivos de datos guardados en formato SYLK (vínculo simbólico), un formato utilizado por algunas aplicaciones de hoja de cálculo.
dBASE (*.dbf). Abre archivos con formato dBASE para dBASE IV, dBASE III o III PLUS, o dBASE II. Cada caso es un registro. Las etiquetas de valor y de variable y las especificaciones de valores perdidos se pierden si se guarda un archivo en este formato.
SAS (*. sas7bdat, *.sd7, *.sd2, *. ssd01. *. ssd04, *.spt). Abre los archivos de las versiones 6-9 del SAS y archivos de transporte SAS. Con la sintaxis de comandos, también puede leer etiquetas de valor de un archivo de catálogo de formato SAS.
Stata (*.dta). Abre archivos de las versiones 4–8 de Stata.
Texto (*.txt, *.dat. *.csv). Abre archivos del tipo texto: delimitado por tabulaciones (*.txt), delimitado por comas (*.csv).

Todos los archivos (*.*). Nos presenta todos los tipos de archivos que se encuentran en la dirección señalada.
Importación de datos.-Lo más recomendable **no es colocar** los datos directamente en el **Editor de datos** (Vista de datos del SPSS), sino más bien emplear una **hoja de cálculo** para introducir los datos y posteriormente **importarlos al Editor de datos del SPSS**. Para la importación de datos se debe tener en cuenta que: las dimensiones de la base de datos en SPSS son el número de **filas x el número de columnas**. **No deben existir celdas vacías** dentro de la matriz de datos de filas x columnas**, todas las celdas deben de tener un valor incluso si están en blanco**. Se aplican las siguientes reglas:
-Comenzar colocando los datos desde la primera celda de la hoja de cálculo (en el **Excel** es la **celda A1**).
-Los valores de la primera fila del archivo serán leídos como **nombres de las variables.**
-El número de variables **lo determina la última columna c**on al menos una celda no en blanco del archivo y lo mismo para el número de filas.
-El tipo de datos y el ancho de columna se determinan automáticamente dependiendo si se trata de variables numéricas o cadena.
-**Las celdas en blanco de la matriz** si corresponden a variables numéricas no tratadas como ***missing*** **(perdidos)**, si corresponden a variables categóricas son consideradas como una categoría más (esto generalmente se presenta cuando en la hoja de cálculo, borramos los datos, lo recomendable es que si no necesitamos cierta variable (columna) o datos (fila) es mejor eliminar y no borrar).

Suponga que tiene un cuestionario capturado en una Tabla de Excel y que la requiere importar a **SPSS**. **Ver Figura 1.16.**

Figura 1.16. Cuestionario CKM_MKT_Digital

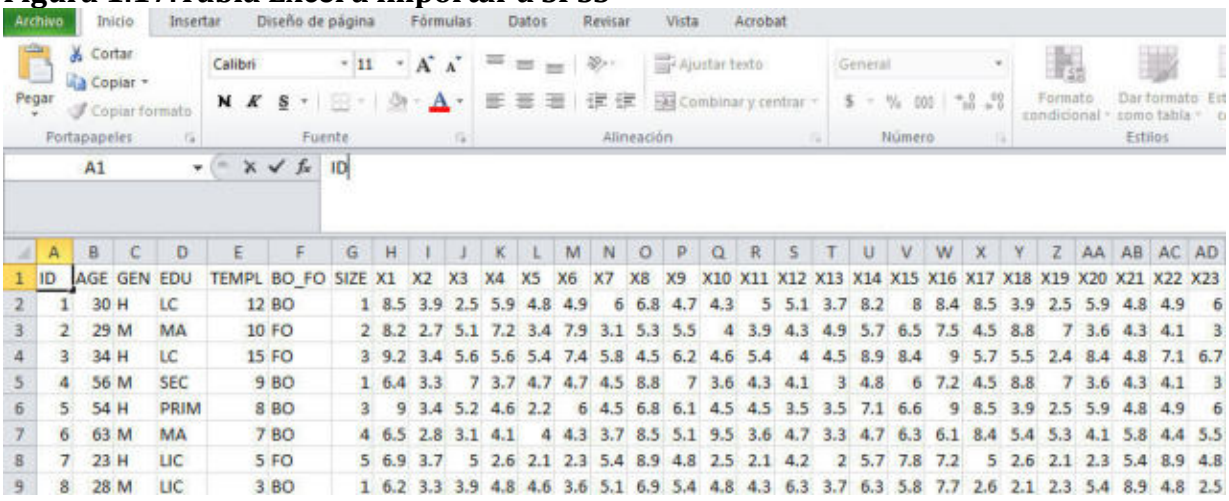
VARIABLE	INDICATOR	AUTHOR
X_1	Customer is a Resource of NPD ideation; Customer Driven-Innovation (Innovation from Customers). Mutual Innovation.	Nambisan (2002); Desouza (et al., 2007); Gibbert y Probst,2002
X	Strategy of close collaboration with customers. Communities of creation.	Nambisan (2002); Gibbert y Probst,2002)
X_3	Customer as a User collaborates intensively in the product testing and support. Customer Focused Innovation (Innovation for Customers)	Nambisan (2002); Desouza (et al., 2007)
X_4	Customer as a Co-creator helps over NPD design and development; Customer Centered Innovation (Innovation with Customers); Prosumerism; Team-Based-CoLearning. Joint Intellectual Property	Nicolai (et al., 2011); Desouza (et al., 2007); Gibbert y Probst,2002
X_5	The firm is warned about the dependence on customer´s personality	Kausch (et al. 2014)
X_6	The firm is warned about the dependence on customer´s experience	
X_7	The firm is warned about the dependence on customer´s point of view	
X_8	The firm is warned about to choose the wrong customer	
X_9	The firm is warned about the risk to integrate the	

	customer to the company´s side	
X_{10}	Tolerance of Failure	Gloet y Samson (2013)
X_{11}	Rewards and Recognition	
X_{12}	Exchange the knowledge between employees across departments	Nicolai (et al., 2011); OECD (2003)
X_{13}	Communication among employees and management	
X_{14}	Internal Sources of Knowledge are more form Back Office Departments	Baker y Hart (2007); Garcia-Murillo y Annabi (2002)
X_{15}	Internal Sources of Knowledge are more from Front Office Departments	
X_{16}	-Internal Sources of Knowledge are more from Front Office Departments	
X_{17}	Supplier	Baker y Hart (2007); Garcia-Murillo y Annabi (2002)
X_{18}	Scientist, Universities, Patents, Exhibitions Technological Consultant	
X_{19}	Competitor	
X_{20}	More Service systems produce a better Customer Retention	Garcia-Murillo y Annabi (2002)
X_{21}	More CRM systems produce a better Customer Satisfaction	
X_{22}	More CKM systems produce more New Customers)	
X_{23}	You have KPI of CKM	

Fuente: propia

La **Figura 1.17**. Representa parcialmente los datos para análisis, capturados del cuestionario **CKM_MKT_Digital.sav** mostrado anteriormente, codificados y en escala 1-5 de Likert

Figura 1.17.Tabla Excel a importar a SPSS



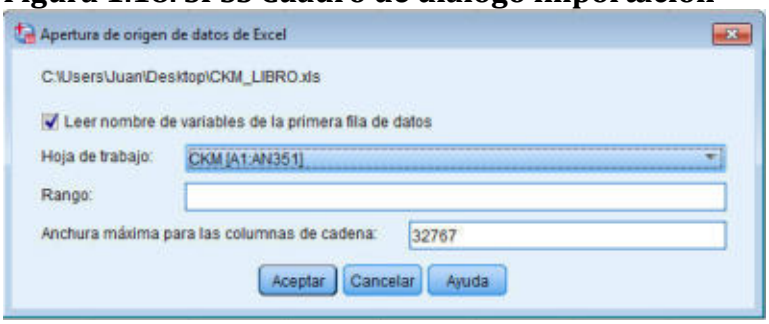
ID	AGE	GEN	EDU	TEMPL	BO_FO	SIZE	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23
1	30	H	LC	12	BO	1	8.5	3.9	2.5	5.9	4.8	4.9	6	6.8	4.7	4.3	5	5.1	3.7	8.2	8	8.4	8.5	3.9	2.5	5.9	4.8	4.9	6
2	29	M	MA	10	FO	2	8.2	2.7	5.1	7.2	3.4	7.9	3.1	5.3	5.5	4	3.9	4.3	4.9	5.7	6.5	7.5	4.5	8.8	7	3.6	4.3	4.1	3
3	34	H	LC	15	FO	3	9.2	3.4	5.6	5.6	5.4	7.4	5.8	4.5	6.2	4.6	5.4	4	4.5	8.9	8.4	9	5.7	5.5	2.4	8.4	4.8	7.1	6.7
4	56	M	SEC	9	BO	1	6.4	3.3	7	3.7	4.7	4.7	4.5	8.8	7	3.6	4.3	4.1	3	4.8	6	7.2	4.5	8.8	7	3.6	4.3	4.1	3
5	54	H	PRIM	8	BO	3	9	3.4	5.2	4.6	2.2	6	4.5	6.8	6.1	4.5	4.5	3.5	3.5	7.1	6.6	9	8.5	3.9	2.5	5.9	4.8	4.9	6
6	63	M	MA	7	BO	4	6.5	2.8	3.1	4.1	4	4.3	3.7	8.5	5.1	9.5	3.6	4.7	3.3	4.7	6.3	6.1	8.4	5.4	5.3	4.1	5.8	4.4	5.5
7	23	H	LIC	5	FO	5	6.9	3.7	5	2.6	2.1	2.3	5.4	8.9	4.8	2.5	2.1	4.2	2	5.7	7.8	7.2	5	2.6	2.1	2.3	5.4	8.9	4.8
8	28	M	LIC	3	BO	1	6.2	3.3	3.9	4.8	4.6	3.6	5.1	6.9	5.4	4.8	4.3	6.3	3.7	6.3	5.8	7.7	2.6	2.1	2.3	5.4	8.9	4.8	2.5

Fuente: Excel 10 Microsoft

Posteriormente procedemos a abrir los datos desde el SPSS (en formato Excel, este archivo deberá estar previamente cerrado), así **teclee: Archivo->Abrir-> Datos…**
En el **cuadro de dialogo de Abrir datos:**
-Buscar en: Disco Local (C:)

-Nombre de archivo: ejemplo
-Archivos tipo: Excel (*.xls, *.xlsx, *.xlsm)
-Abrir
En el cuadro de dialogo de Apertura de origen de datos de Excel (**Figura 1.18**)

Figura 1.18. SPSS Cuadro de diálogo importación



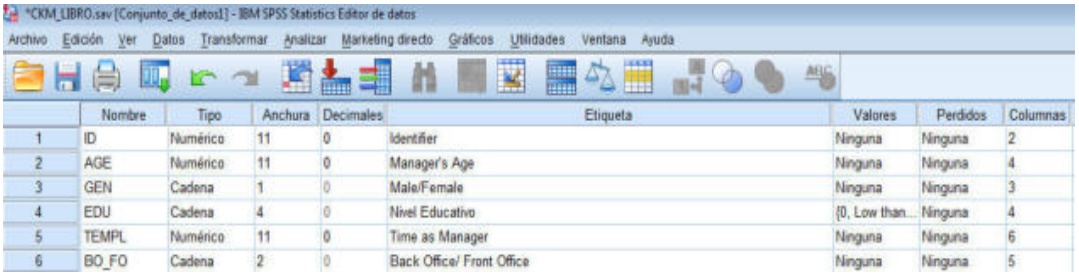
Fuente: SPSS 20 IBM

En la Ventana de Apertura de origen de datos de Excel
-Leer nombre de variables de la primera fila de datos: Leerá la primera fila como nombre de las variables (tienen que estar marcadas, siempre que la primera fila corresponda al nombre de las variables).
- Hoja de trabajo: Seleccionamos la hoja donde esta nuestros datos (nos llega a mostrar las hojas que tienen datos).
-Aceptar
Apreciaremos los datos en el Editor de datos del SPSS, la **Vista de datos** y **Vista de variables (que se deberá actualizar según sea el caso)**. Ver **Figura 1.19.**

Figura 1.19. SPSS Vista de datos

	ID	AGE	GEN	EDU	TEMPL	BO_FO	SIZE	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23
1	1	30	H	LC	12	BO	1	8.5	3.9	2.5	5.9	4.8	4.9	6.0	6.8	4.7	4.3	5.0	5.1	3.7	8.2	8.0	8.4	8.5	3.9	2.5	5.9	4.8	4.9	6.0
2	2	29	M	MA	10	FO	2	8.2	2.7	5.1	7.2	3.4	7.9	3.1	5.3	5.5	4.0	3.9	4.3	4.9	5.7	6.5	7.5	4.5	8.8	7.0	3.6	4.3	4.1	3.0
3	3	34	H	LC	15	FO	3	9.2	3.4	5.6	5.6	5.4	7.4	5.8	4.5	6.2	4.6	5.4	4.0	4.5	8.9	8.4	9.0	5.7	5.5	2.4	8.4	4.8	7.1	6.7
4	4	56	M	SEC	9	BO	1	6.4	3.3	7.0	3.7	4.7	4.7	4.5	8.8	7.0	3.6	4.3	4.1	3.0	4.8	6.0	7.2	4.5	8.8	7.0	3.6	4.3	4.1	3.0
5	5	54	H	PRIM	8	BO	3	9.0	3.4	5.2	4.6	2.2	6.0	4.5	6.8	6.1	4.5	4.5	3.5	3.5	7.1	6.6	9.0	8.5	3.9	2.5	5.9	4.8	4.9	6.0
6	6	63	M	MA	7	BO	4	6.5	2.8	3.1	4.1	4.0	4.3	3.7	8.5	5.1	9.5	3.6	4.7	3.3	4.7	6.3	6.1	8.4	5.4	5.3	4.1	5.8	4.4	5.5

Vista de variables



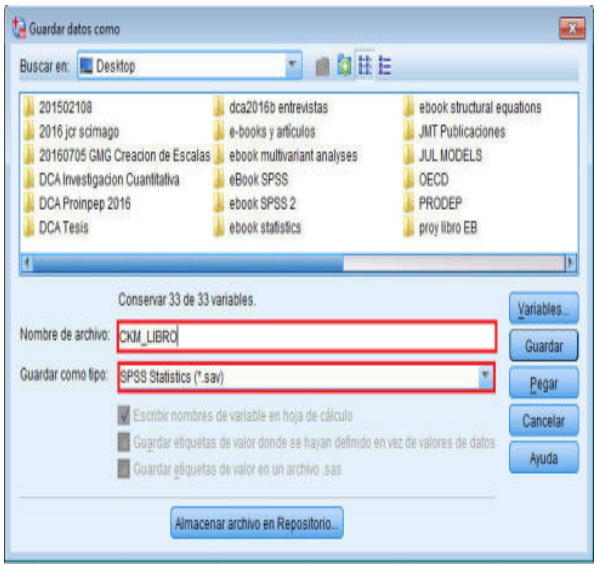
	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas
1	ID	Numérico	11	0	Identifier	Ninguna	Ninguna	2
2	AGE	Numérico	11	0	Manager's Age	Ninguna	Ninguna	4
3	GEN	Cadena	1	0	Male/Female	Ninguna	Ninguna	3
4	EDU	Cadena	4	0	Nivel Educativo	{0, Low than...	Ninguna	4
5	TEMPL	Numérico	11	0	Time as Manager	Ninguna	Ninguna	6
6	BO_FO	Cadena	2	0	Back Office/ Front Office	Ninguna	Ninguna	5

Fuente: SPSS 20 IBM

Para guardar los cambios realizados, **Teclear: Archivo ->Guardar**
Especificar el nombre, será guardado en formato SPSS (***.sav**)
O teclear: Archivo->Guardar como… y especificar el nombre y formato.
Ver Figura 1.20.

Nota: Formatos disponibles:
-SPSS Statistics (*.sav)
-SPSS 7.0 (*.sav)
-SPSS/PC+ (*.sys)
-ASCII en formato fijo (*.dat)
-Excel 2.1 (*.xls)
-Excel 97 a 2003 (*.xls)
-Excel 2007 a 2010 (*.xlsx)
-dBASE IV (*.dbf)
-dBASE III (*.dbf)
-SAS v6 para Windows (*.sd2)
-SAS v6 para UNIX (*.ssd01)
-SAS v6 para Alpha/OSF (*.ssd04)
-Versión 9+ de SAS para Windows (*.sas7bdat)
-Versión 9+ de SAS para UNIX (*.sas7bdat)
-Transporte de SAS (*.xpt)
-Stata versión 8 Intercooled (*.dta)
-Stata versión 8 SE (*.dta)

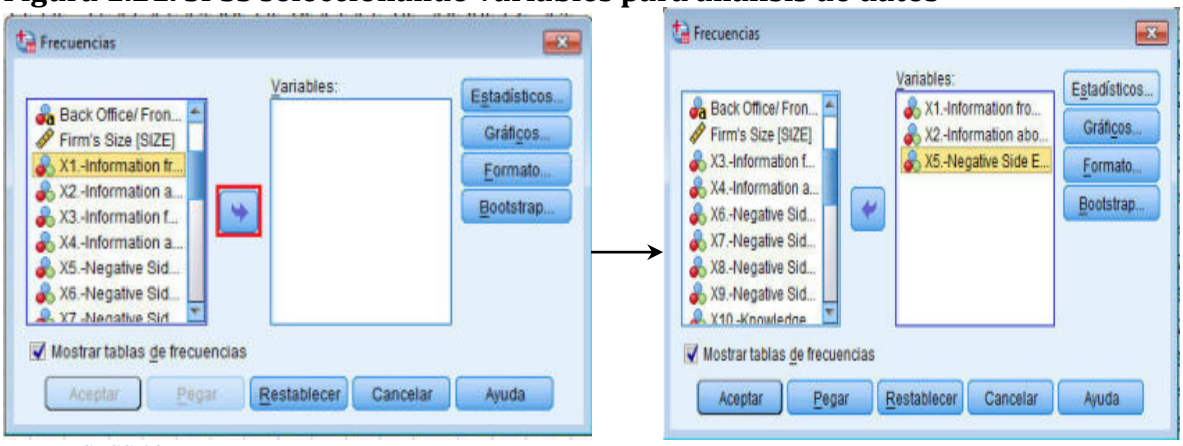
Figura 1.20. SPSS Guardando archivo importado



Fuente: SPSS 20 IBM

Trabajando con el SPSS. El SPSS tiene una forma de trabajo intuitiva. Por ej. Teclee **Analizar->Estadísticos Descriptivos** y escoja las variables a analizar. **Figura 1.21.**

Figura 1.21. SPSS seleccionando variables para análisis de datos



Fuente: SPSS 20 IBM

Una vez enviada la variable seleccionada al lado derecho el botón de selección invierte su dirección señalando al lado izquierdo , si se desea que la variable sea retirada de la selección.

Guardar Resultados como archivo SPSS. Esta forma de almacenamiento de los Resultados, se da en formato SPSS el cual tendrá la extensión ***.spv**, la ventaja de guardar de esta forma es que se puede editar tanto los cuadros de resultados como los gráficos, estos últimos solo se pueden editar si se los guarda de esta forma.

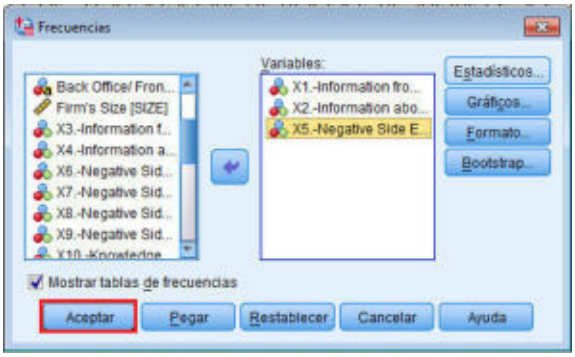
Teclear: Archivo -> Guardar

-Especificar la dirección donde será guardado.
-Nombre de archivo: (colocar el nombre)
-Guardar como tipo: Archivo del visor (*.spv)
-Guardar
-Especificar la dirección donde será guardado.

O teclee Archivo ->Guardar como…

-Nombre de archivo: (colocar el nombre)
-Guardar como tipo: Archivo del visor (*.spv)
-Guardar. Ver **Figura 1.22.**

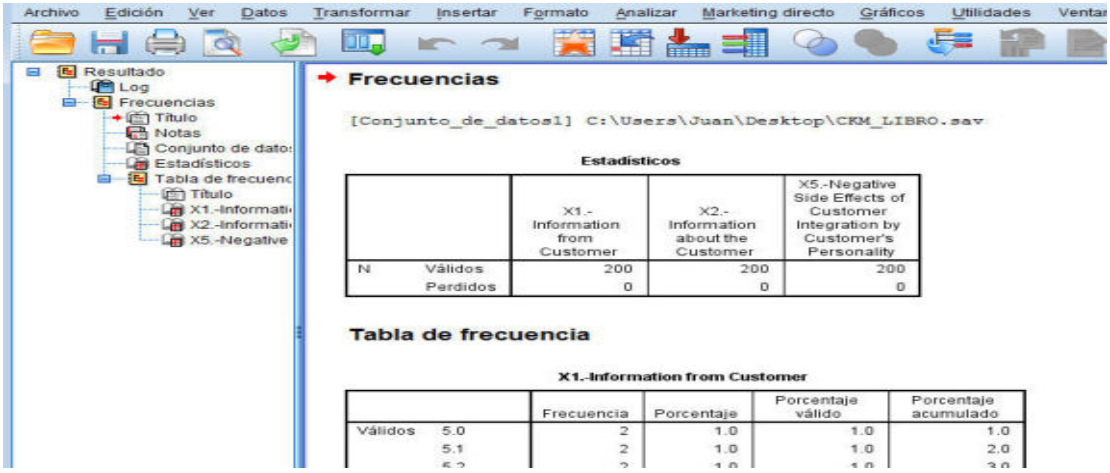
Figura 1.22. SPSS aceptar análisis de datos



Fuente: SPSS 20 IBM

Lo que arroja los resultados, de la **Figura 1.23.**

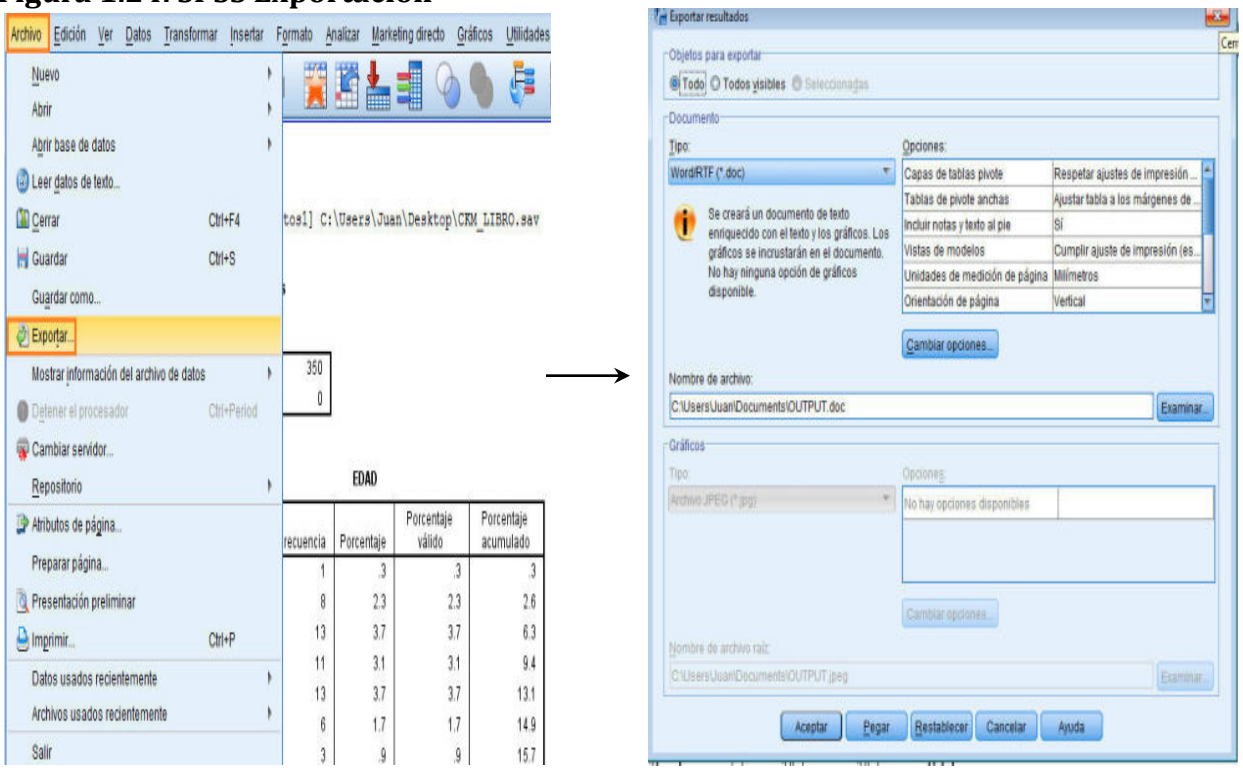
Figura 1.23. SPSS resultados análisis de datos



Fuente: SPSS 20 IBM

Exportación.-Esta opción nos permite **Exportar los resultados** a diferentes formatos, presentándonos la siguiente ventana de exportación de resultados.
Así, **Teclear: Archivo -> Exportar...En la ventana de Especificar la dirección donde será guardado.**
-Objeto a exportar: Todos;
-Tipo: Word/RTF (*.doc);
-Nombre de Archivo, Examinar.
En la ventana de Guardar archivo, seleccionamos la dirección donde se guardara.
-Nombre de archivo: (colocamos el nombre); **->Guardar; - Aceptar**. Ver **Figura 1.24.**

Figura 1.24. SPSS Exportación



Fuente: SPSS 20 IBM

Nos permite guardar en diferentes formatos, como:
-Excel. En diferentes versiones.
-HTML (*.htm). En formato de página web, el cual se puede abrir no solo como página web, sino también se puede abrir con Word y Excel.
-Informe web (*.htm o *.mht). Estos dos formatos ya se mencionaron anteriormente.
-Formato de documento portátil (*.pdf)
-PowerPoint (*.ppt). En formato de presentación PowerPoint.
-Texto – Sin formato (*.txt). Guarda el archivo como texto sin formato y el grafico lo guarda por separado.
-Word/RTF (*.doc). La salida de los resultados se almacena en el procesador de texto de Word.
-Ninguno (solo gráficos). Guardara solo los gráficos y no los cuadros de resultados.

1.12. SPSS. Transformación de datos

El SPSS nos permite crear, transformar o agrupar las variables, en el caso de la agrupación nos referimos a generar grupos dentro de los valores de las variables, estos grupos formados pueden ser recodificadas en una nueva o la misma variables en función a las ya existentes, con la previa carga de una base de datos (en nuestro ejemplo: **CKM_MKT_Digital.sav.**) (IBM, 2011b).

Transformación de datos. Esta opción nos permite transformar una variable, de acuerdo se requiera. Por ejemplo, en la Normalización de una variable, se tiene a partir de **Logaritmo Neperiano.**

Así, se deberá **teclear: Transformar->Calcular Variable:**

-Ingresar Variable de destino: por ej. ***LNX***

-Asignar Expresión numérica

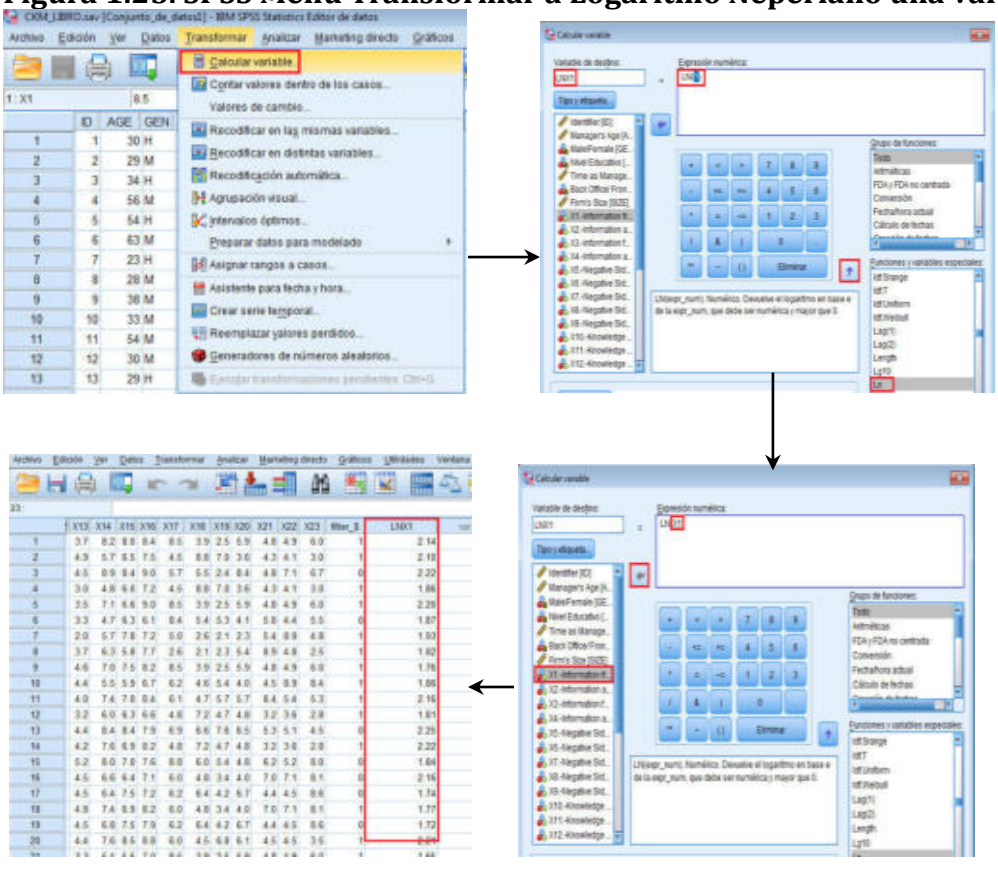
-Buscar en Grupo de funciones: Todo

-En Funciones y variables especiales, escoger **Logaritmo Neperiano**; con la flecha ubicada al lado izquierdo del cuadro de Funciones y variables especiales

-Marcar para subir la función; en el paréntesis de la función, señalar e ingresar la variable de la base de datos, con la flecha ubicada al lado derecho del cuadro donde se aprecian las variables de la base de datos.

-Oprimir la tecla aceptar, tras la cual, se podrá verificar el ingreso de una nueva columna en la base de datos **CKM_MKT_Digital.sav** con la nueva variable $\boldsymbol{LNX_1}$ normalizada con **Logaritmo Neperiano**. Ver **Figura 1.25.**

Figura 1.25. SPSS Menú Transformar a Logaritmo Neperiano una variable



Fuente: SPSS 20 IBM

Referencias

IBM (2011a). *IBM SPSS Statistics Base 20.* EUA. .Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Base.pdf

IBM (2011b).*Guía breve de IBM SPSS Statistics 20.* EUA.Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Brief_Guide.pdf

IBM (2011c). IBM SPSS Missing Values 20. EUA. .Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Missing_Values.pdf

Capítulo 2. Técnicas Multivariantes

2.1. Análisis Multivariante. Antecedentes

- Desde fines del siglo XX a la fecha, se ha percibido una gran necesidad por conocer y practicar las técnicas estadísticas multivariantes en todos los campos de la investigación científica, particularmente en las Ciencias de la Administración. Existen diversas razones, siendo las más importantes:
- Dado el requerimiento de estudiar lo complejo de la realidad en la mayoría de las investigaciones científicas, se tiene la necesidad de analizar **relaciones simultáneas** entre tres o más variables que describen a dichos fenómenos. ***"A menos que el problema sea tratado como un problema multivariante, está tratado superficialmente"*** (Hair, 1999)
- El desarrollo del poder de cómputo con mayor capacidad de procesamiento y almacenamiento de datos es cada vez mayor, acompañados de programas informáticos cada vez más fáciles de utilizar por cualquier tipo de persona
- **De forma general** se refiere a **todos los métodos estadísticos que analizan simultáneamente medidas múltiples de cada individuo u objeto sometido a investigación.** El análisis multivariante puede considerarse como un **análisis simultáneo de más de dos variables.**
- **De forma estricta**, muchas técnicas multivariantes son extensiones del **análisis univariante** (análisis de distribuciones de una sola variable) y del **análisis bivariante** (correlaciones que incluyen varias variables predictor, clasificaciones cruzadas, análisis de la varianza y regresiones simples. Por ejemplo, la variable dependiente que se encuentra en el análisis de la varianza se extiende para incluir múltiples variables dependientes en el análisis multivariante de la varianza. Otras técnicas, por cierto, sí están diseñadas exclusivamente para tratar con problemas multivariantes, por ejemplo,

el análisis factorial que sirve para identificar la estructura subyacente de un conjunto de variables o el análisis discriminante que sirve para diferenciar entre grupos basados en un conjunto de variables.

Una de las razones de la dificultad por **definirlo es que el término multivariante no se usa de la misma forma en la literatur**a. Para saber más, consulte: IBM, 2011a; IBM, 2011b; IBM, 2011c.
Para algunos investigadores, **multivariante** significa **simplemente examinar relaciones entre más de dos variables.**
Otros usan el término sólo para problemas en los que se supone que todas las variables múltiples tienen una **distribución normal multivariante**.
Sin embargo, para ser considerado verdaderamente multivariante**:**
"*Todas las variables deben ser aleatorias y estar interrelacionadas de tal forma que sus diferentes efectos no puedan ser interpretados separadamente con algún sentido*"
El propósito del análisis multivariante es **medir, explicar y predecir** el grado de relación de los "***valores teóricos***" (combinaciones ponderadas de variables). Así, el **carácter multivariante** reside en los **múltiples valores teóricos (combinaciones múltiples de variables)** y no sólo en el número de variables u observaciones.
La estadística **univariante** y **bivariante**, son la base del análisis multivariante. Para comprenderlo, se debe entender conceptualmente el elemento básico del análisis multivariante (**valor teórico**), los **tipos de escalas de medida** utilizadas, los **resultados estadísticos de los test de significación** y los intervalos de confianza. Cada concepto juega un papel importante en la correcta aplicación de cualquier técnica multivariante.
Valor teórico, es una **combinación lineal de variables con ponderaciones determinadas empíricamente.** Así, el investigador deberá especificar las variables, mientras que las **ponderaciones** son objeto específico de determinación por parte de la **técnica multivariante**. Un valor teórico de n variables ponderadas ($\mathbf{X_1}$ **a** $\mathbf{X_n}$) puede expresarse: **Valor Teórico=**$w_1X_1+w_2X_2+w_3X_3+...w_nX_n$

Nota: las **X** son las variables observadas y $\mathbf{w_n}$ es la ponderación determinada por la técnica multivariante.
El resultado es un valor único que representa una combinación de todo el conjunto de variables que mejor se adaptan al objeto del análisis multivariante específico**.** Ver **Figura 2.1.**

Figura 2.1.-El valor teórico de acuerdo a la técnica multivariante

Técnica Multivariante	Valor Teórico
Regresión múltiple	El valor teórico se determina de tal forma que guarde la mejor correlación con la variable que se está prediciendo
Análisis discriminante	El valor teórico se forma de tal manera que produzca resultados para cada observación que diferencien de forma máxima entre grupos de observaciones.
Análisis factorial	Los valores teóricos se forman para representar mejor las estructuras subyacentes o la dimensionalidad de las variables tal y como se representan en sus intercorrelaciones.

Fuente: propia

Se debe entender no sólo su impacto conjunto para lograr cumplir el objetivo de cada técnica, sino también **la contribución de cada variable** separada al efecto del valor teórico en su conjunto.

2.2. Tipos de escala de medida

El análisis de datos involucra la separación, identificación y medida de la variación en un conjunto de variables, tanto entre ellas mismas como entre una variable dependiente y una o más variables independientes. La palabra **medida**, es la clave para entender que el investigador no es capaz de separar o identificar una variación a menos que pueda **ser medible.** Así, la medida mes vital para representar el concepto buscado y crucial en la selección del método multivariante adecuado.

Los tipos básicos de datos:

-No métricos (cualitativos).-Son atributos, características o propiedades categóricas que identifican o describen a un sujeto y llegan a ser únicas. Por ej. , si uno es hombre, no puede ser mujer. No hay cantidad de género.

-Métricos (cuantitativos).- Los sujetos pueden ser identificados por diferencias entre grado o cantidad. Por ej. Nivel de calidad, o satisfacción.

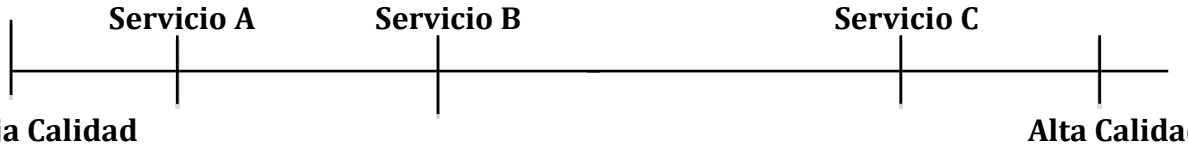
2.2.1. Tipo de escalas de medida no métricas

Son de tipo:

- **Escala nominal.-**Este tipo asigna números que se usan para etiquetar o identificar sujetos u objetos; se les conoce también como **escalas de categoría,** ya que proporcionan el número de ocurrencias en cada clase o categoría de la variable en estudio. **Así, letras, números o símbolos asignados a los objetos no tienen más significado cuantitativo que indicar la presencia o ausencia del atributo bajo investigación.** Por ejemplo: género, religión, tipo de sangre, partido político de afiliación de una persona. Para manipular los datos, un analista asigna números o símbolos a cada categoría, por ejemplo: 1 religión católica, 2 religión protestante, 3 religión judía o A mujeres, b hombres, Los números, letras, símbolos sólo representan a las categorías o clases y no implican cantidades de un atributo o características.
- **Escala ordinal.-**Son el siguiente nivel de precisión de medida que las nominales. Así, **las variables son susceptibles de ser ordenadas o clasificadas en relación a la cantidad del atributo poseído.** Cada subclase puede ser comparada con otra en

términos de una relación **mayor qué o menor qué.** Por ejemplo diferentes niveles de calidad percibidos por el consumidor con diferentes servicios. Ver **Figura 2.2**

Figura 2.2. Ejemplo de escala ordinal



Nota. Los números utilizados en este tipo de escala, No son Cuantitativos, sólo indican posición relativa ya que no hay una medida de cuánta calidad recibe el consumidor en términos absolutos; es más, el investigador no conoce la distancia entre los puntos de la escala de calidad. Muchas de las escalas de las ciencias del comportamiento caen dentro de este tipo de escala.

2.2.2. Tipo de escalas de medida métricas

Son de tipo:

- **Escala de intervalos (tiene cero arbitrario).-**Representa el siguiente nivel de precisión de medida, y que ya permite realizar operaciones matemáticas. Tiene unidades constantes de medida, de forma que las diferencias entre puntos adyacentes de cualquier parte de la escala son iguales. Esta escala se dice que tiene **cero arbitrario** que **no significa que exista una cantidad cero o ausencia de la medida ya que incluso hay medidas debajo de cero.** Por ejemplo, los grados Celsius o Fahrenheit de temperatura. Así también, cualquier punto situado en la escala, sea un múltiplo de otro situado en la misma escala. Por ejemplo, si se tiene un día soleado de 26 grados Celsius, no implica que exista el doble de calor de 13 grados Celsius.
- **Escala de razón (tiene cero absoluto).-**Representa el nivel máximo de precisión de medida, representando todas las ventajas de las escalas anteriores. Tiene unidades constantes de medida, de forma que las diferencias entre puntos adyacentes de cualquier parte de la escala son iguales. Esta escala se dice que tiene **cero absoluto** que **significa que existe una cantidad cero o ausencia de la medida.** Por ejemplo, las medidas de peso, de velocidad, ingreso, etc. Así también, cualquier punto situado en la escala, es un múltiplo de otro situado en la misma escala. Por ejemplo, si tiene un ingreso de 50,000 usd. mensuales registrado para los managers, significa que es el doble de los 25,000 usd ganados que se han registrado de los supervisores.

2.3. Por qué entender las escalas de medida

Esto es porque el investigador está obligado a:

- Identificar la escala de medida de cada variable empleada de tal manera de no utilizar datos métricos como no métricos y viceversa
- La escala de medida es crucial para determinar la técnica multivariante, es decir, las propiedades métricas o no métricas de las variables dependientes y/o independientes son los determinantes en la selección de la técnica multivariantes.

2.4. Error de medida y medidas multivariantes

El uso de múltiples variables así como la dependencia de su combinación (el **valor teórico**), en las técnicas multivariantes, dirige la atención al **error de medida,** la cual es el grado en que los valores observados **no son representativos de los valores "verdaderos".** Fuentes posibles y muy probables, son:

-Errores en la entrada de datos

-Imprecisión en la medición (por ejemplo, escalas de puntuación de siete puntos a la actitud medida cuando el investigador sabe que los encuestados sólo pueden responder con precisión a una puntuación de tres puntos)

-Incapacidad de los encuestados a proporcionar información precisa (por ejemplo, las respuestas a la renta de una economía familiar pueden ser razonablemente precisas pero rara vez lo son completamente).

Así, se debe asumir que **todas las variables usadas en las técnicas multivariantes tienen algún grado de error de** medida, y su impacto es traducirlo como añadir **"desviaciones"** a las variables medidas u observadas. Por tanto, el valor observado obtenido representa tanto el nivel **"*verdadero*"** como su **"*desviación*".** Cuando se calculan correlaciones o medias, normalmente el efecto **"*verdadero*"** está parcialmente camuflado por el **error de medida,** causando bajas correlaciones y pérdida de precisión de las medias. El impacto específico del error de medida en las relaciones de dependencia se trata con más detalle en las **Ecuaciones Estructurales**

El objetivo del investigador **es reducir el error de medida,** a partir de la **validez** y **fiabilidad** de la medida.

- **La validez** es el grado en que **la medida representa** con precisión lo que se supone que representa a partir de un conocimiento profundo de lo que se va a medir y sólo entonces realizar la medida tan "***correcta***" y precisa como sea posible. Sin embargo, **la precisión no asegura la validez.** Por ejemplo, la validez de constructo.
- La **fiabilidad** o **grado en que la variable observada mide el valor "*verdadero*"** y está **"*libre de error*"**; **es lo opuesto al error de medida**. Si la misma medida se realiza repetidas veces, por ejemplo, las medidas más fiables mostrarán una mayor consistencia que las medidas menos fiables. El investigador deberá valorar siempre las variables que están siendo usadas y si se pueden encontrar medidas alternativas **válidas**, elegir la variable con la **mayor fiabilidad.**

Existe la opción de desarrollar **mediciones multivariantes,** llamadas también **escalas sumadas,** donde diversas variables se unen en una **medida compuesta** para representar un concepto (por ejemplo, una escala de liderazgo de entrada múltiple o puntuaciones sumadas de un servicio). **El objetivo es evitar usar sólo una única variable para representar un concepto**, y en su lugar utilizar varias variables como **indicadores,** representando todos ellos diferentes facetas del concepto para obtener una perspectiva más completa. El uso de **indicadores** múltiples permite al investigador llegar a una especificación más precisa de las respuestas deseadas y no deja la fiabilidad plena a una única respuesta sino en la **respuesta "*media*" o "*típica*"** de un conjunto de respuestas relacionadas**.** Por ejemplo, al medir la calidad de servicio, uno podría preguntar una única

cuestión, **" *¿cuál es su grado de calidad de servicio?",*** y basar el análisis en una única respuesta.
O se podría desarrollar una escala aditiva que combinara varias respuestas de calidad de servicio, en diferentes formatos de respuesta y áreas de interés, que contemple la **calidad de servicio total**. La premisa básica es que las respuestas múltiples **reflejan con mayor precisión la respuesta "*verdadera*" que la respuesta única.** Es de suma utilidad conocer **compilaciones de escalas** que proporcionan una escala **"*lista para ser empleada"* con una fiabilidad demostrada [**Bearden et al. 1993**,** Brunner et al.1993**]**. Las repercusiones del **error de medida** y **baja fiabilidad** no pueden ser observadas directamente ya que se encuentran en las variables observadas. Así, se deberá, trabajar siempre para **aumentar la validez y la fiabilidad**, lo que al final llevará a un **descripción más detallada** de las variables de interés. Aunque los malos resultados no siempre se deben al error de medida, su sola presencia es garantía de **distorsión en las relaciones observadas** y hace menos poderosas las técnicas multivariantes. Reducir el error de medida, aunque implique esfuerzo, tiempo y recursos adicionales, puede mejorar resultados débiles o marginales, así como fortalecer los resultados probados.

2.5. Pruebas estadísticas

Una vez ingresados los datos en **SPSS** y producido la estadística descriptiva, por lo general desearemos analizar los datos para probar ciertas hipótesis. Por ejemplo, al ingresar los datos de un cuestionario sobre la eficiencia en el trabajo. Ahora queremos probar si las personas en los puestos de relaciones con los clientes son más eficientes que las que están en el soporte operativo. Además, es posible que desee examinar la relación entre los niveles de eficiencia y rendimiento de trabajo. Por lo tanto, necesitamos saber qué pruebas estadísticas llevar a cabo. Debemos estar conscientes de qué tipo de prueba estadística queremos realizar en nuestros datos, por ejemplo:

1. Un número de pruebas estadísticas examinará las **diferencias entre las muestras**: es decir, **compararán muestras para inferir si las muestras provienen de la misma población o no** (**por ejemplo, pruebas *t***).
2. **Otras pruebas examinan la asociación entre muestras,** tal como una **correlación** (por ejemplo, **la correlación de *Pearson***) o **pruebas de independencia** (por ejemplo: **prueba de *Chi-cuadrada***).

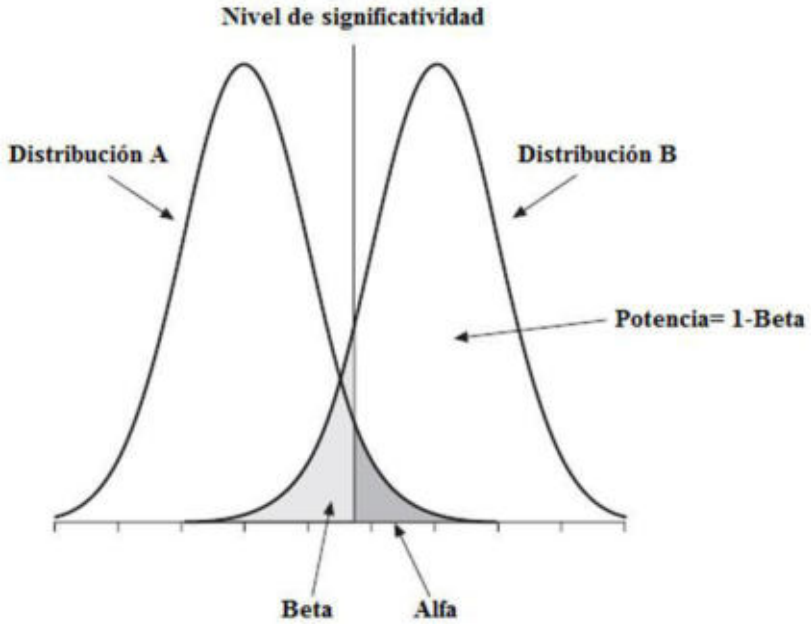
Antes de seleccionar, vale la pena considerar como se construye la significación estadística para dichas pruebas.

2.6. Introducción a las pruebas paramétricas

A continuación se ofrece una breve descripción de la **lógica de las pruebas de significatividad**. El objetivo es sólo introductorio como un recordatorio de las pruebas que se pueden realizar con **SPSS**. Así como existe una serie de diferentes pruebas estadísticas disponibles, también existe una lógica común para las pruebas de significatividad, particularmente **para pruebas que comparan muestras**. Por ejemplo, el caso de la lógica de una **prueba una cola (*one-tailed test*)**. La explicación incluye la razón por la cual se llama así.
Al realizar una prueba de significatividad, calculamos el valor de una estadística particular (*t, F, r,* etc.) basado en nuestras muestras. Para un tamaño de muestra dado, **la**

distribución estadística parte de si las muestras fueron de la misma población, como un primer supuesto**.** Esta es la distribución de la estadística cuando la hipótesis nula ($\boldsymbol{H_0}$) es verdadera. Esto se llama distribución A en la **Figura 2.3.**

Figura 2.3. Prueba estadística de una cola



Fuente: Hair et al. (1999)

2.7. Significatividad estadística y potencia estadística

Cuando calculamos el valor de la estadística para nuestras **muestras tenemos que decidir si pertenece a la Distribución A**, o a una segunda distribución (**Distribución B**). La **Distribución B** corresponde a la estadística de cuando **las muestras provienen de diferentes poblaciones (y la hipótesis nula es falsa)**. Para tomar esta decisión que **elegimos un punto de corte** en la escala (**llamado Nivel de significatividad**). Esto es mostrado por la línea vertical de la **Figura 2.3.**

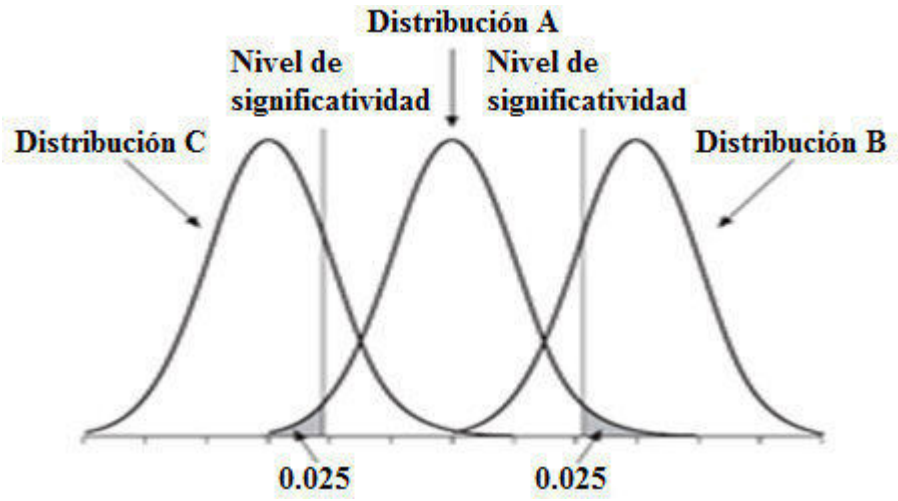
Si el valor calculado del estadístico cae a la izquierda del nivel de significatividad en la Figura 2.3 entonces decimos que NO es posible rechazar la hipótesis nula y afirmar que el valor pertenece a la distribución A. Si el valor de la estadística cae a la derecha del nivel de significatividad en la Figura 2.3, entonces rechazamos la hipótesis nula y afirmamos que el valor pertenece a la Distribución B. Como las distribuciones se superponen, no podemos seleccionar un nivel de significatividad que separe completamente. Esto significa que a veces vamos a afirmar que la estadística pertenece a una distribución cuando realmente pertenece a la otra. Convencionalmente, elegimos al **nivel de significatividad para que el valor de α = 0.05.** Esto coloca el **95% de la Distribución A** la izquierda del nivel de significatividad, con sólo el **5% de la Distribución A**, del lado "***equivocado***" de la línea. Esto configura **el riesgo de cometer un error Tipo I**, **reclamando una diferencia cuando No hay uno, con una**

probabilidad de 0.05 (p = 0.05). Esto nos da un riesgo de sólo 5 en 100 de alegar falsamente un resultado estadísticamente significativo.

Habiendo establecido el nivel de significatividad usando la **Distribución A**, podemos ver que esto deja partes de la **Distribución B** en ambos lados del nivel de significatividad. Esto significa que también podemos cometer un **error de Tipo II, no reclamando ninguna diferencia cuando realmente hay una diferencia en las poblaciones de las que proceden las muestras.** La probabilidad de cometer un **error de Tipo II** se muestra por **β (Beta)** en la **Figura 2.3**. Si el valor de nuestra estadística pertenece a la **Distribución B**, pero se encuentra dentro del área marcado por **β (Beta) reclamaremos** equivocadamente que viene de la **Distribución A** como el valor que está a la izquierda del nivel de significatividad.

El ejemplo anterior se conoce como una **prueba unilateral (o de una cola, *one-tailed*)** cuando tomamos nuestra decisión sobre la significatividad de nuestra estadística en un solo extremo, o cola, de la **Distribución A**. Con una **prueba de dos colas (*two-tailed*)** se argumenta que una diferencia podría surgir cuando el valor de nuestra estadística proviene ya sea de una **Distribución B**, que se superpone al extremo superior de la **Distribución A**, o de una tercera **Distribución C**, que se superpone con el extremo inferior de la **Distribución A**. Esto se muestra en la **Figura 2**.4 donde, tenemos que establecer un nivel de significatividad en ambas colas de **Distribución A.** Para dar un valor global de **α = 0.05**, el tamaño de cada una de las colas de la Distribución A tendría un corte del nivel de significatividad de **0.025.**

Figura 2.4. Prueba estadística de dos-colas



Fuente: Hair et al. (1999)

Si volvemos a la **Figura 2.3** podemos ver que, si toda el área bajo una curva de distribución **se establece en 1**, entonces la cantidad de **Distribución B** a la derecha del nivel de significación es **1 - β (Beta).** Esta es la probabilidad de reclamar una diferencia estadísticamente significativa entre nuestras muestras cuando realmente es uno. Aquí estamos haciendo una afirmación correcta de significación estadística. Esto se conoce como la **potencia de la prueba estadística**. Así, se deduce que **una prueba unilateral (una cola, *one-tailed*) es más poderosa que una prueba bilateral (de dos colas, *two-tailed*)**. Además, podemos ver que una prueba estadística con muy poca potencia es poco probable que detecte la significatividad estadística, incluso cuando las muestras provienen de diferentes poblaciones. Esto es porque β (Beta) será grande y la mayor parte de la **Distribución B** será el lado "***equivocado***" del nivel de significatividad.
El nivel convencional de significatividad se especifica como ***p* = 0.05** (**que establece el riesgo de ser rechazada falsamente la hipótesis nula al 5 por ciento, o 1 en 20**). Así que cuando el valor calculado de una estadística tiene una probabilidad de menos de **0.05** escribimos esto como ***p* <0.05**. Otros niveles significativos son posibles de ser escogidos para indicar que un valor es altamente significativo. Así que a veces empleamos los niveles ***p* = 0.01** o incluso ***p* = 0.001** de significatividad. Cuando un resultado tiene una probabilidad menos que estos valores escribimos el hallazgo como ***p* <0.01** o ***p* <0.001**, según corresponda.

A excepción del **análisis clúster y el análisis multidimensional** todas las técnicas multivariantes, se basan en la **inferencia estadística de los valores de una población** o **la relación entre variables de una muestra escogida aleatoriamente de esa población.** Si estamos realizando **un censo** de toda la población, **entonces la inferencia estadística no es necesaria,** porque cualquier diferencia o relación, por pequeña que sea, **es "*verdad*" y existe**. Pero rara vez, **casi nunca, se realiza un censo**; por tanto, **en un estudio se está obligado a deducir inferencias de una muestra** y para interpretarlas se deben especificar los **niveles aceptables de error estadístico**.
El modo más común es determinar el **nivel de error de Tipo 1**, también conocido como **Alfa** y es **la probabilidad de rechazar correctamente la hipótesis nula cuando es cierta** (**"*positivo falso*"**).
Al especificar un nivel alfa, el investigador fija los márgenes admisibles de error con una probabilidad de concluir que **la significación existe cuando en realidad no existe.**
Por tanto, la potencia es la probabilidad de que la inferencia estadística se indique cuando esté presente. La relación de las diferentes probabilidades de error se muestra a continuación en el hipotético planteamiento de la evaluación de la diferencia entre dos medias. **Ver Figura 2.5**

Figura 2.5. Probabilidades de error de la evaluación caso hipotético: diferencia entre dos medias:

		Realidad	
		Ho=Cierta	Ha: Falsa
Decisión Estadística	Ho= Aceptar	I-Alfa	Beta o Error Tipo II
	Ha= No Aceptar	Alfa o Error Tipo I	1-Beta o Potencia del test de inferencia estadística

Fuente: Hair et al. 1999

Aunque **Alfa** establece el nivel de significación estadística aceptable, es **el nivel de potencia** el que dicta la **probabilidad de "*éxito*"** en la búsqueda de las diferencias, de existir si es que existen.
No se plantean niveles aceptables tanto de alfa como de beta, debido a que **los errores de Tipo I y Tipo II están inversamente relacionados,** esto es que a medida que el **error tipo Alfa** se hace más restrictivo (cercano a cero), el **error tipo B aumenta**. Al disminuir el error de Tipo I también se reduce el poder de la prueba estadística. Por tanto, se debe con seguir un equilibrio entre el nivel de **alfa** y la **potencia resultante;** de hecho, ésta no es sólo una función de **Alfa,** sino que está determinada por tres factores:
1.-Efecto tamaño-La probabilidad de conseguir **significación estadística** se basa no sólo en consideraciones estadísticas sino también en la **magnitud real del efecto que nos interesa** en la población, denominado **efecto tamaño.** Así, **un efecto grande es más probable de encontrar que un efecto pequeño** por lo que se afecta a la **potencia de la prueba estadística**. Por lo tanto, para evaluar la potencia de cualquier prueba estadística,

se debe entender primero el efecto examinado. **Los efectos de tamaño** se miden en términos **estandarizados** para facilitar la comparación. Las **diferencias respecto de la media se determinan en términos de desviaciones estándar**, así que **un efecto tamaño de 0.5** indica que **la diferencia respecto de la media es la mitad de la desviación estándar.** Para las correlaciones, el efecto tamaño se basa en la correlación efectiva entre las variables.

Al determinar el **nivel de error de Tipo l**, también se determina un error asociado, denominado **error de Tipo II o Beta**, que es **la probabilidad de rechazar la hipótesis nula cuando es realmente falsa**.

Una probabilidad más interesante es **1 - Beta**, denominado la **potencia del test de inferencia estadística,** la cual es **la probabilidad de rechazar correctamente la hipótesis nula cuando debe ser rechazada**. Por tanto, la potencia es la probabilidad de que la inferencia estadística se indique cuando esté presente. Ver **Figura 2.6**

Figura 2.6. Nivel de error y significado

Nivel de error	Significado
I o Alfa	Probabilidad de rechazar correctamente la hipótesis nula cuando es cierta (**"*positivo falso*"**).
II o Beta	Probabilidad rechazar incorrectamente la hipótesis nula cuando es realmente falsa
1-Beta	Probabilidad de rechazar correctamente la hipótesis nula cuando debe ser rechazada

Fuente: Hair et al. 1999

2.-Alfa.-A medida que alfa se vuelve más restrictivo, la potencia decrece, lo que significa que como el analista reduce la oportunidad de encontrar un efecto incorrecto significativo, **la probabilidad de encontrar correctamente un efecto también disminuye**.

Las directrices convencionales sugieren **niveles alfa de 0.05 o 0.01.** Pero se deberá considerar el impacto de esta decisión **sobre la potencia** antes de seleccionar el **nivel Alfa.**

3. El tamaño de la muestra. En cualquier nivel de alfa dado, **al aumentar la muestra siempre existirá una mayor potencia del test estadístico**, que genera **"*exceso*"** de potencia, es decir, se observará **que efectos cada vez más y más pequeños serán significativos** hasta que, **para muestras muy grandes** casi cualquier efecto es significativo. Se debe tener siempre presente que **el tamaño de la muestra puede afectar a la prueba estadística tanto por hacerlo insensible (para muestras muy pequeñas) o demasiado sensible (para muestras muy grandes).**

Las relaciones entre **alfa, tamaño de la muestra, efecto tamaño y potencia** son bastante complicadas, pero se pueden encontrar ciertos puntos de partida. (Cohen, 1977) **ha examinado la potencia para la mayor parte de las pruebas de inferencia estadística** y ha proporcionado pautas para los niveles aceptables de potencia, sugiriendo que los estudios deben diseñarse para conseguir **niveles de alfa de al menos 0.05 con niveles de potencia del 80 por ciento.** Para conseguir dichos niveles, **deben considerarse simultáneamente los tres factores**. Para ilustrar mejor lo anterior, tenemos 2 ejemplos:

Ejemplo 1.-la comprobación de la **diferencia entre las puntuaciones medias de dos grupos**. Suponga que el efecto tamaño sea entre **pequeño (0.2)** y **moderado (0.5**);

-Problema 1: el investigador debe **determinar el nivel alfa y el tamaño de muestra necesario de cada grupo**. La **Figura 2.7** muestra el impacto tanto del tamaño de la muestra como del nivel alfa sobre la potencia. Así, **la potencia llega a ser aceptable para tamaños de muestra de 100 o más en situaciones con un efecto tamaño moderado para ambos niveles de alfa.** Pero cuando ocurre un **efecto tamaño pequeño**, las pruebas estadísticas tienen poca potencia, incluso con niveles de alfa expandidos a muestras de **200 o más**. Por ejemplo, **una muestra de 200 en cada grupo con un alfa de 0.05 todavía tiene un 50 por ciento de posibilidades de encontrarse diferencias significativas si el efecto tamaño es pequeño**. Esto sugiere se debe anticipar que los efectos van a ser pequeños, debe diseñar el estudio con muestras mucho mayores y/o niveles de alfa menos restrictivos (**0.05 o 0.10**).

Figura 2.7. Niveles de potencia entre dos medias comparadas:
Variaciones por el tamaño de la muestra, el nivel de significación y el efecto tamaño

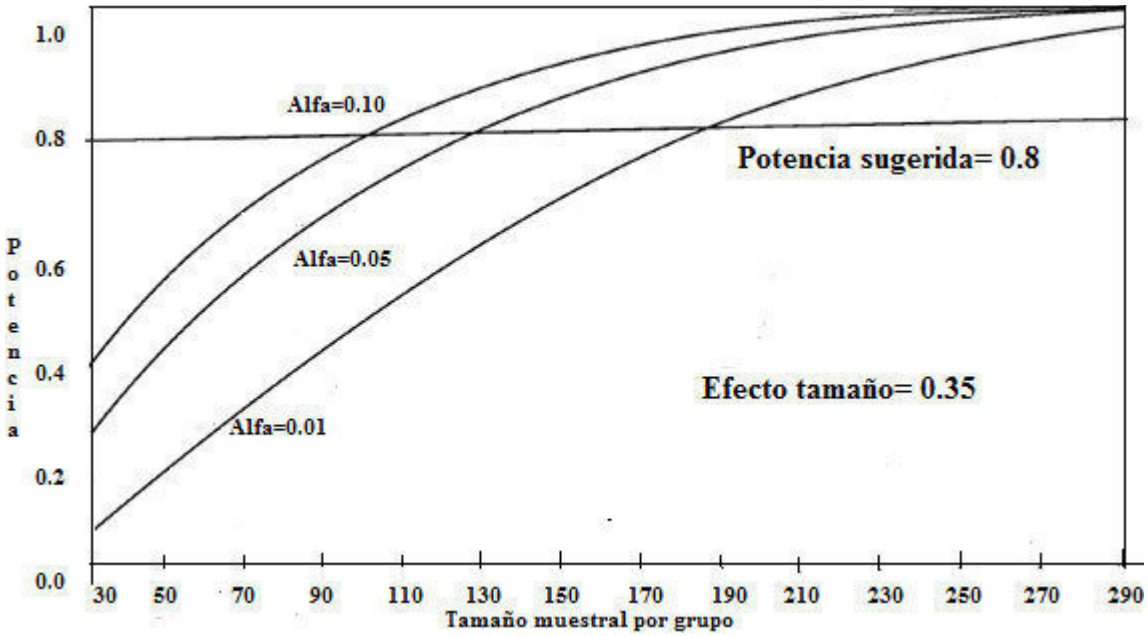
	Efecto Tamaño con Alfa= 0.05		**Efecto Tamaño con Alfa=0.01**	
Tamaño de muestra	Pequeño (0.2)	Moderado (0.5)	Pequeño (0.2)	Moderado (0.5)
20	0.095	0.338	0.025	0.144
40	0.143	0.598	0.045	0.349
60	0.192	0.775	0.067	0.549
80	0.242	0.882	0.092	0.709
100	0.290	**0.940**	0.120	**0.823**
150	0.411	0.990	0.201	0.959
200	**0.516**	**0.998**	**0.284**	**0.992**

Fuente Cohen (1977)

Ejemplo 2.-la **Figura 2.8** representa gráficamente la potencia para niveles de significación de 0.01; 0.05 y 0.10 con tamaños de muestra de 20 a 300 por grupo, cuando el efecto tamaño (0.35) se ubica entre pequeño y moderado.

-Problema 2: el investigador debe determinar el nivel alfa y el tamaño de muestra necesario de cada grupo. La especificación de un nivel de significación de un 0.01 requiere una muestra de 200 por grupo para conseguir el **nivel deseado de potencia del 80 por ciento**. Pero si se relaja el nivel **Alfa**, se alcanza la potencia del 80 por ciento para muestras de 130 para un nivel **Alfa** 0.05 y muestras de 100 para un nivel de significación de un 0.10.

Figura 2.8. Impacto del tamaño de muestra en la potencia de algunos niveles alfa con efecto tamaño de 0.35



Fuente: Hair et al. (1999)

Como se observa, se debe planificar la investigación, estimando el efecto tamaño esperado para seleccionar entonces **el tamaño de la muestra y el nivel alfa** y conseguir **el nivel de potencia deseado** que se utilizará también para determinar la **potencia real conseguida**, de tal forma que los resultados puedan ser correctamente interpretados.

¿Qué tanto de los resultados se deben al efecto tamaño, tamaño muestra o niveles de significación? Los analistas pueden evaluar cada uno de estos factores por su impacto sobre la significatividad o no significatividad de los resultados.

Se recomienda referirse a estudios publicados donde se analicen los detalles concretos de la determinación de la potencia **[Cohen, J. 1977]** o acudir a varios programas de computadora personal que incluso **asisten a los estudios de planificación para conseguir la potencia deseada o calcular la potencia de los resultados reales [BMDP Statistical Software 1991, Brent, Edward E., et al 1991]**.

Se sugiere revisar por cierto, las técnicas de **regresión múltiple** y **análisis multivariante de la varianza** donde es posible que pueda discutir con más detalle las aplicaciones más comunes del **análisis de potencia.**

2.8. Requisitos adicionales a considerar

La lógica anterior de las pruebas de significatividad nos obliga a hacer ciertos supuestos acerca de nuestros datos. La razón de esto es que usamos la información de nuestras muestras para estimar los valores de la població n (**denominados parámetros**, ya que por ejemplo, usamos las medias y las desviaciones estándar de las muestras para calcular las medias y desviación estándar de la població n de las que fueron tomadas). Por lo tanto, esta serie de pruebas se denominan **pruebas paramétricas** de las que se debe considerar:

1. **¿Las muestras presentan sesgos?** esencialmente esto significa que la muestra en sí ya representaría adecuadamente a la población. Sin embargo, nosotros argumentamos que la muestra debe ser seleccionada aleatoriamente de la población a fin de evitar el sesgo. No obstante, s la muestra se encuentra sesgada valores que se obtengan no serán buenos estimadores de la población.
2. **¿En qué escala de medida están sus datos?** Las **pruebas paramétricas** requieren datos de **intervalo**, donde los números consecutivos en el escala son a **intervalos iguales** (recuerde que los datos de intervalo medidos en una escala con un **cero genuino** son denominados **datos de razón**. La velocidad es un ejemplo de dicha escala. Las **escalas de temperatura** son intervalo pero **no de razón** ya que el cero en la escala no significa que no exista la temperatura). También asumimos que las escalas **son continuas**: es decir, que **no hay espacios** o **rupturas** en ellos (**no hay datos discretos**). Por ejemplo, el **tiempo, la distancia y la temperatura** son escalas de intervalo. Sin datos de intervalo no podríamos calcular estadísticas significativas como la **media y desviación estándar.**
3. **¿Las puntuaciones de cada muestra se toman de poblaciones normalmente distribuidas?** Algunas de nuestras pruebas estadísticas nos obligan a hacer esta inferencia. La prueba estadística asume que sea el caso y si no, entonces el resultado o puede subestimar o sobrestimar el valor de la estadística. Esto puede comprobarse trazando sus datos en un **histograma o diagrama de caja (*boxplot*)** o por más precisión mediante la realización de una muestra de ***Kolmogorov-Smirnov***, que prueba estadísticamente la normalidad de los datos.
4. **¿Los datos cumplen con el supuesto de homocedasticidad?** Para que las pruebas funcionen apropiadamente, asumimos que cualquier manipulación que realicemos (como el efecto de los estímulos económicos sobre el rendimiento) afecta a cada miembro de la población en la misma medida y, por lo tanto, NO afecta a la distribución general (variación) o la forma de la distribución de las puntuaciones de la población. Por lo tanto, **las varianzas de población deben ser las mismas,** y nuestras muestras deben tener variaciones similares como las usamos para estimar los valores de su población. A veces esto se vuelve bastante complejo para los elementos de medidas repetidas y es posible que desee considerar el uso de **una prueba no paramétrica** si tiene inquietudes sobre su datos, tales como si tiene:

   • **Datos ordinales** donde las puntuaciones proporcionan un orden pero la escala no tiene intervalos iguales, tal como una escala de calificación, o

   • **La escala no es continua**, o

   • **Los datos no se distribuyen normalmente**, o

   • **Los datos que violan el supuesto de la homocedasticidad**

   Un ejemplo de una **prueba no paramétrica** es la **prueba *U* de *Mann-Whitney*, un equivalente no paramétrico de la prueba *t* independiente.**

   Normalmente preferiríamos **realizar una prueba paramétrica ya que este tipo de análisis es más potente y utiliza las puntuaciones reales**; mientras que **las pruebas no paramétricas hacen menos suposiciones s**obre los datos, ya que generalmente realizan a cabo un análisis sobre la clasificación de las que en las puntuaciones más que de los puntajes en sí mimos.

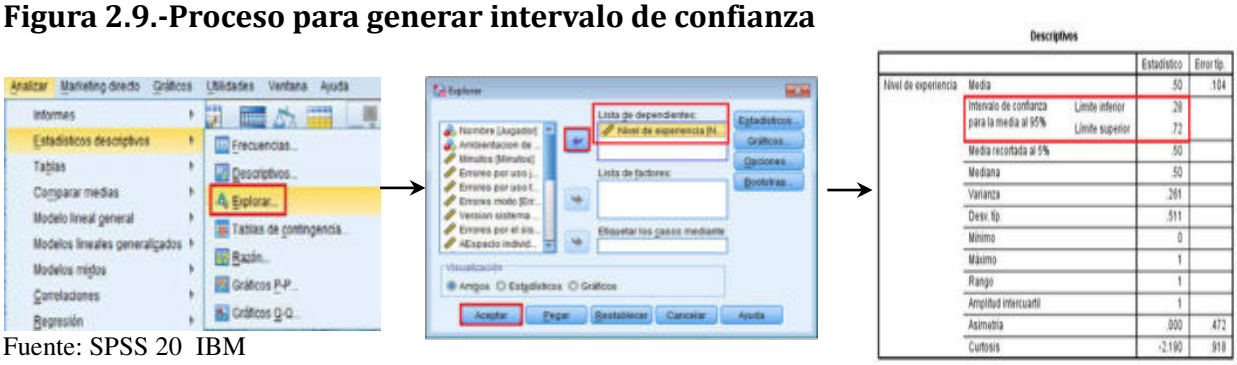
5. **¿Cómo llevar a cabo las pruebas de significatividad?** Aquí se advierte el estar conscientes de las limitaciones ya que prevalece el debate sobre las pruebas de significatividad estadística. Existen argumentos que afirman que las pruebas de significatividad no siempre son la mejor manera de hacer inferencias de los datos. También existe la preocupació n de que la significatividad estadística no es lo más importante para la explicació n de los datos de una investigació n. En este caso, considere dos estudios con datos casi idénticos.
   -Suponga que realiza una **prueba *t* en ambos casos**; en el primer caso, el **valor *t* calculado tiene una probabilidad de 0.049 (de ocurrir cuando la hipótesis nula es verdadera). Aquí decimos que el resultado es estadísticamente significativa ya que la probabilidad es menor que el nivel de significancia de 0.05.**
   -En el segundo estudio tiene una **probabilidad de 0.051**. Aquí decimos que el resultado no es estadísticamente significativo ya que la **probabilidad es mayor que .05.** Aunque los conjuntos de datos son casi idénticos estamos haciendo conclusiones opuestas. Por lo tanto, los reportes de **'significativo' o 'No significativo**' son un problema aquí. Una respuesta es reportar los valores reales de la probabilidad, en lugar de só lo decir que un resultado es o no es estadísticamente significativo. De esta manera, el lector puede comparar el valor de probabilidad al nivel de la significatividad. De hecho, **SPSS** le reporta el valor de probabilidad en lugar del nivel significativo. A pesar de que a menudo encabeza una columna en una tabla con el término "***Nivel sig*.**" presenta la probabilidad real, como p = 0.028, en lugar de reportar p <0.05. **Recomendamos que reporte la probabilidad en términos del nivel de significatividad, por ejemplo, *p* <0.05 o *p*> 0.05**. (Puede haber casos en que los valores estén próximos a la significatividad, por ejemplo p = 0.049, donde desea indicar esto en su informe, por lo que para estos casos particulares, puede que desee incluir el valor de probabilidad real, así como su significado.).
   Nota: en ocasiones SPSS da un valor de probabilidad de ***p* = 0.000**. En realidad, el valor de probabilidad no puede ser cero. Este valor significa un **valor de *p* ha sido redondeado hasta tres decimales.** Como no debemos informar erró neamente un valor ***p* de 0.000 podemos cambiar el último cero a uno y declarar que *p* <0.001.** Alternativamente, puede obtener el valor ***p* real** haciendo ***click*** una vez sobre la tabla, coloque el mouse sobre el **0.000** y luego haga doble ***click*** para obtener el valor de probabilidad. Sin embargo, reportar p <0.001 suele ser suficiente.
6. **¿Cuál es el intervalo de confianza?** Una alternativa para dar la significatividad estadística de un hallazgo es reportar **confianza de intervalo** en cuanto éstos proporcionen una conclusión más apropiada para el análisis. Un **intervalo** de **confianza (IC)** define un rango de valores dentro de los cuales estamos seguros (con un cierta probabilidad) que se encuentra nuestro valor de población, por ejemplo, si tenemos una muestra que puede calcular el intervalo de confianza de la media. Este es un estimado de la media de la población. Con una media de 4.00, nuestro **intervalo de confianza podría ser de 3.50 a 4.50 con *p* = 0.95.** Esto nos indica que, para **95 de cada 100 muestras**, **el intervalo de confianza** contiene la media de la población. Así que podemos usar esto como una estimación de la media de la población. Una forma de reportar este valor es el siguiente: Media = 4.00 (**IC** del 95%: 3.50 a 4.50) o **IC** del 95% de la media = 4.00 ± 0.5.

Recuerde que las **pruebas de significatividad tienen que ver con la estimación de los valores de la población y el intervalo de confianza es una forma diferente de hacer esto**. Es claro por tanto, que, **un valor estrecho de intervalo de confianza nos reporta un mejor estimado de la media de la población que un valor amplio de intervalo de confianza**. En una **prueba de diferencia**, como la **prueba *t***, la diferencia en las **medias muestrales** es un estadístico importante. Así **SPSS** reporta el **intervalo de confianza de la diferencia en las medias de la muestra**, reportando por tanto, una **estimación de la diferencia en valores de la población**. Considere **una prueba *t*** para muestras en par, con una **diferencia en las medias de +3.00**. El **valor *t*** calculado (de **1.588**, **gl** = 19, **p** = 0.129, prueba de dos colas) **no sea significativo al nivel de significación de 0.05**. **Los Intervalo de confianza** de esta **diferencia de medias, +3,00** (**IC del 95%: -0.954 a 6.95**4), indica que la diferencia en los medias para las poblaciones podría ser casi tan alta como **+7** pero podría ser casi tan baja como **-1**, por lo que **NO podemos rechazar la posibilidad de que sea cero**. Por lo tanto, **el intervalo de confianza apoya los resultados de la prueba de significatividad.**

Siempre es una buena idea observar los intervalos de confianza en los reportes de **SPSS,** así como verificar la significatividad estadística de sus resultados. Los reportes de **SPSS** incluirán **intervalos de confianza** como parte de la tabla de resultados.

Nota: puede generar el **intervalo de confianza de una media** de un conjunto de datos y sus descriptivos mediante **teclear: Analizar->Estadísticos descriptivos->Explorar->Lista de variables dependientes (seleccionar la variable métrica de preferencia: Nivel de experiencia). ->Aceptar. Ver Figura 2.9**

Figura 2.9.-Proceso para generar intervalo de confianza



Fuente: SPSS 20 IBM

7. **¿Qué hay de la potencia estadística?** Esta preocupación surge cuando se produce una estadística con **una probabilidad mayor que el nivel de significatividad**, por ejemplo **p> 0.05**. En respuesta a esto, a menudo se informa que '**aceptamos la hipótesis nula**', lo que indica que nuestras muestras provienen de la misma población**. Sin embargo, esto no es verdad**. El fracaso por lograr la **significatividad estadística** puede ser debida a varias razones:

   - Una de ellas puede ser que nuestra prueba esté carente de **potencia estadística,** incluso cuando nuestras muestras provengan de diferentes poblaciones.

   -Otra razón es que se tengan muy pocos participantes en el estudio, lo que reducirá el poder de la prueba.

-Alternativamente, podemos estar buscando un efecto muy pequeño y nuestro análisis es incapaz de detectarlo.

-En nuestras pruebas puede **existir diferencia de potencia estadística**. Muchos investigadores realizan pruebas estadísticas sin comprobar previamente la potencia estadística circunstancia que no es recomendable dejar de hacer. Como una regla siempre a seguir es que: una prueba que tiene una potencia de **1 - β = 0.8** se ve como una prueba de **alta potencia**; **0.5 como de potencia media y 0.2 como de baja potencia**. **Desafortunadamente, SPSS no puede resolver el poder de su prueba antes de que Usted lo realice,** sin embargo, una vez realizadas, en unas cuantas pruebas puede apreciar la potencia de su prueba o el efecto del tamaño a nivel ***post hoc*** (al final del análisis). La potencia estadística depende del **tamaño de la muestra** y del **tamaño del efecto**. Así, **cuanto mayor sea el tamaño de la muestra más potente la prueba**. Además, **cuanto mayor sea el tamaño del efecto (por ejemplo, la diferencia entre las medias de población) más probable es que lo encontremos.**

Como lo informamos, desafortunadamente, **SPSS NO realiza un análisis de potencia estadística antes de que inicie cualquier análisis de datos**. Sin embargo, hay programas disponibles que pueden hacer esto para Usted. Sólo toma unos momentos para llevar a cabo el análisis de potencia y el resultado le reporta a Usted qué tamaño de la muestra necesita para la prueba que requiera hacer.,

8. **¿Cómo proceder con otro software para calcular la potencia estadística?** Deberá:

• Establecer la potencia de su prueba. Usted puede elegir una potencia alta (**0.8**) si desea aumentar las posibilidades de encontrar un efecto

• Establecer el nivel de significatividad (α). ¿Va a realizar prueba de significatividad en **p 0.01**, que permite reduce el riesgo de un **error de Tipo I**, **pero que también reduce la potencia estadística**? o ¿realizará un convencional **p = 0.05**?

• Seleccione la prueba que desea realizar, por ejemplo, **prueba *t* de medidas independientes.**

• Estime **el tamaño del efecto** que está buscando. Apóyese en estudios anteriores de su estado del arte, o incluso una prueba piloto puede proporcionar esta información, forzándolo a aprender sobre las puntuaciones a esperar en un estudio sobre su tema.

• Ajuste **el tamaño de la muestra** para la potencia que desea. El resultado de un análisis de potencia estadística le indicará cuántos participantes necesita en las muestras para alcanzar este nivel de potencia estadística. Esta es una información muy útil, ya que no sólo le impide seleccionar una cantidad extremadamente baja de participantes así como una extremadamente alta. A menudo existe la creencia de que "***cuanto más mejor***" en el análisis estadístico, pero sólo necesita un tamaño de muestra lo suficientemente grande para hacer el trabajo.

2.9. Conclusiones de las pruebas de significatividad

Las pruebas de significatividad estadística nos proporcionan información muy útil, por lo que se ofrecen algunos consejos sobre la interpretación de resultados, como sigue:

1. Un resultado estadísticamente significativo. Cuando el resultado de nuestra prueba es estadísticamente significativo, rechazamos la hipótesis nula. Señalamos que las diferencias muestrales son lo suficientemente grandes como para indicar diferencias en las poblaciones que estamos examinando o que el nivel de asociación indica una asociación en

las poblaciones. Además, es posible considerar **los riesgos de tipo I y Errores de tipo II** por un momento para considerar:
-¿Cuál es el valor de probabilidad real que ha producido?, ¿cuán importante es su valor de |probabilidad?, ¿qué tamaño de efecto se ha encontrado en otros estudios sobre este tema?, el hallazgo puede ser estadísticamente significativo, pero ¿es importante?, ¿es este efecto grande o pequeño de acuerdo a la literatura sobre este tema?,
-SPSS generalmente nos proporciona **intervalos de confianza** con nuestras pruebas de significatividad. Esto nos da una estimación de los valores de la población. Examine estos valores para ver el rango de la estimación.
-Suponga que encontró una diferencia en su muestra de 6 y los intervalos de confianza fueron de 5 a 7, esto significa que estamos **95 por ciento seguros** de que el valor de la población se encuentra Entre los límites superior e inferior, de modo que **en el peor de los casos nuestra diferencia sería todavía ser 5 que todavía podría ser una gran diferencia.**
-Si el intervalo fuera grande, por ejemplo, entre **0.5 y 6.5**, y nuestro valor es **3.5, s**ería menos seguro en cuanto al verdadero valor de la población, ya que podría estar en cualquier lugar entre estos límites. De hecho, existe la posibilidad de que el valor de la población sea tan bajo como 0.5, y aunque el resultado de la prueba puede ser significativo esta diferencia puede no ser de ningún valor práctico.

2. Un resultado estadísticamente NO significativo. Cuando los hallazgos **no son significativos** (por ejemplo ***p*> 0.05**), informamos **que no hemos encontrado evidencia para indicar una diferencia (o asociación)** en las poblaciones. Esto suele ser denominado "***aceptar la hipótesis nula***". Podemos considerar el resultado un poco más cuidadosamente. **No queremos hacer un error de Tipo II**. Así que cuando obtenemos un resultado no significativo Considerar dos cosas:
-¿**Realizamos una prueba lo suficientemente potente para encontrar un efecto**? ¿Estábamos buscando un **gran efecto?,** ¿tenemos suficientes participantes para que sea una prueba lo suficientemente potente? En el análisis de varianza **SPSS** le permite comprobar el tamaño del efecto y la potencia estadística después de haberla realizado.
-Observe su intervalo de confianza para verificar si apoya los resultados de la prueba de significatividad. En una **prueba *t*** esperamos que el intervalo de confianza esté alrededor del valor cero para un hallazgo no significativo.

2.10. Tipos de Técnicas Multivariantes

A continuación, mencionamos las más relevantes, despelgadas en la **Figura 2.10**:

1.El análisis factorial, con variaciones tales como: **el análisis de componentes y el análisis factorial común**. Se usa para el análisis de interrelaciones entre un gran número de variables y explicarlas en términos de sus dimensiones subyacentes comunes (**factores**). El objetivo es hacer la reducción de datos contenida en un número de variables originales, en un conjunto más pequeño de variables (**factores**) con mínima pérdida de información. Al proporcionar una estimación empírica de la estructura de las variables consideradas, ésta técnica se convierte en una base sólida para la creación de escalas aditivas.

2.-La regresión múltiple, técnica adecuada cuando el investigador incluye sólo una **variable métrica dependiente** que se asume está **relacionada con una o más variables métricas independientes**. Su objetivo es la **predicción de cambios en la variable dependiente** en respuesta a **cambios las variables independientes**. Esto se consigue muy a menudo a través de **la regla estadística de los mínimos cuadrados.**

Esta técnica es útil siempre se esté interesado en **predecir** la cantidad o la magnitud de la **variable dependiente**. Por ejemplo, se puede hacer la predicción nuevos servicios (**variables dependientes**) con información referente a características del mercado como ingreso, preferencias, atributos deseados, etc. (**variables independientes**).

3.- Análisis discriminante múltiple. Se aplica considerando que la **variable dependiente es dicotómica** (es decir, innovador-no innovador) o **multidicotómica** (es decir, calidad: alta-medio-bajo) y por tanto **no métrica**. Las **variables independientes son métricas**. Es útil en situaciones donde **la muestra total puede dividirse en grupos** basándose en una **variable dependiente** caracterizada por **varias clases conocidas**. Los objetivos primarios de ésta técnica son **entender las diferencias de los grupos y predecir la verosimilitud** de que la entidad (persona u objeto) pertenezca a una clase o grupo particular basándose a partir de varias **variables métricas independientes**. Por ejemplo, puede usarse para distinguir innovadores de no innovadores de acuerdo a sus perfiles demográficos, de ingresos, psicográficos., preferencias, etc. Entre sus aplicaciones más difundidas se encuentran las agencia tributarias que la usan para comparar las declaraciones seleccionadas con las devoluciones compuestas hipotéticas de un contribuyente normal (para distintos niveles de ingresos) a fin de identificar las devoluciones y áreas más prometedoras para la auditoría

4.-Análisis multivariante de la varianza y covarianza (MANOVA) es una técnica que se usa simultáneamente para explorar las relaciones entre diversas categorías de variables independientes (**tratamientos**) y **dos o más variables métricas dependientes**. Es una extensión del análisis univariante de la varianza (**ANOVA**). El análisis multivariante de la covarianza (**MANCOVA**) puede usarse en conjunción con **MANOVA para eliminar (después del experimento) el efecto de cualquier variable independiente no controlada sobre las variables dependientes.** El procedimiento es similar al de la **correlación parcial bivariante. MANOVA** es útil cuando **se diseña una situación experimental** (con variables de **tratamiento no métricas**) para comprobar hipótesis de la varianza de respuestas de grupos sobre **dos o más variables métricas dependientes.**

5.-Análisis conjunto. Es una técnica de **dependencia** emergente que introduce sofisticación en la evaluación de objetos, sean nuevos productos, servicios o ideas. La aplicación más directa está en el desarrollo de nuevos productos o servicios, permitiendo la

evaluación de productos complejos mientras que mantiene un contexto de decisión realista para el encuestado. El analista de mercado es capaz de evaluar la importancia de atributos así como los niveles de cada atributo mientras que los consumidores evalúan sólo los perfiles de unos pocos productos, que son combinaciones de niveles de producto. Por ejemplo, un concepto de un servicio que tiene tres atributos (precio, rapidez y accesibilidad), cada uno de los cuales a tres niveles (por ejemplo calidad alta, media, baja). En lugar de evaluar 27 combinaciones posibles (3 X 3 X 3), es posible evaluar un subconjunto (9 o más) por su atractivo para los consumidores, con la ventaja que el investigador sabe no sólo cuál es la importancia de cada atributo sino también la importancia de cada nivel (el atractivo de la alta calidad frente a las calidades media y baja). Más aún, cuando se completan las evaluaciones del consumidor, pueden usarse los resultados del análisis conjunto en simuladores del diseño del producto y/o servicio, que mostrarán la aceptación del cliente para cualquier número de formulaciones de producto y ayudar en el diseño del producto/ servicio óptimo.

6.-Correlación canónica. Es la extensión lógica del **análisis de regresión múltiple** dado que implica **una única variable dependiente métrica** y **varias variables métricas independientes**. En ésta técnica **el objetivo es correlacionar simultáneamente varias variables dependientes métricas y varias variables métricas independientes.** Mientras que la regresión múltiple implica una única variable dependiente, **la correlación canónica implica múltiples variables dependientes.** El principio subyacente es desarrollar una combinación lineal de cada conjunto de variables (**tanto independientes como dependientes**) para maximizar la correlación entre los dos conjuntos. El procedimiento implica obtener un conjunto de **ponderaciones para las variables dependientes e independientes** que proporcione la **correlación única máxima** entre el conjunto de **variables dependientes y e independientes.**

7.-Análisis clúster. Es una técnica analítica para desarrollar subgrupos significativos de individuos u objetos. El objetivo es clasificar una **muestra de entidades (personas u objetos)** en un número pequeño de grupos mutuamente excluyentes **basados en similitudes entre las entidades.** A diferencia del análisis discriminante, **los grupos no están predefinidos. Por consiguiente, se usa la técnica para identificar los grupos.** Usualmente implica al menos dos etapas:

-**La primera** es la medida de alguna forma de similitud o asociación entre las entidades para determinar cuántos grupos existen en realidad en la muestra.

-La segunda etapa es describir las personas o variables para determinar su composición. Este paso se realiza aplicando **el análisis discriminante** a los grupos identificados por la técnica **clúster**.

8.-Análisis multidimensional, donde el objetivo es transformar los juicios de los consumidores de similitud o preferencia (por ejemplo, preferencias por tiendas o marcas. comerciales) en distancias representadas **en un espacio multidimensional**. Si los objetos A y B son en opinión de los encuestados **más similares** que el resto de los pares posibles de objetos, las técnicas de análisis multidimensional situarán a los objetos A y B de tal forma que **la distancia entre ellos, en un espacio multidimensional es menor que la distancia entre cualquier otro par de objetos**. Los mapas perceptuales muestran el posicionamiento relativo entre los objetos, pero es necesario un análisis adicional para evaluar qué atributos predicen la posición de cada objeto.

9.-Análisis de correspondencias. Es una técnica recientemente desarrollada de **interdependencia** que **facilita la reducción dimensional** de una **clasificación** de objetos sobre un conjunto **de atributos** y el **mapa perceptual** de objetos relativos a estos atributos. Además, se debe considerar la necesidad de **"*cuantificar datos cualitativos*"** que se encuentran en las variables nominales. Esta técnica difiere de otras técnicas de interdependencia en su capacidad para **acomodar tantos datos no métricos como relaciones no lineales.**

Emplea una **tabla de contingencia** que es **la tabulación cruzada de dos variables categóricas**. A continuación **transforma los datos no métricos en un nivel métrico y realiza una reducción dimensional** (**muy similar al análisis factorial**) y un **mapa perceptual** (similar al **análisis multidimensional**). Por ejemplo, las preferencias por atributos de servicio de los encuestados pueden ser tabuladas de forma cruzada con variables demográficas (por ejemplo, género, categorías de renta, ocupación). La técnica permite la asociación o **"*correspondencia*"** de marcas y las características de aquellos que prefieren cada marca para mostrarlos en **un mapa bi o tri dimensional**, relacionando las marcas con características de los encuestados. Las marcas percibidas como similares se localizan en una cercana proximidad unas de otras. De la misma forma, las características más distintivas de los encuestados que prefieren cada marca están determinadas también por la proximidad de las categorías de las variables demográficas respecto de la posición de la marca. La técnica aporta una representación multivariante de la interdependencia de datos no métricos que no es posible realizar con otros métodos.

10.- Modelos de probabilidad lineal. Se le conoce como **análisis *}***, consistiendo en una **combinación de regresión múltiple y análisis de discriminante múltiple**. Esta técnica es **similar al análisis de regresión múltiple** en que **una o más variables independientes** se usan para **predecir una única variable dependiente**. Lo que distingue al modelo de probabilidad lineal de la regresió n mú ltiple es que la **variable dependiente es no métrica, como en el análisis discriminante**. La escala **no métrica** de la **variable dependiente** requiere diferencias en el método de estimació n y supuestos sobre el **tipo de distribución subyacente**, siendo en la mayoría de sus otras facetas similar **a la regresión múltiple**. Así, al tener la variable dependiente especificada correctamente **se emplea la técnica de estimación apropiada,** usando igualmente los supuestos básicos considerados en la regresió n mú ltiple. Esta técnica se distingue del **análisis discriminante** en que acomodan todos los tipos de variables independientes (**métricas y no métricas**) y **no requieren el supuesto de normalidad** multivariante. Sin embargo, en muchos casos, particularmente con más de dos niveles de la variable dependiente, el análisis discriminante es la técnica más apropiada.

11.-Modelos de ecuaciones estructurales. Denominado simplemente como LlSREL (el nombre de uno de los paquetes informáticos más populares), es una técnica que permite separar las relaciones para cada conjunto de variables dependientes. Esta técnica proporciona una estimación más adecuada y eficiente para series con ecuaciones simultáneas mediante regresiones mú ltiples. Se caracteriza por dos componentes básicos: **(Bearden et al. 1993) el modelo estructural o modelo guía y (BMDP Statistical Software 1991) el modelo de medida.**

-El modelo **"*guía*"**, relaciona variables independientes y variables dependientes. En estos casos, **la teoría, antes que la experiencia u otras directrices,** permitirán que el

investigador distinga qué **variables independientes** predicen cada variable dependiente. Los modelos discutidos que incluyen múltiples variables dependientes **-análisis multivariante de la varianza y correlación canónica-no son apropiados en esta situación**, dado que permiten sólo una única relación entre variables dependientes e independientes.

-El modelo de medida permite al investigador usar varias variables (**indicadores**), para una **única variable dependiente o independiente**. En este caso, el investigador puede evaluar la contribución de cada ítem de la escala así como incorporar cómo la escala mide el concepto (**fiabilidad**) en la estimación de las variables dependientes e independientes. Este procedimiento es similar al desarrollo del **análisis factorial** de los ítems de la escala y utiliza las **cargas factoriales en la regresión.**

12. Técnicas multivariantes emergentes. El uso generalizado de las herramientas informáticas ayudó a iniciar la era del análisis multivariante tal y como lo conocemos hoy, con un número de técnicas especializadas que se pueden aplicar a una gama amplia de situaciones. No obstante, ahora nos encontramos al **principio de una era en la cual el análisis multivariante incorpora nuevos enfoques para identificar y representar las relaciones multivariantes.**

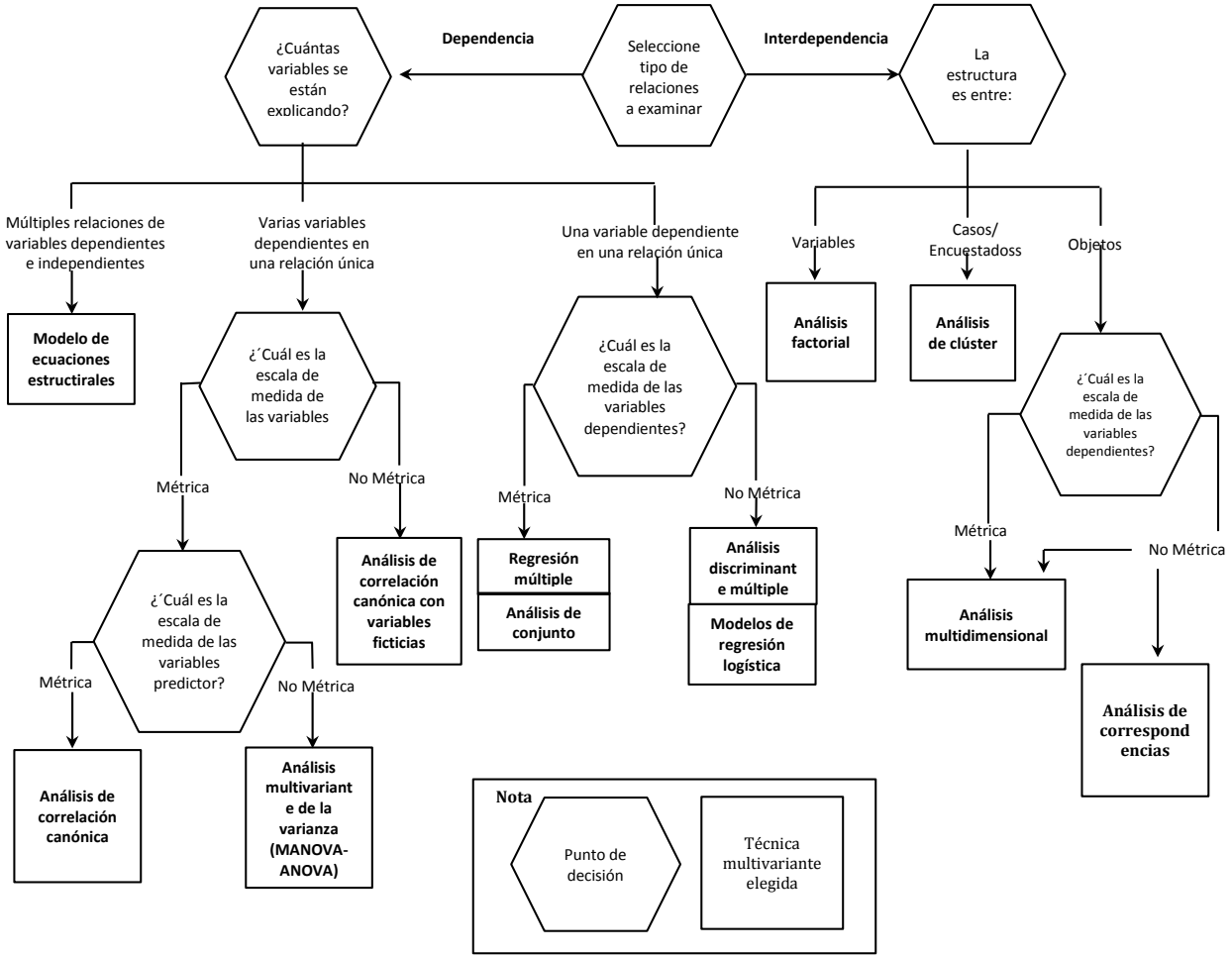
Un área de desarrollo en es la **búsqueda de datos y las redes neuronales**.

-La búsqueda de datos es el intento de cuantificar las relaciones entre grandes cantidades de información con una especificación previa mínima de la naturaleza de las relaciones. Una técnica que se usa muchas veces junto con la búsqueda de datos son las re des neuronales, una técnica de análisis flexible que es capaz de llevar a cabo una identificación-de relaciones (parecida a la regresión múltiple o al análisis discriminante) o la reducción de datos y el análisis estructural (semejante al análisis factorial o clúster).

-Las redes neuronales son diferentes a las técnicas multivariantes más tradicionales tanto en la formulación del modelo como en los tipos de relaciones más complejos que se pueden formular.

-La técnica de la muestra repetida o "*arranque*". Elimina la necesidad de cumplir determinados supuestos estadísticos (**como la normalidad**), mediante el uso del ordenador para **replicar una "*muestra repetida*" de la muestra original**, con el reemplazo y la generación de una estimación empírica de la distribución muestral.

Figura 2.10. Diagrama de Flujo para selección de técnica multivariante



Fuente: Hair et al. (1999)

2.11. Análisis de dependencia y selección de la técnica multivariante

Para una mejor selección de las técnicas multivariantes, se sugiere practicar con **3** de los juicios que un investigador debe realizar sobre el objeto de estudio y la naturaleza de los datos:

1.- ¿Pueden dividirse las variables en: **dependientes o independientes** basándose la clasificación en alguna teoría? La respuesta a esta cuestión indica si se debería utilizar un análisis de dependencia o interdependencia

2.- Si puede hacerse, ¿cuántas de estas variables son tratadas como dependientes en un análisis simple?

3.- ¿Cómo son las variables medidas?
Nótese que en la **Figura 2.10**, las **técnicas de dependencia** están en el lado izquierdo y las **técnicas de interdependencia** están a la derecha.
Un **análisis de dependencia**, es aquel en el que una variable o conjunto de variables es identificado como la **variable dependiente** y que va a ser explicada por otras variables conocidas como **variables independientes.** Por ejemplo el **análisis de regresión múltiple**. Un análisis de **interdependencia** es aquel **en que ninguna variable o grupo de variables es definido como independiente o dependiente,** y es un procedimiento que implica el análisis de todas las variables del conjunto simultáneamente. Otro ejemplo: El **análisis factorial**.
Así, se recomienda en la selección del método multivariante:

1. El **análisis de interdependencia**
2. El **análisis de dependencia** y después referirse la clasificación de la **Figura 2.10**

Los métodos del **análisis de dependencia** pueden ser a su vez divididos en dos tipos, según:

- **El número de variables dependientes.** Puede clasificarse como de una variable dependiente única, como varias variables dependientes o incluso varias relaciones de dependencia/independencia.
- **El tipo de escalas de medida empleadas para las variables.** Con variables métricas (numéricas/cuantitativas) o no métricas (cualitativas/categóricas).Si implica una única variable dependiente que es métrica, la técnica apropiada es tanto el **análisis de regresión múltiple** como el **análisis conjunto**. Éste último es un caso especial ya que se trata de un procedimiento de dependencia que puede tratar la variable dependiente como métrica o no métrica, en función de las circunstancias. Si la única variable dependiente es **no métrica (categórica)**, entonces la técnica apropiada es, o **análisis discriminante múltiple**, o los **modelos de probabilidad lineal**. Cuando el problema implica varias variables dependientes, hay otras **4** técnicas estadísticas apropiadas:

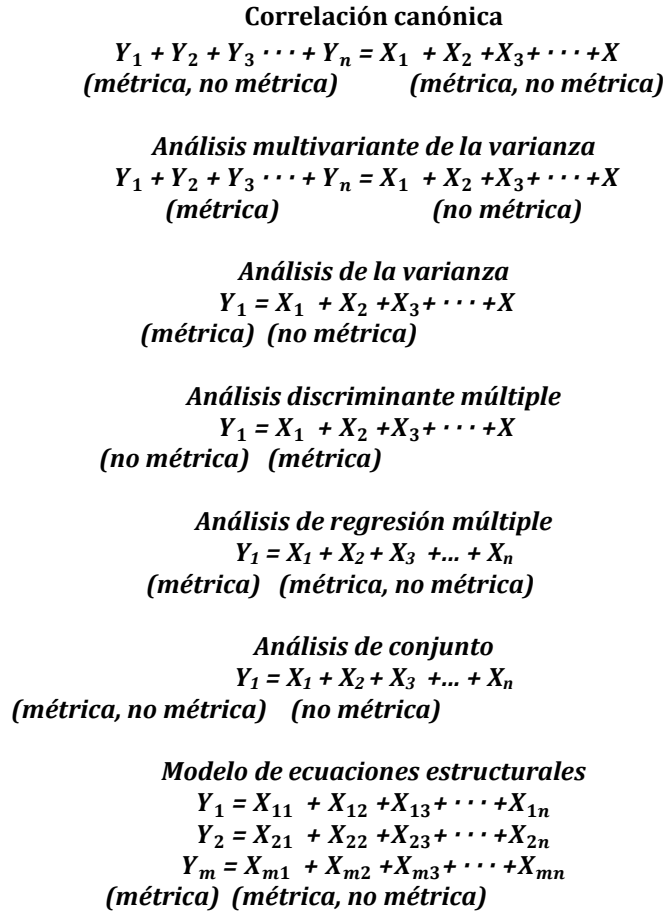
-Si varias variables dependientes son **métricas**, debemos entonces apuntar a las variables independientes.
-Si las variables independientes son **no métricas**, debemos elegir **análisis de la varianza.**
-Si las variables independientes son **métricas**, la apropiada es la **correlación canónica**
-Si varias variables dependientes **son no métricas**, entonces pueden transformarse a través de una variable ficticia de código (**0-1**) y puede utilizarse también el análisis canónico (la codificación de la variable ficticia **es una manera de transformar datos no métricos en datos métricos**. La creación de variables ficticias, en las cuales se asignan unos y ceros al sujeto, dependiendo de si cuenta o no con cierta característica. Por ejemplo, si un sujeto es masculino se le asigna un 0 y si el sujeto es femenino se le asigna un 1, o al contrario).
Finalmente, si se postula un conjunto de relaciones de variables dependientes/independientes, entonces el modelo de ecuaciones estructurales es el apropiado. Para la selección de la técnica multivariante se recomienda ver tanto la **Figura 2.10** como el **Apéndice: Matriz de pruebas estadísticas sugeridas.**

2.12. Relaciones de los métodos multivariantes

En el **análisis de interdependencia** no pueden ser clasificadas las variables como dependientes o independientes. Las variables son analizadas simultáneamente a fin de encontrar una estructura subyacente para el conjunto total de variables o sujetos. Si realiza análisis de la **estructura de las variables**, entonces **el análisis factorial** es la técnica apropiada. Si los **casos o los encuestados** se van a agrupar para representar una estructura, entonces seleccionaremos el **análisis clúster**. Si el interés está en la **estructura de objetos**, se debe aplicar las técnicas de **análisis multidimensional**. Como ocurre con el análisis de dependencia, deberían considerarse las propiedades de las técnicas de medición. Generalmente, el **análisis factorial** y el **análisis clúster** se consideran **análisis de interdependencia métricos**. Sin embargo, los **datos no métricos** se pueden transformar a través de **variables ficticias codificadas** para usarlos con **análisis factorial** y **análisis clúster.** Existen desarrollos tanto de las aproximaciones métricas como de las no métricas al **análisis multidimensional**. Si va a analizar las interdependencias entre objetos medidos por datos **no métricos**, la técnica a utilizar será **análisis de correspondencias.**

La **Figura 2.11** muestra varios de los **análisis de dependencia multivariante** en términos de la naturaleza y número de las variables dependientes e independientes.

Figura 2.11. Relaciones entre métodos de dependencia multivariante



Fuente: Hair et al. (1999)

Como se aprecia, la **correlación canónica** es posible considerarla como el modelo general en el cual se basan otras muchas técnicas multivariantes, dado que sitúa la mínima restricción respecto al **tipo y número de variables** tanto de valor teórico **dependiente como independiente.** Como las restricciones se basan en valores teóricos, es posible alcanzar conclusiones más precisas con la escala específica empleada en la medición de los datos. Estas técnicas multivariantes parten desde el método general del **análisis canónico** hasta el más especializado método de **modelización de ecuaciones estructurales**.

2.13. Recomendación de cómo usar

El análisis multivariante es una herramental poderoso pero que necesariamente debe estar basado en un sólido constructo teórico conceptual; inútil es practicarlo si no se cuenta con esto último debido a que la complejidad de las relaciones entre las variables, exige un conocimiento profundo de los modelos bajo análisis para su explicación. Así, se sugiere para su mejor aplicación:

1. **Determinar significación práctica y estadística.** La potencia de ejecución y los resultados que arrojan las diversas técnicas multivariantes, debe ser realizada con precaución y **evitar la miopía al solamente tomar en cuenta la significación conseguida por los resultados sin entender sus interpretaciones**, a favor o en contra de lo esperado. Así, de los resultados aparte de entender la significación estadística también debe hacerlo a su significación práctica, que se refiere a la pregunta ***" ¿y para qué?"***. Es decir, para cualquier aplicación, **los resultados deben tener un efecto demostrable que justifique la acción,** con implicaciones teóricas y sustantivas, que en la mayoría de las veces, se deducen de su **significación práctica.** Por ejemplo, **un análisis de regresión** que haga la predicción de las **diversas tecnologías que soportan a la mercadotecnia digital** en los próximos **5 años**, medidas como la probabilidad entre **0 y 1** entre ellas, con nivel de significación establecido de **0.05.** Los ejecutivos aceptan los resultados, los analizan y modifican la estrategia de la empresa. Sin embargo, **lo que no se ha tenido en cuenta es que mientras la relación era significativa, la capacidad predictiva posiblemente era baja**, tan baja que la estimación de la posibilidad de la entrada de ciertas tecnologías podría variar tanto como un **20% al nivel de significación del 0.05**. ¡La relación de la **"*significación estadística*"** podría entonces tener un **rango de error de 40 %!** A una firma del cual se predice que tiene una oportunidad de introducir una nueva tecnología de **50/50** podría realmente tener probabilidades del **30 al 70%** representando niveles inaceptables sobre los cuales actuar. Los investigadores y los gerentes no han probado la significación práctica o de gestión de los resultados, olvidando que **la relación todavía necesitaba un posterior refinamiento.**
2. **El tamaño muestral afecta a todos los resultados.** La discusión de la **potencia estadística** demuestra que el impacto sustancial del **tamaño muestra** opera en la consecución de la **significación estadística**, tanto en tamaños muestrales grandes como pequeños. Para **muestras pequeñas**, la sofisticación y la complejidad del análisis multivariante fácilmente resulta tanto en:
   -Muy poca potencia estadística de la prueba para identificar de forma realista resultados significativos **o**

-Fácilmente un **"*sobre aprovechamiento*"** de los datos de tal forma que **sean artificialmente buenos** porque se ajustan muy bien a la muestra, aunque no sean generalizables.

Lo mismo ocurre para **muestras grandes** que, pueden hacer a los **test estadísticos altamente sensibles**. Siempre que los tamaños muestrales excedan los **200 o 400 encuestados**, el investigador debería examinar todos los resultados significativos para asegurarse que tienen significación práctica debido al **aumento de la potencia estadística como consecuencia del tamaño muestra**. Los tamaños muestrales también afectan a los resultados cuando los análisis implican **grupos de encuestados**, como ocurre en el **análisis discriminante** o en **MANOVA**. Tamaños muestrales **desiguales** entre los grupos influencian a los resultados y requieren un análisis y/o interpretación adicional. Por tanto, el investigador o usuario del análisis multivariante debería siempre valorar los resultados a la luz de la muestra utilizada.

3. **Conozca los datos.** Las técnicas, por su propia naturaleza, identifican relaciones complejas difíciles de representar de forma simple. Como resultado, **la tendencia es aceptar los resultados sin el típico examen que uno emprende en los análisis univariante y bivariante** (por ejemplo, gráfico de dispersión de correlaciones y ***boxplots*** de comparaciones de media). Estos **"*atajos*"** pueden ser el preludio del complicaciones fatales ya que el análisis multivariante **requiere un examen más riguroso de los datos porque la influencia de atípicos, violaciones de los supuestos y la pérdida de datos** puede agravarse a través de varias variables y tener efectos sustancialmente diferentes. Para servirse de todos los beneficios, el analista debe también **"*saber dónde mirar*"** con formulaciones alternativas del modelo original, tales como **relaciones no lineales e interactivas**. Además, cuenta con un conjunto de técnicas de diagnóstico en continua expansión que **permiten que estas relaciones multivariantes sean descubiertas por medios similares a los métodos univariantes y bivariantes**. El investigador de un problema multivariante debe tomarse su tiempo en utilizar estas medidas de diagnóstico para un mayor entendimiento de los datos y de las relaciones básicas que existen.
4. **Verifique continuamente la parsimonia del modelo.** Las técnicas se diseñan para acomodar las variables en el análisis. Esto sin embargo, **no debe sustituir el desarrollo de modelos conceptuales antes de que se apliquen las técnicas multivariantes**. Aunque es siempre importante evitar omitir una variable predictor crítica, denominada **error de especificación**, por varias razones deberá también **evitar insertar variables indiscriminadamente** debido a que las **variables irrelevantes** aumentan la capacidad del análisis para ajustar la muestra de datos pero a costa de **sobre ajustar los datos** y hacerlos **menos generalizables** para la población. Así, las **variables irrelevantes no sesgan, regularmente las estimaciones de las variables relevantes, pero pueden enmascarar los efectos verdaderos debido a la multicolineadad**. La **multicolinealidad** es el grado en el que cualquier efecto de una variable **puede ser prevista o explicada por las otras variables del análisis**. **A mayor multicolinealidad, la capacidad para definir el efecto de cualquier variable disminuye,** por lo que **incluir variables que no son relevantes** conceptualmente atrae varios **efectos potencialmente dañinos**, incluso si las variable adicionales no sesgan directamente los resultados del modelo.

5. **Atienda los errores.** A pesar de la potencia del análisis multivariante, Difícilmente se consigue la mejor predicción en el primer análisis, por lo que se be partir de la cuestión, **" *¿adónde podemos ir desde aquí?*".** La mejor respuesta es considerar a los **errores en la predicción**, tanto si son los **residuos del análisis de regresión, la ausencia de clasificación de observaciones en el análisis discriminante o los atípicos del análisis clúster**.En cada caso, se deben utilizar los **errores de predicción** no como una medida de error o como algo meramente a eliminar, sino como un punto de partida para **diagnosticar la validez de los resultados obtenidos** y como una indicación de las relaciones que quedan sin explicar.
6. **Validar los resultados.** Una de las mayores capacidades del análisis multivariante es **identificar interrelaciones complejas**, las cuales implican que puede darse el caso de que **los resultados sean específicos sólo para la muestra y no generalizables a la población**. Así, se debe siempre asegurar que existen observaciones suficientes por parámetro estimado y **evitar el *"sobreajuste"*** de la muestra..
7. **Validez.** Se realiza por diferentes métodos, que incluyen:
   -**División de la muestra** y el uso de una **submuestra** para estimar el modelo y usar una segunda submuestra para estimar la **precisión predictiva**,
   -Empleo de ana lisis de **"*bootstrapping*" (Mooney & Duval 1993 o Brent et al.,1993)** incluso conseguir una muestra distinta para asegurar que los resultados son apropiados para otras muestras.

Cualquiera que sea la técnica multivariante empleada, el investigador debe centrarse **no sólo en estimar un modelo significativo** sino también en **asegurar que es representativo de la población en su conjunto.** Recuerde: el objetivo n**o es encontrar el mejor "*ajuste*"** sólo para la muestra sino **desarrollar el modelo que mejor describa a la población en su conjunto.** Tanto las numerosas técnicas multivariantes disponibles y la extensa cantidad de supuestos que implica su aplicación, hace muy evidente que para finalizar con éxito un análisis multivariante se tiene algo más que la selección del método correcto. Esto va desde definir el problema hasta el diagnóstico crítico de los resultados. Una propuesta para la aplicación de los métodos multivariantes, es la de **7 pasos** (Hair, 1999), a manera de recomendaciones:

Paso 1: Objetivos.

El punto de partida para cualquier análisis multivariante es definir el problema de investigación, los objetivos analíticos de forma conceptual, **antes de especificar** cualquier variable o medida. No debe exagerar el papel del desarrollo del modelo conceptual o teoría. No importa si la investigación es práctica o académica, se debe ver en primer lugar el **problema en términos conceptuales**, a partir de los conceptos e identificar las relaciones fundamentales a investigar. **Desarrollar un modelo conceptual no es dominio exclusivo de los académicos; también se ajusta a la aplicación a la experiencia del mundo real.** Un **modelo conceptual** no debe ser complejo y detallado, sino una simple representación de relaciones a estudiar. Si se propone una **relación de dependencia** como objetivo de investigación, debe especificar los conceptos **dependientes e independientes**. **Nota: se define un concepto, más que una variable**. Para la aplicación de una **técnica de interdependencia**, se deben determinar las **dimensiones** de la **estructura** o **similitud**. En situaciones de **dependencia** como de **interdependencia**, debe identificar **primero las**

ideas o temas de interés en lugar de fijarse en las medidas a utilizar. Esto minimiza la posibilidad de que conceptos relevantes sean omitidos en el esfuerzo por desarrollar medidas y definir los detalles del diseño. Un buen caso de desarrollo de modelos conceptuales son las ecuaciones estructurales. Con los **objetivos y el modelo conceptual especificados**, ya es posible elegir la técnica multivariante apropiada.
Después de haber seleccionado un **método de dependencia o interdependencia**, la última decisión consiste en **elegir la técnica específica basada en las características de medición de las variables dependientes e independientes**. Se pueden especificar las variables antes del estudio o se pueden definir después de recoger los datos.

Paso 2: Diseño
A partir del modelo conceptual establecido, ponga en práctica de la técnica elegida. Para cada técnica, debe desarrollar un plan de análisis específico que dirija el conjunto de supuestos que subyacen en la aplicación de la técnica, que van desde consideraciones generales de **tamaños de muestra mínimos o deseados, tipos de variables** (**métricas vs. no métrica**s) y **métodos de estimación** (tipo de medida de asociación usada en el análisis multidimensional, la estimación de los resultados agregados o desagregados en el conjunto o el uso de formulaciones especiales de variables para representar efectos interactivos o no lineales en la regresión). Estos supuestos resuelven los detalles específicos y finalizan la formulación del modelo y los requisitos del esfuerzo de recogida de datos

Paso 3: Supuestos de aplicabilidad
Evaluación de los supuestos básicos de la técnica multivariante
A partir de la recogida de datos, **el primer análisis no consiste en estimar el modelo multivariante,** sino en **evaluar los supuestos subyacentes**. Todas las técnicas multivariantes las tienen, tanto **estadísticos como conceptuales**, que afectan las relaciones multivariantes. Para las técnicas basadas en la **inferencia estadística** se deben tener en cuenta los supuestos de **normalidad multivariante, linealidad, independencia de los términos de error e igualdad de las varianzas en una relación de dependencia**.
Cada técnica tiene una serie de **supuestos conceptuales** que tratan sobre asuntos como la **formulación de modelos** y los **tipos de representaciones.** Antes de intentar cualquier estimación del modelo, asegúrese de que se encuentran cubiertos los supuestos estadísticos y los conceptuales.

Paso 4: Estimación y ajuste del modelo
Ya planteados los modelos, se realiza la **estimación efectiva del modelo multivariante** y una **valoración global del ajuste del modelo**. En el proceso de estimación, tiene distintas opciones para **elegir las características específicas de los datos** (por ejemplo, uso de covarianzas en **MANOVA**) o **maximizar el ajuste de los datos** (por ejemplo, **rotación de los factores o funciones discriminantes**). Después de estimar el modelo, se **evalúa el ajuste** para averiguar si consiguen **niveles aceptables sobre los criterios estadísticos** (por ejemplo, nivel de significación), identificar las relaciones propuestas y conseguir la **significación práctica**. El proceso es iterativo por lo que es muy probable que **deberá re-especificar el modelo** para **mejorar** los niveles de ajuste y/o explicación, en la que determine si los resultados están excesivamente afectados por un único o pequeño conjunto

de observaciones que indican que los resultados tiendan a ser inestables. Estos esfuerzos aseguran que los resultados sean **"*robustos*" y estables** al aplicarlos razonablemente a todas las observaciones de la muestra. Ajustes inadecuados a las observaciones pueden identificarse como **atípicas**, observaciones influyentes u otros resultados dispersos (por ejemplo, **conglomerados** de un único miembro o casos seriamente desclasificados en el **análisis discriminante**).

Paso 5: Interpretación
Con un nivel aceptable de ajuste del modelo, al interpretar los valores teóricos se revela la naturaleza de las relaciones multivariantes. La interpretación de los efectos para variables individuales se realiza examinando los coeficientes estimados (**ponderaciones**) para cada variable en el valor teórico (por ejemplo, ponderaciones de regresión, cargas de los factores o utilidades conjuntas). Algunas técnicas también estiman los valores teóricos múltiples que representan las dimensiones subyacentes de la comparación o asociación (por ejemplo, **funciones discriminantes o componentes principale**s). La interpretación puede conducir a re-especificaciones adicionales de las variables y/o formulación del modelo, se estima de nuevo y se interpreta una vez más. El objetivo es identificar la evidencia empírica de las relaciones multivariantes de los datos muestrales que pueden generalizarse para el total de la población.

Paso 6: Validación
Antes de dar por bueno los resultados, deberá someterlos a un conjunto final de diagnósticos que aseguran el **grado de generalidad** de los resultados por los **métodos de validación disponibles.** Validar el modelo se refiere a la **demostración de la generalidad de los resultados al conjunto de la población**. Los diagnósticos añaden poco a la interpretación de los resultados pero **sirven para asegurar los resultados más descriptivos de los datos y su generalización al conjunto de la población**.

Para cada técnica, el uso de los pasos anteriores de construcción de un modelo multivariante se indicará en un diagrama de flujos de decisiones dividido en dos secciones.
La primera sección (**los pasos 1 a 3**) se refiere a los temas abordados con la preparación para la propia **estimación de modelos** (es decir, los objetivos de investigación, consideraciones para el diseño de la investigación y el ensayo para las suposiciones).
La segunda sección (**los pasos 4 a 6**) se refiere a las cuestiones del **modelo de estimación, interpretación y validación**. El diagrama de flujos de decisiones proporciona un método simplificado pero sistemático para la aplicación de organizada del diseño de modelos multivariantes cuando se aplica cualquier de sus técnicas.

Referencias

Bearden, W. 0., Netemeyer, R. G. y Mobley M. F. (1993), Handbook of Marketing Sea/es, *Multi-Item Measures for Marketing and Consumer Behavior*. Newbury Park, Calif.: Sage.

BMDP Statistical Software, Inc. (1991), SOLO Power Analysis. Los Angeles.

Brent, E., Edward J. Mirielli, y Thompson A. (1993), Ex-Sample1M: *An Expert System to Assist in Determining Sample Size*, Version 3.0. Columbia, Mo.: Idea Works.

Brent, E., et al. (1991), *Statistical Navigator Professional™*: An Expert System to Assist in Selecting Appropriate Statistical Analyses, Version 1.0. Columbia, Mo.: Idea Works.

Brunner, G. C., y Paul J. H. (1993), *Marketing Scales Handbook*, A Compilation of Multi-Item Measures. Chicago: American Marketing Association.

Cohen, J. (1977), *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

Hair, J.F.; Anderson, R.E.; Black, W.C. (1999). *Análisis Multivariante*. 5a. Ed. España:Prentice Hall

IBM (2011a). *IBM SPSS Statistics Base 20.* EUA. .Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Base.pdf

IBM (2011b).*Guía breve de IBM SPSS Statistics 20.* EUA.Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Brief_Guide.pdf

Mooney, C. Z., y Duval R. D. (1993), Bootstrapping: *A Nonparametric Approach to Statistical inference*. Beverly Hills, Calif.: Sage.

Capítulo 3. Análisis de Datos

3.1. Análisis de datos y su importancia

Para trabajar con cualquier técnica multivariante, se debe realizar primero, un examen de datos, el cual es una etapa necesaria y vital, que ciertamente es muy demandante en tiempo, y que es común que **habitualmente se descuida por parte de los analistas de datos**. El llevarlo a cabo de manera cuidadosa y detallada conduce a una mejor predicción, evaluación y hace más fácil explicar las dimensionalidades, a partir de tablas y gráficas que los sistemas de software producen del análisis de datos (Anderson, 1969),. Estas tablas y gráficas proporcionan al analista un **conjunto de formas simple y completa**, para examinar tanto las variables individuales como las relaciones entre ellas. Obstáculos adicionales que se deben resolver son la **evaluación** y **solución** de los problemas del diseño de la investigación así como la recolección de datos. Son de particular interés 3 análisis de datos:

1.-La **evaluación de los datos ausentes**, Los datos ausentes son una molestia para los investigadores. Pueden ser producto de errores en la introducción de los datos o de la omisión de respuestas por parte de los encuesta dos. En este capítulo se discutirá la clasificación de los datos ausentes y los procesos o razones que explican su presencia.

2.-La **identificación de casos atípicos** y o respuestas extremas, pueden influenciar indebidamente el resultado de un análisis multivariante, por lo que se deben utilizar métodos para evaluar su impacto.

3.-La **comprobación de los supuestos subyacentes**., donde debe evaluar el **ajuste de la muestra** de datos con los **supuestos subyacente**s en la técnica multivariante. Por ejemplo, los investigadores que desean analizar el **análisis de regresión** estarán interesados en

evaluar los **supuestos de normalidad, homocedasticidad, independencia del error y linealidad.** Cada una de estas cuestiones deberían ser abordadas en cierta medida para cada aplicación de la técnica multivariante. Adicionalmente, deberá considerar los **métodos para incorporar las variables no métricas** en aplicaciones que requieren **variables métricas** mediante la creación de un **tipo de variable métrica** especial conocida como variable **ficticia.** La aplicabilidad del uso de las variables ficticias varía con cada proyecto de análisis de datos. Para saber más, consulte: IBM 2011a; IBM 2011b; IBM, 2011c.

3.2. Análisis de datos y su importancia

Las técnicas multivariantes por su gran poder analítico, también crean una gran carga de actividades adicionales, ya que debe asegurarse de que exista congruencia entre los constructos teóricos y estadísticos sobre las que se basan. Al hacer análisis de datos en las técnicas multivariantes, se obtiene (IBM,2011):

- Una comprensión del enlace **datos-relaciones entre las variables-resultados esperados** cuya complejidad está en continuo aumento y que por ello, permite el refinamiento del modelo multivariante y proporcionar perspectivas claras y razonables para la interpretación de los resultados.
- La demanda de análisis cada vez más, de enormes cantidades de datos, que implican que la **potencia estadística** requiera de supuestos más complejos que los que encontramos en los análisis univariantes.
- La complejidad analítica necesaria para asegurar los requerimientos estadísticos de aplicación de la técnica multivariante elegida, obliga usar una serie de **técnicas de examen de los datos que en muchas ocasiones rivaliza en complejidad con la propia técnica multivariante.**
- Los **efectos de los datos ausentes,** los cuales por definición no se representan directamente en los resultados, **pueden ser sustanciales** dado el impacto que tienen sobre la naturaleza y carácter de las variables

El examen de los datos, típicamente se basa desde el **proceso simple de inspección visual de los gráficos** al **proceso estadístico multivariante** que incluye: **análisis de datos ausentes y a la comprobación de los supuestos subyacentes en todos los métodos multivariantes**. Por esto la importancia de detenerse en esto antes de entregarse a las técnicas multivariantes. No debe considerarse que **malgasta el tiempo, el esfuerzo y los recursos dedicados al proceso de examen de los datos**, debe ver estas técnicas como una **"*inversión en un seguro multivariante*".** Cabe advertir que aunque una técnica sirva para hacer una estimación adecuada y obtener resultados, existen problemas **"*ocultos*" a prever,** que surgen de las cuestiones antes dichas con un potencial de problemas catastróficos y que es posible evitar.

El examen de datos sugerida incluye **fases**, que son:

1. **Examen gráfico** de la naturaleza de las variables a analizar y las relaciones que forman las bases del análisis multivariante,
2. **Proceso de evaluación** para entender el impacto que pueden tener los **datos ausentes** sobre el análisis, y una serie de alternativas para casos reiterados de ausencia de datos en el análisis

3. **Identificación de casos atípicos**, con técnicas que por su **singularidad** pueden distorsionar las relaciones sobre una o más variables estudiadas y
4. **Los métodos analíticos de evaluación** de la capacidad de los datos para cumplir los supuestos estadísticos específicos de las técnicas multivariantes.
5. Introducción y evaluación de técnicas para **incorporar variables no métricas cuando se requieren variables métricas** creando una serie de **variables métricas de reemplazo** para representar **las categorías de las variables no métricas.**

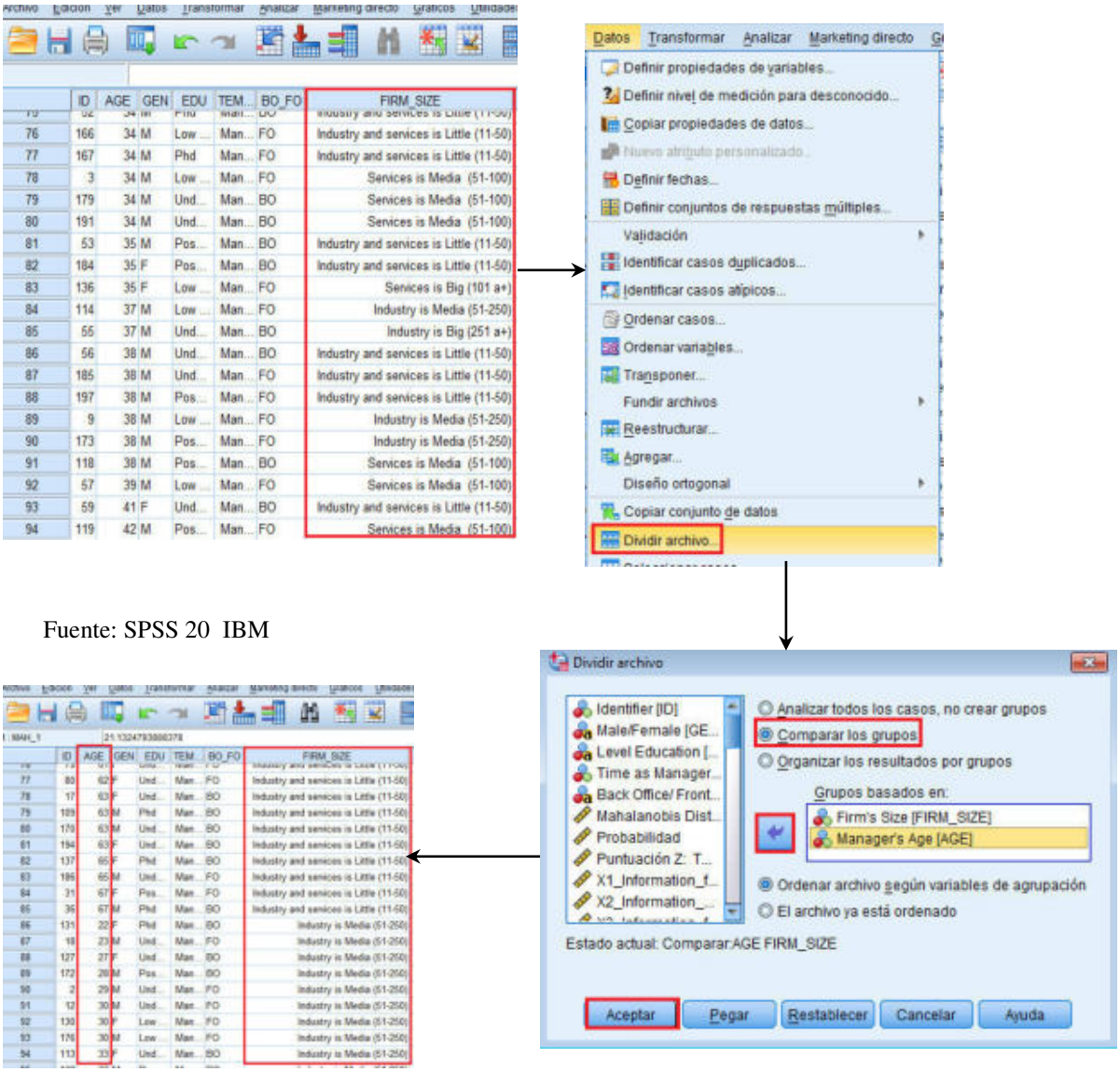
Siempre debe prevalecer el objetivo de **entender, evaluar e interpretar los resultados** más complejos, por lo que se debe comprender las características básicas de los datos y sus relaciones subyacentes, destacando que al considerar un análisis univariante, el nivel de comprensión es muy simple, particularmente con ayuda de software especializado como el SPSS.

3.2.1. Análisis de la forma de la distribución

Es el punto de partida para entender la naturaleza de cualquier variable se basa en **caracterizar la forma de su distribución**, de la que es básica el obtener una perspectiva adecuada de la variable a través de un **histograma.**(representación gráfica de los datos que muestra la frecuencia de los casos y valores) en categorías de datos, **tecleando: Analizar->Estadísticos descriptivos->Explorar->Selección variables->Gráficos y en descriptivos, seleccionar: De tallo y hojas así como Histograma** (ver **Figura 3.1**). Las frecuencias representan la forma de la distribución de respuestas para su examen. Si el rango de respuestas va de **1 a 10**, investigador puede construir un histograma contando el número de respuestas que fueron **1, 2**, etc. En el caso de **variables continuas** se forman categorías, dentro de las cuales la **frecuencia de los valores de datos está tabulada.** La altura de las barras representa la frecuencia de los valores de los datos en cada categoría. Si el examen de la distribución tiene como objetivo **evaluar su normalidad** (**tecleando: Analizar->Frecuencias->Selección variables->Gráficos, en Histogramas seleccionar: Mostrar curva normal en el histograma**) es posible **superponer la curva normal sobre la distribución**, como se ha hecho en la **Figuras: 3.2**. **y 3.3** El **histograma** puede utilizarse para examinar cualquier tipo de variable, desde los valores originales a los residuos de una técnica multivariante. Una variante del **histograma** es el **diagrama de tallo y hojas** que presenta el mismo cuadro gráfico pero que también proporciona una enumeración de los valores de los datos reales.

-Problema 1: Dividir el archivo **CKM_MKT_Digital.sav** de acuerdo a los diferentes tipos de empresa.
-Teclear: Datos->Dividir archivo; Selección Comparar los grupos; Selección de variable (Firm's size) ->Aceptar.

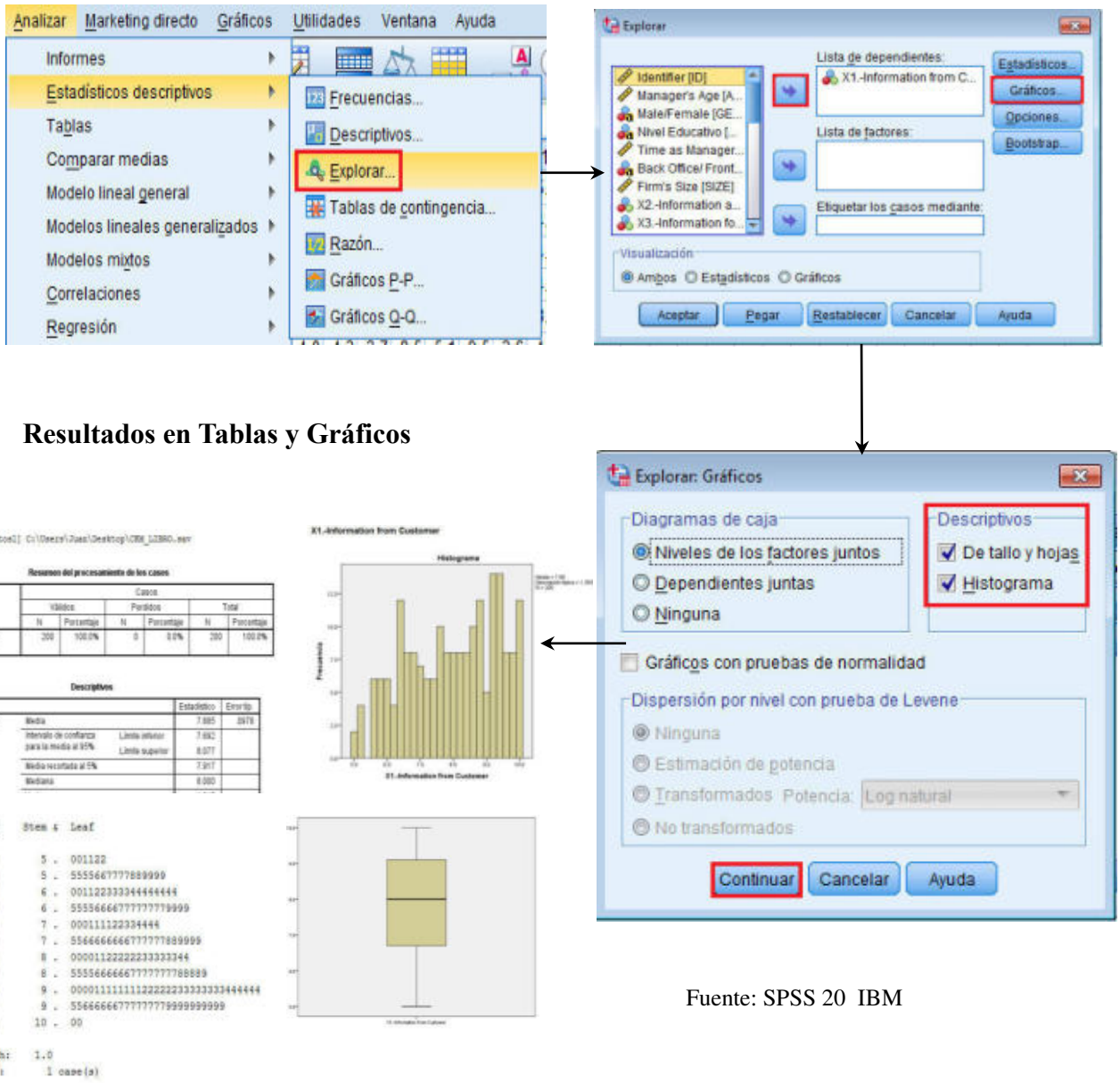
Figura 3.1. Proceso para ordenar la base de datos



Fuente: SPSS 20 IBM

Resultado: Se obtiene la base de datos ordenada de acuerdo a lo requerido

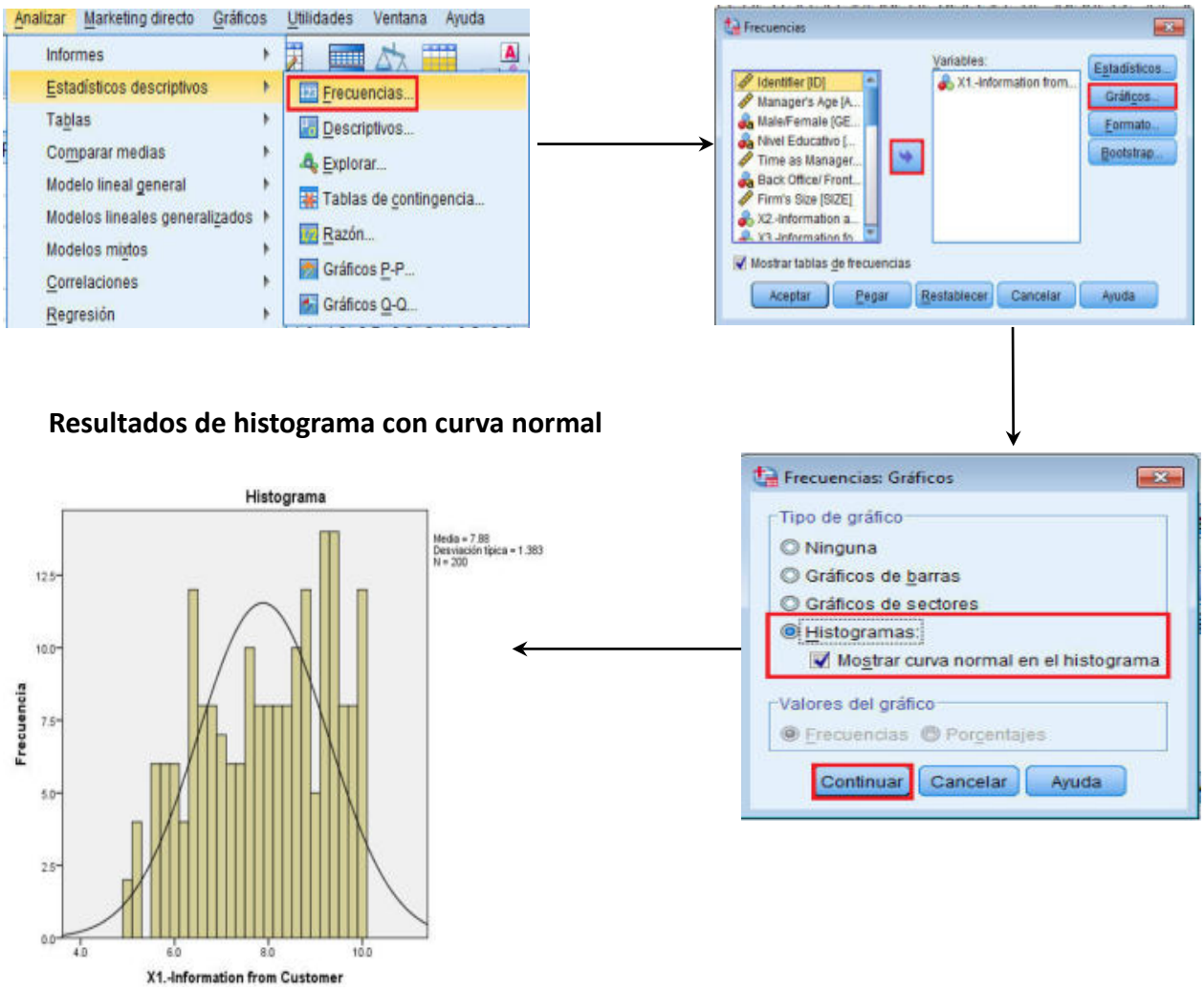
Figura 3.2.- Histograma de la variable X_1 de la base de datos CKM_MKT_Digital.sav



Resultados en Tablas y Gráficos

Fuente: SPSS 20 IBM

Figura 3.3.- Histograma de la variable X_1 de la base de datos CKM_MKT_Digital.sav



Fuente: SPSS 20 IBM

Siempre debe prevalecer el objetivo de entender, evaluar e interpretar los resultados más complejos, por lo que se debe comprender las características básicas de los datos y sus relaciones subyacentes, destacando que al considerar un análisis univariante, el nivel de comprensión es muy simple, particularmente con ayuda de software especializado como el SPSS. (IBM,2011).

3.2.2.Análisis de la forma de la distribución

Es el punto de partida para entender la naturaleza de cualquier variable se basa en **caracterizar la forma de su distribución**, de la que es básica el obtener una perspectiva adecuada de la variable a través de un **histograma.**(representación gráfica de los datos que muestra la frecuencia de los casos y valores) en categorías de datos, **tecleando: Analizar->Estadísticos descriptivos->Explorar->Selección variables->Gráficos y en descriptivos, seleccionar: De tallo y hojas así como Histograma** (ver **Figura 3.3**). Las frecuencias representan la forma de la distribución de respuestas para su examen. Si el rango de respuestas va de **1 a 10**, investigador puede construir un histograma contando el número de respuestas que fueron **1, 2**, etc. En el caso de **variables continuas** se forman categorías, dentro de las cuales la **frecuencia de los valores de datos está tabulada.** La altura de las barras representa la frecuencia de los valores de los datos en cada categoría. Si el examen de la distribución tiene como objetivo **evaluar su normalidad** (**tecleando: Analizar->Frecuencias->Selección variables->Gráficos, en Histogramas seleccionar: Mostrar curva normal en el histograma**) es posible **superponer la curva normal sobre la distribución**, como se ha hecho en la **Figura 3.3**. El **histograma** puede utilizarse para examinar cualquier tipo de variable, desde los valores originales a los residuos de una técnica multivariante. Una variante del **histograma** es el **diagrama de tallo y hojas** que presenta el mismo cuadro gráfico pero que también proporciona una enumeración de los valores de los datos reales.

3.2.3. Análisis de relación entre variables

Aparte del examen de la distribución de una variable se encuentra el examen de las relaciones entre dos o más variables. Uno de los métodos más aplicados para el **análisis de las relaciones bivariantes** es el **gráfico de dispersión**, basado puntos de datos de **2** variables. Una variable se presenta en el eje horizontal y la otra en el vertical. Las variables pueden ser valores: **observados, esperados o residuos.** Los puntos del gráfico muestran un **patrón** que representa la relación entre las variables.

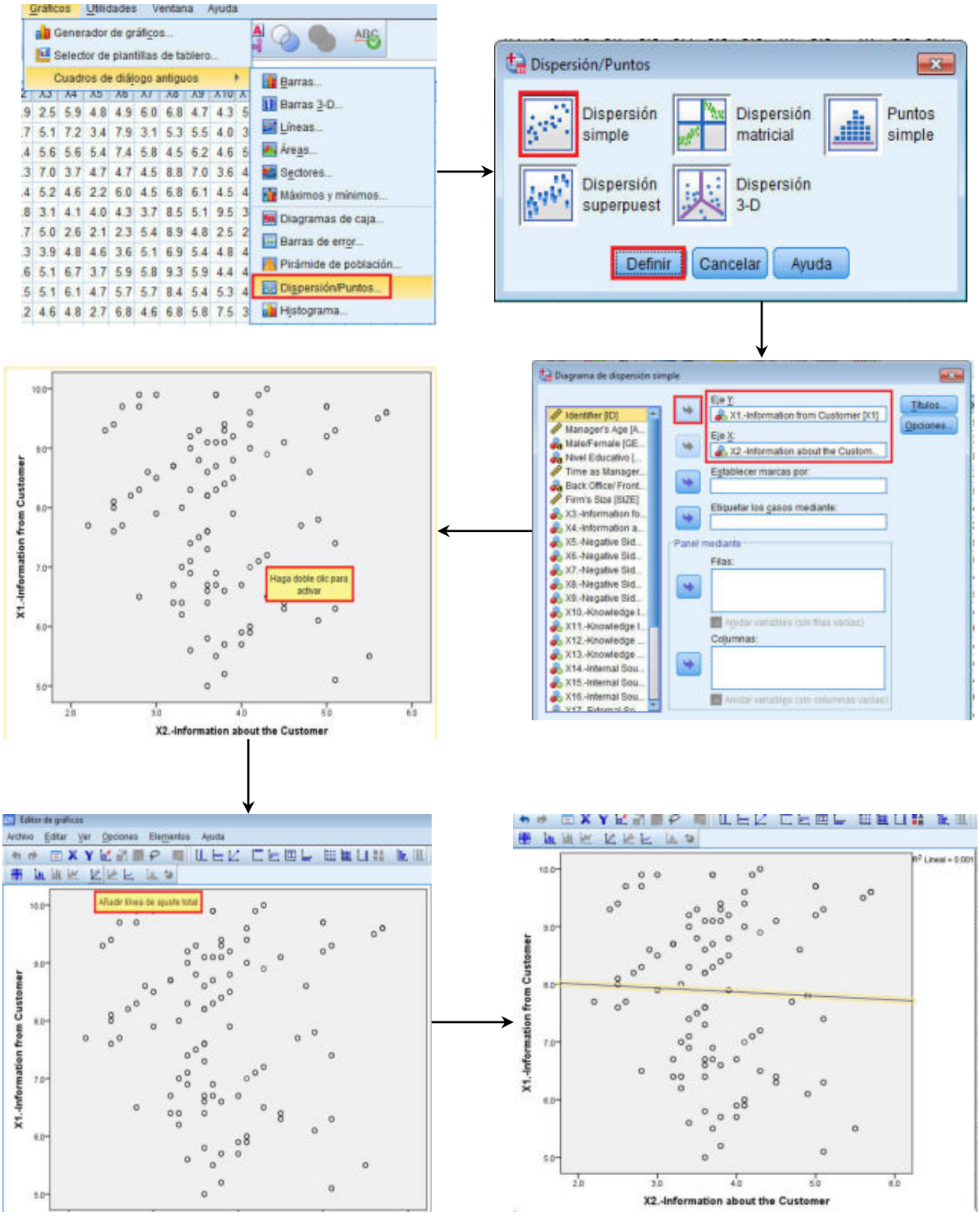
Así, cuando los puntos se organizan:

- A lo largo de una **línea recta** se dice que se obtiene **una relación lineal de correlación**.
- Un conjunto de **puntos curvados puede indicar una relación no lineal**, que se puede tratar de varias formas de acuerdo al punto de vista con el que se quiera tratar su linealidad.
- Si no existen patrones, es decir, sólo se muestra un conjunto de puntos aparentemente aleatorios, en este caso, **no hay relación**.

Para lograr los gráficos, **teclee:**

Gráficos->Cuadro de diálogo antiguos->Dispersión puntos->Dispersión simple->Definir->Selección de variables (X_1 vs X_2) Eje y; Eje x; Títulos; Aceptar->Hacer doble click en gráfico resultante->En modo de Editor gráficos, señalar Añadir línea de ajuste total->Aplicar…Ver **Figura 3.4.**

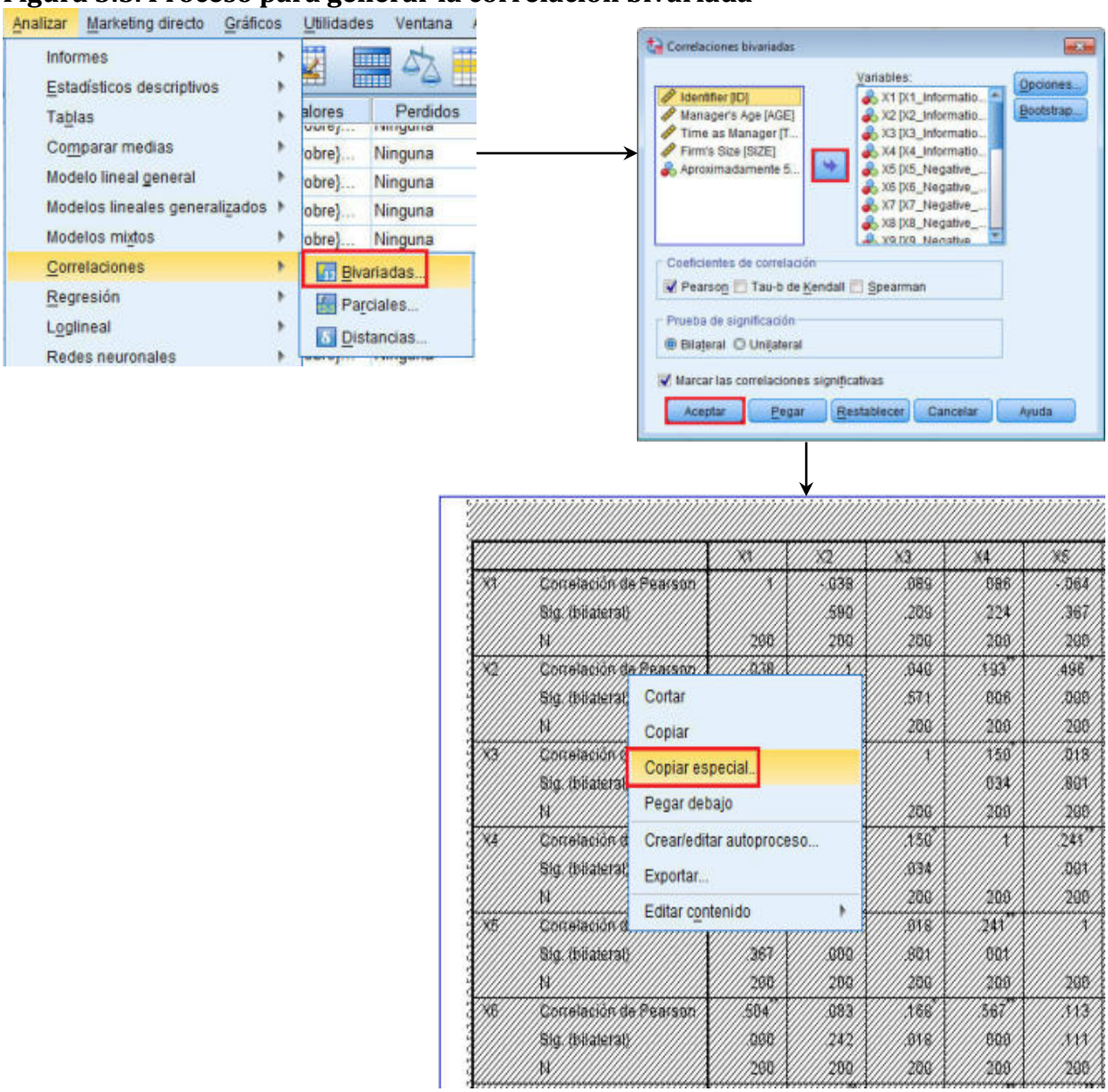
Figura 3.4. Gráficos de dispersión de variables métricas



Fuente: SPSS 20 IBM

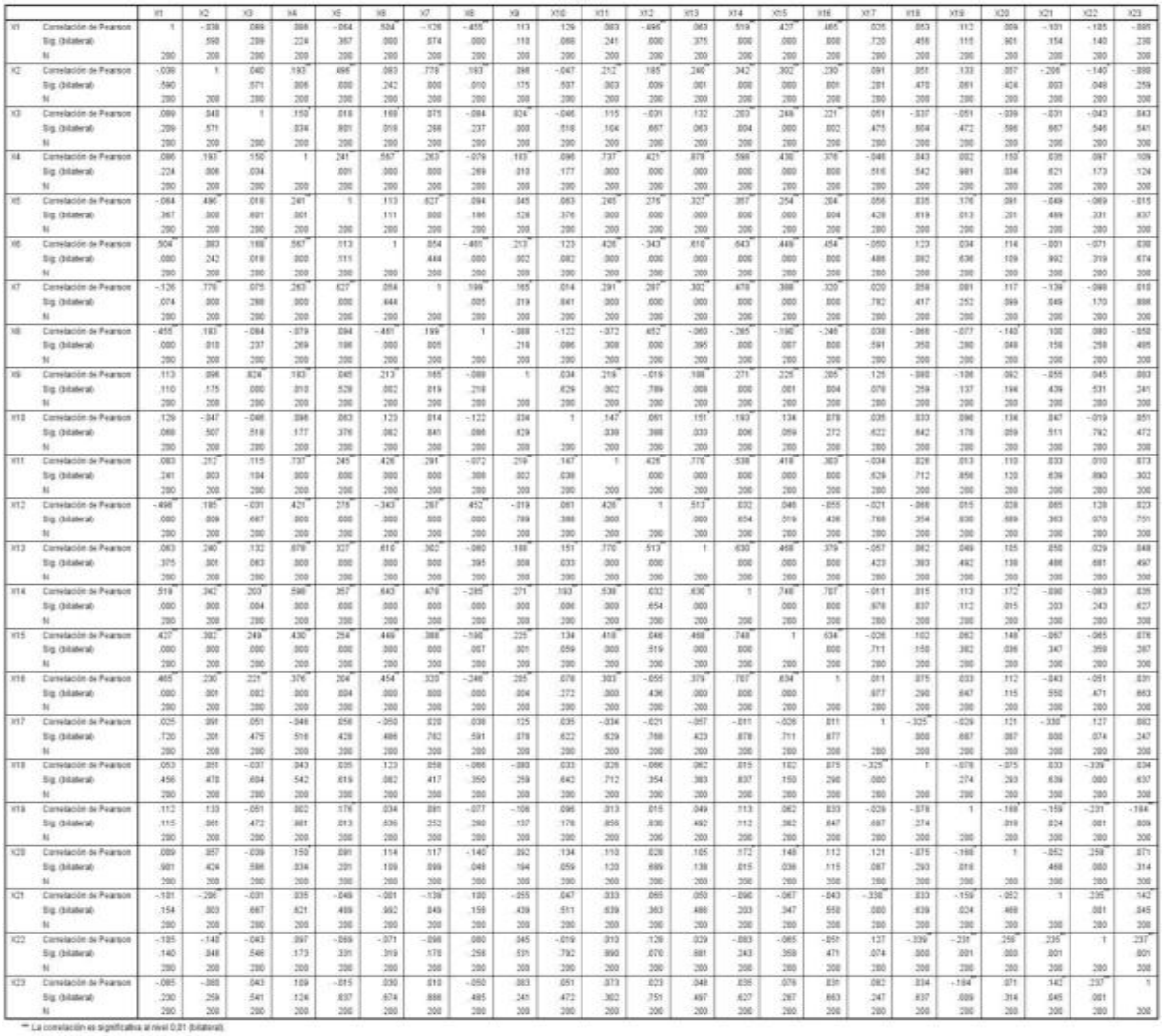
Para lograr los valores de correlación, teclee: **Analizar->Correlaciones->Bivariadas; Seleccionar variables; Aceptar; Copiar especial**, se producirá la Matriz de Correlación Ver **Figura 3.5 y 3.6**

Figura 3.5. Proceso para generar la correlación bivariada



Fuente: SPSS 20 IBM

Figura 3.6 Matriz de correlación



Correlaciones

		X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23
X1	Correlación de Pearson	1	-,038	,089	,086	-,064	,504**	-,426**	-,455**	,113	,129	,083	-,496**	,063	,519**	,427**	,485**	,025	,053	,112	,009	-,101	-,185**	-,095
	Sig. (bilateral)		,590	,209	,224	,367	,000	,000	,000	,110	,068	,241	,000	,375	,000	,000	,000	,720	,456	,115	,901	,154	,140	,238
	N	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280
X2	Correlación de Pearson	-,038	1	,040	,193**	,496**	,083	,778**	,193**	,098	-,047	,252**	,185**	,246**	,342**	,302**	,230**	,091	,051	,133	,057	-,206**	-,140**	-,098
	Sig. (bilateral)	,590		,571	,806	,000	,242	,000	,010	,175	,507	,063	,009	,001	,000	,000	,001	,281	,470	,061	,424	,003	,048	,259
	N	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280
X3	Correlación de Pearson	,089	,040	1	,150*	,018	,169**	,075	-,084	,024	-,046	,115	-,031	,132*	,283**	,248**	,221**	,051	-,037	-,051	-,039	-,031	-,043	,043
	Sig. (bilateral)	,209	,571		,034	,801	,019	,208	,237	,688	,518	,166	,667	,063	,804	,000	,002	,475	,604	,472	,596	,667	,546	,541
	N	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280
X4	Correlación de Pearson	,086	,193**	,150*	1	,241**	,567**	,263**	-,079	,193**	,096	,737**	,421**	,878**	,598**	,438**	,376**	-,048	,043	,032	,102	,035	,097	,109
	Sig. (bilateral)	,224	,006	,034		,001	,000	,000	,289	,010	,177	,000	,000	,000	,000	,000	,000	,516	,542	,981	,034	,621	,173	,124
	N	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280
X5	Correlación de Pearson	-,064	,496**	,018	,241**	1	,113	,627**	,094	,045	,063	,245**	,276**	,327**	,357**	,254**	,264**	,056	,035	,176**	,095	-,049	-,069	-,015
	Sig. (bilateral)	,367	,000	,801	,001		,111	,000	,196	,528	,376	,000	,000	,000	,000	,000	,000	,404	,428	,019	,013	,281	,489	,331	,837
	N	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280
X6	Correlación de Pearson	,504**	,083	,169**	,567**	,113	1	,054	-,465**	,253**	,123	,426**	-,343**	,610**	,643**	,449**	,454**	-,050	,123	,034	,114	-,001	-,071	,038
	Sig. (bilateral)	,000	,242	,019	,000	,111		,444	,000	,002	,082	,000	,000	,000	,000	,000	,000	,486	,082	,636	,109	,992	,319	,674
	N	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280
X7	Correlación de Pearson	-,126*	,778**	,075	,263**	,627**	,054	1	,199**	,165**	,014	,291**	,287**	,392**	,678**	,388**	,320**	,020	,058	,091	,117	-,139*	-,099	,010
	Sig. (bilateral)	,074	,000	,208	,000	,000	,444		,005	,019	,841	,000	,000	,000	,000	,000	,000	,782	,417	,252	,099	,049	,170	,868
	N	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280
X8	Correlación de Pearson	-,455**	,193**	-,084	-,079	,094	-,465**	,199**	1	-,088	-,122*	-,072	,652**	-,060	-,265**	-,196**	-,246**	,038	-,068	-,077	-,140*	,100	,080	-,052
	Sig. (bilateral)	,000	,010	,237	,269	,196	,000	,005		,218	,096	,308	,000	,395	,000	,067	,000	,591	,358	,280	,049	,158	,259	,495
	N	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280
X9	Correlación de Pearson	,113	,098	,024	,193**	,045	,213**	,165**	-,088	1	,034	,219**	-,019	,198**	,271**	,226**	,285**	,125*	-,068	-,106	,062	-,055	,045	,093
	Sig. (bilateral)	,110	,175	,688	,010	,528	,082	,019	,218		,629	,002	,789	,008	,000	,001	,000	,079	,259	,137	,194	,439	,531	,241
	N	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280
X10	Correlación de Pearson	,129*	-,047	-,046	,096	,063	,123*	,014	-,122*	,034	1	,147*	,091	,151*	,193**	,134*	,078	,035	,033	,096	,136*	,047	-,019	,051
	Sig. (bilateral)	,069	,507	,518	,177	,376	,082	,841	,096	,629		,038	,388	,033	,006	,059	,272	,622	,642	,179	,059	,511	,762	,472
	N	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280
X11	Correlación de Pearson	,083	,252**	,115	,737**	,245**	,426**	,291**	-,072	,219**	,147*	1	,426**	,770**	,539**	,418**	,383**	-,034	,026	,013	,110	,033	,010	,073
	Sig. (bilateral)	,241	,003	,104	,000	,000	,000	,000	,308	,002	,038		,000	,000	,000	,000	,000	,629	,712	,856	,120	,639	,890	,302
	N	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280
X12	Correlación de Pearson	-,496**	,185**	-,031	,421**	,276**	-,343**	,287**	,652**	-,019	,091	,426**	1	,513**	,632**	,046	-,055	-,021	-,068	,015	,028	,065	,128*	,023
	Sig. (bilateral)	,000	,009	,667	,000	,000	,000	,000	,000	,789	,388	,000		,000	,654	,519	,436	,768	,254	,830	,689	,363	,070	,701
	N	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280
X13	Correlación de Pearson	,063	,246**	,132*	,678**	,327**	,610**	,392**	-,060	,198**	,151*	,770**	,513**	1	,630**	,468**	,379**	-,057	,062	,049	,105	,050	,029	,048
	Sig. (bilateral)	,375	,001	,063	,000	,000	,000	,000	,395	,008	,033	,000	,000		,000	,000	,000	,423	,303	,492	,139	,486	,681	,497
	N	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280
X14	Correlación de Pearson	,519**	,342**	,283**	,598**	,357**	,643**	,678**	-,285**	,271**	,193**	,539**	,632**	,630**	1	,748**	,787**	-,041	,015	,113	,172**	-,096	-,083	,035
	Sig. (bilateral)	,000	,000	,004	,000	,000	,000	,000	,000	,000	,006	,000	,000	,000		,000	,000	,678	,837	,112	,015	,203	,243	,627
	N	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280
X15	Correlación de Pearson	,427**	,302**	,248**	,438**	,254**	,449**	,388**	-,196**	,226**	,134*	,418**	,046	,468**	,748**	1	,634**	-,026	,102	,062	,149*	-,067	-,065	,076
	Sig. (bilateral)	,000	,000	,000	,000	,000	,000	,000	,001	,000	,059	,000	,519	,000	,000		,000	,711	,158	,382	,036	,347	,359	,267
	N	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280
X16	Correlación de Pearson	,485**	,230**	,221**	,376**	,264**	,454**	,320**	-,246**	,285**	,078	,383**	-,055	,379**	,787**	,634**	1	,051	,075	,033	,112	-,043	-,051	,031
	Sig. (bilateral)	,000	,001	,002	,000	,004	,000	,000	,000	,000	,272	,000	,436	,000	,000	,000		,677	,290	,647	,115	,550	,471	,663
	N	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280
X17	Correlación de Pearson	,025	,091	,051	-,048	,056	-,050	,020	,038	,125*	,035	-,034	-,021	-,057	-,041	-,026	,051	1	-,325**	-,029	,121	-,336**	,127*	,092
	Sig. (bilateral)	,720	,201	,475	,516	,428	,486	,742	,591	,078	,622	,629	,768	,423	,678	,711	,677		,000	,687	,087	,000	,074	,247
	N	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280
X18	Correlación de Pearson	,053	,051	-,037	,043	,035	,123*	,058	-,068	-,068	,033	,026	-,068	,062	,015	,102	,075	-,325**	1	-,078	-,075	,033	-,339**	,034
	Sig. (bilateral)	,456	,473	,604	,542	,619	,082	,417	,350	,259	,642	,712	,354	,383	,837	,150	,290	,000		,274	,293	,639	,000	,637
	N	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280
X19	Correlación de Pearson	,112	,133*	-,051	,032	,176**	,034	,091	-,077	-,106	,096	,013	,015	,049	,113	,062	,033	-,029	-,078	1	-,168**	-,159**	-,231**	-,184**
	Sig. (bilateral)	,115	,061	,472	,681	,013	,636	,252	,280	,137	,178	,856	,830	,492	,712	,282	,647	,687	,274		,010	,024	,001	,009
	N	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280
X20	Correlación de Pearson	,009	,057	-,039	,102	,095	,114	,117	-,140*	,062	,136*	,110	,028	,105	,172**	,149*	,112	,121	-,075	-,168**	1	-,052	,258**	,071
	Sig. (bilateral)	,901	,424	,596	,034	,231	,109	,049	,049	,194	,059	,120	,689	,139	,015	,036	,115	,087	,293	,010		,468	,000	,314
	N	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280
X21	Correlación de Pearson	-,101	-,206**	-,031	,035	-,049	-,001	-,139*	,100	-,055	,047	,033	,065	,050	-,096	-,067	-,043	-,336**	,033	-,159**	-,052	1	,235**	,142*
	Sig. (bilateral)	,154	,003	,667	,621	,409	,992	,049	,158	,439	,511	,639	,363	,486	,203	,347	,550	,000	,639	,024	,468		,001	,045
	N	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280
X22	Correlación de Pearson	-,185**	-,140*	-,043	,097	-,069	-,071	-,099	,080	,045	-,019	,010	,128*	,029	-,083	-,065	-,051	,127*	-,339**	-,231**	,258**	,235**	1	,237**
	Sig. (bilateral)	,140	,048	,546	,173	,331	,319	,170	,258	,531	,762	,890	,070	,681	,243	,359	,471	,074	,000	,001	,000	,001		,001
	N	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280
X23	Correlación de Pearson	-,095	-,098	,043	,109	-,015	,038	,010	-,052	,093	,051	,073	,023	,048	,035	,076	,031	,092	,034	-,184**	,071	,142*	,237**	1
	Sig. (bilateral)	,230	,259	,541	,124	,837	,674	,868	,495	,241	,472	,302	,701	,497	,627	,287	,663	,247	,837	,009	,314	,045	,001	
	N	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280	280

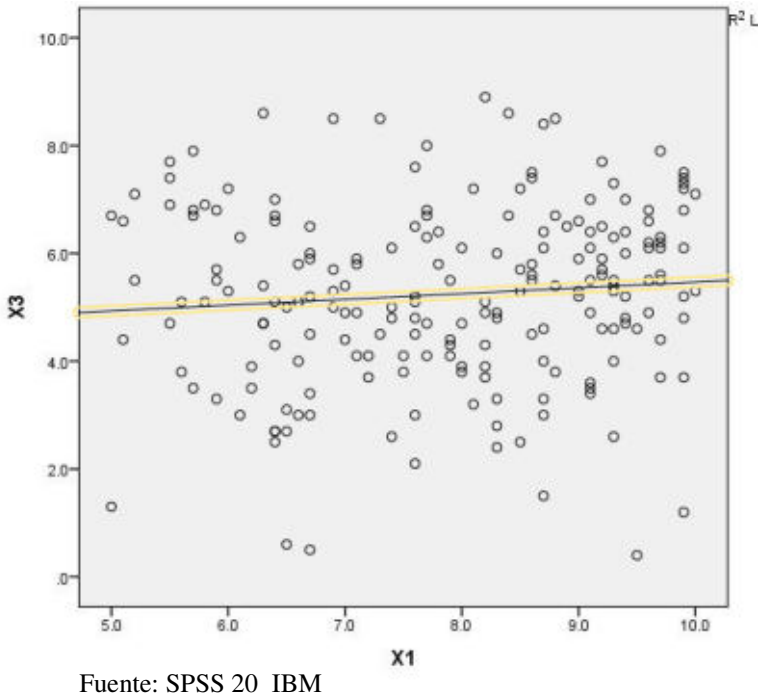
** La correlación es significativa al nivel 0,01 (bilateral).

* La correlación es significante al nivel 0,05 (bilateral).

Fuente: SPSS 20 IBM

La matriz de correlaciones tiene la gran ventaja de apreciar todos los resultados en una sola vista ya que el realizar tablas de dispersión por cada uno delos cruces, resulta muy tedioso para el estudio. Por ejemplo, **teclee: Gráficos->Cuadro de diálogo antiguos->Dispersión puntos->Dispersión simple->Definir->Selección de variables (X_1 vs X_3) Eje y; Eje x; Títulos; Aceptar->Hacer doble click en gráfico resultante->En modo de Editor gráficos, señalar Añadir línea de ajuste total->Aplicar…**Ver **Figura 3.7**

Figura 3.7. Gráfico de dispersión con línea de ajuste total variable X_1 vs. X_3



Fuente: SPSS 20 IBM

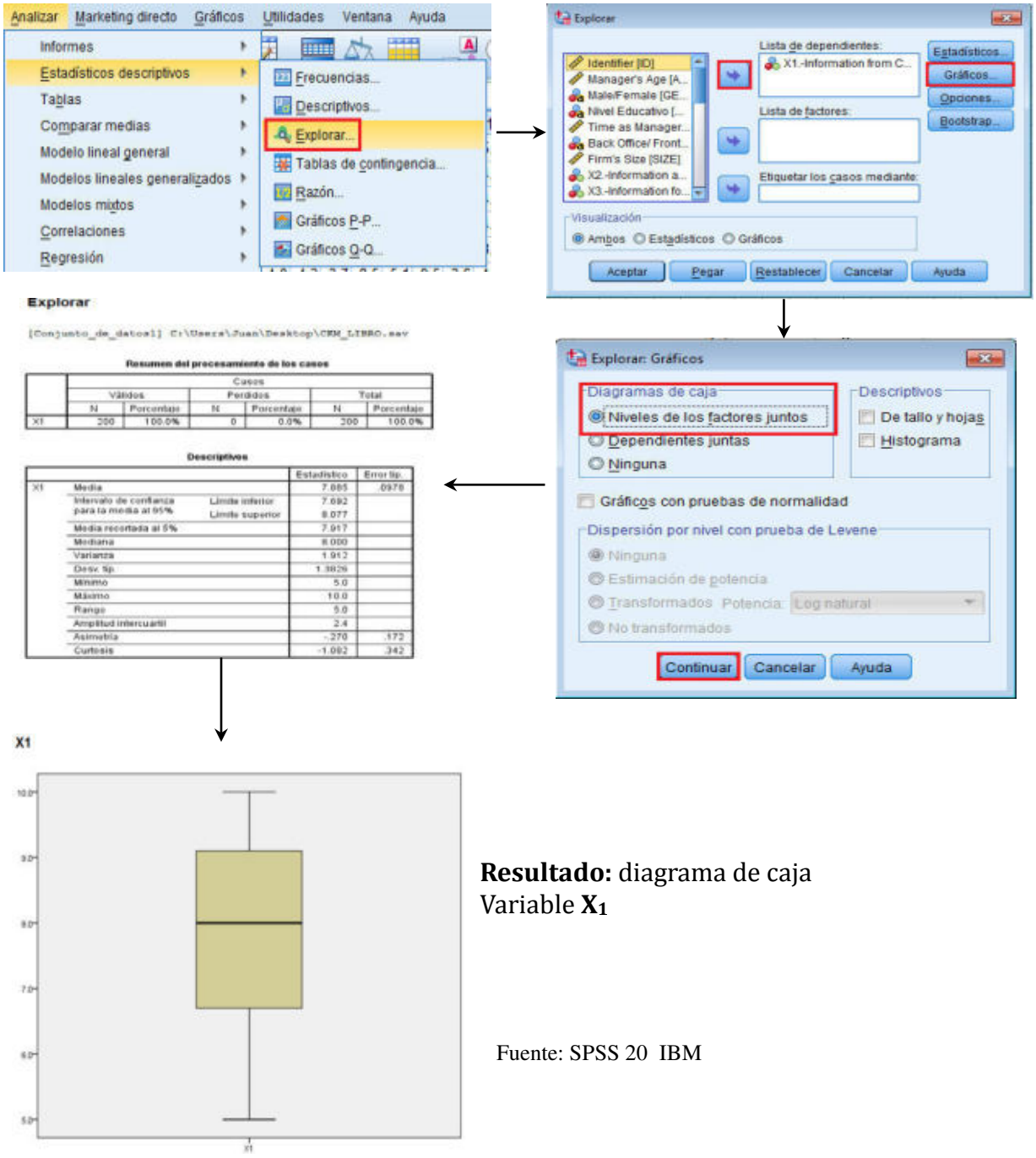
3.2.4. Análisis de las diferencias entre grupos

Otro escenario que deberá enfrentar en su proceso de investigación es el entender el carácter y la diferencia entre dos o más grupos de una variable para **dos o más variables métricas**, tal y como ocurre en el **análisis discriminante, análisis de la varianza y análisis multivariante de la varianza**. En estos casos es importante entender **cómo se encuentran distribuidos los valores para cada grupo y determinar si son suficientes las diferencias entre ellos como para tener significación estadística.**

Por otro lado, se deben identificar los **casos atípicos que pueden resultar ser aparentes sólo cuando los valores de los datos se separan en grupos**. Un método gráfico utilizado y muy popular es el de **cajas *(boxplot*)**, a nivel de distribución de los datos. Los **límites superior e inferior** de la caja marcan los **cuartiles superior e inferior de la distribución de los datos.** Por tanto, la longitud de la caja es la **distancia entre el primer y el tercer cuartil**, de forma que la caja contiene el **50 por ciento de los datos centrales de la distribución**. La línea dentro de la caja señala la posición de la **mediana**. Si ésta cae cerca del final de la caja, se indica la presencia de **asimetría**. Cuanto mayor sea la caja, mayor es la extensión de las observaciones. Las líneas que se extienden desde cada caja (**bigotes**) representan la distancia entre la mayor y la menor de las observaciones que están a menos de un cuartil de la caja y se marcan con una **X**. Los **casos atípicos** son observaciones que se sitúan entre **1.0 y 1.5 cuartiles** fuera de la caja. Los **valores extremos** son aquellas **mayores que 1.5** cuartiles fuera de los límites de la caja. (IBM,2011).

-Problema 2: genere gráfico de cajas para la variable **X_1,** de la base de datos de **CKM_MKT_Digital.sav.**
-Teclear: Analizar->Estadísticos descriptivos->Explorar->Selección variables (X_1) ->Gráficos->En Diagramas de caja seleccionar Niveles de los factores juntos; Continuar; Aceptar (ver **Figura 3.8**).

Figura 3.8. Proceso para generación de gráfico de caja variable X_1



Resultado: diagrama de caja
Variable **X_1**

Fuente: SPSS 20 IBM

3.3. Datos ausentes

Son muy habituales en el análisis multivariante (IBM,2011) ; rara vez el investigador evita enfrentarse a este problema. El desafío consiste en enfrentarse a los resultados producidos por esta causa en los procesos de estimación y que **afecten a la generalidad de los resultados**. Para hacer esto, la ocupación inicial básica a realizar **es determinar las razones que subyacen en el dato ausente**, siendo en muchos casos la extensión de la ausencia de datos una **cuestión secundaria**. Es muy importante detectar la causa de la ausencia de datos a fin de tomar las decisiones pertinentes para su corrección. Los **datos ausentes** son generados por:

- **Acciones externas al encuestado** por diversos motivos tales como: errores en la captura de los datos, problemas en su recolección, el encuestador no es claro en la toma de datos, etc. Cuando se descubren con oportunidad, existe un cierto control de los efectos que producen por parte del investigador.
- **Acciones internas por parte del encuestado,** tales como: rehusarse a contestar, sesgar la respuesta de la encuesta, o mostrar respuestas sin mayor compromiso que librarse de la encuesta, motivaciones diversas, etc. Son difíciles de detectar por parte del investigador por lo que los efectos no son previsibles, por lo que las acciones de corrección, se limitan a identificar patrones en los datos ausentes que caracterizarían dicho proceso, a través de plantear cuestiones como:
  -¿Se pueden identificar uno o distintos patrones?
  -¿Están los datos ausentes distribuidos de forma aleatoria dentro del cuestionario?
  -¿En qué medida son relevantes?

Si se determinan pautas y la extensión de los datos ausentes es suficiente como para garantizar un curso de acción, entonces se asume que está operando algún proceso de ausencia de datos y que algunos de los resultados estadísticos podrían estar sesgados en la medida en que las variables incluidas en el análisis están influidas por los procesos de pérdida de datos.

Las causas de los procesos tanto de la **ausencia de datos como la ausencia de respuestas** en la recolección son de similar interés, ya que cabe preguntarse:

- ¿Son diferentes las personas que no respondieron de las personas que sí lo hicieron? Si es así:

-¿Qué impacto producirán las diferencias en el análisis, los resultados o su interpretación?

Como se observa el impacto de los **datos ausentes** es perjudicial no sólo por sus potenciales ***sesgos "ocultos"*** sino también por su **efecto en el tamaño de la muestra** para el análisis. **De no aplicarse soluciones corregirlo, ninguna observación con datos ausentes deberá incluirse en el análisis**. Esto reduce la muestra a grado tal que la puede hacer inadecuada. De ser así, el investigador deberá buscar observaciones adicionales o encontrar una solución para la ausencia de datos en la muestra original. Estas acciones, aunque prácticas están soportadas mínimamente en pocas guías para el diagnóstico y solución de la ausencia de datos. Por esta razón, se presentarán algunos tipos de procesos de ausencia de datos, métodos para identificar su naturaleza y las soluciones existentes para dar cabida a la ausencia de datos en el análisis multivariante. La **Figura 3.9** contiene un ejemplo sencillo de datos ausentes entre **10 casos**.

Figura 3.9.-Tabla con datos ausentes

Caso	**Variables**					**Datos Ausentes por Caso**	
ID	**X_1**	**X_2**	**X_3**	**X_4**	**X_5**	**No.**	**%**
1	9.2	2.2	3.3	4.3	5.6	0	0
2	8.7	1.5				3	60
3		2.5			6.4	3	60
4	3.3	5.5		4.9	9.6	1	20
5	4.5	1.9			3.3	2	40
6	4.6.	9.5	4.6	8.2	9.3	0	0
7		6.7		8.8	8.3	2	40
8	5.6	9.7	5.5	9.8		1	20
9	9.8				8.4	3	60
10	3.7	3.9	9.4	9.6	6.7	0	0
No.	2	1	6	4	2	**15**	Total Valores Ausentes
%	20	10	60	30	20	**30**	
Datos Ausentes por Variable							

Fuente: propia

En la investigación basada en encuestas, el número de datos ausentes varía mucho entre los **casos** y las **variables.** En el ejemplo, podemos ver que todas las **variables** (**X_1 a X_5**) tienen algunos datos ausentes, en **X_3**, falta más de la mitad (**60%**) de todos los valores. Así también **3 casos** (**2,3 y 9** tienen más del **50 %** de datos ausentes (**60%**) y sólo **3** casos tienen datos completos. En conjunto, un **30%** de los valores están ausentes. Si se hiciera un análisis multivariante que necesitara datos completos, los datos se verían reducidos a solamente **3 casos,** que para cualquier tipo de análisis es muy pobre en aportación. Este nivel de reducción en los casos disponibles es frecuente en muchas aplicaciones y se considera como primera opción. Las soluciones más sofisticadas para se abordarán más adelante y con detalle. No obstante, otra siguiente opción es la **eliminación de las variables y/o casos**. En el ejemplo, si suponemos que el constructo no se altera sustancialmente con la **supresión de una variable**, la eliminación de **X_3** es una manera de reducir el número de datos ausentes. Con esto incrementaría a **5 casos** con información completa. Si se eliminan los tres casos (**2, 3 y 9**) con cantidades de datos ausentes excepcionalmente altas, el número total de datos ausentes se reduce ahora a solamente **6 casos**, No obstante, **4/6** datos ausentes de **X_3** están en **X_4** presentes en **X_2** , del que se puede establecer un patrón de datos de valores bajos posibles a sustituir. Esta asociación sistemática entre los **datos ausentes** y **datos válidos** tiene un impacto directo sobre cualquier análisis en los que se incluyen **X_4** y **X_2** para evaluar el impacto del proceso de datos ausentes sobre los resultados. Antes de planear cualquier solución contra la ausencia de datos, debe diagnosticar y comprender los procesos que los motivan y que subyacen en este fenómeno. En las ocasiones en que estos procesos se encuentran bajo control del investigador y se identifican explícitamente. Se les denomina prescindible, lo que significa que no se necesitan

soluciones específicas para la ausencia de datos dado que los límites de la ausencia de los datos son inherentes a la técnica usada **[**Little et al. 1987**].**

3.3.1. Datos ausentes prescindibles

Ejemplos de **proceso de datos ausentes prescindibles son** (IBM, 2011) **:**

1. **El dato ausente** de aquellas observaciones de una población no incluidas en la muestra. Recuerde que el propósito de la técnica multivariante es la generalización de las observaciones de la muestra al conjunto de la población; esto implica un gran esfuerzo por salvar los datos ausentes de las observaciones que no están en la muestra. Debe hacer prescindibles estos datos ausentes mediante el uso de una muestra probabilística de los encuestados seleccionados ya que esto permite especificar los procesos de datos ausentes causantes de las observaciones omitidas de forma aleatoria y que los datos ausentes pueden explicarse como un error muestral de los procedimientos estadísticos. Así, los **"*datos ausentes*"** de las observaciones no seleccionadas son prescindibles.
2. **Los datos están censurados**, los cuales son **observaciones incompletas** como consecuencia de su etapa en el proceso de ausencia de datos, como el caso del **análisis de las causas de fallecimiento**. Es claro que los encuestados que todavía viven no pueden proporcionar información completa (es decir, la causa de su muerte) y por tanto están censurados.

   Justificar los datos ausentes como prescindibles implica que el proceso está operando aleatoriamente y que los efectos son identificables y explícitamente ajustados a la técnica usada. Sin embargo, en muchos casos, esto no es así por lo que debe evaluar la medida y el impacto en que los datos ausentes determinan si es un proceso aleatorio o, en caso contrario, si se puede remediar con alguna de las soluciones existentes.

3.3.2. Más tipos de procesos de ausencia de datos

La ausencia de datos (IBM. 2011) ocurre por muchas razones y en muchas situaciones, tales como:

1. El que ocurre en cualquier situación y se debe a **factores de procedimiento**, tales como:

   -Errores en la entrada de datos que crean códigos inválidos,

   -Restricciones de representatividad como, los datos de las poblaciones pequeños en censos mundiales

   -Fallos al completar el cuestionario o incluso la morbidez del encuestado. En estas situaciones, el investigador tiene escaso control sobre los procesos de ausencia de datos, aunque pueden aplicarse ciertas soluciones si se encuentra que los datos ausentes son de **carácter aleatorio**.
2. Otro tipo ocurre cuando **la respuesta es inaplicable**, como las preguntas en relación a los años de matrimonio para adultos que nunca han estado casados. Así, los análisis pueden ser específicamente formulados para acomodar a estos encuestados.
3. Otros tipos se identifican y manipulan con menos facilidad, siendo la mayoría relacionados directamente con el encuestado, tales como el **desistimiento o renuncia** del mismo a responder a ciertas **preguntas de carácter sensible** como el nivel y/o fuente principal de ingresos, o cuando el encuestado no tiene opinión o conocimiento suficiente para contestar la pregunta.

De esta manera, se debe anticipar para minimizarlos en el diseño de la investigación y en la recopilación de datos. Sin embargo, puede ocurrir muy bien que el investigador deba enfrentarse con los datos ausentes resultantes que si ocurren siguiendo una pauta aleatoria, existen soluciones para mitigar sus efectos.

3.3.3. Examen de los tipos de datos ausentes

A fin de determinar el tipo de solución para la ausencia de datos (IBM, 2011), en primer lugar, **debe averiguar el grado de aleatoriedad** de los datos ausentes, de acuerdo a los siguientes pasos :

1. **Datos ausentes que provienen de un proceso no aleatorio.** Esto sucede cuando se tienen dos variables ($\mathbf{X_1}$ y $\mathbf{X_2}$), con $\mathbf{X_1}$ sin datos perdidos, $\mathbf{X_2}$ con algunos datos ausentes. Si en el análisis de las variables se encuentra **un proceso de datos ausentes** entre ambas, donde existen diferencias significativas para casos de $\mathbf{X_2}$ con datos válidos y datos ausentes en función de los valores de $\mathbf{X_1}$.Cualquier análisis tiene que comprobar explícitamente los procesos de datos ausentes entre $\mathbf{X_1}$ y $\mathbf{X_2}$ o si no se introduce sesgo en los resultados.
2. **Datos ausentes que provienen de un proceso aleatorio (MAR= *missing completely at random*).** Si los valores ausentes de $\mathbf{X_2}$ dependen de $\boldsymbol{X_1}$, pero no en $\mathbf{X_2}$. Esto es, los valores observados de $\mathbf{X_2}$ representan una muestra de los valores reales de $\mathbf{X_2}$ para cada valor de $\boldsymbol{X_1}$, pero los datos observados para $\mathbf{X_2}$ no representan necesariamente una **muestra verdaderamente aleatoria** para todos los valores de $\mathbf{X_2}$. Aunque el proceso de datos ausentes es aleatorio en la muestra, **sus valores no son generalizables para la población**. Por ejemplo, supongamos que conocemos tenemos las respuestas de 2 tipos de gerentes, tanto de la empresa pública como privada ($\mathbf{X_1}$), al cuestionarles sobre los ingresos financieros anuales ($\mathbf{X_2}$), al requerirles informen de las cantidades asignadas encontramos que los datos ausentes son aleatorios para ambos sexos pero que ocurren con mayor frecuencia para los de la empresa privada que la pública ($\mathbf{X_1}$). Mientras que el proceso de ausencia de datos está operando de forma aleatoria, cualquier solución aplicada a los datos ausentes debe tener en cuenta tipo de gerencia de empresa de los encuestados porque afecta a la distribución definitiva de los valores de ingresos financieros ($\mathbf{X_2}$).
3. Un mayor nivel de aleatoriedad ya es considerado **proceso completamente aleatorio (MCAR = *missing completely random*)**, por lo que los valores observados de $\mathbf{X_1}$ son verdaderamente una muestra aleatoria de todos los valores de $\mathbf{X_1}$, sin un proceso subyacente que tienda a sesgar los datos observados. En el ejemplo anterior, esto se mostraría por el hecho de que los datos para los ingresos financieros estén aleatoriamente ausentes en la misma proporción tanto gerentes públicos como privados. Si esta es la forma del proceso de ausencia de datos, cualquier solución se podría aplicar sin tener en cuenta el impacto de cualquier otra variable o proceso de datos ausentes.

3.3.4 El diagnóstico de la aleatoriedad en el proceso de pérdida de observaciones

Se recomienda considerar **3** métodos para realizar el diagnóstico:

1. Valorar los datos ausentes para una única variable $\mathbf{X_1}$ al formar dos grupos de observaciones con datos ausentes para $\mathbf{X_1}$ y aquellos con valores válidos de $\mathbf{X_1}$. Se realizan entonces los test para determinar si existen diferencias significativas entre los

dos grupos sobre otras variables de interés. Si se encuentran patrones de diferencias significativas, indicaría que existe un **proceso de pérdida de datos no aleatorio**. De nuestro ejemplo, en primer lugar formaríamos **2** grupos de encuestados, aquellos con datos ausentes en la pregunta sobre ingresos financieros y aquellos que responden a la pregunta. Se procede a entonces a comparar los porcentajes de acuerdo al tipo de gerencia. Si el tipo de gerencia (empresa privada) se encontrara en mayor proporción en el grupo de datos ausentes, sospecharíamos que el proceso **no ha operado de forma aleatoria**. Si la variable que se está comparando fuese **métrica** (por ejemplo, una actitud o percepción) en lugar de **categórica** (género), entonces el apropiado es la **prueba *t***. Así, Usted deberá examinar un número de variables para ver si surge cualquier tipo de **patrón consistente**. Se debe tomar en cuenta y recordar, que ciertas diferencias u ocurren por azar, o por una serie de diferencias con un patrón subyacente.

2. Otra aproximación, consiste en utilizar las **correlaciones dicotomizadas** para evaluar la correlación de los datos ausentes en cualquier par de valores. Los valores válidos, de cada variable se representan por el valor **1**, mientras que los datos ausentes son reemplazados por valores de **0**, los cuales entonces se correlacionan. Éstas indican el grado de asociación entre los valores perdidos sobre cada par de variables. Bajas correlaciones implican aleatoriedad en el par de variables. No existen guías o recomendaciones que identifiquen el nivel de correlación que indique si los datos ausentes no son aleatorios, los test de significación estadística de las correlaciones proporcionan una estimación conservadora del grado de aleatoriedad. Si la aleatoriedad es indicativa para todos los pares de variables, entonces puede suponer que los datos ausentes pueden clasificarse como **proceso completamente aleatorio.** Si existen correlaciones significativas entre algunos pares de variables, entonces puede suponer que los datos son sólo **MAR (*missing completely at random*)** y estas relaciones deben ser tenidas en cuenta en cualquier solución que se quiera aplicar.
3. Finalmente, es posible realizar un test conjunto de aleatoriedad que determine si los datos ausentes pueden ser clasificados como **MCAR (*missing completely at random*)** que analiza el patrón de datos ausentes sobre todas las variables y las compara con el patrón esperado para un proceso de datos ausentes aleatorio. Si no se encuentran diferencias significativas, los datos ausentes pueden ser clasificados como **MCAR (*missing completely at random*)**. Si se encuentran diferencias significativas, sin embargo, el investigador debe usar las aproximaciones descritas más arriba para identificar los procesos específicos de datos ausentes que no son aleatorios.

3.4. Aproximaciones al tratamiento de datos ausentes

Es importante seguir las recomendaciones (Hair et al., 1999), los cuales son:

3.4.1.El diagnóstico de la aleatoriedad en el proceso de pérdida de observaciones

Se clasifican en **4** categorías y están basadas en la aleatoriedad de los procesos de datos ausentes, en función del método empleado para estimarlos **[**Little et al. 1987**].** Si se encuentran procesos de datos ausentes **MAR (*missing completely at random*)** o no aleatorios, deberá aplicar sólo el método diseñado específicamente para este proceso **[**Little et al. 1987.**]**. La aplicación de cualquier otro método introduce sesgos en los resultados. Sólo

si el investigador de termina que el proceso de ausencia de datos puede clasificarse como **MCAR (*missing completely at random*)** pueden utilizarse todas las aproximaciones discutidas en las siguientes secciones.
No obstante, muchas veces los investigadores evalúan la aleatoriedad de los datos ausentes antes de aplicar una de las soluciones de datos ausentes. Incluso si la solución es la apropiada, el investigador debe tener en cuenta los impactos específicos de los resultados asociados con ella. Con mucha frecuencia, se aplica una solución sin una evaluación de los procesos de ausencia de da tos, la conveniencia de la solución seleccionada o las consecuencias que tendrá. En tal caso, nunca se dará cuenta de los efectos porque están camuflados bajo los resultados generales:

3.4.2. Utilizar sólo aquellas observaciones con datos completos

El tratamiento más simple y directo es la de incluir **sólo aquellas observaciones con datos completos**, (**aproximación de casos completos)**. Se encuentra en todos los programas estadísticos y es el método por defecto en muchos programas. No obstante, esta aproximación debería usarse sólo si los datos ausentes son **MCAR *(missing completely at random*)** ya que **los datos ausentes que no lo son tienen elementos no aleatorios que sesgarían los resultados**. Por tanto, incluso aunque sólo se usen observaciones válidas, **los resultados no son generalizables para la población**. Más aún, en muchas situaciones, el **tamaño de la muestra** resultante queda reducida a una **muestra inapropiada** para el análisis. Ésta aproximación se ajusta mejor a casos en los que la **extensión de la ausencia de datos es pequeña,** en los que la **muestra es suficientemente grande** para permitir la supresión de los casos con los datos ausentes y en los que las relaciones entre los datos son tan fuertes que no pueden verse afectadas por cualquier proceso de datos ausentes.

3.4.3.Supresión de caso(s) y/o variable(s)

Es otra solución simple que consiste en **suprimir el caso(s) y/o variable(s) que peor se comporta(n)** respecto a los datos ausentes. En ésta categoría se determina la extensión de los datos ausentes sobre cada caso y variable y entonces **destruye** los casos y variables **que exceden el nivel especificado**. En muchos casos donde se presenta un **patrón de datos no aleatorio**, puede constituir la **solución más eficiente.** Usted encontrará que los datos ausentes están concentrados en un pequeño subconjunto de **casos y/o variables**, con cuya exclusión se reduce sustancialmente la extensión de los datos ausentes. Al momento, no existen guías suficientes para el nivel de exclusión necesario, pero cualquier decisión deberá basarse tanto en consideraciones **empíricas como teóricas**. Si se encuentran los valores ausentes para lo que será una **variable dependiente** en el análisis propuesto, **habitualmente se excluye el caso**. Con esto, se evita cualquier **incremento artificial** en el poder explicativo del análisis que pudiera ocurrir cuando estime en primer lugar los datos ausentes para la **variable dependiente** por uno de los **procesos de imputación** a describir a continuación y después usar los valores estimados en el análisis de las relaciones de dependencia. Si una variable que **no sea la dependiente** tiene valores ausentes y es una candidata a la eliminación, el investigador debe asegurarse de que **existan variables alternativas**, con la expectativa de que estén altamente correlacionadas, para representar la intención de la variable original. Deberá siempre considerar lo que se

gana al eliminar una fuente de datos ausentes y lo que se pierde al no contar con una determinada variable en el análisis multivariante.

3.4.4.Métodos de imputación

Ésta categoría se realiza a través de uno de los muchos **métodos de imputación** (Hair et al. 1999, IBM, 2011), que es el proceso de estimación de valores ausentes **basado en valores válidos de otras variables y/o casos de la muestra**. El objetivo se consigue empleando relaciones conocidas que se identifican en los valores válidos de la muestra y que ayudan en la estimación de valores ausentes. Sin embargo, deberá considerar cuidadosamente el uso de la imputación en cada instancia, dados sus **potenciales impactos sobre el análisis [**Dempster y Rubín, 1983**].** Se dice que es de doble filo su implementación ya que por un lado existe la posibilidad de llevarlo a creer que los datos están completos después de todo, y es por el otro lado, tiene también la posibilidad de unir situaciones donde el problema es suficientemente menor con situaciones donde los estimadores estándar aplicados a los datos reales e imputados tienen sesgos sustanciales.

Los métodos siguientes, se recomienda utilizarlos con **variables métricas** debido a:

1. Para **variables métricas**, se pueden hacer estimaciones de los datos ausentes con valores como una **media de todos los valores válidos.**
2. Para **variables no métricas** se requiere una estimación de un valor específico en vez de una estimación en una escala continua. Existe mucha diferencia entre estimar un valor ausente para una **variable métrica**, tal como la disposición por crear innovaciones o una percepción del uso de la tecnología e incluso los ingresos por innovaciones, que estimar el género del encuestado cuando este dato está ausente. Por tanto, las **variables no métricas no se logran típicamente mediante el proceso de imputación**, sino que requieren la aproximación de **la modelización** específica abordada en la siguiente sección o se omiten por estar ausentes.

Los métodos de **imputación** pertenecen a dos tipos:

1. El uso de toda la información disponible a partir de un subconjunto de casos para generalizar sobre la muestra entera, o

2. Como métodos para estimar valores de reemplazo para los datos ausentes que, de esta forma, se analizan mediante técnicas multivariantes estándar.

3.4.5.El uso de toda la información disponible como técnica de imputación

El primer tipo de método de imputación **no reemplaza** los datos ausentes sino que **imputa las características** de distribución (por ejemplo, la desviación media o estándar) o las relaciones (por ejemplo, correlaciones) de todos los valores válidos disponibles. Se le conoce también como **enfoque de disponibilidad completa**, este método (opción ***PAIRWISE*** en SPSS) se usa principalmente para **estimar correlaciones** y maximizar la información disponible en la muestra. La característica diferencial de esta aproximación es que cada correlación se basa en un conjunto de observaciones potencialmente único y que el número de observaciones empleadas en los cálculos puede variar en cada correlación. **El proceso de imputación no consiste en reemplazar los datos ausentes** por el resto de los casos, sino en utilizar **las correlaciones obtenidas como representantes** para la muestra entera. Se puede comparar esta aproximación al **enfoque de disponibilidad completa** mencionada anteriormente, que usa solamente **datos de**

observaciones que no tienen datos ausentes. Cualquiera de las dos aproximaciones puede introducir sesgos si el proceso de datos ausentes no es **MCAR (*missing completely at random*).**Aunque el método de disponibilidad completa maximiza los datos utilizados y salva el problema de los datos ausentes de una única variable eliminando un caso del análisis entero, pueden también surgir muchos problemas de esta aproximación. En primer lugar, las correlaciones pueden calcularse "***fuera de rango***" y de forma inconsistente con otras correlaciones de la matriz de correlación. Cualquier correlación entre $\mathbf{X_1}$ y $\mathbf{X_2}$ queda restringida por su correlación con una tercera variable $\mathbf{X_3}$**.** De acuerdo a la fórmula:

$$\text{Rango de } r_{xy} = r_{xz}r_{yz} +/- \sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}$$

La correlación entre **x e y** puede variar sólo de **+1 a -1** si tanto **x** e **y** tienen una correlación cero con todas las otras variables de la matriz de correlación. Es muy raro que se dan correlaciones con otras variables distintas de cero. En la medida en que las correlaciones con otras variables aumenten, el rango de posibles correlaciones entre **x e y** disminuye, aumentando de esta forma la posibilidad de que la correlación en un único conjunto de casos sea inconsistente con las correlaciones derivadas de otros conjuntos de casos. Por ejemplo, si **x** e **y** tienen correlaciones de 0.6 y 0.4 respectivamente con $\mathbf{X_3}$, entonces el rango de correlaciones posibles entre X e Y es 0.24 ± 0.73, o de -0.49 a 0.97. **Cualquier valor fuera de este rango es matemáticamente inconsistente**, aunque podría ocurrir si se obtiene la correlación con un número y conjunto de casos diferentes para las dos correlaciones en el enfoque de disponibilidad completa. Un problema asociado es que los auto valores de la matriz de correlación pueden llegar a ser negativos, alterando así las propiedades de varianza de la matriz de correlación. Aunque la matriz de correlación puede ajustarse para eliminar este problema, muchos programas no incluyen este programa de ajuste. En casos extremos, la matriz estimada de varianzas/covarianzas no es positiva definida. Todos estos problemas deben ser considerados al seleccionar esta aproximación, frente a excluir casos con datos ausentes.

3.4.6.Sustitución de datos ausentes

La segunda forma de imputación consiste en el método efectivo de sustitución de los datos ausentes por valores estimados sobre la base de otra información existente en la muestra. Esta medida puede llevarse a cabo de muchas maneras, que van desde una sustitución directa de valores, a procesos de estimación basados en relaciones entre variables. La exposición siguiente se centrará en los métodos más ampliamente utilizados, aunque existen otras muchas formas de imputación **[**Little et al. 1987**].**

-Sustitución de caso

En este método, las observaciones con datos ausentes se sustituyen con otras observaciones no muestrales. Un ejemplo común es reemplazar un hogar que está en la muestra pero con el que no se puede contactar o que tiene gran cantidad de datos ausentes con otro hogar que no está en la muestra, preferiblemente muy similar al de la

observación original. Este método es el que más se utiliza para sustituir las observaciones con datos ausentes completos, aunque también puede emplearse para reemplazar observaciones con menores cantidades de datos ausentes.

3.4.7.Sustitución por la media

Uno de los métodos más empleados consiste en sustituir los valores ausentes por una variable cuyo valor medio se calcula sobre todas las respuestas válidas. De esta forma, las respuestas de la muestra válida se usan para calcular el valor de sustitución. La lógica de esta aproximación es que la media es el mejor valor de sustitución. Esta aproximación, aun que es extensamente utilizada, tiene **3** desventajas:

1. Invalida las estimaciones de la varianza derivadas de las fórmulas estándar de la varianza para conocer la verdadera varianza de los datos.
2. La distribución real de los valores se encuentra distorsionada por la sustitución de los datos ausentes por la media.
3. Este método modifica la correlación observada porque todos los datos ausentes tendrán un valor único constante. Sin embargo, tiene la ventaja de que se puede llevar a cabo fácilmente y de proporcionar una información completa para todos los casos.

3.4.8.Sustitución por valor constante

En este método, Usted sustituye los datos ausentes por un valor constante derivado de fuentes externas o investigación previa. Su naturaleza es similar al método de sustitución de la media, que difiere sólo en la fuente del valor de sustitución. La imputación de valor constante tiene las mismas desventajas que el método de sustitución de la media, y el investigador debe asegurarse que el valor de sustitución de una fuente externa es más válido que el valor generado internamente por la media. Este método puede proporcionar la opción de reemplazar los datos ausentes con un valor que podría ser considerado más válido que la media de la muestra.

3.4.9.-Imputación por regresión

Este método se usa el **análisis de regresión** para predecir los valores ausentes de una variable basándose en su relación con otras variables del conjunto de datos. Al mismo tiempo de usar las relaciones ya existentes en la muestra como base de predicción, también tenemos **varias desventajas** asociadas con este método:

1. Refuerza las relaciones ya existentes en los datos. Conforme aumente el uso de este método, los datos resultantes son más característicos de la muestra y menos generalizable.
2. A menos que se añadan valores estocásticos a los valores estimados, se subestima la varianza de la distribución.
3. Este método supone que la variable con datos ausentes tiene correlaciones sustanciales con otras variables. Si estas correlaciones no son suficientes para producir una estimación significativa, entonces son preferibles otros métodos, como la sustitución por la media.

El procedimiento de regresión no está restringido en las estimaciones que hace. Por tanto, los valores predichos puede que no correspondan a los rangos válidos de las variables

(por ejemplo, predecir un valor de 11 para una escala de 10 puntos), requiriendo por tanto alguna forma de ajuste adicional. Este método es prometedor en aquellos casos donde se presenten niveles moderados de dispersión de los datos ausentes y donde las relaciones entre las variables están lo suficientemente establecidas como para que Usted confíe en que el uso del método no tendrá impacto sobre la generalidad de los resultados.

3.4.10.Imputación múltiple

Este es el último método de imputación y es en realidad una combinación de varios métodos. En esta aproximación, se usan dos o más métodos para derivar una estimación compuesta, usualmente la media de las diversas estimaciones para el dato ausente. La lógica de esta aproximación es que el uso de la aproximación múltiple minimiza los problemas específicos con cualquier método simple siendo su composición la mejor estimación. La elección de esta aproximación se basa fundamentalmente en la concesión mutua entre la percepción del investigador de los potenciales beneficios ponderada y el esfuerzo sustancialmente superior que requiere realizar y combinar las múltiples estimaciones.

3.4.11.Procedimientos basados en el modelo

Esta categoría incorpora explícitamente los **datos ausentes en el análisis**, bien sea a través de un proceso específicamente diseñado para la estimación de datos ausentes, o bien como una porción integral del análisis multivariante estándar. Hay **2 aproximaciones**:

1. Utiliza estimaciones de **máxima verosimilitud** que intentan modelizar los procesos que subyacen en la ausencia de datos y realizar la estimación más precisa y razonable **[**Little et al. 1987**].** Un ejemplo es la aproximación **EM** en **SPSS**. Representa un método frecuente de **2** etapas (las **etapas E.-*Expectation* y M.-*Maximization***) en los que la **etapa E** realiza las mejores posibles estimaciones de los **datos ausentes** y a continuación **la etapa M** realiza estimaciones de los **parámetros (medias, desviaciones típicas o correlaciones)** con la suposición de que se reemplazaron todos los datos ausentes. El proceso continúa con estas dos etapas hasta que el cambio de los valores estimados es despreciable y se reemplazan todos los datos ausentes.
2. **Datos ausentes** directamente en el análisis, definiendo observaciones con datos ausentes como un subconjunto selecto de la muestra. Esta aproximación es más apropiada para tratar con los datos ausentes de las variables independientes cuando hay una relación de dependencia. Los datos ausentes deben ser tratados como un hecho práctico que debe ser investigado, en lugar de como un desastre a ser mitigado. Además, implícita en esta filosofía está la idea de que como otros aspectos de la muestra de datos, los datos ausentes son una propiedad de la población a la que se busca generalizar.

Al tener datos ausentes en una **variable no métrica**, Usted puede definir fácilmente aquellas observaciones como un grupo separado y entonces incluirlas en cualquier análisis, como **ANOVA o MANOVA** o incluso el **análisis discriminante**. Cuando los datos ausentes se encuentran en una **variable independiente métrica** de una relación de dependencia, se ha desarrollado un procedimiento para incorporar las observaciones en el análisis mientras se mantienen las relaciones entre las variables válidas **[**Cohen et al. 1983**]**. Este procedimiento se ilustrar mejor en el con texto de un **análisis de regresión**, aunque también puede utilizarse en otras relaciones de **dependencia**. El primer paso es codificar

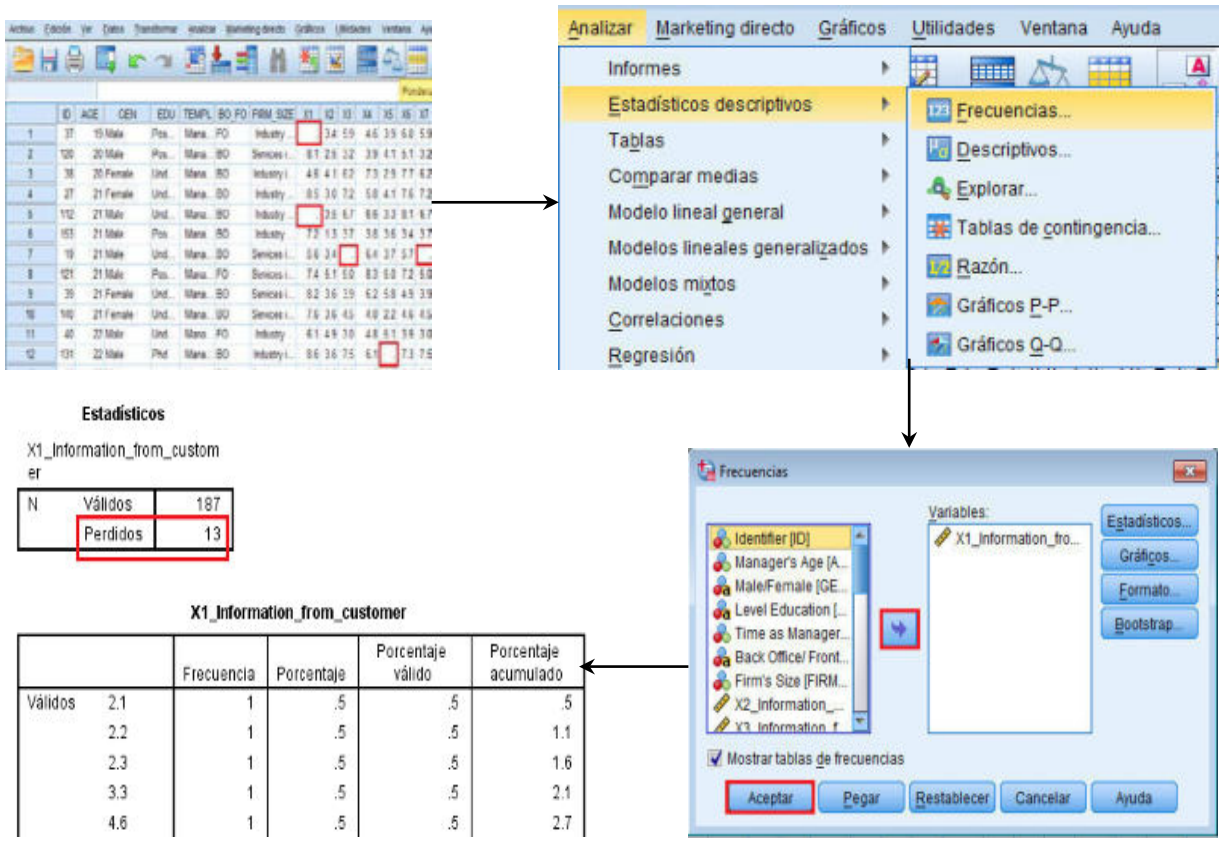
todas las observaciones con datos ausentes como **variables ficticias** (donde los casos con datos ausentes reciben un valor de uno y el resto de los casos un valor de cero). Los valores ausentes se imputan así por el método de **la sustitución por la media**. Finalmente, la relación se estima por **medias normales**. Las **variables ficticias** representan la diferencia en la **variable dependiente** entre aquellas observaciones con datos ausentes y aquellas observaciones con datos válidos. El test del **coeficiente de la variable ficticia** evalúa la significación estadística de esta diferencia. El coeficiente de la variable original representa la relación entre todos los casos con datos no ausentes. Este método permite al analista retener todas las observaciones en el análisis con el fin de mantener el tamaño de la muestra, mientras que también proporciona un test directo de las diferencias entre los dos grupos junto con las relaciones estimadas entre **variables dependientes e independientes**

3.5. Datos perdidos en SPSS

Es posible que quiera distinguir los datos perdidos **porque un encuestado se niegue a responder de los datos perdidos** porque la pregunta afecta a dicho encuestado.

-Problema 3: hacer que los valores perdidos **(.)** de la Variable **X_1cuenten en la estadística descriptiva**.
-Teclear: Analizar->Estadísticos descriptivos->Frecuencias->Selección de variable (X_1) ->Aceptar. Ver Figura 3.10.

Figura 3.10. Base de datos con datos ausentes (.) en variable X_1.



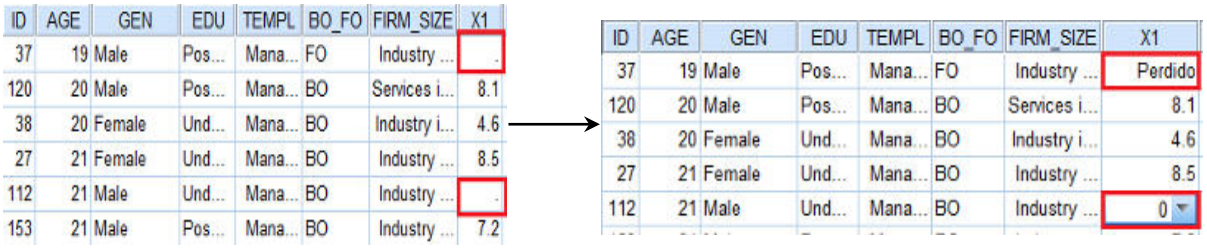
Fuente: SPSS 20 IBM

-Resultado: La contabilización de los datos perdidos así como su ponderación de Porcentaje válido. Persiste el problema de que no se reconocen los valores perdidos como excluidos del conteo; aparecen como si fuera una categoría más.

-Problema 4: Los valores de datos que se especifican como **perdidos** por el usuario aparecen marcados para un tratamiento especial y se excluyen de la mayoría de los cálculos. **Por default valor=0.0=Perdido.**

-Teclear: Analizar->Estadísticos descriptivos->Frecuencias->Selección de variable (X_1) ->Aceptar. Ver Figura 3.11

Figura 3.11. Base de datos con datos ausentes (.) de la variable X_1. Reemplazo de 0 a Perdido

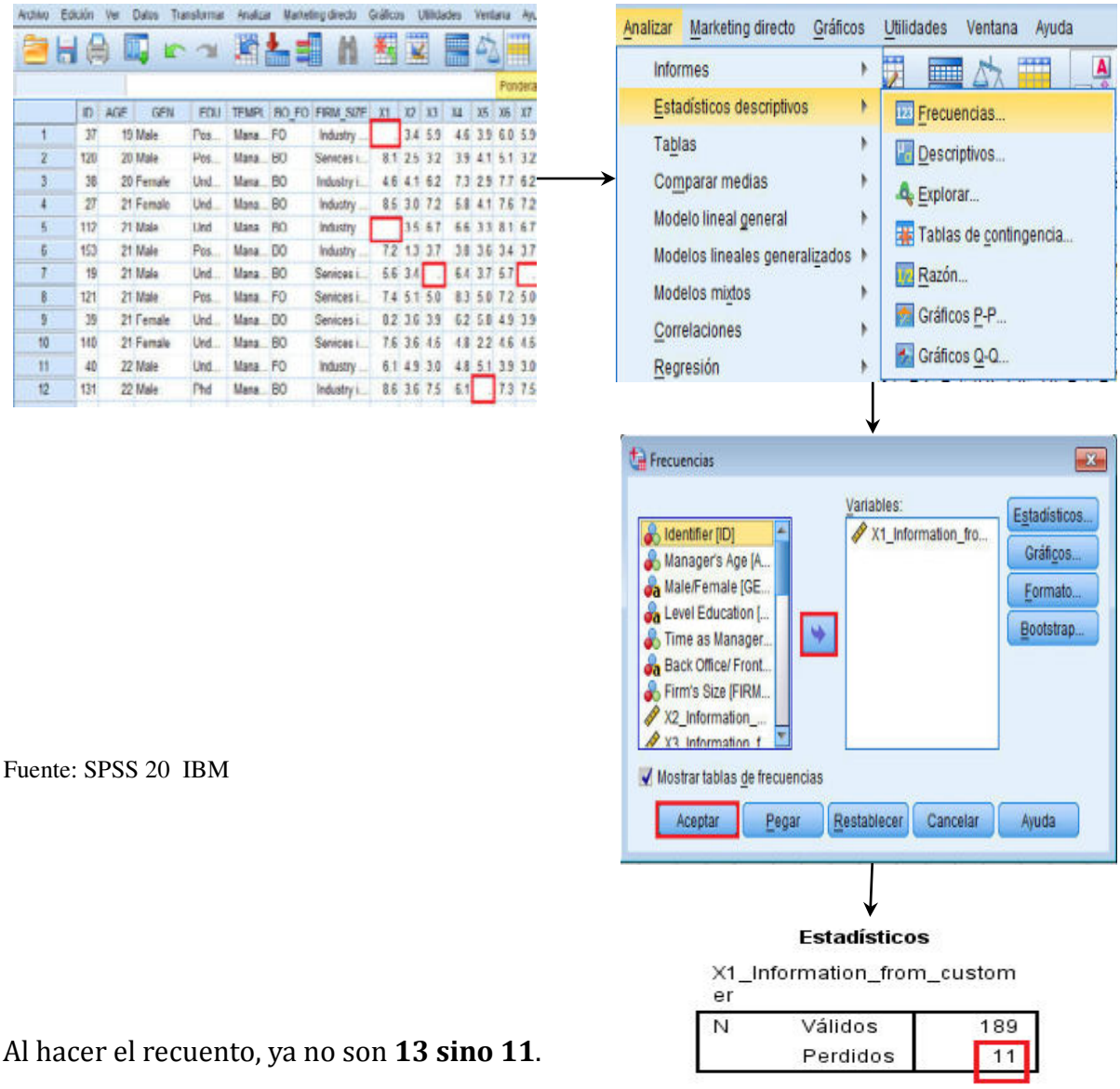


ID	AGE	GEN	EDU	TEMPL	BO_FO	FIRM_SIZE	X1
37	19	Male	Pos...	Mana...	FO	Industry ...	.
120	20	Male	Pos...	Mana...	BO	Services i...	8.1
38	20	Female	Und...	Mana...	BO	Industry i...	4.6
27	21	Female	Und...	Mana...	BO	Industry ...	8.5
112	21	Male	Und...	Mana...	BO	Industry ...	.
153	21	Male	Pos...	Mana...	BO	Industry ...	7.2

ID	AGE	GEN	EDU	TEMPL	BO_FO	FIRM_SIZE	X1
37	19	Male	Pos...	Mana...	FO	Industry ...	Perdido
120	20	Male	Pos...	Mana...	BO	Services i...	8.1
38	20	Female	Und...	Mana...	BO	Industry i...	4.6
27	21	Female	Und...	Mana...	BO	Industry ...	8.5
112	21	Male	Und...	Mana...	BO	Industry ...	0

Fuente: SPSS 20 IBM

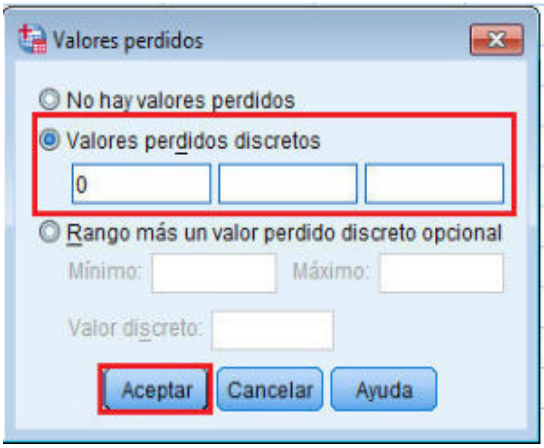
Al cambiar a **0** los valores ausentes aparece la leyenda en el campo: **Perdidos**



Fuente: SPSS 20 IBM

Al hacer el recuento, ya no son **13 sino 11**.

Problema 5: ¿cómo hacer que nuevamente vuelva a contar como valor perdido?
Para resolverlo, es posible utilizar el cuadro de diálogo de valores perdidos:



Fuente: SPSS 20 IBM

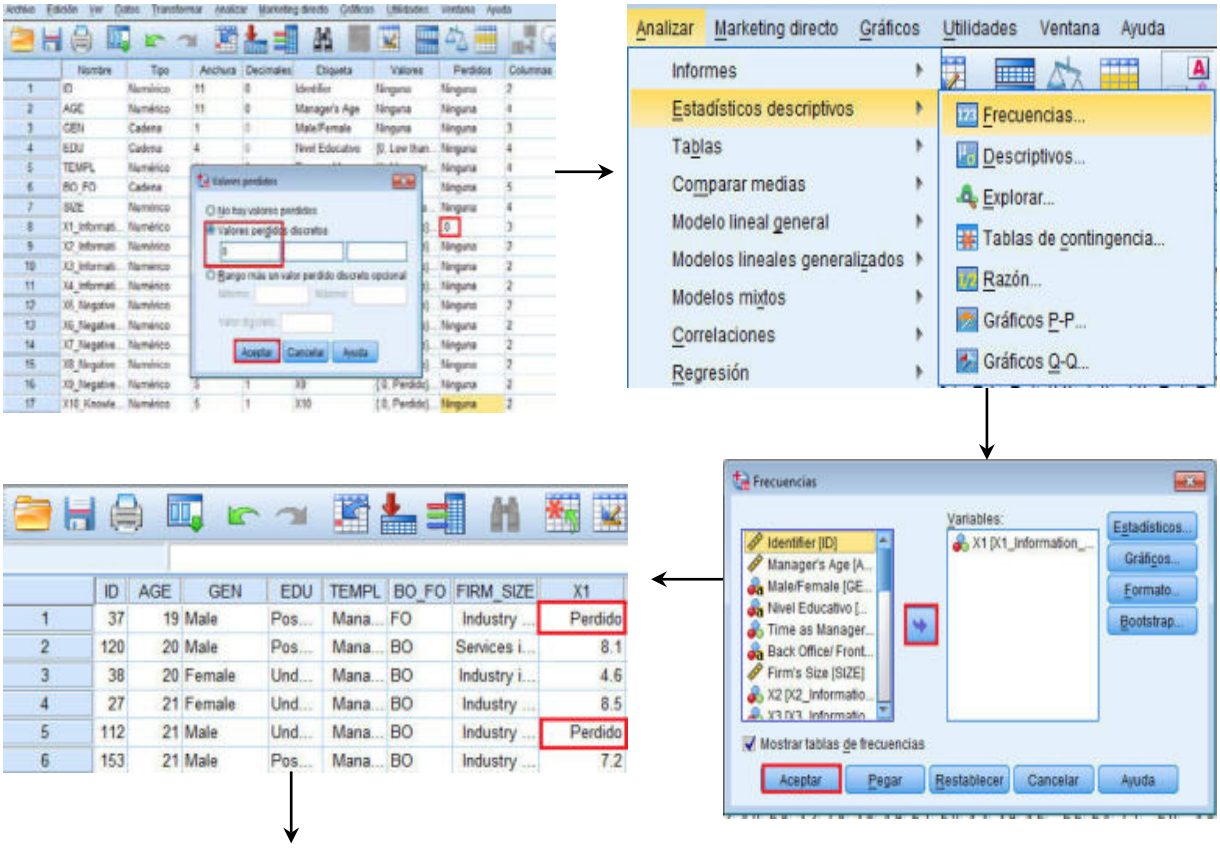
Del que:

- Se pueden introducir hasta tres valores perdidos (individuales) de tipo discreto, un rango de valores perdidos o un rango más un valor de tipo discreto.
- Sólo pueden especificarse rangos para las variables numéricas.
- Se considera que son válidos todos los valores de cadena, incluidos los valores vacíos o nulos, a no ser que se definan explícitamente como perdidos.
- Los valores perdidos de las variables de cadena no pueden tener más de ocho bytes. (No hay ningún límite respecto al ancho definido de la variable de cadena, pero los valores definidos como perdidos no pueden tener más de ocho bytes.)

Para definir como perdidos los valores nulos o vacíos de una variable de cadena, escriba un espacio en blanco en uno de los campos debajo de la selección **Valores perdidos discretos. Ver Figura 3.11**

-Problema 6: hacer que los **valores perdidos** de la variable **X_1 cuenten en la estadística descriptiva**, se deberá codificar dentro del campo **Perdidos** de la base de datos, los casos.
Teclear: Vista de variables->Ubicar cursor en campo perdidos en la variable X_1 y abrir cuadro de diálogo de Valores Perdidos->Seleccionar Valores perdidos discretos, con valor "*0*"->Aceptar->Analizar->Estadísticos descriptivos->Frecuencias->Selección de variable (X_1) ->Aceptar; Ver Figura 3.12.

Figura 3.12. Base de datos con valores perdidos de la variable X_1



	ID	AGE	GEN	EDU	TEMPL	BO_FO	FIRM_SIZE	X1
1	37	19	Male	Pos...	Mana...	FO	Industry ...	Perdido
2	120	20	Male	Pos...	Mana...	BO	Services i...	8.1
3	38	20	Female	Und...	Mana...	BO	Industry i...	4.6
4	27	21	Female	Und...	Mana...	BO	Industry ...	8.5
5	112	21	Male	Und...	Mana...	BO	Industry ...	Perdido
6	153	21	Male	Pos...	Mana...	BO	Industry ...	7.2

Estadísticos

X1_Information_from_custom er

N	Válidos	187
	Perdidos	13

X1_Information_from_customer

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	2.1	1	.5	.5	.5
	2.2	1	.5	.5	1.1

Fuente: SPSS 20 IBM

-Resultado: La **No contabilización** de los datos perdidos así como su ponderación de Porcentaje válido. Ya no persiste el problema de que no se reconozcan los valores´Perdidos como excluídos del conteo; ya no son una Categoría más.

Nota: los datos ausentes por sistema se representan con puntos decimales

-Problema 7: Se requiere conocer cómo se encuentran los datos de la variable **X_1** respecto a la variable **Firm´s Size** (sin ausencia de datos)
-Teclear: Analizar->Estadísticos descriptivos->Tablas de contingencia->selección de variables Fila: X_1; Columna: Firm's Size->Casillas->Fila->Continuar->Aceptar. Ver **Figura 3.13.**

Figura 3.13. Base de datos con valores perdidos de la variable X_1 vs. Variable Firm's Size.

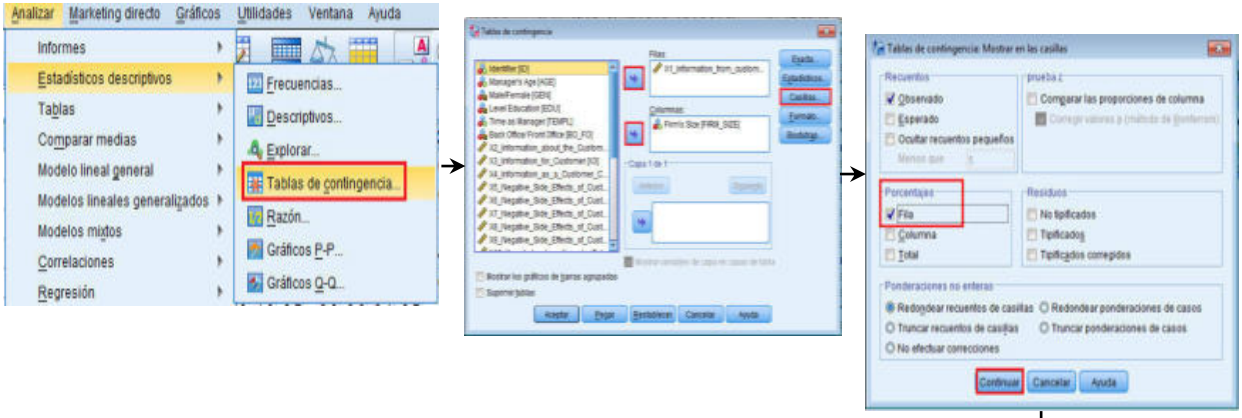


Tabla de contingencia X1_Information_from_customer * Firm's Size

			Firm's Size						
			Industry and services is Micro (1-10)	Industry and services is Little (11-50)	Industry is Media (51-250)	Services is Media (51-100)	Industry is Big (251 a+)	Services is Big (101 a+)	Total
X1_Information_from_customer	2.1	Recuento	0	0	0	1	0	0	1
		% dentro de X1_Information_from_customer	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	100.0%
	2.2	Recuento	0	0	0	0	1	0	1
		% dentro de X1_Information_from_customer	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	100.0%
	2.3	Recuento	0	0	1	0	0	0	1
		% dentro de X1_Information_from_customer	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	100.0%
	3.3	Recuento	0	0	0	0	0	1	1
		% dentro de X1_Information_from_customer	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	100.0%
	4.6	Recuento	0	0	0	0	1	0	1
		% dentro de X1_Information_from_customer	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	100.0%
	9.7	Recuento	0	4	2	1	1	0	8
		% dentro de X1_Information_from_customer	0.0%	50.0%	25.0%	12.5%	12.5%	0.0%	100.0%
	9.9	Recuento	0	4	2	4	1	0	11
		% dentro de X1_Information_from_customer	0.0%	36.4%	18.2%	36.4%	9.1%	0.0%	100.0%
Total		Recuento	12	76	20	39	25	15	187
		% dentro de X1_Information_from_customer	6.4%	40.6%	10.7%	20.9%	13.4%	8.0%	100.0%

Tablas de contingencia

[Conjunto_de_datos1] C:\Users\Juan\Desktop\proy libro mc\CKM_MKT_Digital_OK.sav

Resumen del procesamiento de los casos

	Casos					
	Válidos		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
X1_Information_from_customer * Firm's Size	187	93.5%	13	6.5%	200	100.0%

-Resultado: Se generan datos en matriz ***X_1* vs. Firm's Size** tanto de valores contenidos como perdidos
para mejor presentación de análisis

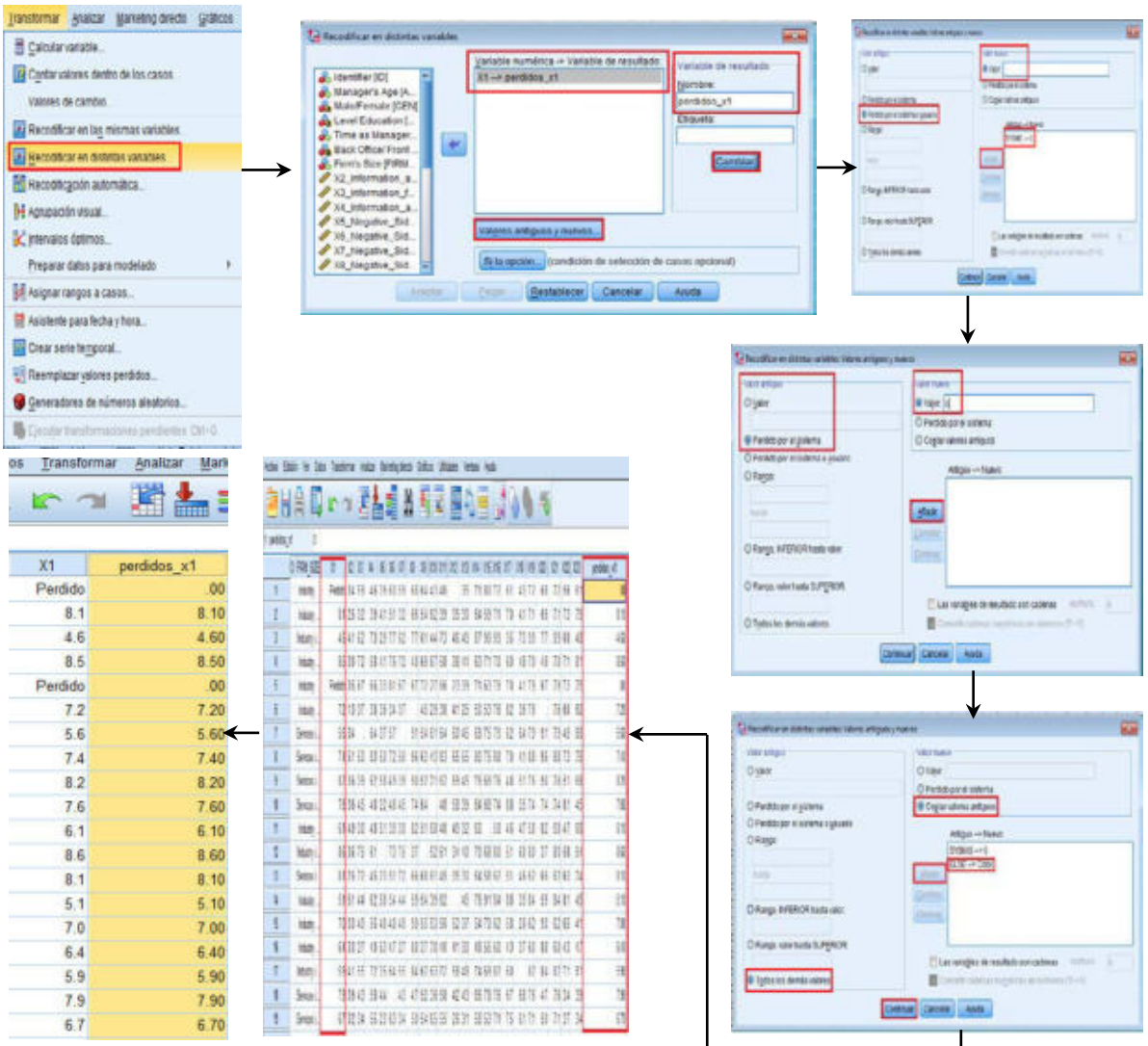
Fuente: SPSS 20 IBM

-Problema 8: se requiere manipular en otro campo, los datos ausentes por sistema **(.) y que éstos tomen su valor 0** de la variable $\mathbf{X_1}$

Fuente: SPSS 20 IBM

-Teclear: Transformar>Recodificar en distinta variable->En Variable numérica->Variable de resultado seleccionar: X_1-> En variable de resultado nombre: perdidos_X_1->Cambiar->Valores antiguos y nuevos->Valor antiguo, seleccionar Perdidos por el sistema; Valor nuevo: 0; Añadir; Copiar valores antiguos; Añadir; Todos los demás valores; Continuar->Aceptar. Ver **Figura 3.14.**

Figura 3.14. Base de datos con valores perdidos en variable X_1 en otra columna



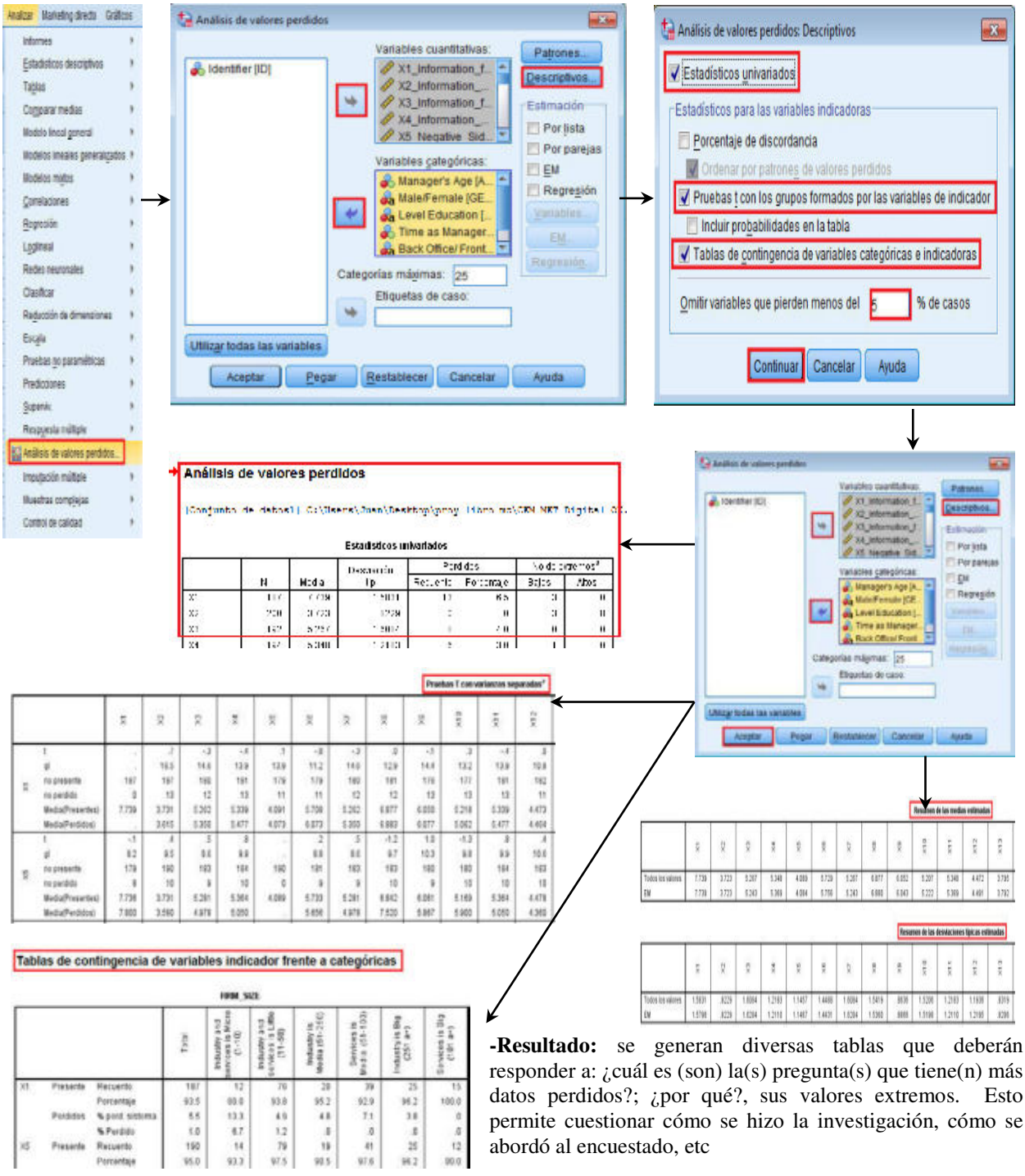
X1	perdidos_x1
Perdido	.00
8.1	8.10
4.6	4.60
8.5	8.50
Perdido	.00
7.2	7.20
5.6	5.60
7.4	7.40
8.2	8.20
7.6	7.60
6.1	6.10
8.6	8.60
8.1	8.10
5.1	5.10
7.0	7.00
6.4	6.40
5.9	5.90
7.9	7.90
6.7	6.70

-Resultado: Se generan tablas que muestran La base de datos que está trabajando con todos los registros que están con **datos ausentes por sistema (.)=0. Es posible mover el nuevo campo a donde lo desee.**

Fuente: SPSS 20 IBM

-Problema 9: identifique cuál es el impacto de los datos perdidos de las variables.
-Teclear: Analizar->Análisis de valores perdidos->selección de variables cuantitativas y cualitativas->Descriptivos->Selección de estadísticos univariados; Pruebas t con los grupos formados por el indicador; Tablas de contingencia de variables categóricas e indicadoras; Continuar->Aceptar ; Ver Figura 3.15.

Figura 3.15. Base de datos con valores perdidos



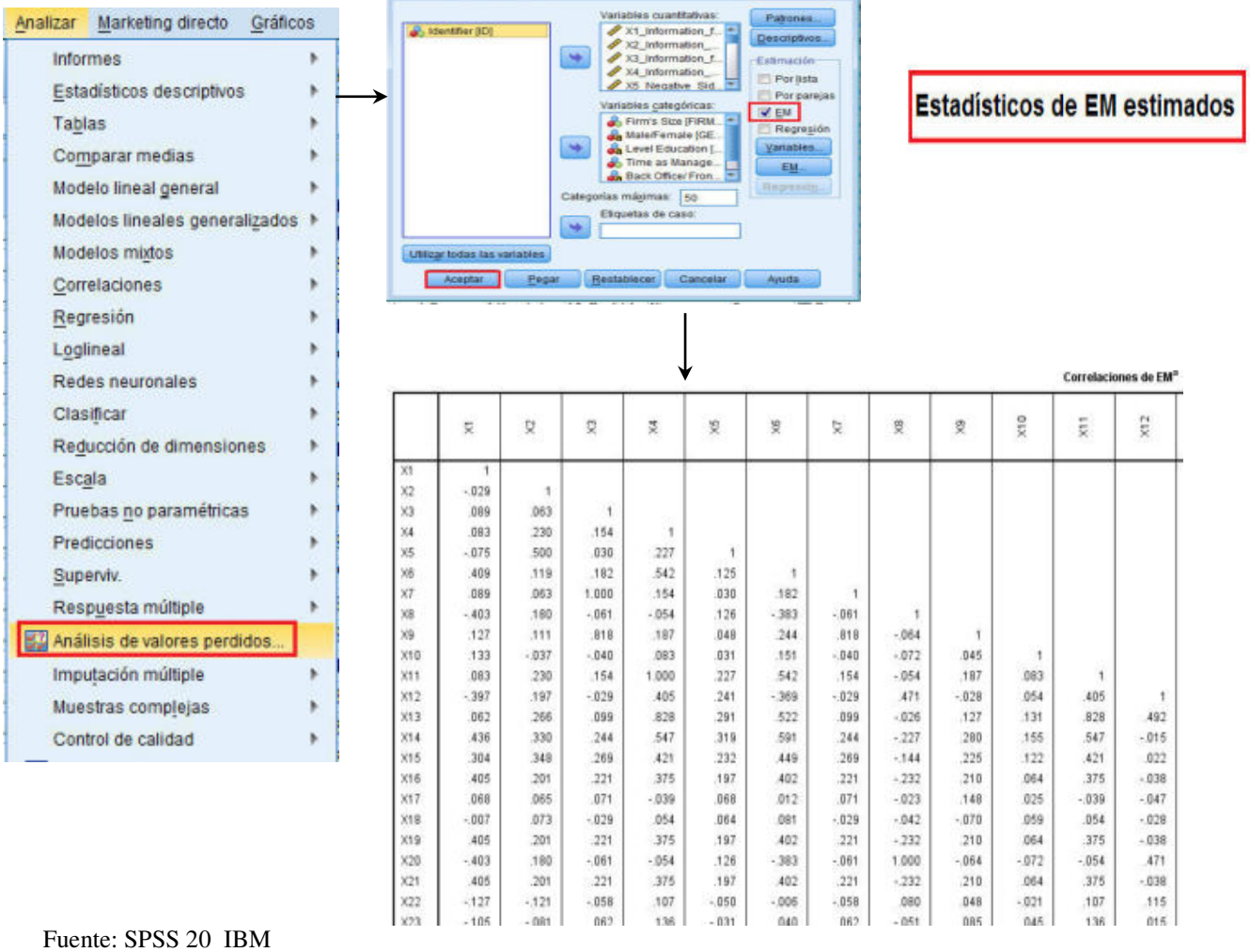
-Resultado: se generan diversas tablas que deberán responder a: ¿cuál es (son) la(s) pregunta(s) que tiene(n) más datos perdidos?; ¿por qué?, sus valores extremos. Esto permite cuestionar cómo se hizo la investigación, cómo se abordó al encuestado, etc

Fuente: SPSS 20 IBM

-Problema 10: identifique si los datos ausentes o perdidos, corresponden o no a un patrón aleatorio.

-Teclear: Analizar->Análisis de valores perdidos->selección de variables cuantitativas y cualitativas->Estimación seleccionar EM->Aceptar Ver Figura 3.16.

Figura 3.16. Base de datos con valores perdidos



	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
X1	1											
X2	-.029	1										
X3	.089	.063	1									
X4	.083	.230	.154	1								
X5	-.075	.500	.030	.227	1							
X6	.409	.119	.182	.542	.125	1						
X7	.089	.063	1.000	.154	.030	.182	1					
X8	-.403	.180	-.061	-.054	.126	-.383	-.061	1				
X9	.127	.111	.818	.187	.048	.244	.818	-.064	1			
X10	.133	-.037	-.040	.083	.031	.151	-.040	-.072	.045	1		
X11	.083	.230	.154	1.000	.227	.542	.154	-.054	.187	.083	1	
X12	-.397	.197	-.029	.405	.241	-.369	-.029	.471	-.028	.054	.405	1
X13	.062	.266	.099	.828	.291	.522	.099	-.026	.127	.131	.828	.492
X14	.436	.330	.244	.547	.319	.591	.244	-.227	.280	.155	.547	-.015
X15	.304	.348	.269	.421	.232	.449	.269	-.144	.225	.122	.421	.022
X16	.405	.201	.221	.375	.197	.402	.221	-.232	.210	.064	.375	-.038
X17	.068	.065	.071	-.039	.068	.012	.071	-.023	.148	.025	-.039	-.047
X18	-.007	.073	-.029	.054	.064	.081	-.029	-.042	-.070	.059	.054	-.028
X19	.405	.201	.221	.375	.197	.402	.221	-.232	.210	.064	.375	-.038
X20	-.403	.180	-.061	-.054	.126	-.383	-.061	1.000	-.064	-.072	-.054	.471
X21	.405	.201	.221	.375	.197	.402	.221	-.232	.210	.064	.375	-.038
X22	-.127	-.121	-.058	.107	-.050	-.006	-.058	.080	.048	-.021	.107	.115
X23	-.105	-.081	.062	.136	-.031	.040	.062	-.051	.085	.045	.136	.015

Fuente: SPSS 20 IBM

a. Prueba MCAR de Little: Chi-cuadrado = 281.133, GL = 979, Sig. = 1.000

-Resultado: Buscar la tabla **Correlaciones de EM**, si en el reporte base de la matriz, se encuentra Prueba **MCAR** de Little con **Sig.>=0.05, entonces los datos ausentes tienen un comportamiento aleatorio**, de lo contrario, deberá realizar los ajustes que considere pertinente para continuar analizando los datos.

3.6. Reemplazar datos perdidos en SPSS

Las observaciones perdidas pueden causar problemas en los análisis y algunas medidas de series temporales no se pueden calcular si hay valores perdidos en la serie. En ocasiones el valor para una observación concreta no se conoce. Además, los datos perdidos pueden ser el resultado de lo siguiente:

- **Cada grado de diferenciación reduce la longitud de una serie en 1.**
- **Cada grado de diferenciación estacional reduce la longitud de una serie en una estación**.
- Si genera una serie nueva que contenga predicciones que sobrepasen el final de la serie existente (al pulsar en el botón **Guardar** y realizar las selecciones adecuadas), la serie original y la serie residual generada incluirán datos perdidos para las observaciones nuevas.
- **Algunas transformaciones (por ejemplo, la transformación logarítmica) generan datos perdidos para determinados valores de la serie original.**
- **Los valores perdidos al principio o fin de una serie no suponen un problema especial;** sencillamente acortan la longitud útil de la serie. **Las discontinuidades que aparecen en mitad de una serie (datos incrustados perdidos) pueden ser un problema mucho más grave**. El alcance del problema depende del procedimiento analítico que se utilice.

El **cuadro de diálogo Reemplazar valores perdidos** crea nuevas variables de series temporales a partir de otras existentes, reemplazando los valores perdidos por estimaciones calculadas mediante uno de los distintos métodos posibles. Los nombres por defecto de las nuevas variables se componen de **los seis primeros caracteres de las variables existentes utilizadas para crearlas, seguidos por un carácter de subrayado y un número secuencial.**

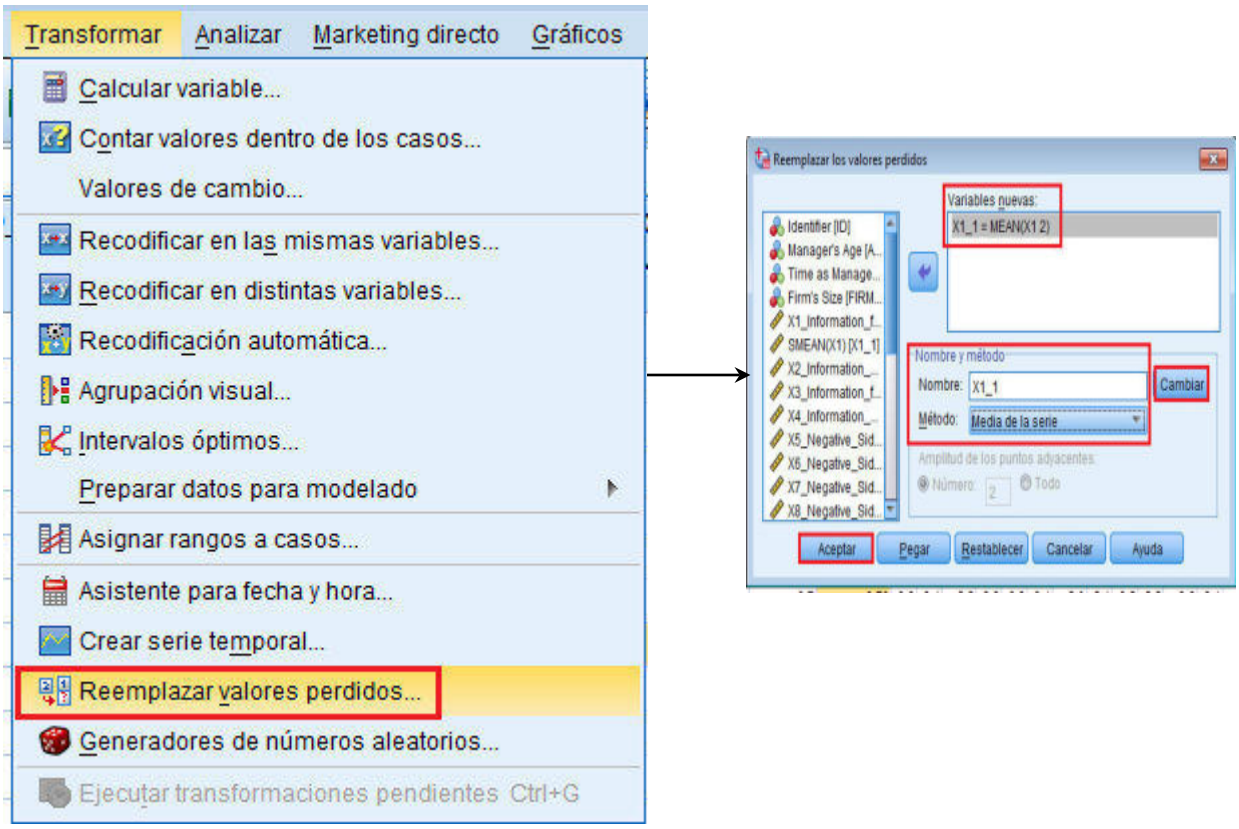
Son métodos de estimación para reemplazar los valores perdidos

1. **Media de la serie.** Sustituye los valores perdidos con la media de la serie completa.
2. Media de puntos adyacentes. Sustituye los valores perdidos por la media de los valores válidos circundantes. La amplitud de los puntos adyacentes es el número de valores válidos, por encima y por debajo del valor perdido, utilizados para calcular la media.
3. **Mediana de puntos adyacentes.** Sustituye los valores perdidos por la mediana de los valores válidos circundantes. La amplitud de los puntos adyacentes es el número de valores válidos, por encima y por debajo del valor perdido, utilizados para calcular la mediana.
4. **Interpolación lineal.** Sustituye los valores perdidos utilizando una interpolación lineal. Se utilizan para la interpolación el último valor válido antes del valor perdido y el primer valor válido después del valor perdido. Si el primer o el último caso de la serie tiene un valor perdido, el valor perdido no se sustituye.
5. **Tendencia lineal en el punto.** Reemplaza los valores perdidos de la serie por la tendencia lineal en ese punto. Se hace una regresión de la serie existente sobre una variable índice escalada de 1 a n.

Los valores perdidos se sustituyen por sus valores pronosticados.

-Problema 11: seleccione opción para reemplazar datos perdidos
-Teclear: Transformar->Reemplazar valores perdidos->seleccionar campos (X_1) ->Seleccionar método: media de la serie o media de puntos adyacentes o mediana de puntos adyacentes o interpolación lineal o tendencia lineal en el punto ->Cambiar->Aceptar. Ver Figura 3.17.

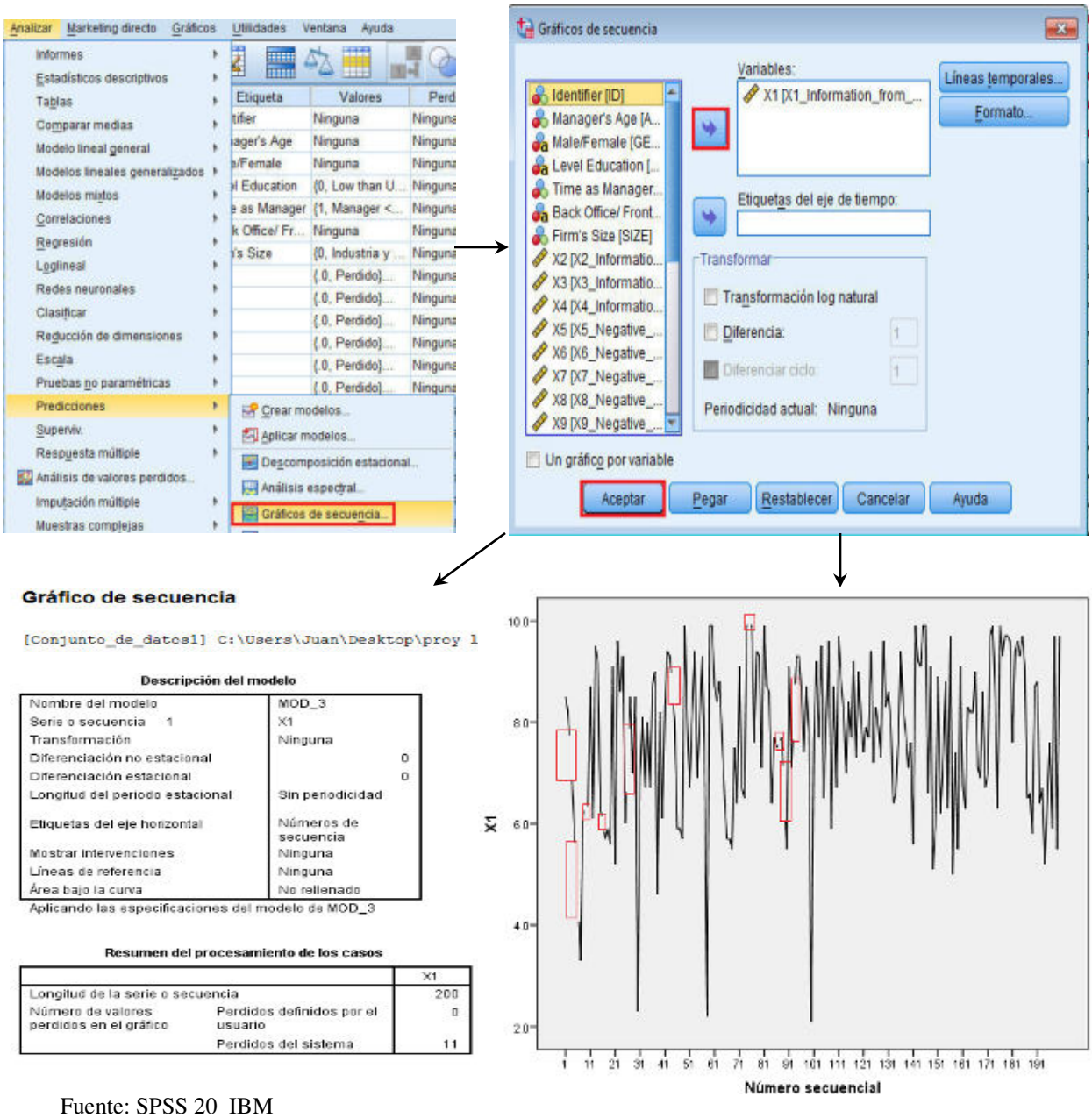
Figura 3.17. Proceso para reemplazar datos perdidos



Fuente: SPSS 20 IBM

-Problema 12: reemplazar los datos perdidos de la variable **X_1**, suponiendo que contienen **(.)** se sugiere realizar un gráfico.
-Teclear: Analizar->Predicciones->Gráficos de secuencia-> Selección de la variable (X_1) Aceptar. Ver Figura 3.18.

Figura 3.18. Proceso realizar gráfico de datos

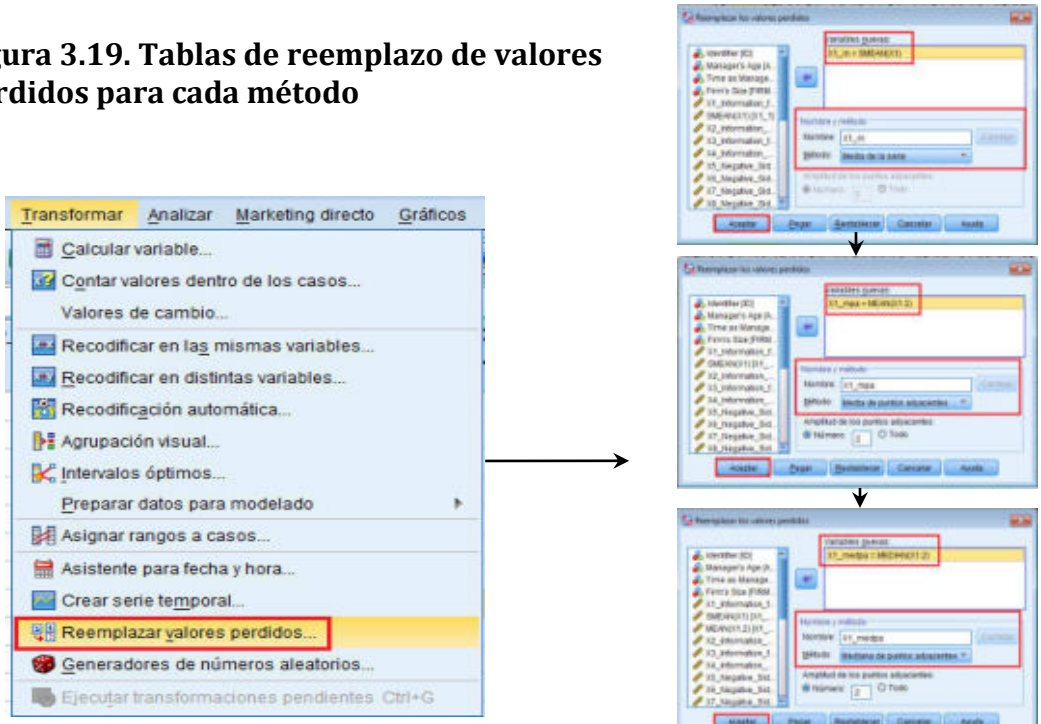


Fuente: SPSS 20 IBM

-Problema 13: muestre en la base de datos cada una de las opciones de reemplazo de los valores perdidos de la variable $\mathbf{X_1}$

-Teclear: arrastrar c/u de los resultados del paso anterior a X_1. Ver Figura 3.19

Figura 3.19. Tablas de reemplazo de valores perdidos para cada método



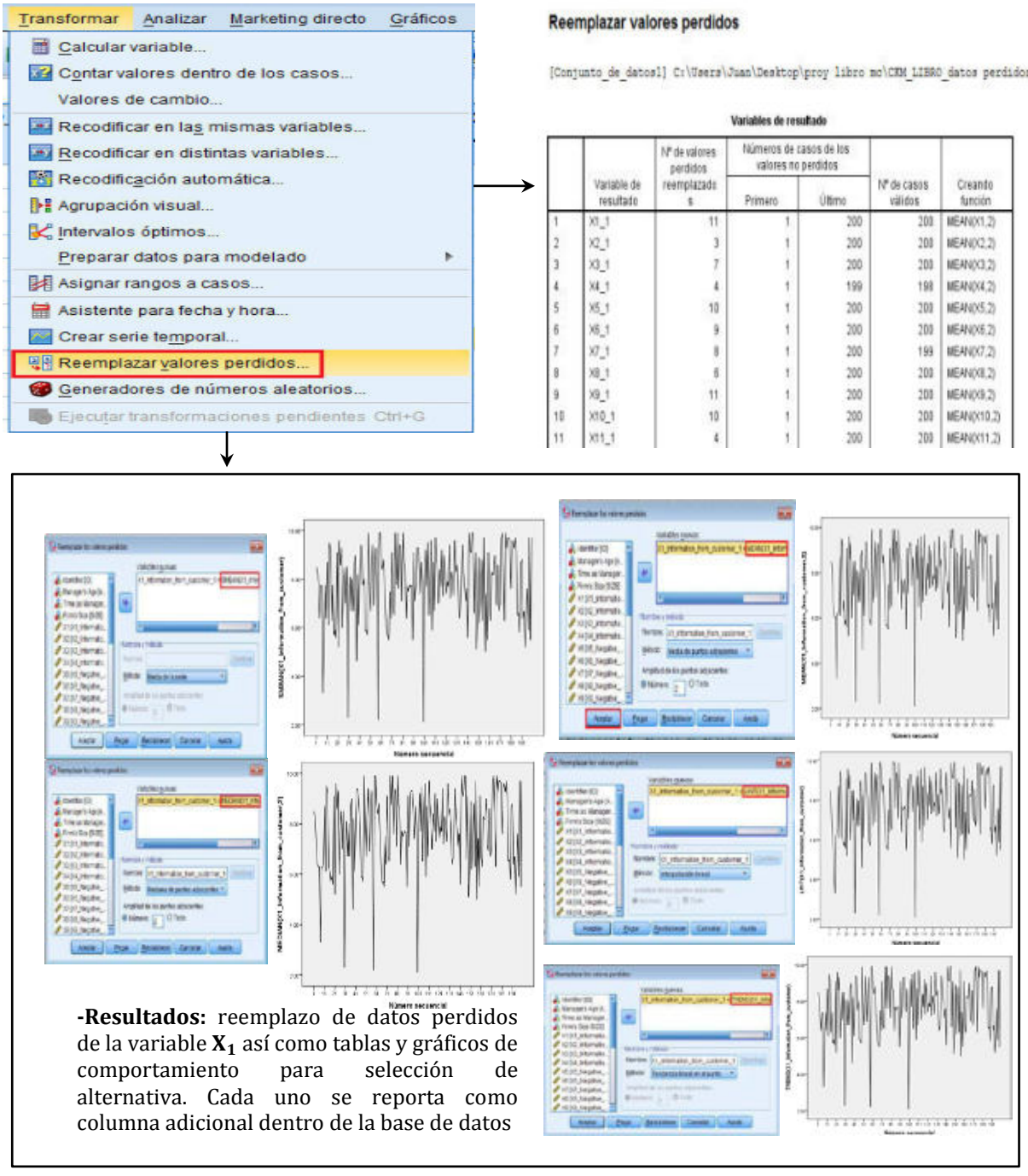
	ID	AGE	GEN	EDU	TEMPL	BO_FO	FIRM_SIZE	X1	X1_m	X1_mpa	X1_medpa	X1_intl	X1_tlp
1	37	19	Male	Pos...	Mana...	FO	Industry ...	Perdido	7.74	.	.	.	7.74
2	120	20	Male	Pos...	Mana...	BO	Industry ...	8.1	8.10	8.10	8.10	8.10	8.10
3	38	20	Female	Und...	Mana...	BO	Industry i...	4.6	4.60	4.60	4.60	4.60	4.60
4	27	21	Female	Und...	Mana...	BO	Industry ...	8.5	8.50	8.50	8.50	8.50	8.50
5	112	21	Male	Und...	Mana...	BO	Industry ...	Perdido	7.74	6.48	6.40	7.85	7.74
6	153	21	Male	Pos...	Mana...	BO	Industry ...	7.2	7.20	7.20	7.20	7.20	7.20

Fuente: SPSS 20 IBM

-Problema 14: genere los gráficos correspondientes a cada técnica de reemplazo de valores perdidos de la variable X_1

-Teclear: y una vez obtenidos los datos graficar las secuencias tecleando: **Analizar->Predicciones->Gráficos de secuencia-> Selección de la variable (X_1) Aceptar. Ver Figura 3.20.**

Figura 3.20. Proceso realizar reemplazo de valores perdidos y sus gráficos, para cada método



-Resultados: reemplazo de datos perdidos de la variable X_1 así como tablas y gráficos de comportamiento para selección de alternativa. Cada uno se reporta como columna adicional dentro de la base de datos

Fuente: SPSS 20 IBM

3.7. Imputación datos perdidos en SPSS

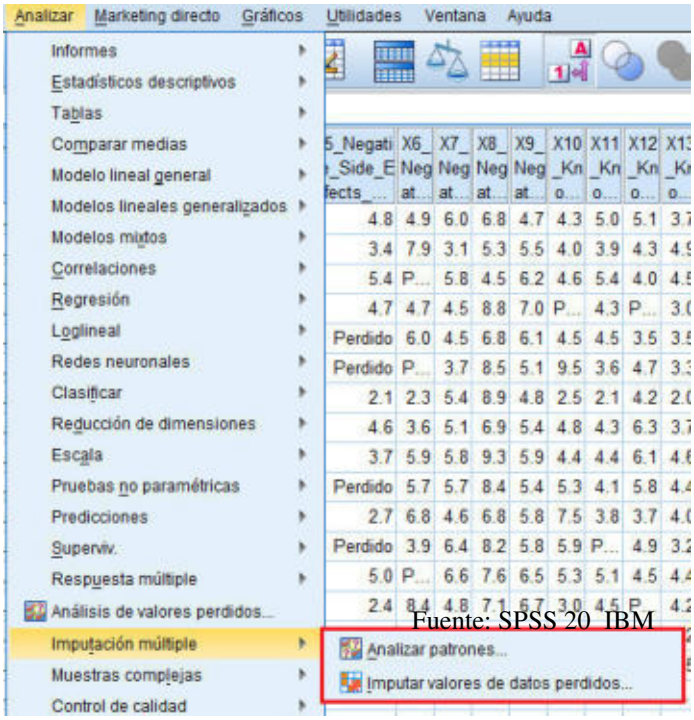
Existen **2 cuadros de diálogo** dedicados a la **imputación múltiple**.

1. Analizar patrones proporciona medidas descriptivas de los patrones de valores perdidos en los datos y puede resultar útil como paso exploratorio antes de la imputación.

2. Imputar valores perdidos se utiliza para generar imputaciones múltiples. Los conjuntos de datos completos pueden analizarse con procedimientos que admiten conjuntos de datos de imputación múltiple.

Ver Figura 3.21.

Figura 3.21. Cuadro de diálogo imputación múltiple



Fuente: SPSS 20 IBM

Fuente: SPSS 20 IBM

3.7.1. Analizar patrones

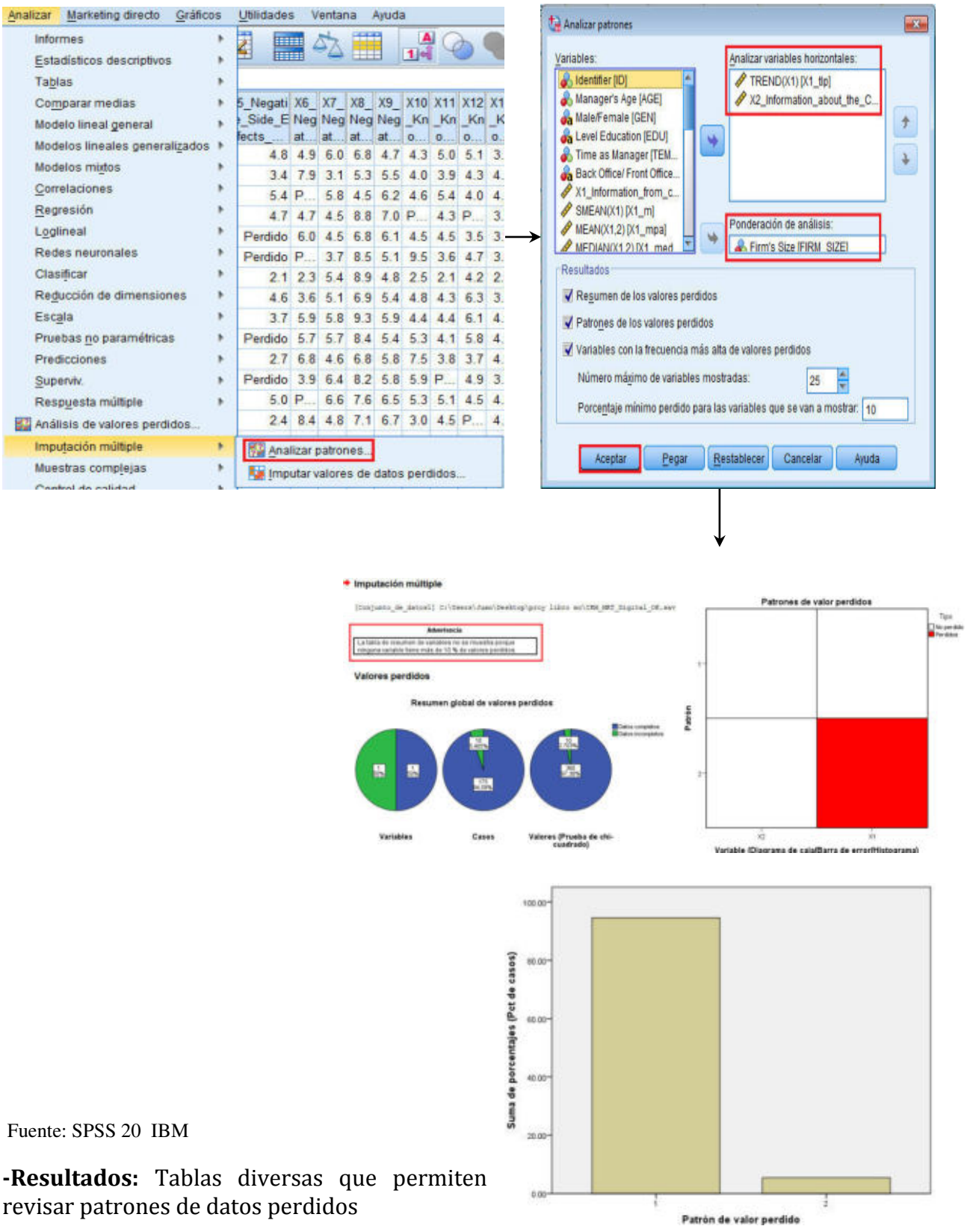
Analizar patrones proporciona medidas descriptivas de los patrones de valores perdidos en los datos y puede resultar útil como paso exploratorio antes de la imputación. Se hace con **3 variables a elección**

-Problema 14: Se requiere comprender mejor los patrones de las variables: **X_1** con datos perdidos, **X_2** y **Firm's size** de **CKM_MKT_Digital.sav**. El análisis de patrones de valores perdidos puede ayudar a determinar los siguientes pasos que se imputarán.

-Teclear: Analizar->Imputación múltiple->Analizar patrones.

Ver Figura 3.22.

Figura 3.22. Cuadro de diálogo imputación múltiple



Fuente: SPSS 20 IBM

-Resultados: Tablas diversas que permiten revisar patrones de datos perdidos

3.7.2. Configuración opcional

Ponderación de análisis. Esta variable contiene ponderaciones de análisis (**regresión o muestra**). El procedimiento incorpora ponderaciones de análisis en resúmenes de valores perdidos. Los casos de ponderaciones de análisis con **valor negativo o cero se excluirán.**

Resultado. Los siguientes resultados opcionales están disponibles:

1. **Resumen de valores perdidos.** Esto muestra un gráfico de sectores con paneles que indica el número y el porcentaje de variables de análisis, casos o datos individuales que tengan uno o más valores perdidos.
2. **Patrones de valores perdidos.** Esto muestra patrones tabulados de valores perdidos. Cada patrón se corresponde con un grupo de casos con el mismo patrón de datos completos e incompletos sobre variables de análisis. Puede utilizar este resultado para determinar si puede utilizar el método de imputación mono tónica para sus datos o, si no, en qué medida se aproximan sus datos a un patrón mono tónico. El procedimiento ordena las variables de análisis para revelar o aproximarse a un patrón mono tónico. Si no hay patrones que no sean mono tónico después de la reordenación, puede llegar a la conclusión de que los datos tienen un patrón mono tónico cuando las variables de análisis se ordenan de tal forma.
3. **Variables con la mayor frecuencia de valores perdidos.** Esto muestra una tabla de variables de análisis ordenadas por el porcentaje de valores perdidos en orden descendente. La tabla incluye estadísticos descriptivos (media y desviación típica) para variables de escala. Puede controlar el número máximo de variables que se mostrará y el porcentaje de ausencia mínimo de una variable para que se incluya en la visualización. Se muestra el conjunto de variables que cumplen ambos criterios. Por ejemplo, si establece el número máximo de variables como 50 y el porcentaje de ausencia mínimo como 25, hará que la tabla muestre un máximo de 50 variables que tengan un mínimo del 25 % de valores perdidos. Si hay 60 variables de análisis pero sólo 15 tienen un porcentaje igual o mayor al 25 % de valores perdidos, el resultado sólo incluirá 15 variables.

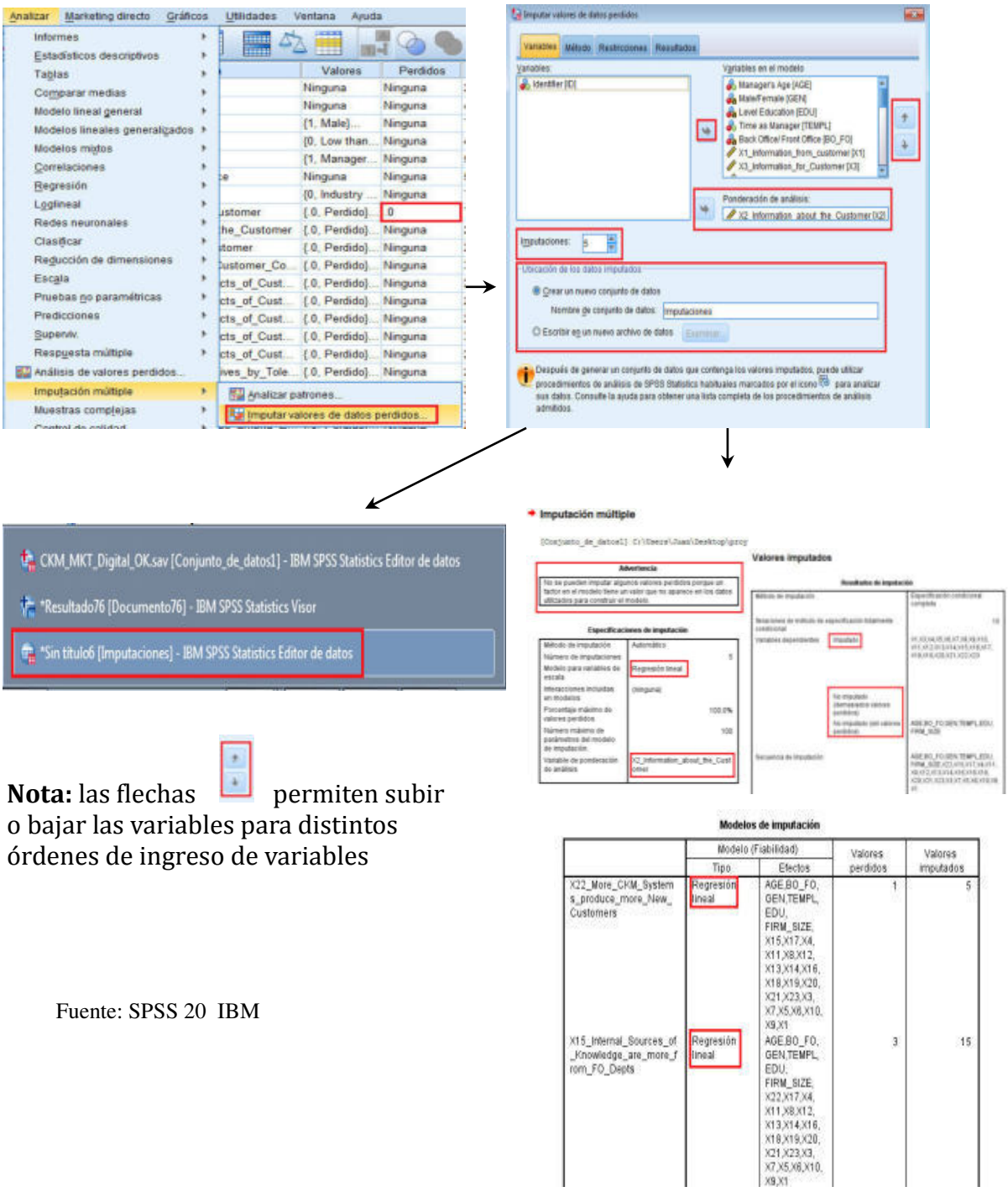
3.7.3. Imputar valores perdidos

Imputar valores perdidos se utiliza para generar imputaciones múltiples. Los conjuntos de datos completos pueden analizarse con procedimientos que admiten conjuntos de datos de imputación múltiple.

-Problema 16: Se requiere calcular las opciones de imputación múltiple de los valores perdidos tomando en cuenta que la variable de ponderación sea métrica y con datos al **100%** (**X_2**) (**se sugiere probar con no métrica y probar**)

-Teclear: Analizar > Imputación múltiple > Imputar valores de datos perdidos->Selección de variables a imputar->Cantidad de imputaciones (1-5) ->Nombre de conjunto de datos (imputaciones) ->Aceptar. Ver Figura 3.23

Figura 3.23. Cuadro de diálogo imputación múltiple



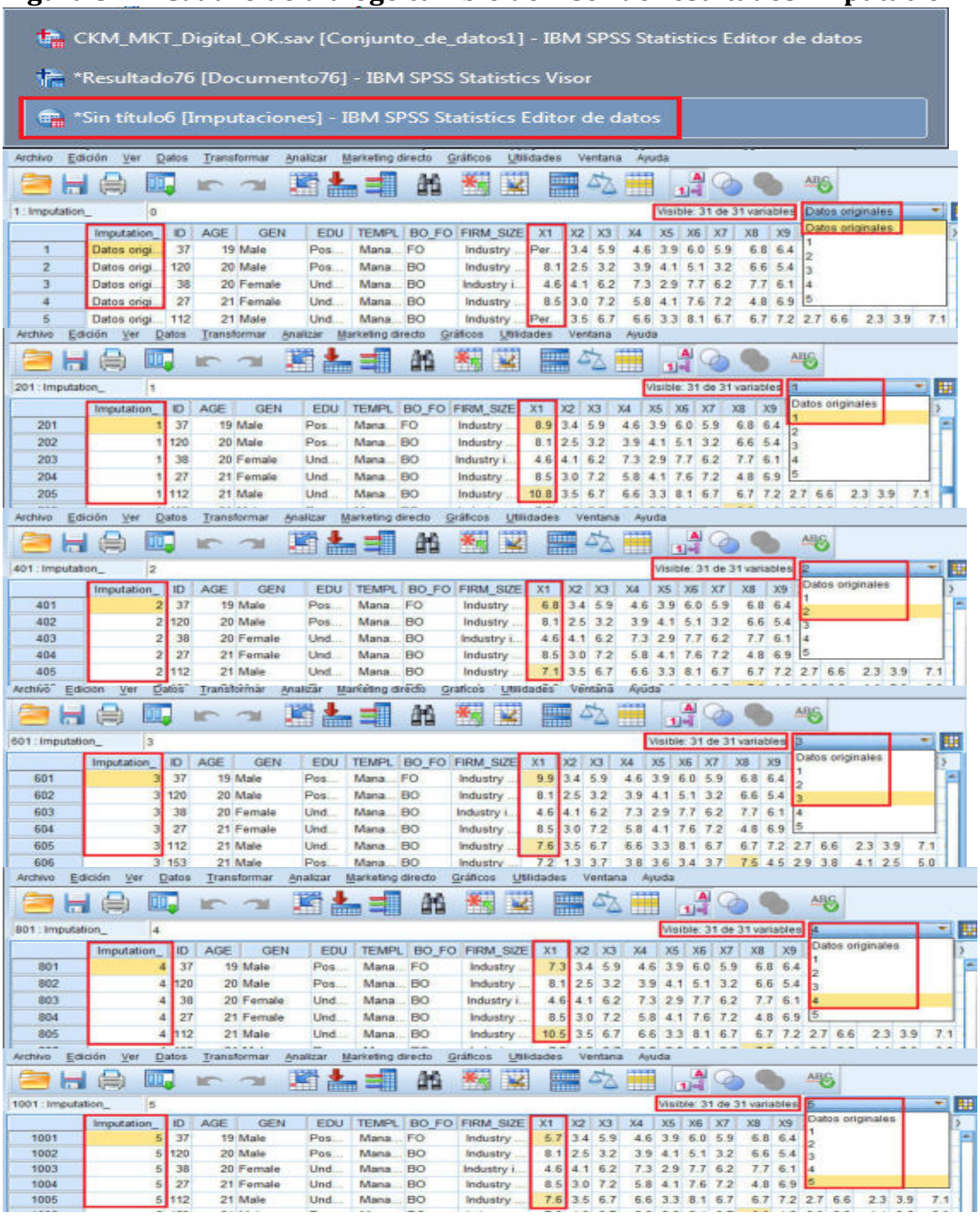
Nota: las flechas permiten subir o bajar las variables para distintos órdenes de ingreso de variables

Fuente: SPSS 20 IBM

-Resultados: diversas tablas que muestran el proceso de imputación mediante regresiones lineales y reporte de ejecución. Se genera **editor de datos alterna**: **imputaciones** con la propuesta de nuevos valores imputados sombreados en amarillo.

-Problema 17: Se requiere mostrar los resultados del **editor de datos** (**imputaciones**) para analizar las opciones previamente seleccionadas (**5** en nuestro caso **+ el original**). Se observarán los diferentes valores imputados, para sus consideraciones. Repetir secuencia de comandos del **problema 16**. **Ver Figura 3.24**

Figura 3.24. Cuadro de diálogo cambio de visor de resultados imputación múltiple



Fuente: SPSS 20 IBM

-Resultados: la **imputación 3,** es la que se elige por presentar los resultados imputados **<=10.** El archivo se puede renombrar como: **CKM_MKT_Digital_imputaciones3.sav**

3.7.4.Casos de datos atípicos (outliers)

Son observaciones con características identificables que les diferencia claramente de las otras observaciones. **No es posible caracterizados categóricamente** y si pudieran ser divididos debido al contexto del análisis y evaluados por los tipos de información que pueden proporcionar, serían clasificados como:

- Benéficos aunque diferentes a la mayor parte de la muestra, tienen el potencial de ser indicativos de las características segmento de la población en el que se descubren, por el curso normal del análisis. P
- Problemáticos, los cuales no son representativos de la población y están en contra de los objetivos del análisis. Pueden distorsionar seriamente los test estadísticos.

Debido a la variabilidad en la evaluación de los casos atípicos, es muy importante que examine los datos en busca de su presencia a fin de averiguar el tipo de influencia e impacto que ejercen. Es una herramienta importante la regresión lineal para su evaluación y a fin de explicarse su causa, se les clasifica en **4 categorías:**

1. Los que son por **error de procedimiento,** tales como la entrada de datos o un error de codificación. Para evitarlos, **deberían identificarse en el nivel de filtrado de datos** y si se pasan por alto, deberían eliminarse o recodificarse como **datos ausentes**.
2. Los que ocurren como consecuencia de un **acontecimiento extraordinario.** Aquí, existe una explicación para la unicidad de la observación. Usted debe decidir si el caso debe o no ser representado en la muestra. Si es así, el caso atípico **debe ser retenido** en el análisis; si no, hay que **suprimirlo**.
3. Los que ocurren debido a que Usted **no tiene explicación**. Son los casos más **apropiados para ser omitidos**, pueden retenerse si cree que representan un segmento válido de la población.
4. Los que ocurren al situarse **fuera del rango ordinario de valores de cada variable** pero que son únicos en su combinación de valores entre las variables. Aquí deberá retener la observación a menos que se disponga de evidencia específica que excluya al caso como un miembro válido de la población.

Son utilizados diversos métodos en la detección de casos atípicos para las situaciones univariantes, bivariantes y multivariantes. Una vez identificados, deben especificarse en una de las cuatro categorías descritas. Finalmente, deberá decidir sobre la retención o exclusión de cada caso atípico, juzgando las características del caso y los objetivos del análisis.

Se identifican desde la perspectiva: **univariante, bivariante o multivariante.** Debe **utilizar cuantas perspectivas sean posibles**, buscando una consistencia entre los métodos de identificación de casos atípicos. Así, tenemos los métodos bajo las perspectivas, siguientes:

1. **Detección univariante.** Esta perspectiva univariante de identificación de casos atípicos examina la distribución de observaciones, seleccionando como **casos atípicos aquellos que caigan fuera de los rangos de la distribución.** Se debe establecer un **umbral para la designación como caso atípico**. Convierte en primer lugar los valores de los datos en **valores estándar**, que tienen una **media cero y una desviación estándar de uno**. Los valores al expresarse en un formato estandarizado, pueden realizar fácilmente comparaciones entre las variables. **Para muestras pequeñas (de 80 o incluso menos observaciones), se sugiere identificar como atípicos aquellos**

casos con valores estándar de 2.5 o superiores. Cuando los tamaños **muestrales son mayores,** las pautas sugieren que **el valor umbral del estandarizado se sitúe entre 3 y 4**. Si no usa los valores estándares, entonces puede identificar los casos que tienen lugar **fuera de las gamas de 2.5 vs. 3 o 4 desviaciones estándares**, lo cual **depende del tamaño muestra**. En cualquier caso, debe darse cuenta que normalmente ocurra que un cierto número de observaciones caigan fuera de esos rangos de la distribución. Deberá esforzarse en identificar sólo aquellas observaciones verdaderamente distintivas y designarlas como casos atípicos.

2. **Detección bivariante.** Además de la evaluación anterior, pueden evaluarse conjuntamente pares de variables mediante un gráfico de dispersión. Casos que se ubiquen manifiestamente fuera del rango del resto de las observaciones pueden identificarse como **puntos aislados** en el gráfico de dispersión **una elipse** que represente un intervalo de confianza especificado (**variando entre 50 y 90 % de la distribución)** para una **distribución normal bivariante.** Esto proporciona una representación gráfica de los **límites de confianza** y facilita la identificación de casos atípicos. **El gráfico de la influencia es otra variante del gráfico de dispersión.** En este caso, cada punto varía en tamaño según su influencia en las relaciones. Estos métodos reportan cierta evaluación de la influencia de cada observación para completar la designación de casos como casos atípicos.
3. **Detección multivariante.** Ésta clasificación implica una evaluación multivariante de cada observación a lo largo de un conjunto de variables. Tomando en cuenta que la mayoría de los análisis multivariantes tienen más de dos variables, necesitará una forma de medición objetiva de la posición multidimensional de cada observación relativa a un punto común. **La medida D^2 de *Mahalanobis* puede usarse con este fin, ya que es una medida de la distancia de cada observación en un espacio multidimensional respecto del centro medio de las observaciones.** Debido a que proporciona una medida común de **centralidad multidimensional,** también tiene propiedades estadísticas que tienen en cuenta las **pruebas de significación.** Por la naturaleza de los test estadísticos, se sugiere que se use un nivel muy conservador, quizá **0.001, como valor umbral para la designación como caso atípico.**
4. **Designación como caso atípico.** Cuando las observaciones estimadas como caso atípico han sido identificadas por métodos **univariantes, bivariantes o multivariantes**, debe entonces seleccionar aquellas observaciones que demuestran **una unicidad real** en comparación con el resto de la población. Debe **abstenerse de designar muchas observaciones como casos atípicos** y **no debe caer en la tentación de eliminar aquellos casos que no son consistentes con los casos restantes, simplemente porque son diferentes.**

3.7.5. Descripción de casos atípicos y especificación

Una vez identificados los casos atípicos, debe generar identificaciones de cada observación atípica y examinar cuidadosamente que los datos de las variables lo sean. Además del examen visual, puede emplear también técnicas multivariantes como el **análisis discriminante** o la **regresión múltiple** para **identificar las diferencias entre los casos atípicos y las otras observaciones**. Deberá continuar este análisis hasta que sea satisfactorio el aspecto de los datos que distinguen el caso atípico del resto de las observaciones. Asigne el caso atípico a uno de los cuatro tipos citados.

3.7.6. Mantenimiento o eliminación de los casos atípicos

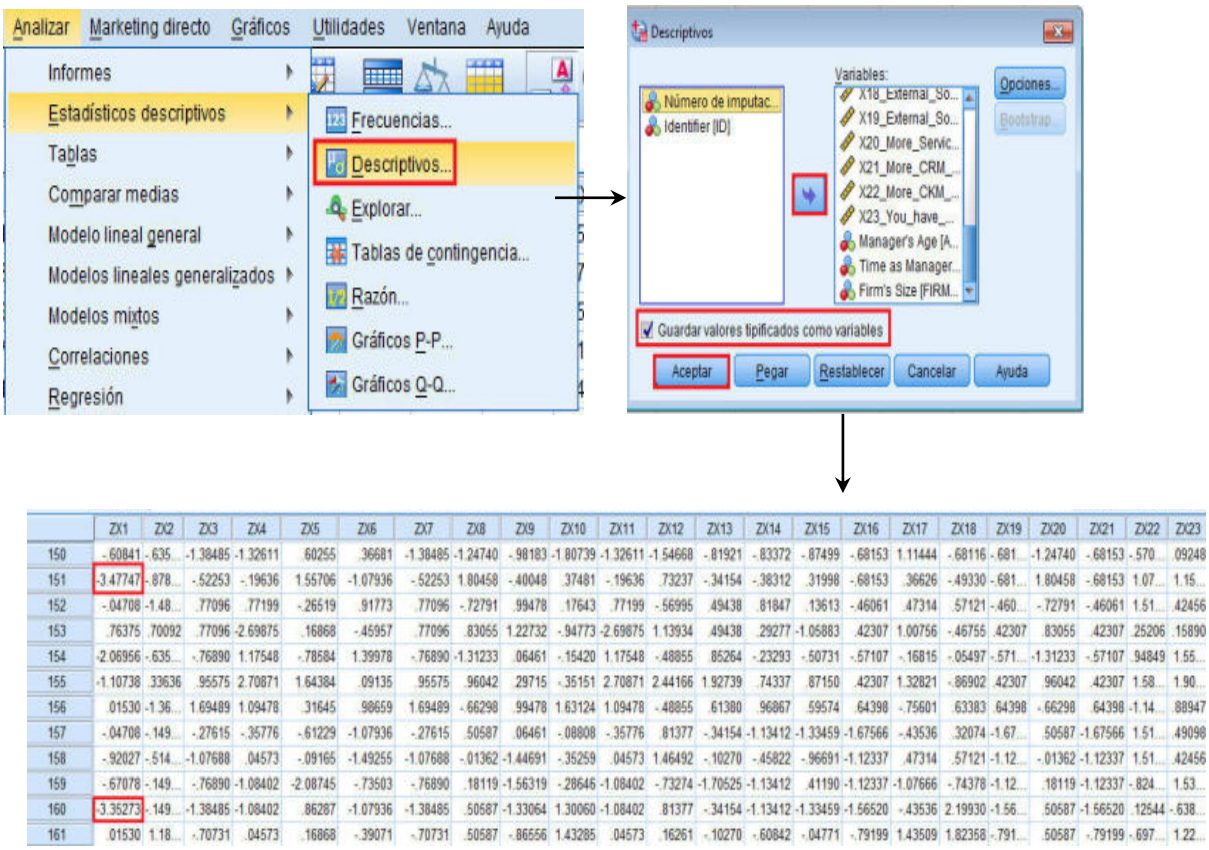
Una vez identificados, especificados y catalogado los casos atípicos, debe decidir entre mantenerlos o destruirlos. Hay muchos supuestos entre los investigadores sobre cómo tratar con los casos atípicos. Nuestro supuesto es que deberían mantenerse a menos que exista una prueba demostrable de que son **verdaderas aberraciones** y **no son representativos de las observaciones de la población**. Pero si representan a un segmento de la población**, las debe retener para asegurar su generalidad al conjunto de la población.** Si se eliminan los casos atípicos, **corre el riesgo de mejorar el análisis pero limitar su generalidad.** Si los casos atípicos son problemáticos en una técnica particular, muchas veces pueden ser manejados de una forma tal que se ajusten al análisis sin que lo distorsionen significativamente.

El análisis de casos atípicos considera un diagnóstico **univariante, bivariante y multivariante**. De encontrarse, serán examinados y se procederá a decidir su mantenimiento o eliminación, mediante las **2 test estadísticos**:

1. **Test detección univariante y bivariante.** El primer paso es examinar las observaciones de cada una de las variables individualmente. Se debe recordar que es una puntuación extrema de una variable continua, que debe identificarse para evitar errores en el análisis estadístico motivado por varias causas ya analizadas. Para detectarlas, hay que medirlas en **puntuaciones (*z*), que miden la distancia de cada punto con respecto a la media en desviaciones típicas.** Se considera que en **muestras pequeñas <80, *z*=>2.5; en muestras grandes >80, *z*>=3.**

   Problema 18: identifique de la base de datos **CKM_MKT_Digital_imputaciones3.sav**, los casos atípicos, por medio de ésta técnica.
   Teclear: Analizar->Estadísticos descriptivos->Seleccionar variable; Guardar valores tipificados como variable (aquí se da conversión a valores z) ->Aceptar. Ver Figura 3.25.

Figura 3.25. Determinación de casos atípicos mediante medición de z



	ZX1	ZX2	ZX3	ZX4	ZX5	ZX6	ZX7	ZX8	ZX9	ZX10	ZX11	ZX12	ZX13	ZX14	ZX15	ZX16	ZX17	ZX18	ZX19	ZX20	ZX21	ZX22	ZX23
150	-.60841	-.635...	-1.38485	-1.32611	.60255	.36681	-1.38485	-1.24740	-.98183	-1.80739	-1.32611	-1.54668	-.81921	-.83372	-.87499	-.68153	1.11444	-.68116	-.681...	-1.24740	-.68153	-.570...	.09248
151	-3.47747	-.878...	-.52253	-.19636	1.55706	-1.07936	-.52253	1.80458	-.40048	.37481	-.19636	.73237	-.34154	-.38312	.31998	-.68153	.36626	-.49330	-.681...	1.80458	-.68153	1.07...	1.15...
152	-.04708	-1.48...	.77096	.77199	-.26519	.91773	.77096	-.72791	.99478	.17643	.77199	-.56995	.49438	.81847	.13613	-.46061	.47314	.57121	-.460...	-.72791	-.46061	1.51...	.42456
153	.76375	.70092	.77096	-2.69875	.16868	-.45957	.77096	.83055	1.22732	-.94773	-2.69875	1.13934	.49438	.29277	-1.05883	.42307	1.00756	-.46755	.42307	.83055	.42307	.25206	.15890
154	-2.06956	-.635...	-.76890	1.17548	-.78584	1.39978	-.76890	-1.31233	.06461	-.15420	1.17548	-.48855	.85264	-.23293	-.50731	-.57107	-.16815	-.05497	-.571...	-1.31233	-.57107	.94849	1.55...
155	-1.10738	.33636	.95575	2.70871	1.64384	.09135	.95575	.96042	.29715	-.35151	2.70871	2.44166	1.92739	.74337	.87150	.42307	1.32821	-.86902	.42307	.96042	.42307	1.58...	1.90...
156	.01530	-1.36...	1.69489	1.09478	.31645	.98659	1.69489	-.66298	.99478	1.63124	1.09478	-.48855	.61380	.96867	.59574	.64398	-.75601	.63383	.64398	-.66298	.64398	-1.14...	.88947
157	-.04708	-.149...	-.27615	-.35776	-.61229	-1.07936	-.27615	.50587	.06461	-.08808	-.35776	.81377	-.34154	-1.13412	-1.33459	-1.67566	-.43536	.32074	-1.67...	.50587	-1.67566	1.51...	.49098
158	-.92027	-.514...	-1.07688	.04573	-.09165	-1.49255	-1.07688	-.01362	-1.44691	-.35259	.04573	1.46492	-.10270	-.45822	-.96691	-1.12337	.47314	.57121	-1.12...	-.01362	-1.12337	1.51...	.42456
159	-.67078	-.149...	-.76890	-1.08402	-2.08745	-.73503	-.76890	.18119	-1.56319	-.28646	-1.08402	-.73274	-1.70525	-1.13412	.41190	-1.12337	-1.07666	-.74378	-1.12...	.18119	-1.12337	-.824...	1.53...
160	-3.35273	-.149...	-1.38485	-1.08402	.86287	-1.07936	-1.38485	.50587	-1.33064	1.30060	-1.08402	.81377	-.34154	-1.13412	-1.33459	-1.56520	-.43536	2.19930	-1.56...	.50587	-1.56520	.12544	-.638...
161	.01530	1.18...	-.70731	.04573	.16868	-.39071	-.70731	.50587	-.86656	1.43285	.04573	.16261	-.10270	-.60842	-.04771	-.79199	1.43509	1.82358	-.791...	.50587	-.79199	-.697...	1.22...

Fuente: SPSS 20 IBM

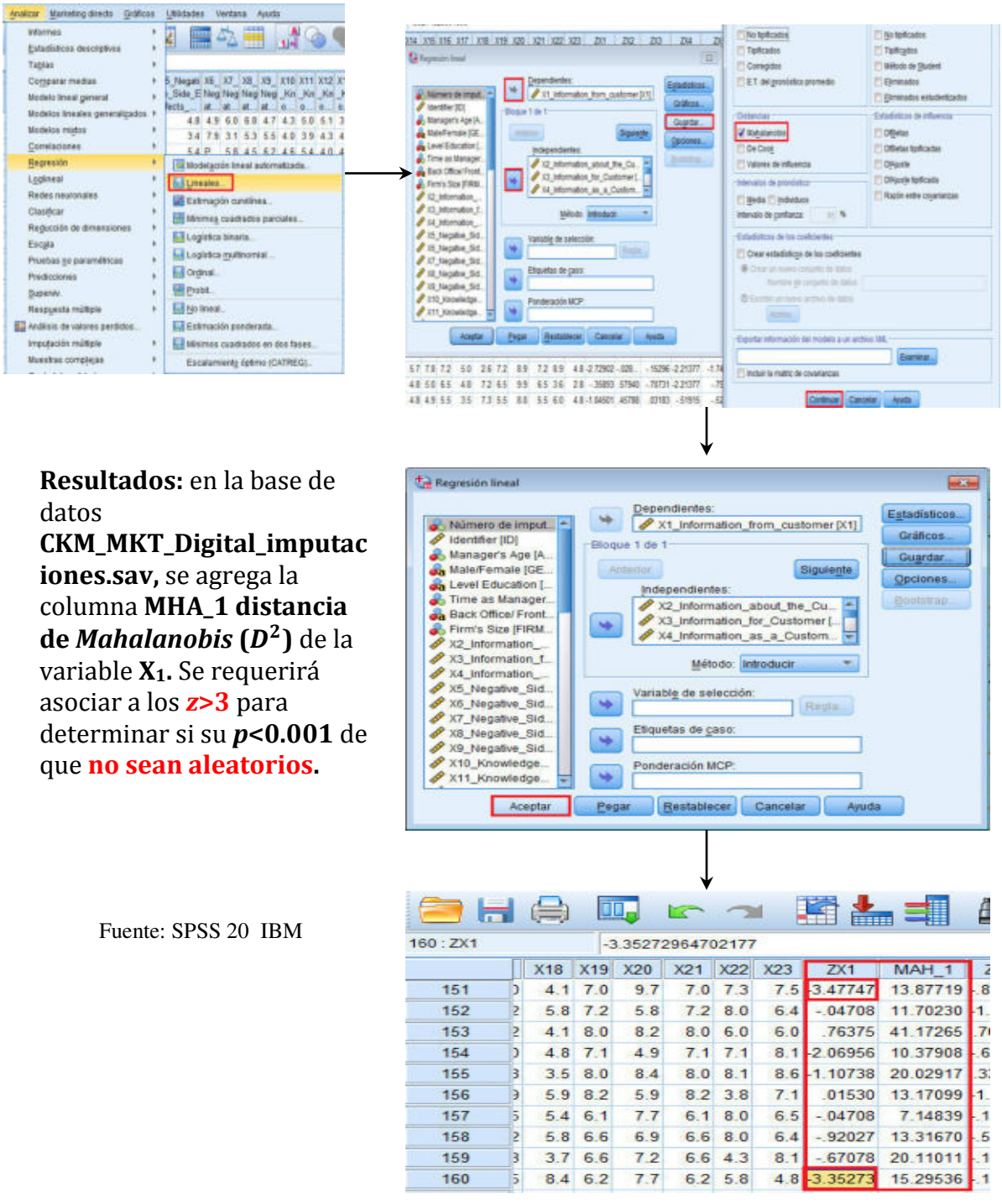
Resultado: al consultar la base de datos **CKM_MKT_Digital_imputaciones3.sav**, se observará que se generará columna $\mathbf{ZX_2}$ el cual arroja los **valores *z*** calculados a fin de determinar aquellos casos atípicos que a nivel magnitud sean ***z*>=3.** El ejemplo del **registro 151 y 160** de $\mathbf{ZX_1}$ **son atípicos.**

2. **Test detección multivariante.** Este método de diagnóstico, **es complementario** y aquí se calcula la probabilidad de que **no sean producidas por el azar** evaluando los casos atípicos multivariantes con la **distancia de *Mahalanobis* (D^2).** Este método evalúa la posición de cada observación comparada con el centro de todas las observaciones de un conjunto de variables. Los **test de significación estadística** con esta medida son muy conservadores (<0.001). con este umbral, se identifican **2 observaciones como significativamente diferentes.** Es interesante resaltar que estas observaciones no sólo en los test multivariantes. **Esto indica que no son únicas en cada variable aislada sino que son únicas en la combinación de variables.**

   -Problema 19: determinar que los casos atípicos de la basa de datos **CKM_MKT_Digital_imputaciones.sav,** tenga una ***p*<0.001** de que **no sean aleatorios.** Usar **distancia de *Mahalanobis* (D^2)**

-Teclear: Analizar-> Regresión -> Lineales -> Selección de variables independiente: X_1; selección de variables independientes: X_2 a X_{23} -> Guardar: Selección: distancia de *Mahalanobis* -> Continuar-> Aceptar. Ver Figura 3.25

Figura 3.26. Proceso para detectar la probabilidad de casos atípicos por distancia de *Mahalanobis*.



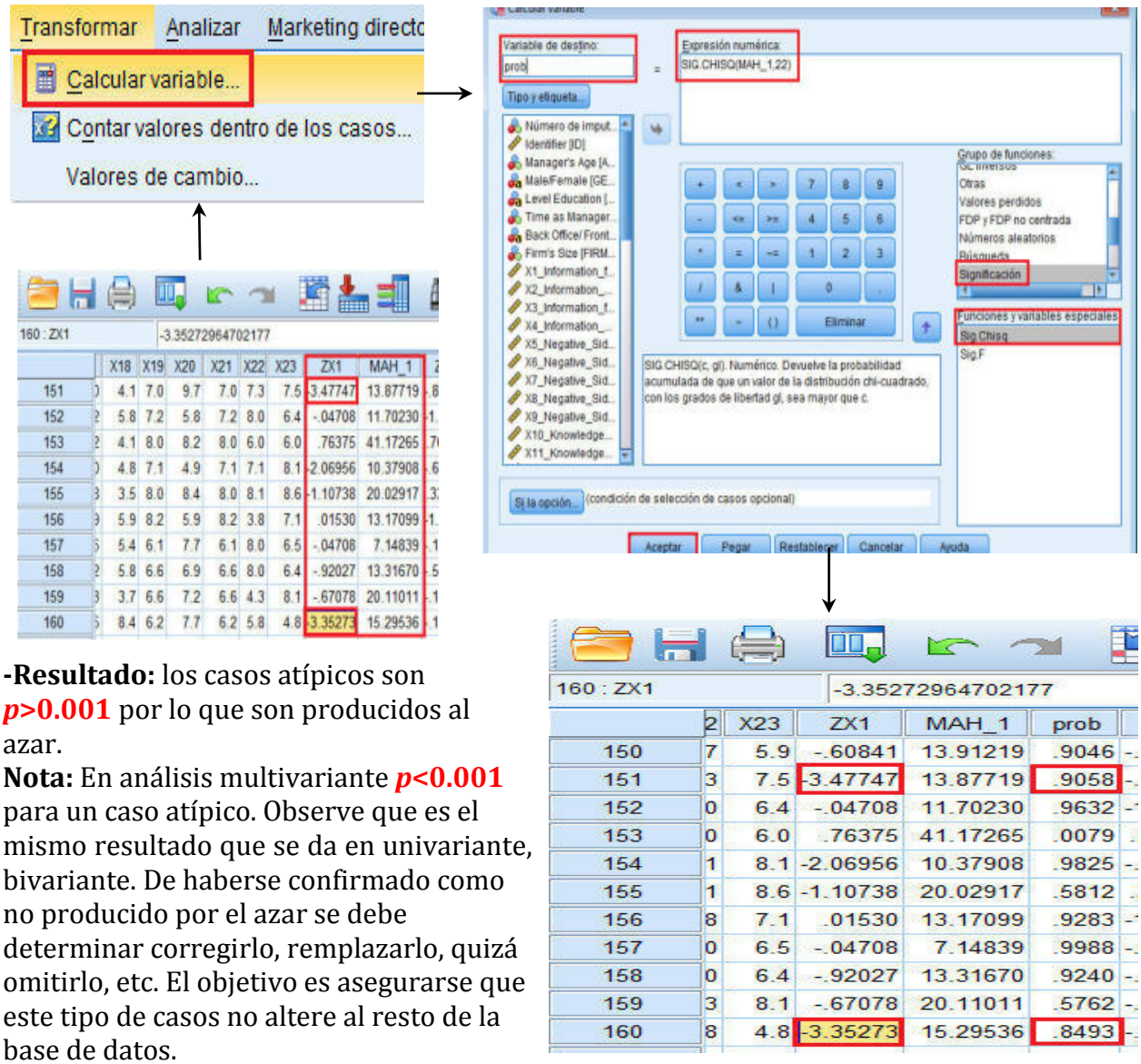
	X18	X19	X20	X21	X22	X23	ZX1	MAH_1
151	4.1	7.0	9.7	7.0	7.3	7.5	-3.47747	13.87719
152	5.8	7.2	5.8	7.2	8.0	6.4	-.04708	11.70230
153	4.1	8.0	8.2	8.0	6.0	6.0	.76375	41.17265
154	4.8	7.1	4.9	7.1	7.1	8.1	-2.06956	10.37908
155	3.5	8.0	8.4	8.0	8.1	8.6	-1.10738	20.02917
156	5.9	8.2	5.9	8.2	3.8	7.1	.01530	13.17099
157	5.4	6.1	7.7	6.1	8.0	6.5	-.04708	7.14839
158	5.8	6.6	6.9	6.6	8.0	6.4	-.92027	13.31670
159	3.7	6.6	7.2	6.6	4.3	8.1	-.67078	20.11011
160	8.4	6.2	7.7	6.2	5.8	4.8	-3.35273	15.29536

Resultados: en la base de datos **CKM_MKT_Digital_imputaciones.sav,** se agrega la columna **MHA_1 distancia de *Mahalanobis* (D^2)** de la variable **X_1.** Se requerirá asociar a los ***z*>3** para determinar si su ***p*<0.001** de que **no sean aleatorios.**

Fuente: SPSS 20 IBM

-Problema 20: sobre calcular la probabilidad de que las **distancia de *Mahalanobis* (D^2) no sean producidas al azar,** es decir **que exista una baja probabilidad.**
-Teclear: Transformar -> Calcular variables -> Nombre variable de destino= probabilidad-> Grupo de funciones seleccionar Significación; Funciones y variables especiales en Significación de Chi-cuadrada; flecha hacia arriba-> seleccionar la variable de distancia de *Mahalanobis* calculada; flecha hacia arriba y dar los grados de libertad en función a las (22) variables métricas calculadas-> Aceptar. Ver Figura 3.27

Figura 3.27. Proceso para detectar la probabilidad de casos atípicos por distancia de *Mahalanobis* (D^2)



	X18	X19	X20	X21	X22	X23	ZX1	MAH_1
151	4.1	7.0	9.7	7.0	7.3	7.5	-3.47747	13.87719
152	5.8	7.2	5.8	7.2	8.0	6.4	-.04708	11.70230
153	4.1	8.0	8.2	8.0	6.0	6.0	.76375	41.17265
154	4.8	7.1	4.9	7.1	7.1	8.1	-2.06956	10.37908
155	3.5	8.0	8.4	8.0	8.1	8.6	-1.10738	20.02917
156	5.9	8.2	5.9	8.2	3.8	7.1	.01530	13.17099
157	5.4	6.1	7.7	6.1	8.0	6.5	-.04708	7.14839
158	5.8	6.6	6.9	6.6	8.0	6.4	-.92027	13.31670
159	3.7	6.6	7.2	6.6	4.3	8.1	-.67078	20.11011
160	8.4	6.2	7.7	6.2	5.8	4.8	-3.35273	15.29536

160 : ZX1 -3.35272964702177

	2	X23	ZX1	MAH_1	prob
150	7	5.9	-.60841	13.91219	.9046
151	3	7.5	-3.47747	13.87719	.9058
152	0	6.4	-.04708	11.70230	.9632
153	0	6.0	.76375	41.17265	.0079
154	1	8.1	-2.06956	10.37908	.9825
155	1	8.6	-1.10738	20.02917	.5812
156	8	7.1	.01530	13.17099	.9283
157	0	6.5	-.04708	7.14839	.9988
158	0	6.4	-.92027	13.31670	.9240
159	3	8.1	-.67078	20.11011	.5762
160	8	4.8	-3.35273	15.29536	.8493

Fuente: SPSS 20 IBM

-Resultado: los casos atípicos son ***p*>0.001** por lo que son producidos al azar.
Nota: En análisis multivariante ***p*<0.001** para un caso atípico. Observe que es el mismo resultado que se da en univariante, bivariante. De haberse confirmado como no producido por el azar se debe determinar corregirlo, remplazarlo, quizá omitirlo, etc. El objetivo es asegurarse que este tipo de casos no altere al resto de la base de datos.

3.8. Mantenimiento o eliminación de datos atípicos (outliers)

Se recomienda seguir, lo siguiente:

3.8.1. Mantenimiento o eliminación de casos atípicos

Como resultado de estos **test de diagnóstico**, ninguna observación parece mostrar las características de un caso atípico que debiera ser eliminado. Cada variable tiene algunas observaciones que son extremas, y que deberían considerarse si se va a utilizar la variable en el análisis. Pero ninguna de las observaciones son extremas sobre un número suficiente de variables como para ser consideradas no representativas de la población. En todos los casos, las observaciones denominadas como casos atípicos, incluso con los test multivariantes, parecen suficientemente similares al resto de las variables como para retenerlas en el análisis multivariante. No obstante, el investigador debería siempre examinar los resultados de cada técnica específica para identificar observaciones que pueden llegar a ser atípicas en esa aplicación particular.

3.8.2. Detección por método multivariante de casos atípicos

Métodos univariados: examine todas las variables métricas para identificar valores únicos y extremos:

- Para muestras **<= hasta 80**, los atípicos son definidos como casos con valores estándar de **>= 2.5.**
- Para muestras más grandes el valor estándar se incrementa hasta **4**
- Si no son usados los valores estándar, identifique aquellos casos que caigan fuera del rango de **2.5 vs. 4** desviaciones estándar, dependiendo del tamaño de la muestra.

Métodos bivariados: enfoque su uso en relaciones específicas de las variables, como variable independiente vs. Dependiente.

- Use gráficos de dispersión con intervalos de confianza en un nivel específico de alfa.

Métodos Multivariantes: muy conveniente en el examen de variabilidad, tales como las variables independientes en regresión o en el caso de análisis factorial.

- Los niveles de umbral para **(D^2) distancia de *Mahalanobis* / grados de libertad (gl),** deben ser conservadores (**0.005 a 0.001**), resultado en valores de **2.5** (pequeñas muestras) vs. **3 o 4** (muestras grandes)

3.9. Supuestos del análisis multivariante

Se recomienda seguir, lo siguiente:

3.9.1. Importancia de los supuestos del análisis multivariante

Ésta es la última etapa del examen de datos y es de relevante importancia debido a:

1. **Que** el **uso habitual de una gran cantidad de variables, hace que las distorsiones y los sesgos potenciales sean más potentes cuando se incumplan los supuestos**. En realidad, las violaciones combinadas llegar a ser incluso más perjudiciales que si se consideran separadamente.
2. La complejidad de los análisis y de los resultados pueden **enmascarar los "*signos*" de las violaciones** de los supuestos que son aparentes en los más sencillos análisis multivariantes.

En casi todos los ejemplos, los procedimientos se estimará el **modelo multivariante** y se producirán resultados incluso cuando los supuestos **se vean severamente incumplidos**. Por tanto, **debe estar atento a cualquier incumplimiento de los supuestos y a las implicaciones que puedan tener para el proceso de estimación o interpretación de los resultados.**

3.9.2. Valoración de las variables individuales frente al modelo univariante

El análisis multivariante requiere que los supuestos subyacentes a las técnicas estadísticas sean contrastados en 2 etapas:

Etapa 1. Para las variables aisladas, semejante a las pruebas de los supuestos del análisis univariante,

Etapa 2. Para el valor teórico del modelo multivariante, que actúa colectivamente sobre las variables a analizar y por tanto debe cumplir los mismos supuestos que las variables individuales.

Esta sección se centra en el examen de las variables univariantes en relación al cumplimiento de los supuestos subyacentes a los procedimientos multivariantes.

3.9.3. Normalidad

Es el supuesto fundamental del análisis multivariante de datos, en referencia al **perfil de la distribución de los datos** para **una única variable métrica y su correspondencia con una distribución normal**, punto de referencia de los métodos estadísticos. Si la variación respecto de la distribución normal es suficientemente amplia, **todos los test estadísticos resultantes no son válidos**, dado que **se requiere la normalidad** para el uso de los **estadísticos de la *t* y de la *F*.** Tanto los métodos estadísticos univariantes como los multivariantes analizados aquí, se basan en el supuesto de **la normalidad univariante**, suponiendo también los multivariantes la **normalidad multivariante**.

La **normalidad univariante** para una única variable es fácil de contrastar, siendo posible varias medidas correctoras, como se muestra más adelante. En otras palabras, la **normalidad multivariante (la combinación de dos o más variables)** implica que las **variables individuales sean normales en un sentido univariante y que sus combinaciones también sean normales.** Por tanto, **si una variable es una normal multivariante, es también normal univariante**. Sin embargo, **lo contrario no es necesariamente cierto (dos o más variables normales univariantes no son necesariamente normal multivariante)**. Por tanto, una situación en la que todas las variables exhiben **normalidad univariante** ayudará a obtener normalidad multivariante, aunque no la garantiza. La **normalidad multivariante** es mucho más difícil de contrastar, aunque existen varios test para situaciones en las que la técnica multivariante se ve particularmente afectada por una violación de los supuestos. Más adelante, nos centraremos en evaluar y **alcanzar la normalidad univariante para todas las variables y acudiremos a la multivariante cuando sea especialmente crítica.** Incluso aunque las muestras grandes tiendan a disminuir los efectos perniciosos de la no normalidad, debe evaluar la normalidad de todas las variables incluidas en el análisis

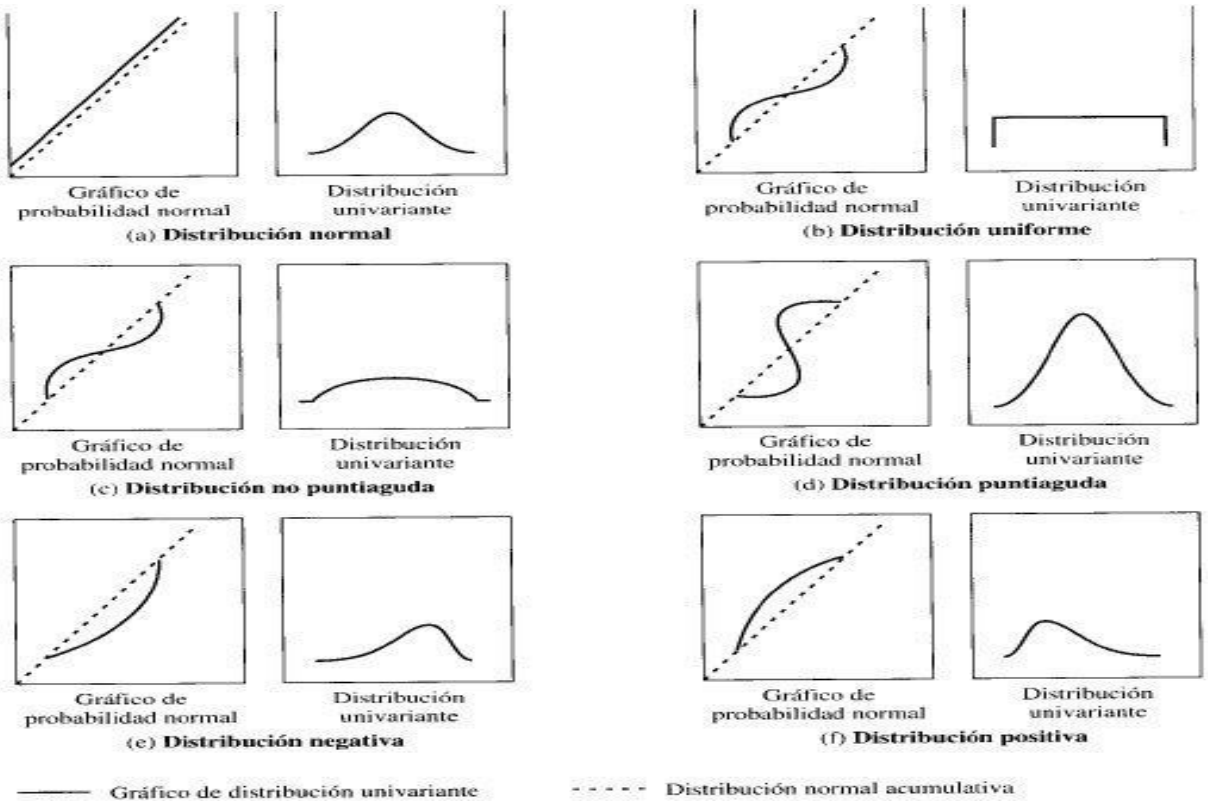
3.9.4. Análisis gráfico de la normalidad

Se puede realizar mediante:

1. El test más simple para diagnosticar la normalidad es una **comprobación visual del histograma** que compare los valores de los datos observados con una distribución aproximada a la distribución normal. Aunque atractivo por ser simple, este método es problemático para muestras pequeñas, donde la construcción del histograma (por ejemplo, el número de categorías o la anchura de las categorías) puede distorsionar la representación visual de tal forma **que el análisis sea inútil**.
2. Una aproximación de **mayor confianza es el gráfico de distribución normal**, que compara la distribución acumulada de los valores reales de los datos con la distribución acumula da de una distribución normal. La **distribución normal** sigue una **línea recta en diagonal**, comparándola con el gráfico de los valores de los datos. **Si una distribución es normal, la línea que representa la distribución real de los datos sigue de cerca a la diagonal.**

La **Figura 3.28** muestra **varios gráficos de distribución normal** y **distribución univariante** de la variable.

Figura 3.28. Gráficos de distribución normal y las correspondientes distribuciones univariantes



Fuente: Hair et al. (1999)

En los **gráficos de distribución normal** se representa una característica del **perfil de la distribución,** la **curtosis** o **"*apuntamiento*"** o **"*llanura*"** de la distribución, comparada con la **distribución normal.** Se crean por tanto las siguientes opciones:

1. **Cuando la línea cae por debajo de la diagonal, la distribución es más llana de lo esperado (*platicúrtica*).**
2. **Cuando la línea cae por encima de la diagonal, la distribución es más puntiaguda que la curva normal (*leptocúrtica*).**

-Problema 21: Comente las curvas mostradas en la anterior figura.
-Resultado: Por ejemplo, en el **gráfico de distribución normal** (**Figura 3.28d**), vemos una curva con un nítido **perfil en S**. Inicialmente la distribución es más plana, **y la línea cae por debajo de la diagonal.** Entonces, la parte **puntiaguda** de la distribución se mueve rápidamente por encima de la diagonal y se desplaza otra vez por debajo de la diagonal a medida que la distribución se aplana. Una distribución **no puntiaguda (*platicúrtica*)** tiene una pauta opuesta (**Figura 3.28c**). Otro modelo común **es un simple arco**, tanto por encima como por debajo de la diagonal, que indica la simetría de la distribución. Una **simetría negativa** (**Figura 3.28e**) se indica mediante un **arco por debajo de la diagonal, mientras que un arco por encima de la diagonal representa una distribución positivamente simétrica (Figura 3.28f).** Una excelente fuente para interpretar los gráficos de distribución normal que muestren los diversos modelos e interpretaciones es (**Daniel y Wood 1980).** Estos modelos específicos no sólo identifican la no normalidad, sino que también nos dicen la forma de la distribución original y la solución apropiada a aplicar.

3.9.5. Test estadístico de normalidad

Además de examinar el gráfico de distribución normal, pueden utilizarse también test estadísticos para evaluar la normalidad. El test más simple es una regla basada en el valor de simetría (disponible como parte de los estadísticos descriptivos básicos para una variable procesada en todos los programas estadísticos). El valor estadístico (z) se calcula como: $z_{simetría}$***=simetría/*** $\sqrt{6/N}$ donde **N** es el tamaño de la muestra. Un valor z también puede ser calculado para el valor de curtosis utilizado la siguiente formula:$z_{simetría}$ ***=curtosis/*** $\sqrt{(24/N)}$
Si el valor calculado de z excede un valor crítico, entonces la distribución es no normal por lo que se refiere a esta característica. El valor crítico es de una distribución z, basada en los niveles de significación que deseemos. Por ejemplo, un valor calculado que exceda: **±2.58** indica que podemos rechazar el supuesto sobre la normalidad de la distribución a un nivel de probabilidad de 0.01. Otro valor crítico habitualmente utilizado es: ± **1.96,** que corresponde a un nivel de error de **0.05.** Los estadísticos son:

1. **Test de *Shapiro-Wilks*, la cual se sugiere como una muestra N <50 observaciones (p>0.05 se acepta Ho)**
2. **Modificación de test de *Kolmogorov-Smirnov*; la cual se sugiere como una muestra N>50 observaciones (p>0.05 se acepta Ho). Se recomienda usarla siempre y cuando no aplique, usar el test *Shapiro-Wilks*.**

Cada uno calcula el nivel de significación para las diferencias respecto a una distribución normal. Se debe recordar que los **test de significación** son menos útiles en

muestras pequeñas **(menores de 30)** y muy sensibles para grandes muestras (**superiores a 1000 observaciones).** Por tanto, el investigador deberá siempre usar tanto los gráficos como cualquier comprobación estadística para evaluar el **grado real de desviación de la normalidad.**

3.9.6. Soluciones para la no normalidad

Existen métodos basados en **transformaciones de datos** para **acomodar las distribuciones no normales (**se exponen más adelante). Esta sección se orienta a la discusión de los test de **normalidad univariante y las transformaciones**. Sin embargo, cuando examinemos los otros **métodos multivariantes**, tales como la **regresión múltiple o el análisis multivariante de la varianza**, y **los test de normalidad multivariante**. Más aún, en muchas ocasiones en que se indica **normalidad** es en realidad un resultado de otras violaciones de los supuestos; por tanto, **remediando los otros incumplimientos eliminamos el problema de la normalidad**. Por esta razón, deber **realizar test de normalidad después o junto con los análisis y soluciones para las otras violaciones.** (Más información de la normalidad multivariante, ver **(**Johnson, R. A., y Wichern D. W. 1982**,** Weisberg, S. 1985**)**

3.9.7. Homocedasticidad

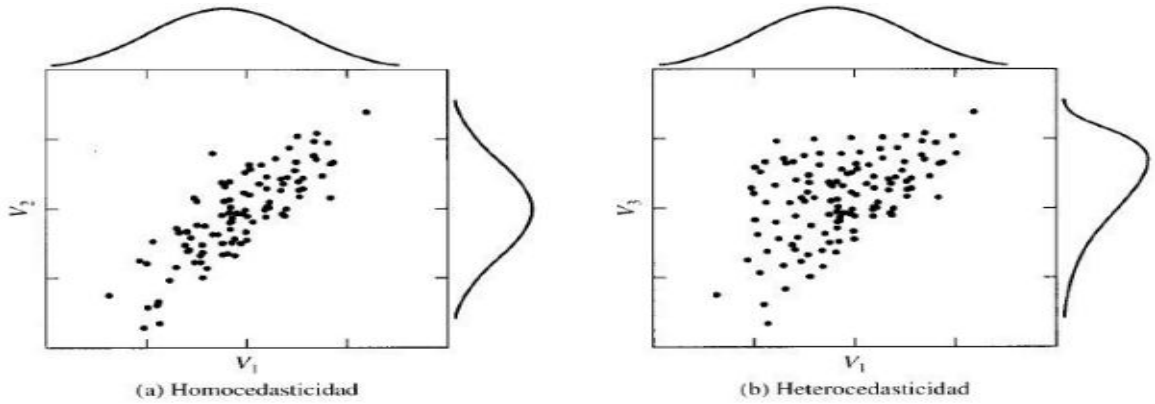
Es un supuesto relativo primordialmente a las relaciones de dependencia entre variables. Se refiere al supuesto de que las **variables dependientes** exhiban **iguales niveles de varianza** a lo largo del rango del predictor de las variables. Es un fenómeno deseable por que la varianza de la variable dependiente que se está explicando en la relación de dependencia no debería concentrarse sólo en un limitado rango de los valores independientes. Aunque **las variables dependientes deben ser métricas**, este concepto de igual **dispersión de la varianza a lo largo de las variables independientes** puede aplicarse cuando las variables son **métricas o no métricas**. Con **variables independientes métricas**, el concepto se basa en la **dispersión de la varianza de la variable dependiente a lo largo del rango de los valores de la variable independiente, que se encuentra en técnicas como la regresión múltiple.**

El mismo concepto se aplica también cuando las **variables independientes son no métricas**. En estos casos, tal y como se encuentran en **ANOVA y MANOVA**, **el centro es ahora la igualdad de la varianza (una variable de pendiente)** o la **matriz de varianza/covarianza (varias variables independientes**) a lo largo de los grupos formados por las **variables independientes no métricas**. La igualdad de las **matrices de varianza/covarianza** se observa también en el **análisis discriminante,** pero en esta técnica el énfasis es en la **dispersión de las variables independientes** a lo largo de los grupos formados por la medida **dependiente no métrica**. En cada uno de estos casos, el propósito es el mismo: **asegurar que la varianza usada en la explicación y predicción se disperse a través del rango de valores, permitiendo así un "*test limpio*" de las relaciones a lo largo de todos los valores de las variables no métricas**.

En la mayoría de las situaciones, tenemos diferentes valores de la variable dependiente para cada valor de la variable independiente. Para que esta relación se capte completamente, la dispersión (varianza) de los valores de la variable dependiente debe ser igual para cada valor de la variable predictor. La mayoría de los problemas con varianzas

desiguales surgen de una de estas dos fuentes. La primera es el tipo de variables incluidas en el modelo. Por ejemplo, a medida que una variable aumenta en valor (es decir, cuando las unidades van desde cero a millones), existe un rango más amplio de respuestas posibles para los valores más elevados. La segunda fuente surge de una distribución simétrica que crea heterocedasticidad. En la Figura **3.29a**, los gráficos de dispersión de puntos de los datos para dos variables (**V_1** y **V_2**) con distribuciones normales exhiben la misma dispersión a lo largo de todos los valores de los datos (es decir, **homocedasticidad**). Sin embargo, en la **Figura 3.29b,** observamos también una dispersión desigual **(heterocedasticidad)** provocada por la simetría de una de las variables (**V_2**). Para diferentes valores de **V_1** , tenemos diferentes pautas de dispersión para **V_2.** Esto provocará que las predicciones sean mejores a ciertos niveles de la variable independiente que a otros. Violando este supuesto a menudo realizamos unos test de las hipótesis muy conservadores o demasiado sensibles.

Figura 3.29. Gráficos de dispersión de relaciones de homocedasticidad y heterocedasticidad.



El efecto de la heterocedasticidad está a menudo también **relacionado con el tamaño de la muestra,** especialmente cuando examinamos la dispersión de la varianza entre grupos. Por ejemplo, en **ANOVA** o **MANOVA**, el impacto de la heterocedasticidad de los test estadísticos depende de los tamaños de la muestra asociados con los grupos de menor o mayor varianza. En el análisis de la regresión múltiple ocurrirán efectos similares en distribuciones altamente simétricas donde existan un número desproporcionado de encuestados en ciertos rangos de la variable independiente.

3.9.8. Test gráfico de igual dispersión de la varianza

La prueba de homocedasticidad de dos variables métricas se evalúa mejor gráficamente. La aplicación más común de esta forma de evaluación se produce en la regresión múltiple, en relación con la dispersión de la variable dependiente a lo largo de las variables independientes métricas. Dado que el eje del análisis de la regresión es el valor teórico, el gráfico de residuos se usa para revelar la presencia de homocedasticidad (o su opuesto, heterocedasticidad, desigual dispersión de la varianza). En análisis de regresión se detallan estos procedimientos en la discusión del análisis de los residuos. Los gráficos de cajas

sirven bien para representar el grado de variación entre los grupos formados por una variable categórica. El largo de la caja y de los bigotes indica la variación de los datos entre este grupo.

3.9.9. Test estadístico de homocedasticidad

Los test estadísticos de igual dispersión de la varianza se refieren a la varianza en grupos formados por variables métricas. El test más común, el **test de *Levene*,** puede usarse para evaluar si las varianzas de una única variable métrica son iguales a lo largo de cualquier cantidad de grupos. Si se está contrastando más de una variable métrica, implicando la comparación de la igualdad de **las matrices de varianzas/covarianzas**, se aplica el ***test M de Box*,** el cual existe tanto en el **análisis multivariante** como en el análisis **discriminante**.

3.9.10. Soluciones para la heterocedasticidad

Los problemas de heterocedasticidad pueden solucionarse a través de transformaciones de datos, similares a las usadas para **conseguir la normalidad**. Como ya se ha mencionado, en muchas ocasiones la heterocedasticidad es el resultado de la **no normalidad de una de las variables**, **y la corrección de la no normalidad resuelve igualmente la dispersión de la varianza**.

3.9.11. Linealidad

Es un supuesto implícito de todas las técnicas multivariantes basadas en **medidas de correlación**, e incluye a: la **regresión múltiple, regresión logística, análisis factorial y los modelos de ecuaciones estructurales**. Dado que las correlaciones representan sólo la **asociación lineal entre variables, los efectos no lineales no estarán representados en el valor de la correlación.** Como resultado, es siempre prudente examinar todas las relaciones para identificar cualquier desplazamiento de la linealidad que pueda impactar la correlación.

3.9.12. Identificación de relaciones no lineales

La forma más común de evaluar la linealidad es examinar los gráficos de dispersión de las variables e identificar cualquier pauta no lineal en los datos. Una aproximación alternativa es ir a un análisis de **regresión múltiple** y **examinar los residuos**. Los **residuos** reflejan **la parte no explicada de la variable dependiente**; por tanto, **cualquier parte no lineal** de la relación quedará reflejada en los residuos. El **examen de los residuos** puede aplicarse a la **regresión múltiple**, donde puede **detectar cualquier efecto no lineal** no representado en el **valor teórico de la regresión.**

3.9.13. Soluciones para la no linealidad

Si se detecta una relación no lineal, la aproximación más directa es transformar una o ambas variables para conseguir la linealidad. Posteriormente en este capítulo, se discutirán unas cuantas transformaciones. Una alternativa a la transformación de los datos es la creación de una nueva variable que represente la parte no lineal de la relación. El proceso de crear e interpretar estos resultados adicionales, que pueden usarse en todas las relaciones lineales, es posible visualizarlo mejor en la **técnica de regresión múltiple.**

3.9.14. Ausencia de errores correlacionados

La predicciones basadas en cualquiera de las técnicas de dependencia **no son perfectas**, y rara vez encontraremos una situación donde lo sean. Sin embargo, debemos asegurar que **cualquiera de los errores de predicción no esté correlacionado con el resto**. Por ejemplo, si se encuentra un indicio que sugiera que **los errores son positivos y negativos alternativamente**, se deberá entender que hay **alguna relación sistemática no explicada de la variable dependiente**. Si existe tal situación, no podemos estar seguros de que nuestros **errores de predicción sean independientes** de los niveles que estamos intentando predecir. Es decir, **existe otro factor que está afectando los resultados**, pero que no está incluido en el análisis.

3.9.15. La identificación de errores correlacionados

Una causa común del incumplimiento de este supuesto se debe al proceso de recolección de datos. Factores análogos **pueden afectar a un grupo y no afectar a otro**. Si se analizan separadamente, los **efectos conjuntos** son constantes y **no influyen** en la estimación de la relación. Pero **si se combinan las observaciones de ambos grupos**, entonces la relación final estimada debe ser un **"*compromiso*"** entre los **dos tipos** de relaciones. Esto provoca un ***"sesgo"*** de los datos porque una causa sin especificar está influyendo en la estimación de la relación. Para identificar los errores correlacionados, **debe identificar las causas posibles.** Siguiendo el ejemplo anterior, la causa podría ser que hay **dos grupos separados** en la recogida de datos. Una vez que la causa potencial haya sido identificada, debe **determinar si existen diferencias entre los grupos**. De encontrar diferencias en los **errores de predicció**n para los dos grupos, se tiene base para determinar que **un efecto no especificado es la "*causa*"** de los errores correlacionados.

3.9.16. Soluciones para los errores correlacionados

Los errores correlacionados **tienen que ser corregidos** con la inclusión del factor causante omitido en el análisis multivariante. En el ejemplo anterior, debe añadir una variable para indicar la clase donde estaban los encuestados. **La solución más común es la adición de una(s) variable(s)** al análisis que representa el factor omitido. La tarea clave a la que se enfrentará no es la solución en sí, sino la identificación del efecto no especificado y una manera de representarlo en el análisis.

3.9.17. Transformaciones de los datos

Estas proporcionan un medio para **modificar variables por una o dos razones**:

1. **Corregir el incumplimiento de los supuestos estadísticos** subyacentes a las técnicas multivariantes.
2. **Mejorar la relación** (**correlación**) entre variables.

La transformación de los datos puede basarse en razones:

1. **"*Teóricas*"** (transformaciones cuya conveniencia se basa en la **naturaleza de los datos**)
2. **"*Derivadas de los datos*"** (donde las transformaciones se sugieren a partir de un examen de los datos).

Así, en cada caso debe proceder muchas veces por **ensayo y error, ponderando** la mejora frente a la necesidad de transformaciones adicionales. Todas las transformaciones descritas

pueden llevarse a cabo fácilmente mediante simples comandos de todos los programas estadísticos (como el **SPSS**) y con más complicados y sofisticados (por ejemplo, véase **Box y Cox 1964**).

3.9.18. Transformaciones de los datos para conseguir la normalidad y la homocedasticidad

Las transformaciones de los datos proporcionan el **medio principal de corregir la no normalidad y heterocedasticidad**. En ambos casos, la **forma** de las variables **sugiere transformaciones específicas**:

Distribuciones no normales, las dos formas más comunes son las distribuciones "***planas***" y las **distribuciones asimétricas:**

1. Para la **distribución plana**, la transformación más común es **la inversa** (es decir, 1/Y) o
2. Las **distribuciones asimétricas** pueden ser transformadas empleando **la raíz cuadrada, logaritmos o incluso la inversa de la variable**.

Normalmente, las **distribuciones negativamente simétricas** se transforman de forma más efectiva empleando **la raíz cuadrada,** mientras que por lo general, el **logaritmo funciona mejor para la simetría positiva.** Se siguiere **aplicar todas las trasformaciones posibles** y seleccionar después, la variable transformada más apropiada.

La **heterocedasticidad** es un problema **asociado a la normalidad,** y en muchos casos la solución del problema tiene que ver también con los **problemas de normalidad.** La **heterocedasticidad** se debe también a la **distribución de la variable(s).** Cuando se examinan los **residuos del análisis de la regresión** buscando **la heterocedasticidad,** se observa que un **indicio de varianzas desiguales es una distribución con perfil de cono de los residuos (vea regresión lineal múltiple).** Si el **cono:**

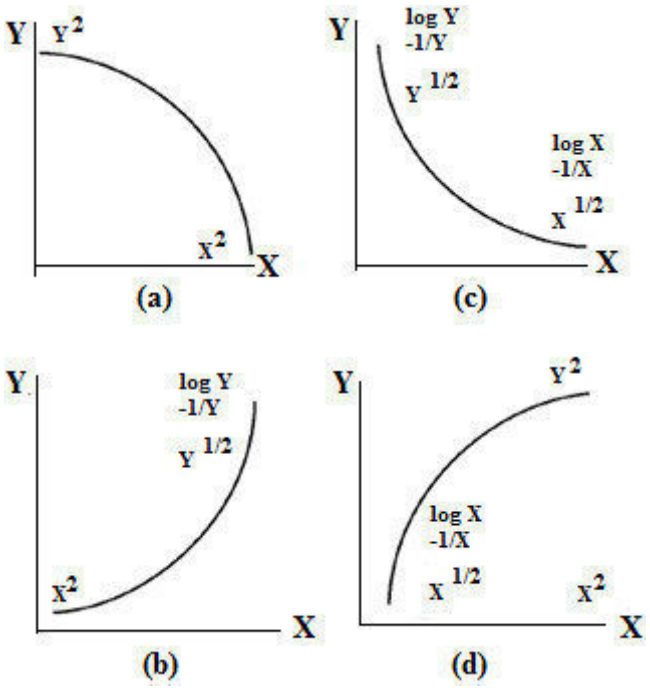
1. **Se abre a la derecha,** escogemos **la inversa;**
2. **Si se abre a la izquierda,** escogemos **la raíz cuadrada.**

Algunas transformaciones pueden asociarse con ciertos tipos de datos. Por ejemplo, **el recuento de frecuencias** sugiere una transformación de **raíz cuadrada;** las proporciones se transforman mejor por la transformación del **arcoseno (XNueva=2 arcoseno$\sqrt{}$Xantigua)** y un cambio proporcional se maneja mejor tomando el logaritmo de la variable. En todos los casos, una vez que se han realizado las transformaciones, los datos transformados deben ser contrastados para ver si se ha logrado la solución deseada.

3.9.19. Transformaciones para conseguir la linealidad

Están disponibles una gran cantidad de procedimientos para **conseguir la linealidad entre dos variables,** pero las **relaciones no lineales** más simples pueden clasificarse en **4** categorías**. Ver Figura 3.30**.

Figura 3.30. Selección de transformaciones para obtener linealidad



Fuente: Mosteller y Tukey (1977).

En cada cuadrante, se muestran las **transformaciones potenciales para variable dependiente e independiente**. Por ejemplo, si las relaciones locales son como las de la **figura 3.30** cuadrante **(a)**, se aplica la **raíz cuadrada** para conseguir la linealidad. Cuando se muestran las posibilidades de **transformación múltiple,** se empieza con el método más adecuado para cada cuadrante para **después bajar** hasta que se consigue la linealidad. Una **aproximación alternativa** consiste en utilizar variables adicionales, denominadas **polinómicas**, que representan los **componentes no lineales**. Este método se aprecia mejor en las **regresiones múltiples lineales.**

3.9.20. Normas generales para las transformaciones

l. Para obtener un efecto perceptible de la transformación, el **ratio entre la medida de la variable y su desviación estándar debe ser < 4,0.**

2. Cuando la transformación puede realizarse sobre una de las dos variables, seleccione la variable con el **ratio más pequeño del ítem l.**

3. Las transformaciones deberán aplicarse a las **variables independientes excepto** en el caso de la **heterocedasticidad.**

4. La **heterocedasticidad** sólo puede solucionarse mediante la **transformación de la variable dependiente en una relación de dependencia.** Si una relación **heterocedástica es además no lineal**, deberían **transformarse la variable dependiente y quizá la independiente.**

5. Las transformaciones pueden cambiar la interpretación de las variables. Por ejemplo, las variables transformadas tomando sus **logaritmos** trasladan la relación en una

medida de cambio proporcional (**elasticidad**). Siempre hay que asegurarse la exploración de todas las posibles interpretaciones de las variables transformadas

Con el fin de ilustrar las **técnicas de contrastación de datos** para conseguir el cumplimiento de los supuestos subyacentes al análisis multivariante y proporcionar un fundamento en el uso de los datos en los es análisis, es posible realizar primero las pruebas a la base de datos de los supuestos de: **normalidad, homocedasticidad y linealidad.** El cuarto supuesto básico, la **ausencia de correlación entre los errores,** sólo puede apreciarse en el contexto de un modelo multivariante específico y por tanto, será cubierto en los últimos capítulos para cada técnica multivariante. Se pondrá **mayor énfasis en las variables métricas**, aunque las variables no métricas serán evaluadas cuando sea apropiado.

3.9.21. Pruebas *Kolmogorov-Smirnov*

La prueba de ***Kolmogorov-Smirnov***, bautizada así en honor de los estadísticos **A. N. *Kolmogorov y N. V. Smirnov*** que la desarrollaron, se trata de un **método no paramétrico sencillo** para probar si existe una diferencia significativa entre una **distribución de frecuencias observada** y una distribución de frecuencias **teórica**. La **prueba de K-S** es, por consiguiente, otra medida de la **bondad de ajuste** de una distribucio n de frecuencia teo rica, como lo es la prueba **Chi-cuadrada.** Sin embargo, la **prueba de K-S** tiene varias ventajas sobre la prueba **Chi-cuadrada**: **es una prueba más poderosa**, y es ma s fa cil de usar, puesto que **no requiere que los datos se agrupen de alguna manera**. El estadístico de **K-S**, D_n, es particularmente útil para juzgar qué tan cerca está la distribución de **frecuencias observada** de la distribución de **frecuencias esperada**, porque la distribución de probabilidad de D_n **depende** del **tamaño de muestra n**, pero es **independiente de la distribución de frecuencias esperada** (D_n **es llamado también un estadístico de "*distribución libre*"**). La **Prueba de *Kolmogorov–Smirnov* de una muestra,** conocida comúnmente como la **prueba K-S,** toma la distribución observada acumulada de los datos y los compara con la distribución acumulada teórica para una población distribuida **normalmente.** El reporte que arroja **SPSS** típico de una variable, es el mostrado en la **Figura 3.31**:

Figura 3.31. Tabla de *Kolmogorov-Smirnov* de una muestra

Pruebas no paramétricas

[imputacion] C:\Users\Juan\Desktop\proy libro mc\CRM_MK

Prueba de Kolmogorov-Smirnov para una muestra

Número de imputación			X1_Information_from_customer
3	N		200
	Parámetros normales[a,b]	Media	7.675
		Desviación típica	1.6033
	Diferencias más extremas	Absoluta	.083
		Positiva	.083
		Negativa	-.078
	Z de Kolmogorov-Smirnov		1.169
	Sig. asintót. (bilateral)		.130

a. La distribución de contraste es la Normal.

b. Se han calculado a partir de los datos.

Fuente: SPSS 20 IBM

Recuerde que:

1. La primera parte de la prueba de una muestra de ***Kolmogorov-Smirnov*** muestra una tabla con el número de elementos **N,** la **Media** y la **Desviación típica** como **parámetros normales.**
2. Dado que el proceso de cálculo se marca como **Distribución de contraste la Normal,** está en la distribución con la que se comparan nuestros datos. Esto es confirmado por en supra índice |**a.** El supra índice |**b** indica que la **distribución observada** corresponde a la **distribución teórica** que es **normalmente distribuida.**
3. Para comprobarlo, cheque que el valor de **Sig. Asintót. (bilateral) debe ser >0.05 que indica que la distribución teórica.** Esto es, que **sus datos NO son significativamente diferentes a la distribución normal en un nivel de significancia de *p*<0.05.**
4. De la tabla anterior se observa que el valor de **Sig. Asintót. (bilateral) >0.130 >0.05,** por lo que **X_1 puede ser asumido que está normalmente distribuido (*p*=0.130).**
5. La parte que dice **Diferencias más extremas,** indica que la diferencia entre la distribución acumulada observada y la distribución acumulada teórica mientras más grande sea, **existe más probabilidades que las distribuciones sean diferentes de una distribución normal.**

1.9. Ejemplo cálculo supuestos del análisis multivariante

Se mostrarán los principales a saber:

3.10. 1.Normalidad de datos de variable métrica respecto a un grupo de casos para prueba de Hipótesis

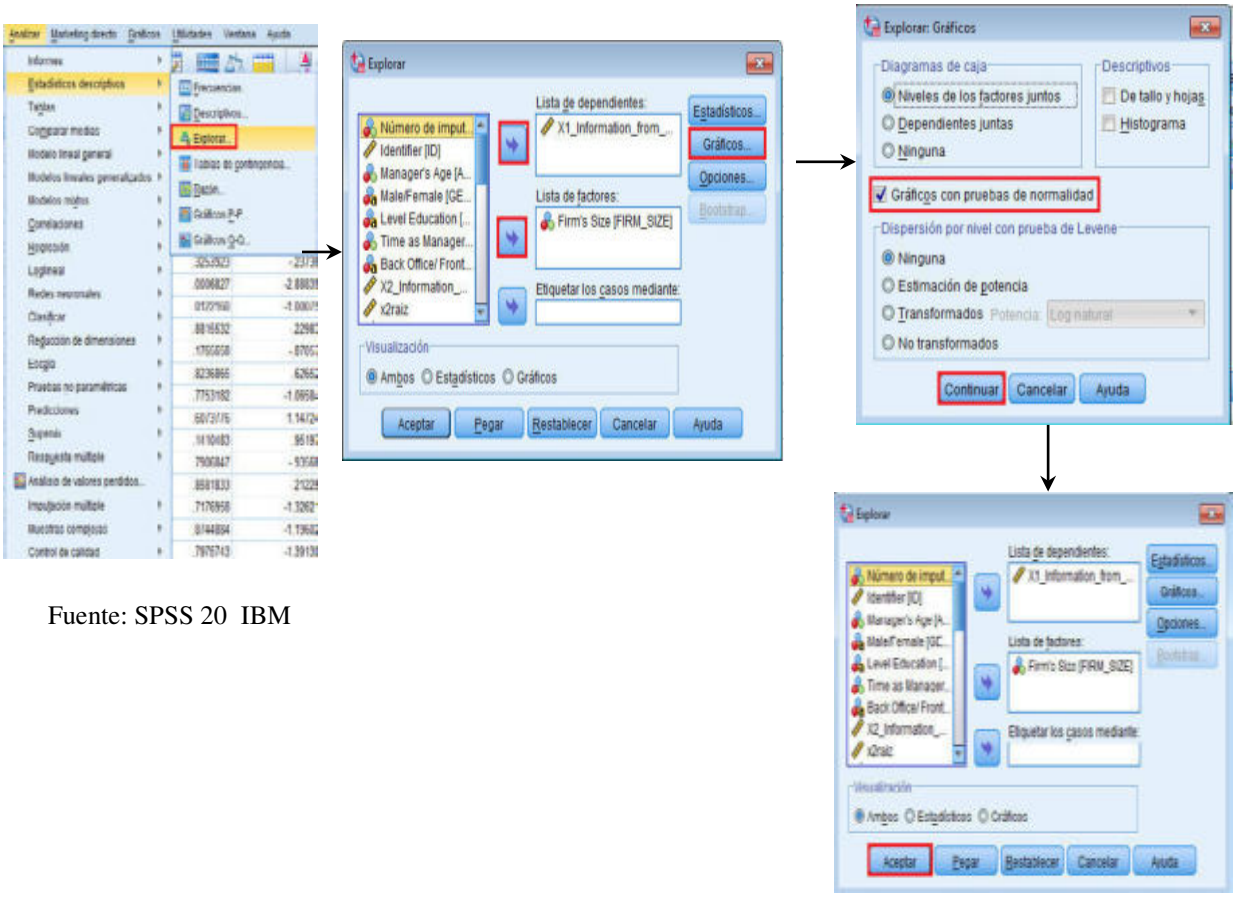
-Problema 22: De la base de datos CKM_MKT_Digital_imputaciones3.sav donde N>50 muestras, pruebe qué Hipótesis es aprobada:

H_0.-La variable **X_1** tiene una población con distribución normal, respecto del tamaño de la Firma (**Firm's size**)

H_1.-La variable **X_1** tiene una población **distinta** de la distribución normal, respecto del tamaño de la Firma (**Firm's size**)

Dado que **N>50 muestras**, la normalidad se analizará de acuerdo a ***Kolmogorov-Smirnov*. -Teclear: Analizar->Estadísticos descriptivos->Explorar->Seleccionar lista de dependientes (X_1.Variable métrica); lista de factores (Firm´s size. Variable nominal) ->Gráficos>Gráficos con pruebas de normalidad->Continuar->Aceptar.** Ver **Figura 3.32.**

Figura 3.32.- Cálculo de la normalidad variable X_1 vs. Firm's size



Fuente: SPSS 20 IBM

-Resultados: Son generados tablas y gráficos diversos. De los más importantes, analice:
1.-La Tabla.-Resumen del procesamiento de los casos, la cual mínimamente debe reportar: suma de los casos válidos =N; casos perdidos= 0. Ver **Figura 3.33**

Figura 3.33. Tabla.-Resumen del procesamiento de los casos

Firm's Size

Resumen del procesamiento de los casos

Número de imputación		Firm's Size	Casos					
			Válidos		Perdidos		Total	
			N	Porcentaje	N	Porcentaje	N	Porcentaje
3	X1_Information_from_customer	Industry and services is Micro (1-10)	15	100.0%	0	0.0%	15	100.0%
		Industry and services is Little (11-50)	81	100.0%	0	0.0%	81	100.0%
		Industry is Media (51-250)	21	100.0%	0	0.0%	21	100.0%
		Services is Media (51-100)	42	100.0%	0	0.0%	42	100.0%
		Industry is Big (251 a+)	26	100.0%	0	0.0%	26	100.0%
		Services is Big (101 a+)	15	100.0%	0	0.0%	15	100.0%

Fuente: SPSS 20 IBM

2.-La Tabla.-**Pruebas de normalidad,** de la que se resalta la de ***Kolmogorov-Smirnov*** (**N>50**), los grados de libertad de c/u de los casos de la variable **Firm's size** y la p<0.05 de la variable $\mathbf{X_1}$ en 1 caso. Ver **Figura 3.34**

Figura 3.34. Pruebas de normalidad

Pruebas de normalidad

Número de imputación		Firm's Size	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
			Estadístico	gl	Sig.	Estadístico	gl	Sig.
3	X1_Information_from_customer	Industry and services is Micro (1-10)	.216	15	.058	.856	15	.021
		Industry and services is Little (11-50)	.118	81	.007	.941	81	.001
		Industry is Media (51-250)	.169	21	.118	.848	21	.004
		Services is Media (51-100)	.099	42	.200*	.930	42	.013
		Industry is Big (251 a+)	.135	26	.200*	.904	26	.019
		Services is Big (101 a+)	.150	15	.200*	.889	15	.065

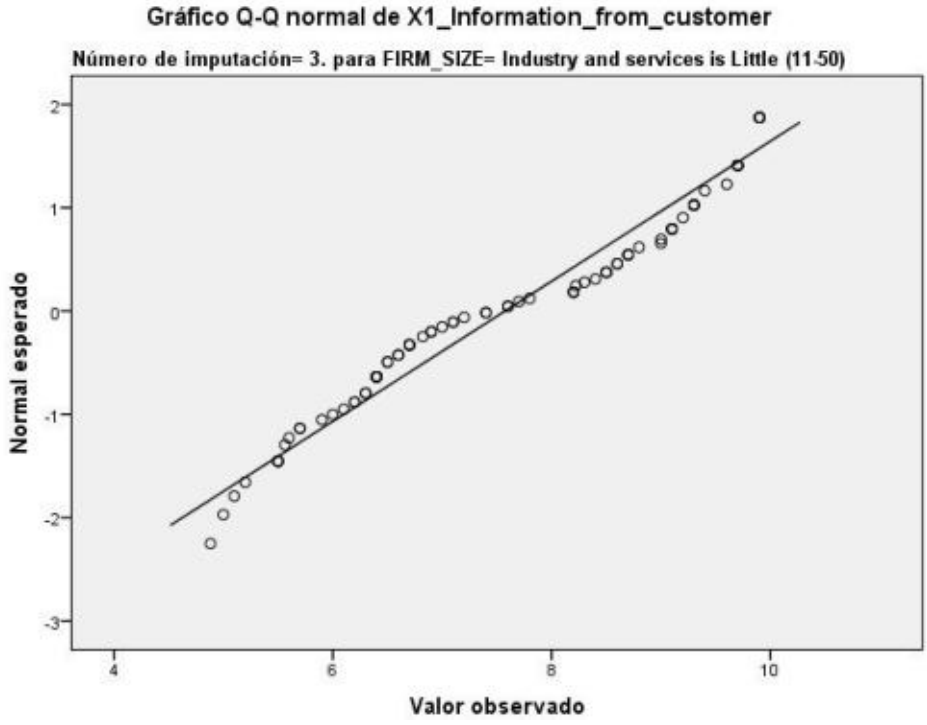
*. Este es un límite inferior de la significación verdadera.

a. Corrección de la significación de Lilliefors

Fuente: SPSS 20 IBM

-Resultados: Son generados gráficos de la recta de dispersión normal por cada caso de la variable **Firm's size** vs. la **variable $\mathbf{X_1}$. Ver Figura 3.35**

Figura 3.35. Gráfico de dispersión normal



A partir de la **Figura 3.34**, podemos determinar que la **Ho** (**Hipótesis nula o teórica**), **NO se acepta (se encuentra en zona de rechazo)** para el grupo de casos donde la variable **Firm's size** (Industry and service is little (11-50)) vs. $\mathbf{X_1}$ (Information from the Customer) ya que ***p*=0.007<0.05**; para el resto de los grupos de casos **SI se acepta ya que p>0.05 (se encuentra en zona de aceptación) *p*<0.05.** La zona de **rechazo de Ho**, quiere decir que hay diferencias entre los puntos a comparar con 95% confianza ***p*>0.05.** La zona de **aceptación de Ho**, quiere decir que hay igualdades entre los puntos a comparar con **95%** confianza.

3.10.2. Normalidad de datos de una variable métrica

-Problema 23: De la base de datos CKM_MKT_Digital_imputaciones3.sav donde N>50 muestras, pruebe qué Hipótesis es aprobada:

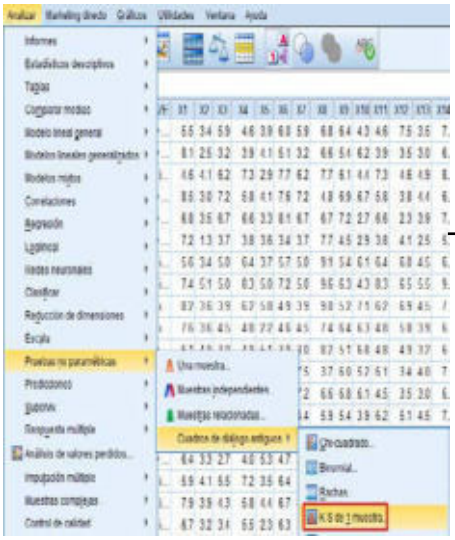
$_0$.-La variable $\mathbf{X_1}$ tiene una población con distribución normal

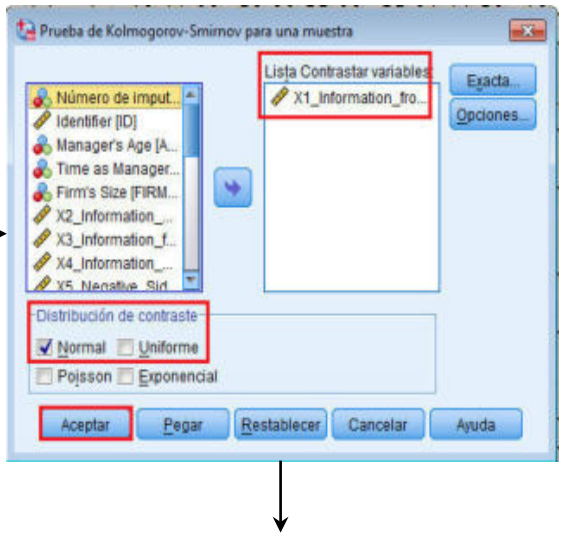
$_1$.-La variable $\mathbf{X_1}$ tiene una población **distinta** de la distribución normal

Dado que **N>50 muestras**, la normalidad se analizará de acuerdo a ***Kolmogorov- Smirnov.***

-Teclear: Analizar->Pruebas no paramétricas->Cuadro de diálogo antiguos->K-S de 1 muestra; Seleccionar lista de variables de prueba (X_1); Seleccionar Distribución de prueba (Normal) ->Aceptar->Continuar->Aceptar. Ver **Figura 3.36.**

Figura 3.36.- Cálculo de la normalidad de datos de la variable X_1





Pruebas no paramétricas

[imputacion] C:\Users\Juan\Desktop\proy libro mc\CKM_I

Prueba de Kolmogorov-Smirnov para una muestra

Número de imputación			X1_Information_from_customer
3	N		200
	Parámetros normales[a,b]	Media	7.675
		Desviación típica	1.6033
	Diferencias más extremas	Absoluta	.083
		Positiva	.083
		Negativa	-.078
	Z de Kolmogorov-Smirnov		1.169
	Sig. asintót. (bilateral)		.130

a. La distribución de contraste es la Normal.

b. Se han calculado a partir de los datos.

Fuente: SPSS 20 IBM

-Resultado: Dado que ***p*=0.130>0.05** SI se acepta la **Ho.-**La variable X_1 en la población tiene distribución normal

-Problema 24: De la base de datos **CKM_MKT_Digital_imputaciones3.sav donde N>50 muestras, pruebe qué Hipótesis es aprobada:**

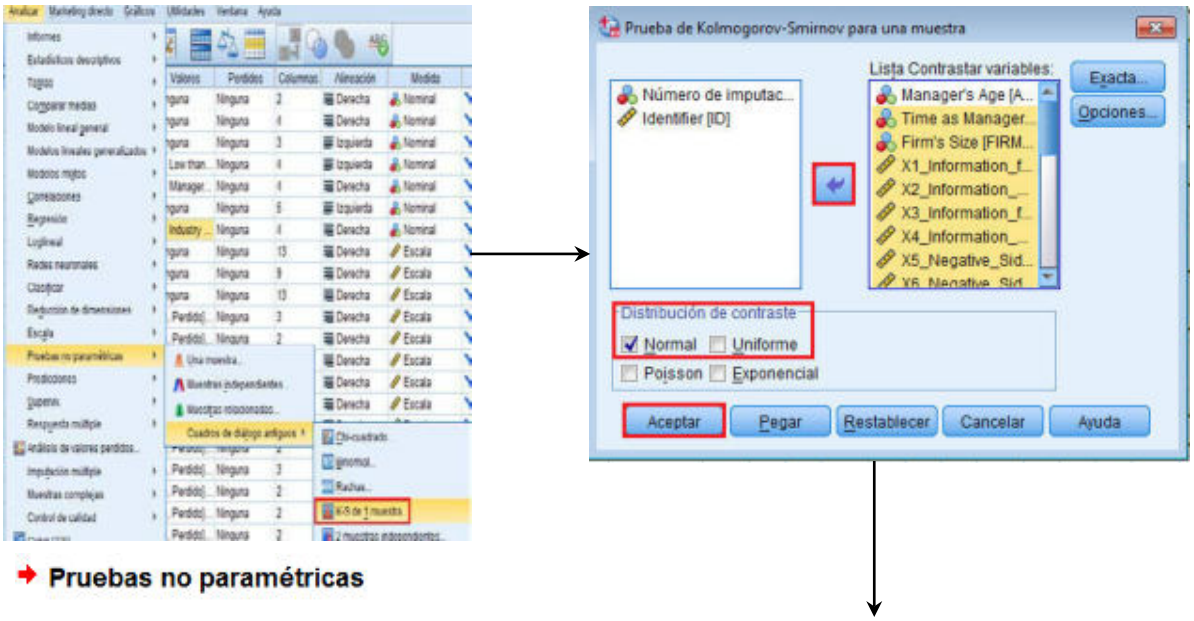
H_0.-Las variables de la base de datos tienen una población con distribución normal

H_1.-Las variables de la base de datos tienen una población **distinta** de la distribución normal

Dado que **N>50** muestras, la normalidad se analizará de acuerdo a ***Kolmogorov-Smirnov.***

-Teclear: Analizar->Pruebas no paramétricas->Cuadro de diálogo antiguos->K-S de 1 muestra; Seleccionar lista de variables de prueba: todas las de la base de datos Seleccionar Distribución de contraste (Normal) ->Aceptar-> Ver **Figura 3.37.**

Figura 3.37.- Cálculo de la normalidad de datos de la base de datos




```
[imputacion] C:\Users\Juan\Desktop\proy libro mc\CKM_MKT_Digital_imputaciones3.sav
```

Número de imputación			Manager's Age	Time as Manager	Firm's Size	X1_Information_from_customer	X2_Information_about_the_Customer
3	N		200	200	200	200	200
	Parámetros normales[a,b]	Media	42.09	2.05	2.14	7.675	3.723
		Desviación típica	13.697	.755	1.449	1.6033	.8229
	Diferencias más extremas	Absoluta	.145	.216	.264	.083	.113
		Positiva	.123	.216	.264	.083	.113
		Negativa	-.145	-.214	-.141	-.078	-.077
	Z de Kolmogorov-Smirnov		2.055	3.060	3.737	1.169	1.596
	Sig. asintót. (bilateral)		.000	.000	.000	.130	.012

a. La distribución de contraste es la Normal.

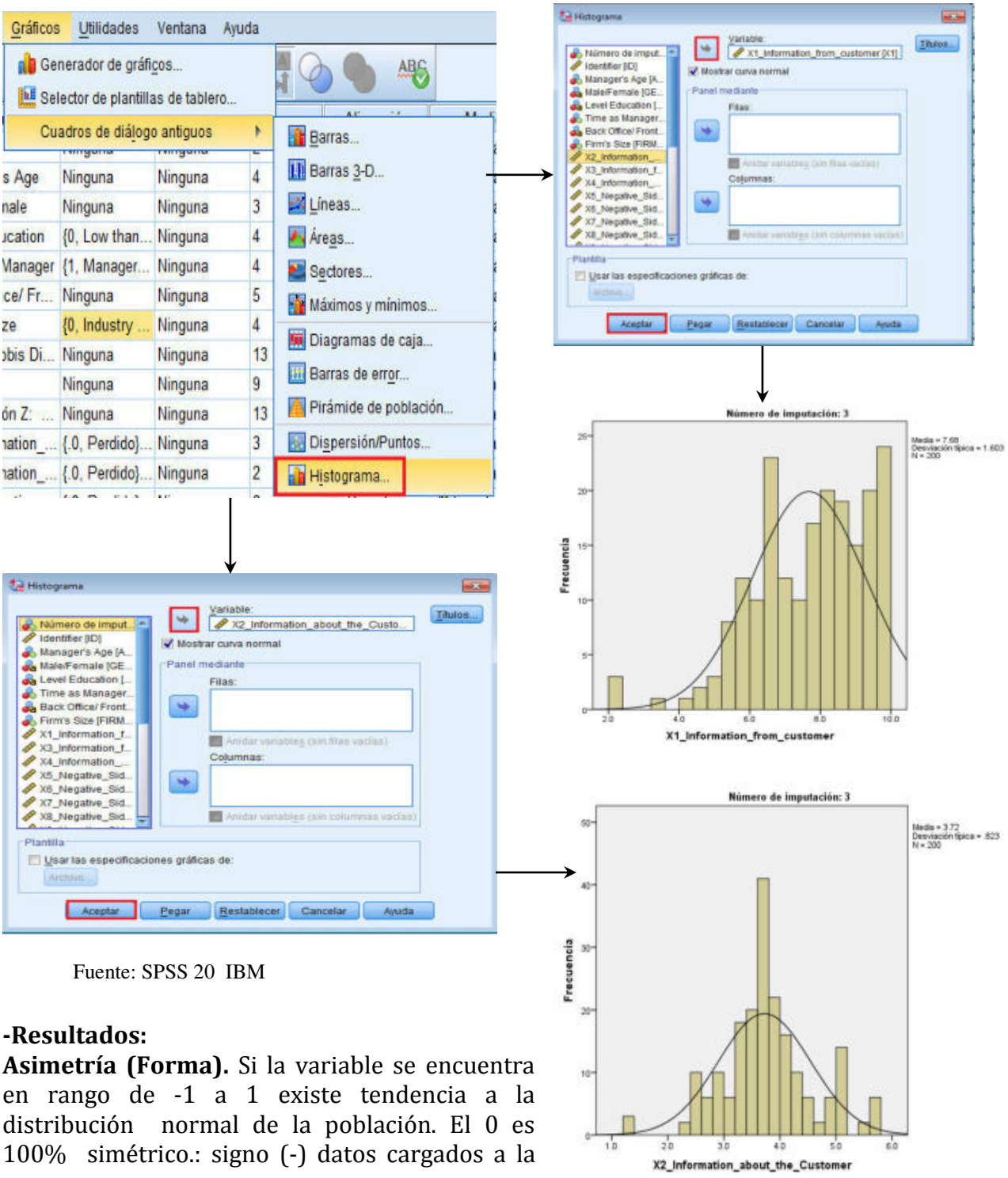
b. Se han calculado a partir de los datos.

Fuente: SPSS 20 IBM

-Resultados: Se desaprueba la **Ho** y aprueba **H_1,** en donde las variables Manager's Age, Time as Manager, **Firm's size** y **X_2 (0.012)** tienen ***p*<0.05, es decir, éstas variables NO tiene distribución normal en su población.**

-Problema 25: Verificar gráficamente para **X_1 y X_2; valorar datos descriptivos. Ver Figura 3.38**
-Teclear: Gráficos->Cuadro de diálogo antiguos ->Histograma. Ver Figura 3.38.

Figura 3.38. Valoración de datos descriptivos.



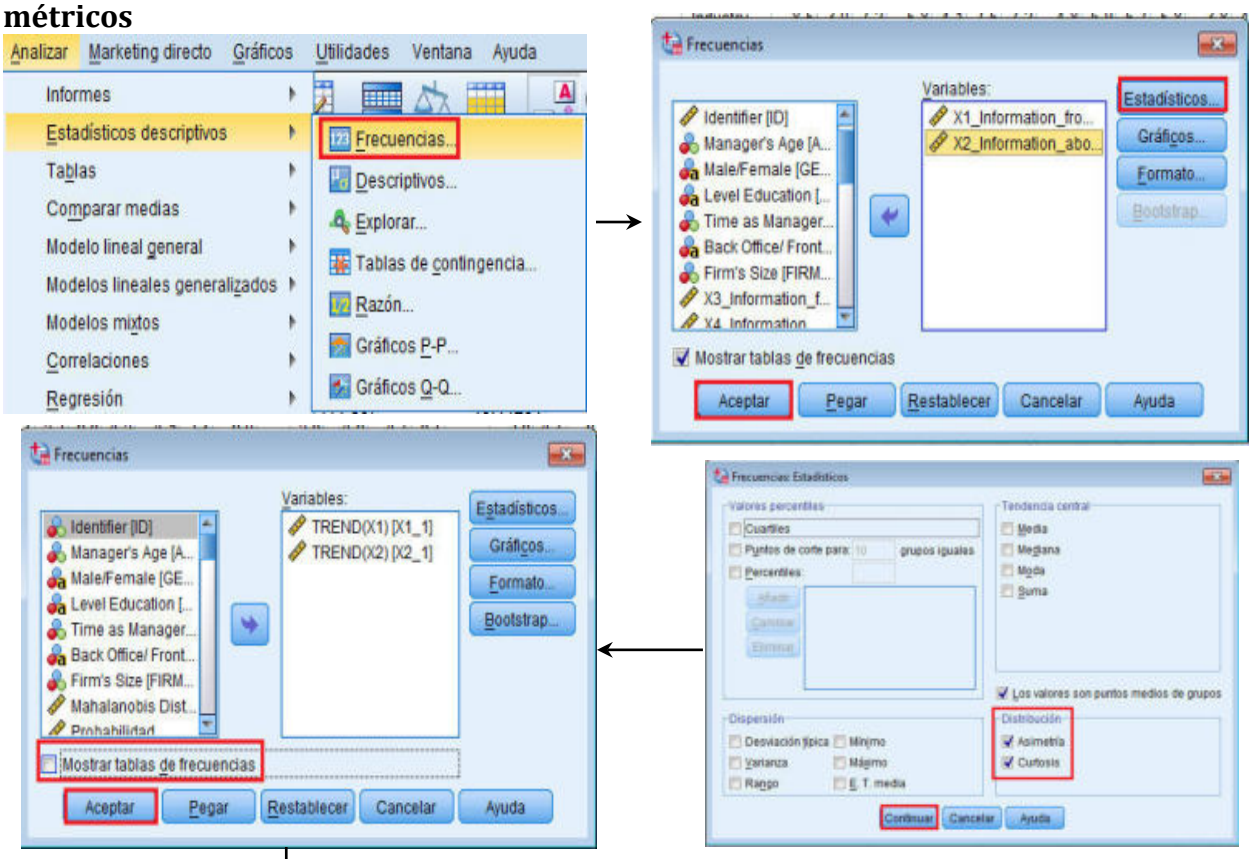
Fuente: SPSS 20 IBM

-Resultados:
Asimetría (Forma). Si la variable se encuentra en rango de -1 a 1 existe tendencia a la distribución normal de la población. El 0 es 100% simétrico.: signo (-) datos cargados a la

derecha; signo (+) datos cargados a la izquierda **Curtosis (Altura).-**Si la variable se encuentra en rango de -1 a 1 existe tendencia a que la altura corresponde a la distribución normal de la población. El 0 es 100% con altura de distribución normal (Mesocúrtica); signo (-); datos más dispersos y la altura es más baja a una distribución normal (Platicúrtica); signo (+) datos más agrupados y la altura de distribución es más alta que en una distribución normal (Leptocúrtica)

-Problema 26: Determinar asimetría y curtosis de la población de las variables **X_1** y **X_2**
-Teclear: Analizar->Estadísticos descriptivos->Frecuencias->Selección de variables (X_1 y X_2) ->Estadísticos->Selección de distribución: Asimetría y Curtosis->Continuar-> Desmarcar Mostrar tabla de frecuencias->Aceptar. Ver Figura 3.39.

Figura 3.39.- Proceso para definir Asimetría y Curtosis de una población de datos métricos



Frecuencias

```
[imputacion] C:\Users\Juan\Desktop\proy libro mc\CKM_
```

Estadísticos

Número de imputación			X1_Information_from_customer	X2_Information_about_the_Customer
3	N	Válidos	200	200
		Perdidos	0	0
	Asimetría		-,750	,148
	Error típ. de asimetría		,172	,172
	Curtosis		,622	-,570
	Error típ. de curtosis		,342	,342
Combinado	N	Válidos	200	200
		Perdidos	0	0

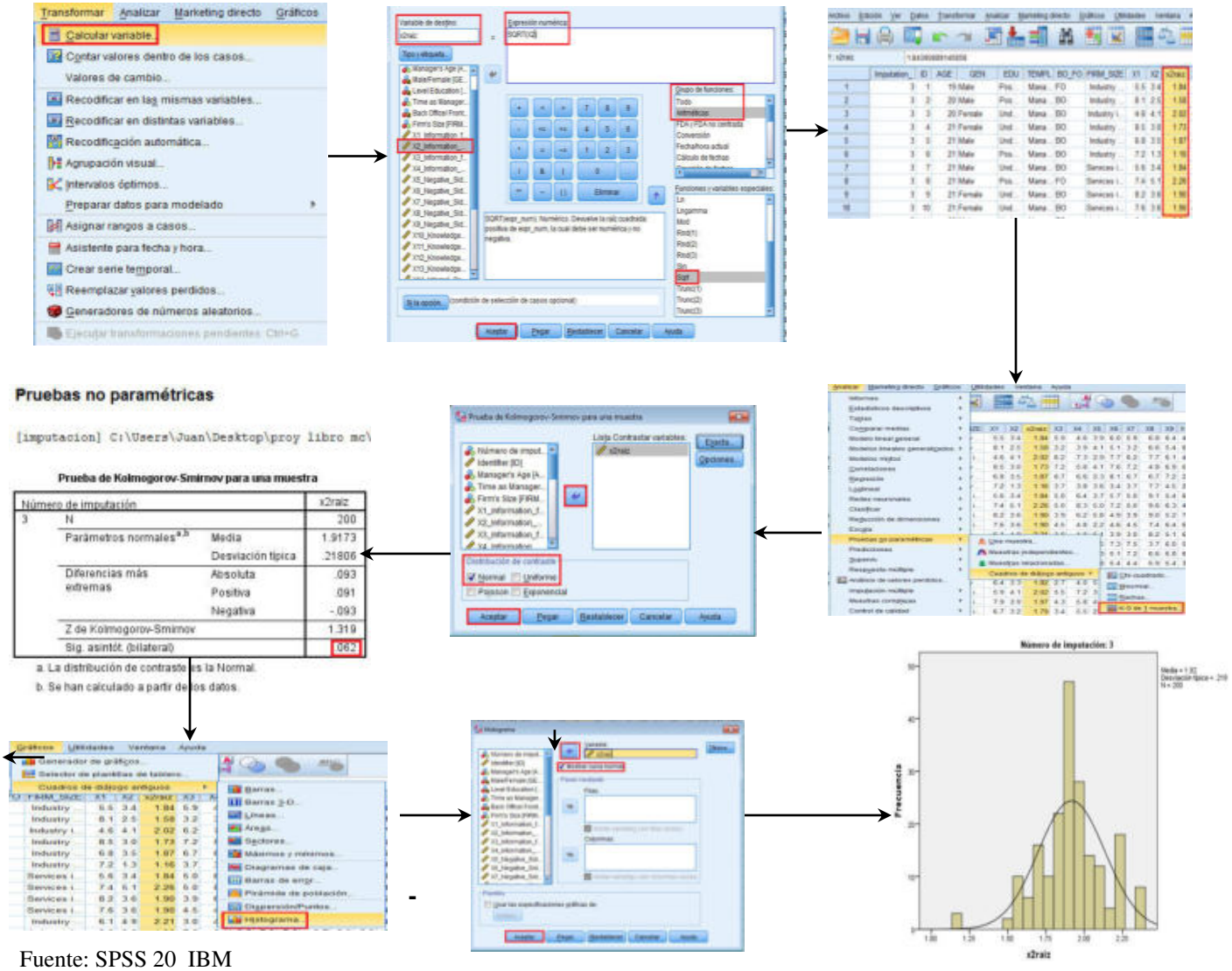
Fuente: SPSS 20 IBM

-Resultados:
Asimetría (Forma). Si la variable se encuentra en rango de -1 a 1 existe tendencia a la distribución normal de la población. El 0 es 100% simétrico.: signo (-) datos cargados a la derecha
; Signo (+) datos cargados a la izquierda
Curtosis (Altura).-Si la variable se encuentra en rango de -1 a 1 existe tendencia a que la altura corresponde a la distribución normal de la población. El 0 es 100% con altura de distribución normal (mesocúrtica); signo (-); datos más dispersos y la altura es más baja a una distribución normal (Platicúrtica); signo (+) datos más agrupados y la altura de distribución es más alta que en una distribución normal (Leptocúrtica)

-Problema 27: De la variable **X_2** (***p*=0.012<0.05**), con población que **NO tiene distribución normal**, normalice dicha variables, Mediante aplicación de **raíz cuadrada** de su población. Muestre curva normal.
-Teclear: Transformar->Calcular variable->Asignar nombre a Variable de destino (X_2raiz);Seleccionar en grupo de funciones (Todo);Seleccionar en Funciones y variables especiales (Sqrt); flecha hacia arriba; Expresión numérica seleccionar variable métrica (X_2) a normalizar->Aceptar y repetir prueba de normalidad de *Kolmogorov-Smirnov* para la variable X2raiz :
-Teclear: Analizar->Pruebas no paramétricas->Cuadro de diálogo antiguos->K-S de 1 muestra; Seleccionar lista de variables de prueba (X_2raiz); Seleccionar Distribución de prueba (Normal) ->Aceptar. Para gráficos:
-Teclear: Gráficos->Cuadro de diálogo antiguos ->Histograma.

Ver Figura 3.40

Figura 3.40.- Proceso para Normalizar una población de datos métricos, mediante la técnica raíz cuadrada



Fuente: SPSS 20 IBM

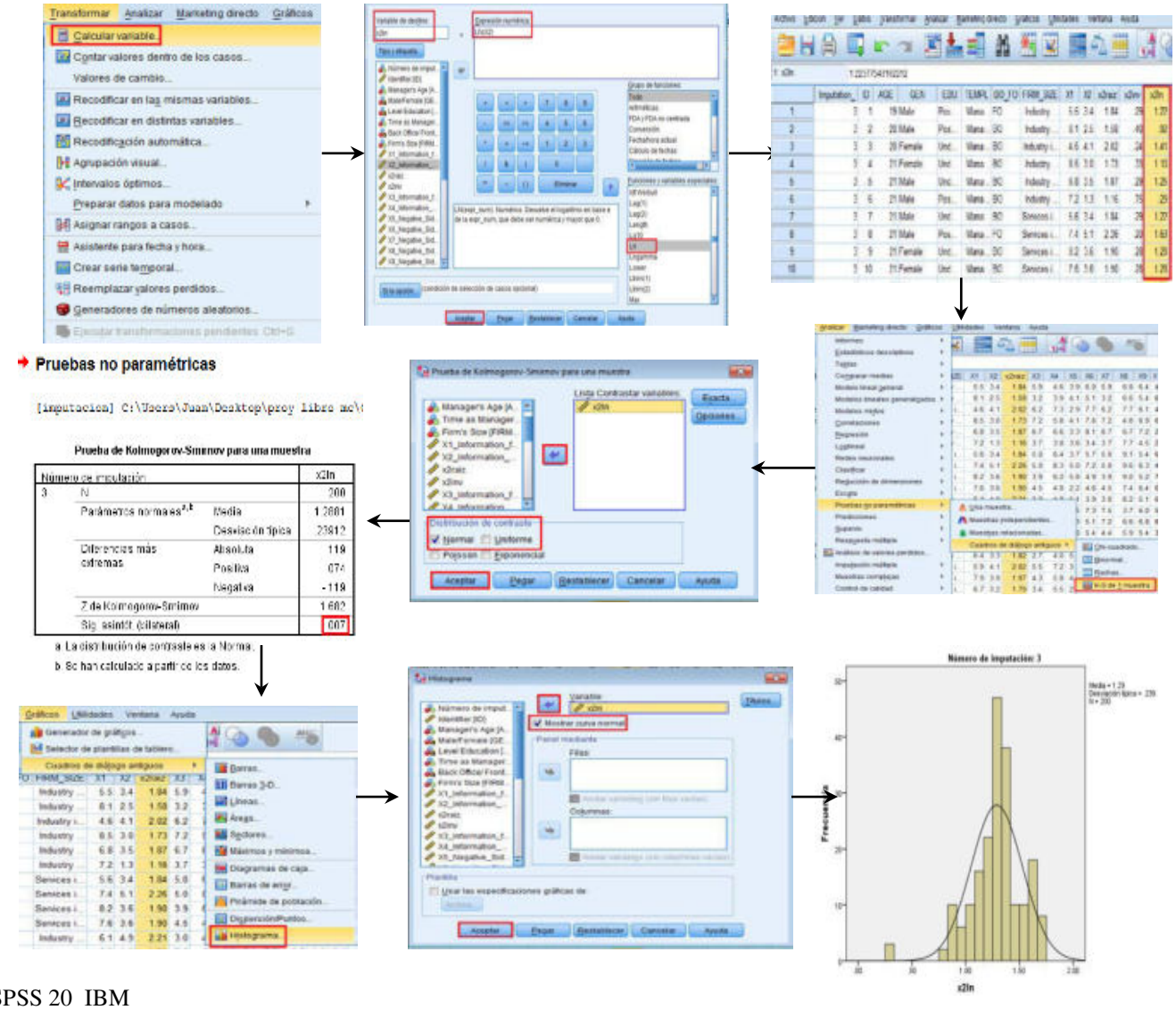
-Resultados: variable X_2 es normalizada ya que tiene p=0.062>0.05

-Problema 28: De la variable X_2 (***p*=0.012<0.05**), con población que **NO tiene distribución normal**, normalice dicha variables, Mediante aplicación de **logaritmo neperiano** de su población. Muestre curva normal.

-Teclear: Transformar->Calcular variable->Asignar nombre a Variable de destino (X_2 ln);Seleccionar en grupo de funciones (Todo);Seleccionar en Funciones y variables especiales (ln); flecha hacia arriba; Expresión numérica seleccionar variable métrica (X2) a normalizar->Aceptar y repetir prueba de normalidad de *Kolmogorov-Smirnov* para la variable X_2ln :

-Teclear: Analizar->Pruebas no paramétricas->Cuadro de diálogo antiguos->K-S de 1 muestra; Seleccionar lista de variables de prueba (X_2ln); Seleccionar Distribución de prueba (Normal) ->Aceptar. Para gráficos:
-Teclear: Gráficos->Cuadro de diálogo antiguos ->Histograma. Ver Figura 3.41.

Figura 3.41.- Proceso para Normalizar una población de datos métricos, mediante la técnica logaritmo neperiano



Fuente: SPSS 20 IBM

-Resultados:
Variable **X_2 NO** es posible normalizar ya que tiene **p=0.007<0.05**

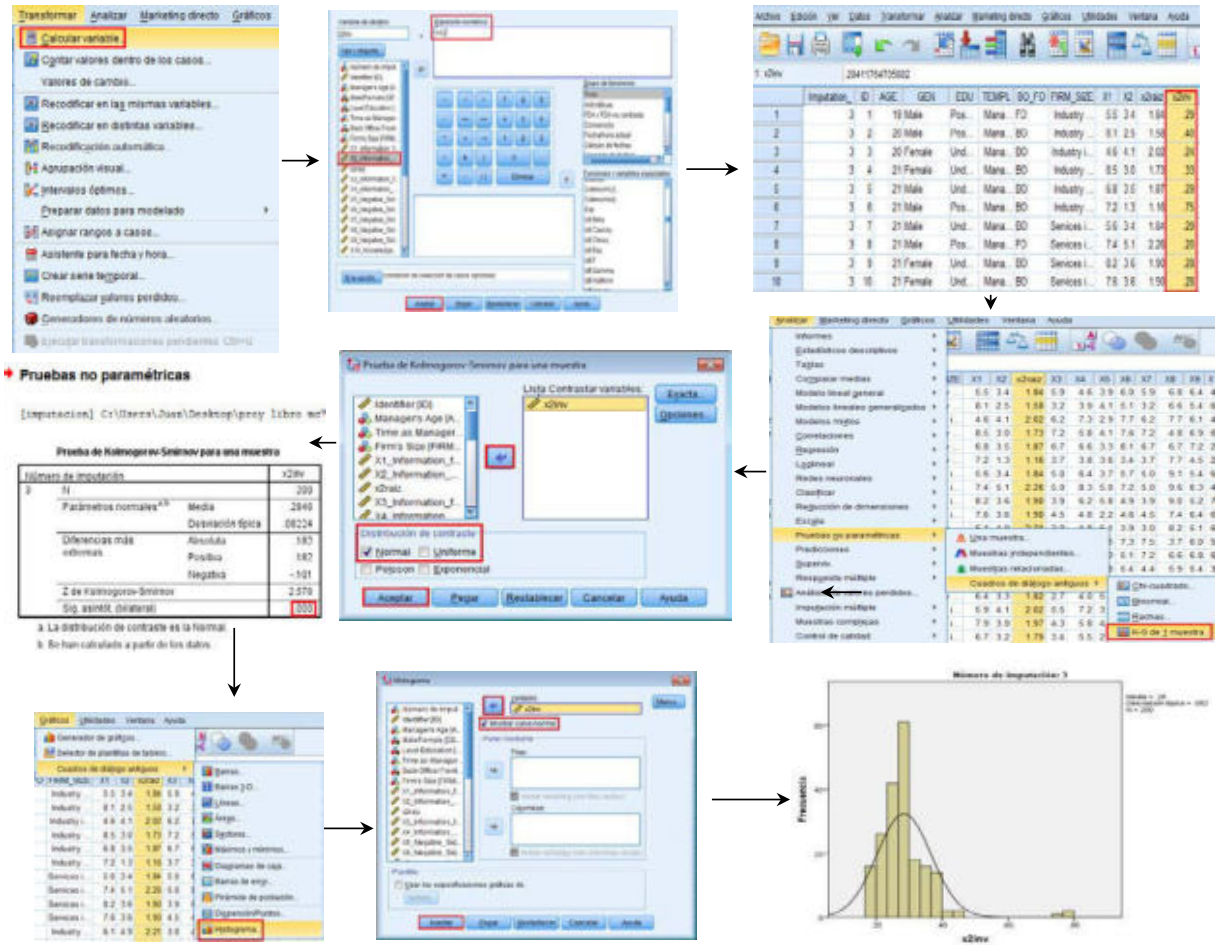
-Problema 28: De la variable $\mathbf{X_2}$ (***p*=0.012<0.05**), con población que **NO tiene distribución normal**, normalice dicha variables, Mediante aplicación de **inversa** de su población. Muestre curva normal.

-Teclear: Transformar->Calcular variable->Asignar nombre a Variable de destino (X_2inv);introducir $1/X_2$; ->Aceptar y repetir prueba de normalidad de *Kolmogorov-Smirnov* para la variable X_2inv:

-Teclear: Analizar->Pruebas no paramétricas->Cuadro de diálogo antiguos->K-S de 1 muestra; Seleccionar lista de variables de prueba (X_2inv); Seleccionar Distribución de prueba (Normal) ->Aceptar. Para gráficos:

-Teclear: Gráficos->Cuadro de diálogo antiguos ->Histograma. Ver Figura 3.42.

Figura 3.42.- Proceso para Normalizar una población de datos métricos, mediante la técnica inversa



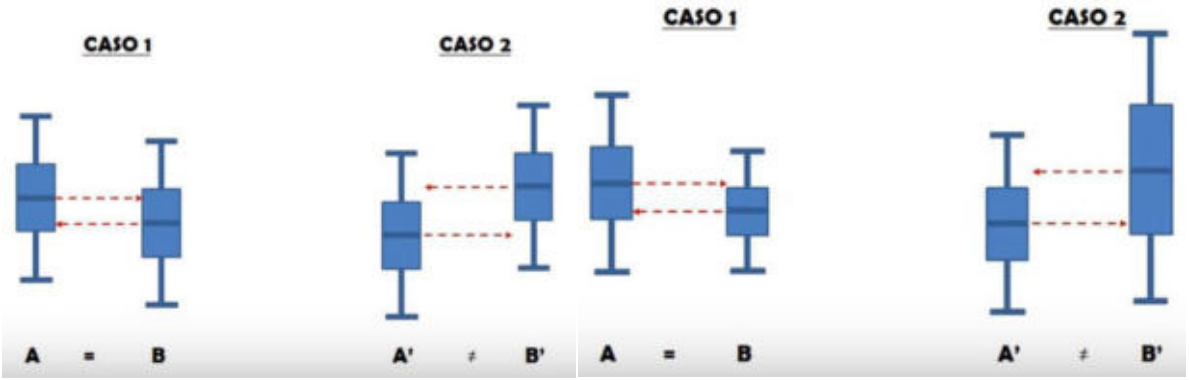
Fuente: SPSS 20 IBM

Resultados: Variable $\mathbf{X_2}$ **NO** es posible normalizar ya que tiene **p=0.000<0.05**

3.10.3. Homocedasticidad

Cuando se quiere comparar grupos, se debe asegurar que la varianza de un grupo sea igual o por lo menos, no sea distinta al otro para proyectar la línea media de un grupo sobre otro y que de estar **dentro** de las cajas, se llegue a la conclusión de que la varianza entre los grupos es igual **(Figura 3.43. CASO 1)** o que si la proyección está por **fuera** de las cajas, se llegue a la conclusión de que la varianza entre los grupos son diferentes **(Figura 3.43. CASO 2)**

Figura 3.43. Varianzas iguales



Cuando las cajas entre los grupos no es igual las proyecciones de las medias entre se muestran dispares en cuanto una de ellas si se proyecta en la otra pero no viceversa (**Figura 3.43. CASO 1 y CASO 2**) .En estos casos no se puede asegurar si hay diferencias o no dado que los tamaños y las varianzas son distintas. Por lo tanto, para comparar se requiere que los tamaños de las cajas sean iguales por lo tanto, las varianzas deben ser iguales, homogéneas, o sea, homocedasticidad.

Es evaluada sobre una base **univariante** (por ej., el **test de *Levene*** en SPSS) donde se **compara la varianza de una variable métrica a lo largo de los niveles de las variables no métricas**. Estos análisis son apropiados en **preparación tanto para el análisis de la varianza (ANOVA) como del análisis multivariante de la varianza (MANOVA)** donde las variables **no métricas** son las **variables independientes**, o el **análisis discriminante** donde las variables **no métricas son las medidas dependientes**.

Los test de homocedasticidad de **dos variables**, se pueden realizar mejor a través del **análisis gráfico, particularmente un análisis de los residuos** (si se comparan 2 variables métricas en regresión lineal)

-Problema 30: De la base de datos **CKM_MKT_Digital_imputaciones3.sav** compruebe las hipótesis:

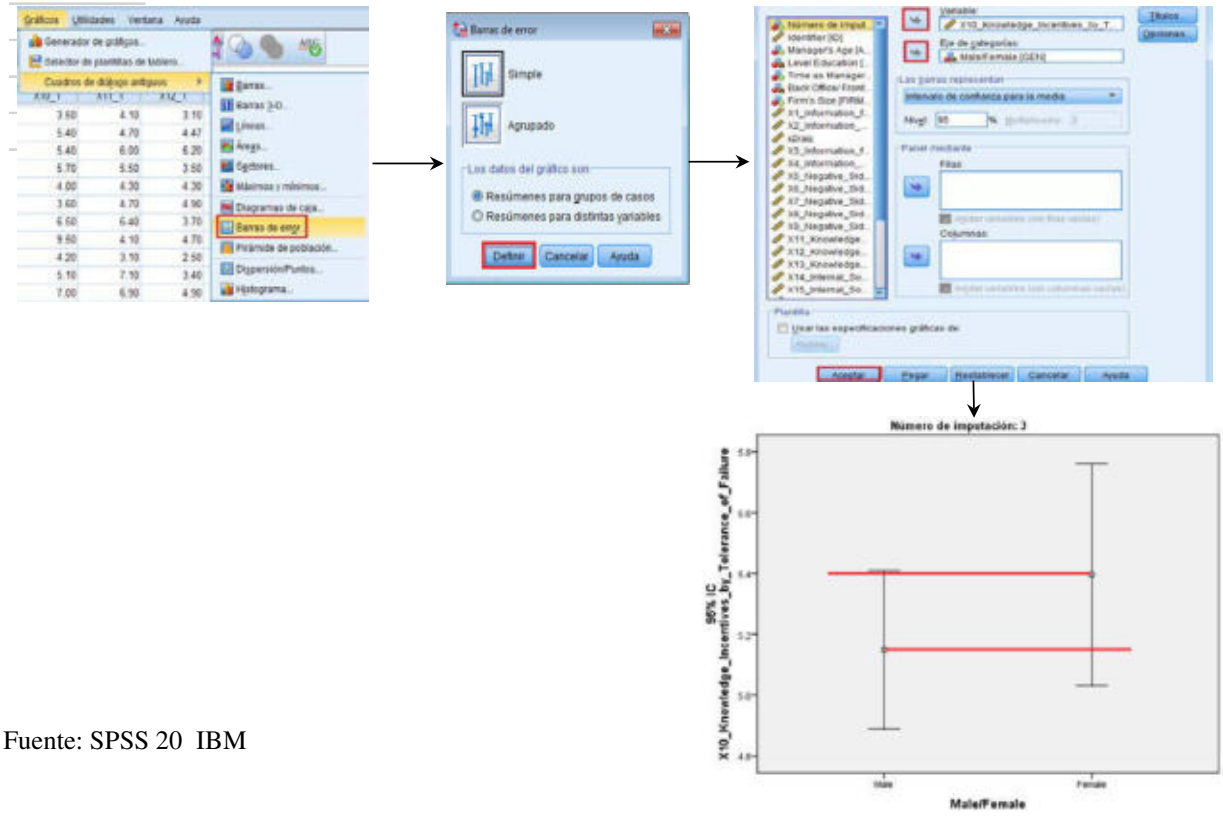
$_0$**: No** existen diferencias significativas entre las varianzas de los grupos de la variable $\mathbf{X_{10}}$ vs. **GEN**

$_1$**: Sí** existen **diferencias** significativas entre las varianzas de los grupos de la variable $\mathbf{X_{10}}$vs. **GEN**

Verificar gráficamente la homocedasticidad.

-Teclear: Gráficos->Cuadro de diálogos antiguos->Barras de error; Seleccione: Simple->Definir->Selección de Variable métrica: X_{10}; Selección de Variable Eje de categorías: GEN->Aceptar. Ver Figura 3.44.

Figura 3.44. Proceso para determinar gráficas de Homocedasticidad



Fuente: SPSS 20 IBM

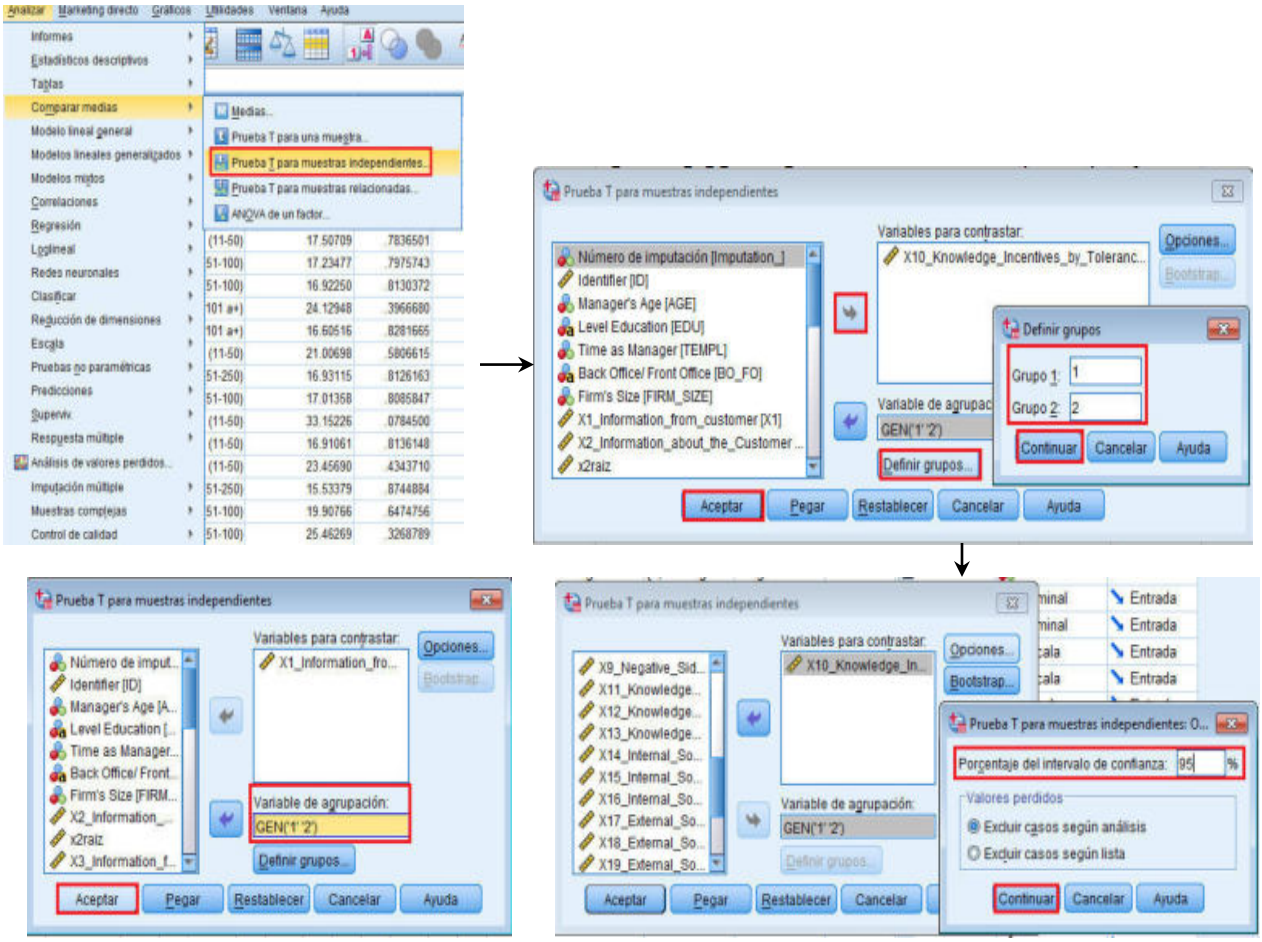
-Resultados: se observa que los grupos de variables **X_{10}** y **GEN**, se encuentran distribuidas sus medias entre cada una de ellas. Sin embargo **NO** es posible asegurar, de manera visual, que sí existe homocedasticidad.

3.10.4. Homocedasticidad (entre 2 Grupos)

-Problema 31: verificar por **test de *Levene***, la homocedasticidad entre las variables $_1$ y **GEN** de la base de datos **CKM_MKT_Digital_imputaciones3.sav.**

-Teclear: Analizar->Comparar medias->Prueba T para muestras independientes->Selección Variables para contrastar métricas: $_{10}$; Selección Variable de agrupación nominal: GEN->Definir grupos: (variable categórica GEN) ->Continuar->Aceptar. Ver Figura 3.45.

Figura 3.45. Prueba de *Levene* para comprobación de Homocedasticidad



Fuente: SPSS 20 IBM

SPSS genera la **tabla Estadísticos de grupo**. **Ver Figura 3.46.**

Figura 3.46. Prueba de *Levene* para comprobación de Homocedasticidad

Estadísticos de grupo

	Male/Female	N	Media	Desviación típ.	Error típ. de la media
X10_Knowledge_Incentives_by_Tolerance_of_Failure	Male	132	5.149	1.5136	.1317
	Female	68	5.396	1.5073	.1828

Variable dependiente

Variable independiente

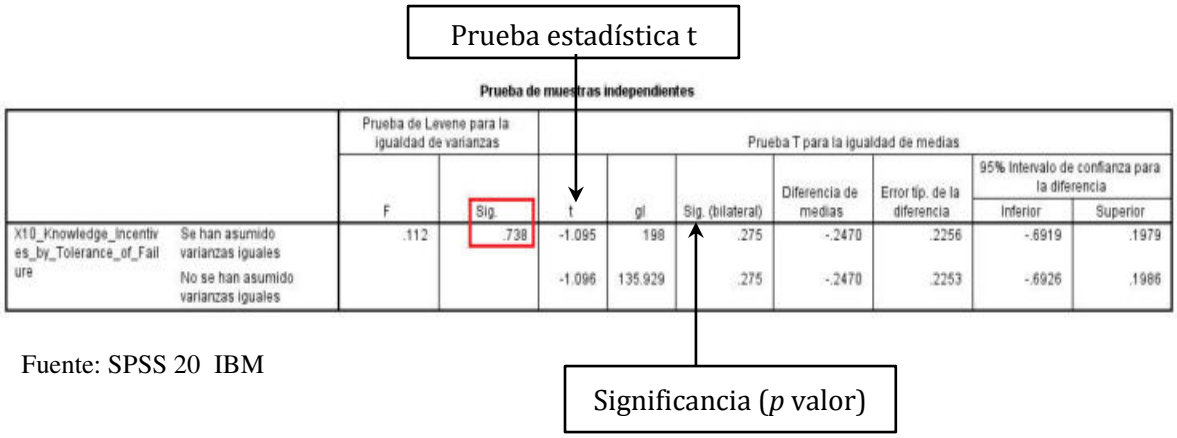
Fuente: SPSS 20 IBM

La cual, nos indica:

- **N** o sea, el número de participantes incluidos con sus resultados de estadística descriptiva.
- Por observación de las **Medias** puede determinarse que las gerentes (female) de **GEN** produjo más incentivos con tolerancia a fallas ($\mathbf{X_{10}}$) que los gerentes (male), pero la diferencia puede no ser tan significante. Para determinar si este resultado **es significante o es debido al azar** es necesario revisar la **tabla de Prueba de muestras independientes.**
- La **Desviación típ.** Muestra que las gerentes (female) de **GEN** tienen una amplia dispersión de puntuaciones que los gerentes (male).
- El **Error típ. De la media** es un estimado de la **Desviación típ.** De la distribución de la media de la muestra basado en la muestra que se está probando. Esto es, que se trata de la distancia estándar o **error de que la media de la muestra provenga de la media de la población.**
- El **Error típ. de la media** es un concepto útil como el usado en el cálculo de las pruebas de confianza de intervalo y las pruebas de significancia, tales como la **prueba *t.***
- En el ejemplo, el **Error típ. de la media** muestra que si se hubieran obtenido todas las medias de cada muestra de las **132 gerentes masculinos (male)** y se analizaran, se estima que la **Desviación típ.** De sus medias sería de **0.1317**. Similarmente si se tomaran todas las muestras de **las 68 gerentes (female)**, se estima que la **Desviación típ.** De sus medias sería **0.1828.**

SPSS produce la **tabla Prueba de muestras independientes. Ver Figura 3.47.**

Figura 3.47. Tabla Prueba de muestras independientes



Fuente: SPSS 20 IBM

Como puede verse de la tabla de la **Figura 3.47**, se generan las líneas: **Se han asumido varianzas iguales/ No se han asumido varianzas iguales**. La columna **prueba de igualdad de varianzas de *Levene*** ayuda a determinar cuál ha de ser usada. Uno de los criterios para usar una **prueba *t* paramétrica** es el supuesto de que ambas poblaciones tienen **varianzas iguales.** Si la **prueba estadística F:**

1. **Si tiene significancia (*p*<=0.05)**, entonces la **prueba de igualdad de varianzas de *Levene*** ha determinado que las **dos varianzas difieren significativamente**, en cuyo caso debemos usar los valores que se muestran en la **parte inferior del recuadro**.

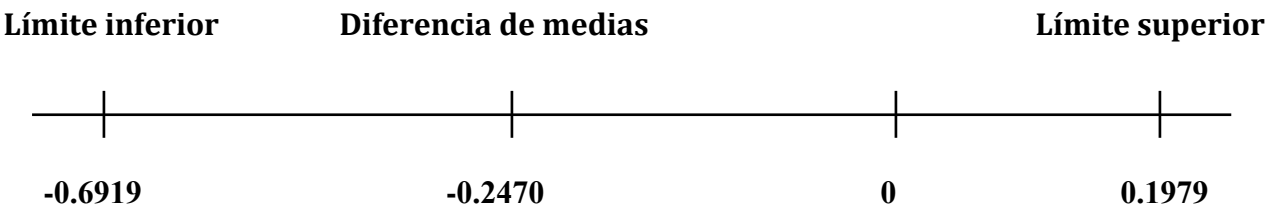
2. Si **NO tiene significancia** (***p*>0.05**) entonces, las **varianzas NO serían significativamente diferentes**, se aceptaría el supuesto de la **igualdad de varianzas** y el uso de los valores que se presentan en **la parte superior del recuadro**.

En el ejemplo **se rechazaría H_1** y se aceptaría **H_0: No** existen **diferencias** significativas entre las varianzas de los grupos de la variable $_{10}$ vs. **GEN, con la siguente explicación:**

- Si la prueba de igualdad de varianzas de ***Levene*** es significante es una materia que se sujeta a juicio académico ya sea que se acepten los valores del fondo del recuadro, o que si tomamos esta violación de los supuestos de las pruebas como una justificante para realizar, en su lugar, la prueba no paramétrica de ***Mann–Whitney U***. Si Usted encuentra una violación inesperada del supuesto de la homogeneidad de la varianza (por ejemplo. Otros estudios de su tipo en su campo que no lo han encontrado), entonces actúe con cuidado al interpretar los resultados de su estudio. Si Usted realiza un reporte basado en el **valor *t*,** asegúrese que el supuesto de valor de varianzas iguales, no sea asumida en la tabla SPSS. Asegure también:
- Para nuestras hipótesis basadas **Sig. (bilateral)** la forma convencional de reportar los hallazgos es el de establecer la **prueba estadística (*t*)** con los grados de libertad entre paréntesis y la probabilidad de este valor con su debido **nivel significancia (*p*)**. Por ejemplo, si se calculó t = -1.095 con 198 grados de libertad y una probabilidad de 0.275, se reporta como: ***t* (198) = -1.095, *p*>0.001.**
- En nuestra prueba se predijo que las gerentes (female) de **GEN** producen más incentivos con tolerancia a fallas ($\mathbf{X_{10}}$) que los gerentes (male), la cual es una hipótesis de una cola (***one-tail hypothesis***). Para hipótesis de bilaterales (***two-tailed hypothesis***) Usted debe establecer que habrá diferencias entre las medias **pero no predecirá la dirección de la diferencia.** Si Usted hace una predicción de una cola (***one-tailed prediction***) necesitará dividir el '**Sig. (bilateral)**' (***p* valor**) a la mitad. Si hay una diferencia significante entre las medias, Usted necesita para determinar si las medias están mostrando la diferencia en la dirección predicha. En el ejemplo, necesitamos asegurarnos que la **puntuación del error medio** para las gerentes (fémale) es de hecho, más grande que la de los gerentes (mal). Nuestra hipótesis es de una cola (***one-tailed hypothesis***) en la medida en la que predecimos una dirección de la diferencia en la media, que haría al valor ***p*= (0.001)/2 = 0.0005**. Como nuestra probabilidad de **0.0005** es más pequeña que nuestro nivel de significancia **0.001,** usaremos el signo menor que **(<)** para indicar **que nuestro resultado es 0.001 nivel de significación.**
- No se preocupe si tiene valores t negativos para hipótesis de 2 colas (***two-tailed hypothesis*).** Ya sea positivo o negativo es dependiente en el grupo de puntuaciones que fuera primeramente ingresado, dentro de las ecuaciones de las **pruebas *t*.** Así se ingresa las gerentes (female) de **GEN con $\mathbf{X_{10}}$ primero** ya que fueron más grandes los valores así que la **t es positivo**
- La **Diferencia de medidas** es la diferencia entre las medidas de nuestros 2 grupos. Imagine que **H_0 es verdadera,** entonces la diferencia real entre las medias de poblaciones **es cero.** Si se seleccionan todas las muestras de tamaño **132** y de **68** y se trabaja la diferencia en sus medias podríamos encontrar que diferencias en las medias serían por el **azar.**

- El **Error típ. De la diferencia** estima la **Desviación típ.** De todas las diferencias en una muestra de medias cuando la **HO hipótesis nula** es verdadera. Esto indica la diferencia en las medias que esperaríamos por el azar si la hipótesis nula es verdadera. En nuestro caso el **Error típ de la diferencia** es estimado para ser **-0.2256.**
- La **prueba *t*** compara la diferencia en las medias con el **Error típ de la diferencia:**

  T=(diferencia de medias/Error típ. De la diferencia)
- De nuestro ejemplo la **Diferencia de medias= -0.2470 es 1.095** más grande que el **Error típ. De la diferencia,** entonces nuestra diferencia de medias es suficientemente grande para ser significante en un nivel de ***p*<0.275.**
- El 95% de confianza de que la población en su diferencia de medias, este entre los límites **inferior y superior.**
- La **prueba *t*** nos dice si nuestra diferencia es significante o no. Sin embargo, el **intervalo de confianza** nos da más información acerca del tamaño de la diferencia.
- El **intervalo de confianza** nos da un estimado de la diferencia real en la población. Al observar nuestras salidas podemos ver que el límite más bajo de **-0.6919** y el **límite más alto 0.1979,** lo cual indica que podemos tener confianza de que la verdadera diferencia media de la población se encuentra entre estos dos valores. Por lo tanto, en el **peor escenario** las gerentes (female) de **GEN** producirán **-0.69** incentivos con tolerancia a fallas **(X_{10})** que los gerentes (male).

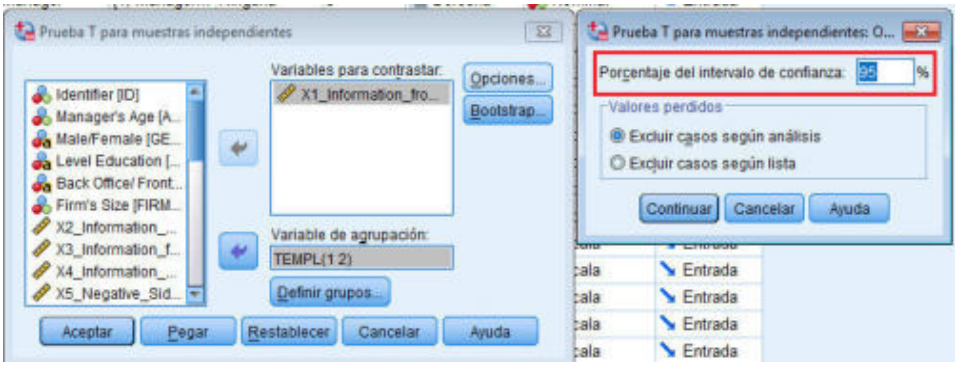


- El **intervalo de confianza** son frecuentemente usados como **indicadores alternativos o suplementarios de significancia estadística.** Aquí un ejemplo de la notación a usar para su reporte:

  Diferencia de medias= -0.2470 (95% IC: -0.6919 a 0.1979)

  Se destaca que el valor por omisión del intervalo de confianza dado por SPSS es: **95 %**, pero puede ser cambiado como se puede apreciar en el cuadro de diálogo abajo anexo. Ver **Figura 3.48.**

Figura 3.48. Cuadro de diálogo para seleccionar el procentaje de intervalo de confianza



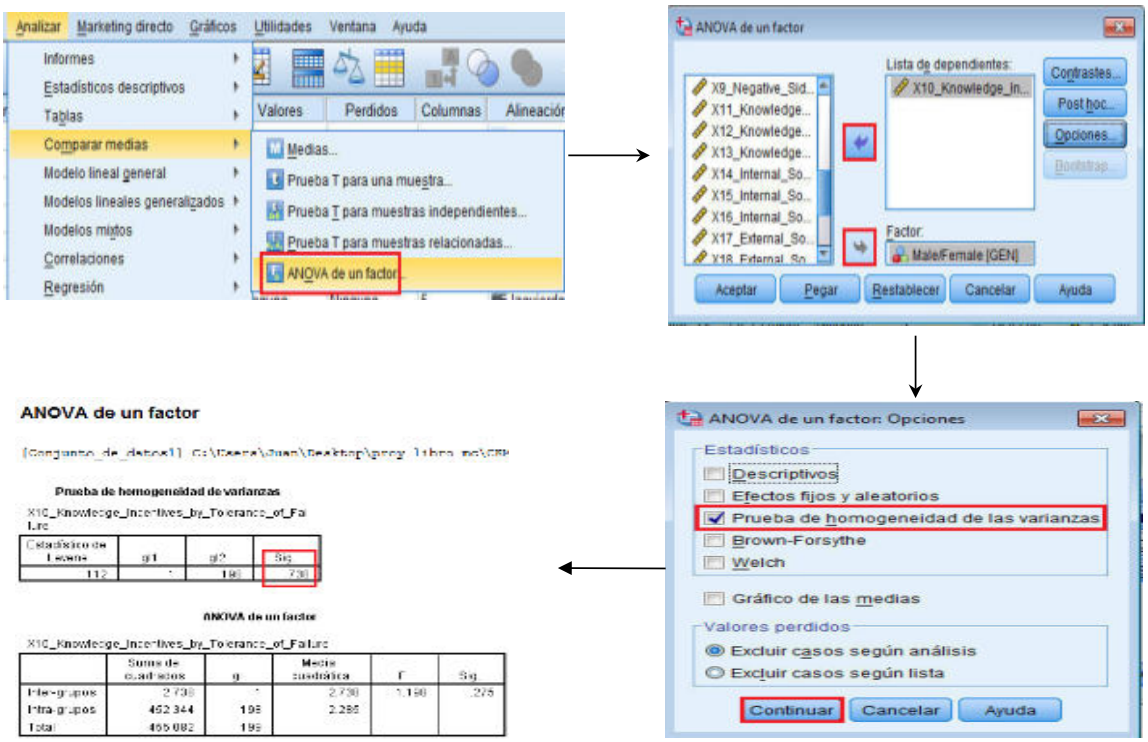
Fuente: SPSS 20 IBM

3.10.5. Homocedasticidad (ANOVA)

-Problema 32: verificar por **test de Levene**, la homocedasticidad (varianzas iguales de las variables dependientes) de todo el grupo de tamaño de empresas (**Firm's size**) de la base de datos **CKM_MKT_Digital_imputaciones3.sav. Nota:** evitar variables de cadena; declarar mejor Tipo: Numérico; Medida: Nominal como caso variable **GEN.**

-Teclear: Analizar->Comparar medias->ANOVA de un factor->Seleccionar en Lista de dependientes: Selección de Variable métrica: X_{10}; Selección de Variable Factor: GEN->Opciones->Selección: Prueba de homogeneidad de las varianzas->Continuar->Aceptar. Ver Figura 3.49.

Figura 3.49. Test de *Levene*

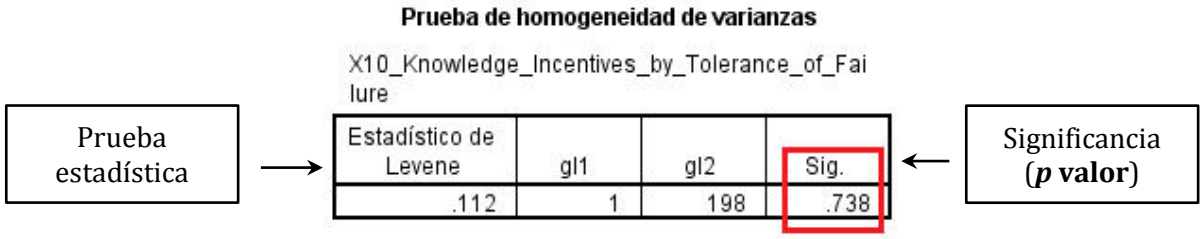


Fuente: SPSS 20 IBM

Resultados: SPSS genera la tabla **Prueba de homogeneidad de variación** es el mismo valor de **p=0.738>0.005** obtenido de la **Prueba T para muestras independientes,** por lo que **se acepta** la **Ho: No existen** diferencias significativas entre las varianzas de los grupos de la variable Firm's size.

SPSS produce la **Tabla prueba de homogeneidad de varianzas de *Levene*. Ver Figura 3.50**

Figura 3.50. Tabla Prueba de homogeneidad de varianzas



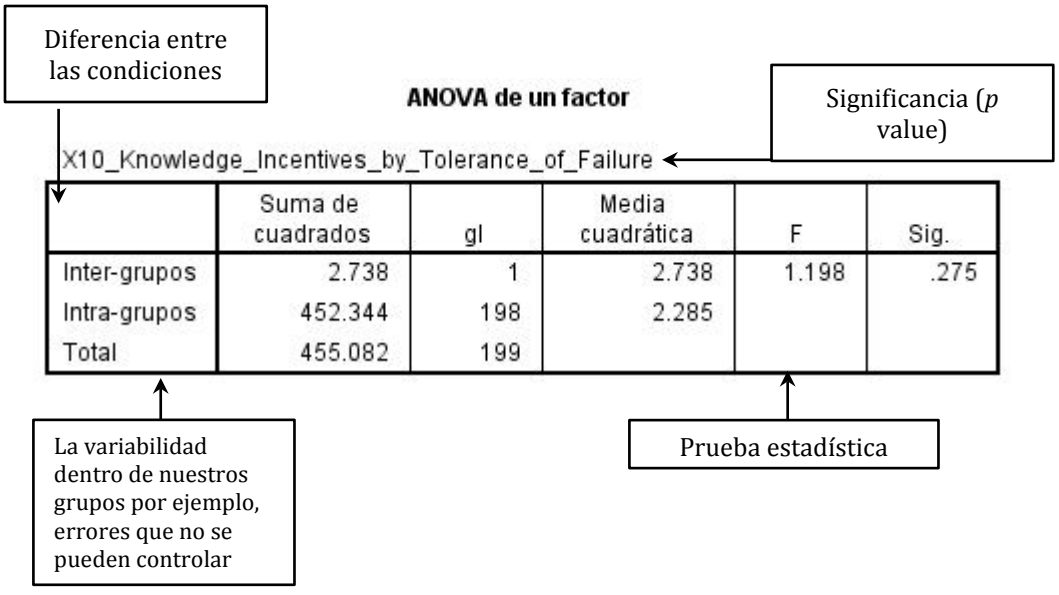
Fuente: SPSS 20 IBM

La cual, explica:

- Si hemos encontrado que nuestro segundo supuesto: "*los grupos tienen aproximadamente iguales varianzas sobre la variable dependiente*".
- **Si la prueba de *Levene* resulta que no es significante (p >0.05), las varianzas son aproximadamente iguales.**Aquí, se observa que **p=0.738>0.05**, por lo que se asume que las varianzas son aproximadamente iguales.
- **Si la prueba de *Levene* resulta que sí es significante (p < 0.05), entonces las varianzas son significativamente diferentes.** Si este fuera el caso, se deberá considerar la transformación que permita hacer que sus varianzas sean más homogéneas

SPSS también genera la **tabla ANOVA de un factor** contiene información clave tomando en cuenta valores calculados de **estadística *F*. Ver Figura 3.51.**

Figura 3.51. Tabla ANOVA de un factor



Fuente: SPSS 20 IBM

La cual explica:

- Los grados de libertad (**gl**) se reportan. En ANOVA se generan 2 valores: uno para el factor **Inter-grupos** y otro factor para el **error** (**Intra-grupos**), en nuestro caso **gl= (1,198)**
- Si el **SPSS** establece que la probabilidad (**Sig.**) es **0.000**, significa que el SPSS ha redondeado (arriba/abajo) la cantidad al número más cercano a 3 decimales. Sin embargo, desearíamos siempre redondear el último 0 a 1, así que ***p* < 0.001.**
- La forma convencional de reportar los hallazgos es establecer la **prueba estadística (F)**, grados de libertad (**gl**) y la probabilidad (**Sig.**), por lo que la notación para reportar es:

  F (1,198) =1.198; p=0.275>0.0
- Como **p>0.001**, significa que **NO** hay una alta diferencia significante entre los grupos

3.10.6. Linealidad

Es el supuesto final a examinar. En el caso de **variables individuales**, se relacionan las pautas de **asociación** entre **cada par de variables** y **la capacidad del coeficiente de correlación** para representar adecuadamente la relación. Si hay un indicio de **relaciones no linea**les, entonces él debe **transformar una o ambas variables para conseguir la linealidad**, como **crear variables adicionales** para representar los **componentes no lineales**. Deberá apoyarse en la **inspección visual** de las relaciones para determinar si están presentes relaciones no lineales, a través de los **gráficos de dispersión** para todas las variables métricas en el conjunto de datos. Este método, suele ser poco confiable ya que es posible considere que las transformaciones no sean necesarias, así que se deberá proceder comprobarlo también en el modelo multivariante entero al llevarse a cabo el **examen de los residuos en la regresión múltiple. Se recomienda realizar primero el AFE para su evaluación.**

3.10.7. Aspecto para reporte de prueba estadística de normalidad

Prueba de gráficos.- Visual e imprecisa
Prueba mediante estadísticos.-Por asimetría y curtosis. Nivel intermedio de precisión
Prueba de hipótesis estadística. Mayor precisión siendo la Ho. Los datos están distribuidos en forma normal, Uso de test ***Kolmogorov-Smirnov*** y / o test de ***Shapiro-Wilks***.
1.-Que pretende probar (por ej. normalidad) y qué técnica aplicó. Por ej. Prueba de ***Kolmogorov-Smirnov***
2.-Qué es lo que obtuvo como resultado.(Estadísticos/gl/sig.)
3.-La p o significancia
4.- Cómo interpreta la significancia (aceptación/rechazo de Ho)
Redacción: la prueba de normalidad KS indica que se cumple/no se cumple el supuesto de normalidad en ciertas variables, con Estadísticos, gl y sig. de acuerdo a la variable $\mathbf{X_1}$...$\mathbf{X_{23}}$ y por lo tanto se rechaza/no se rechaza la Ho. Así también se aprueban/no se aprueban las $\mathbf{_0/H_1}$ debido a que la Homoedasticidad presenta/no presenta varianza/variabilidad igual/diferente dado $\boldsymbol{p}$>0.05

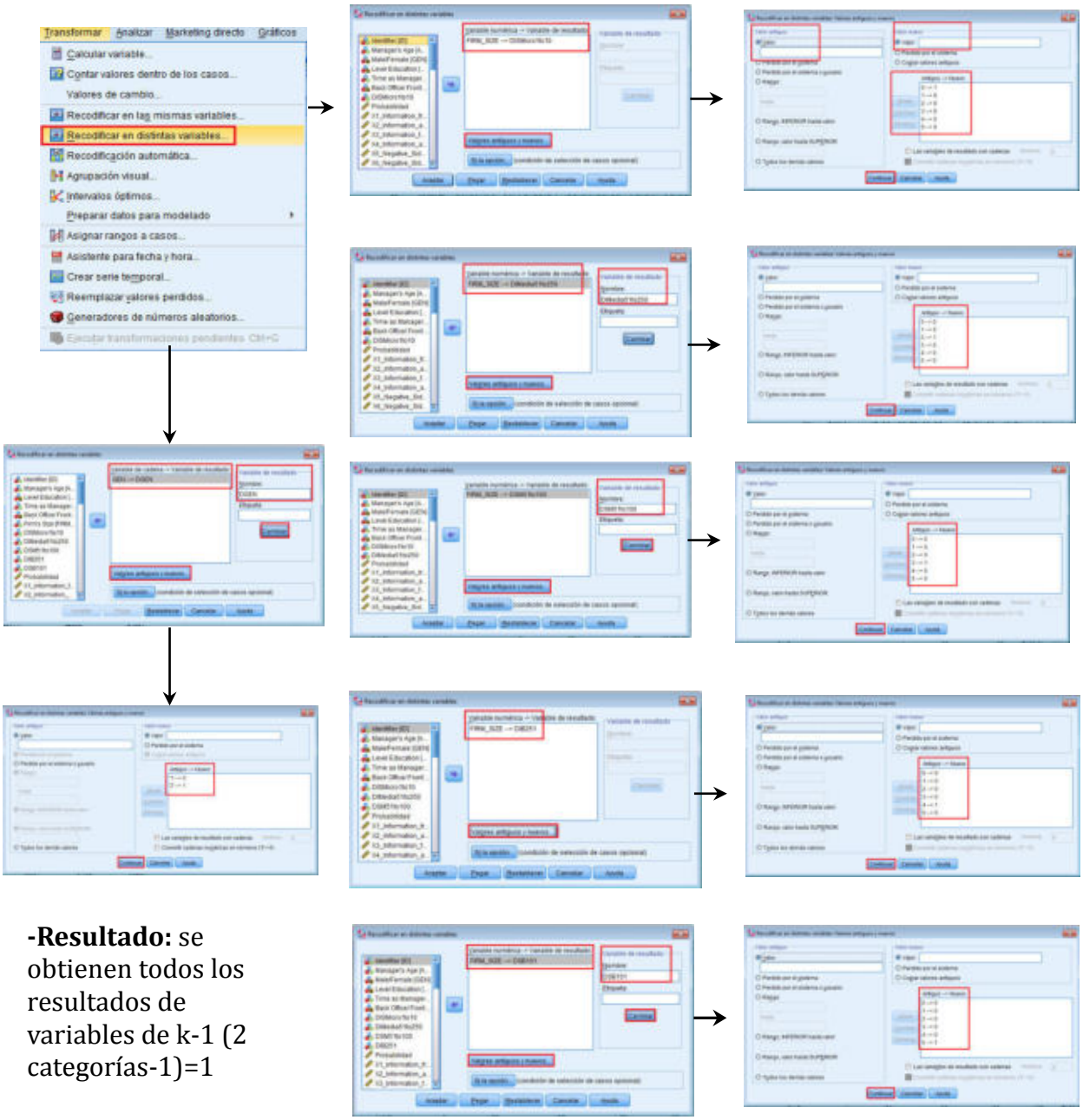
3.11. Datos No métricos con variables ficticias (Dummies)

Un factor crítico en la elección y aplicación de la técnica multivariante correcta es la medición de las propiedades de las variables dependientes e independientes. Algunas

técnicas, tales como el **análisis discriminante** o el **MANOVA**, requieren específicamente **datos no métricos** como **variables dependientes o independientes**. En muchos casos, las **variables métricas** tienen que ser utilizadas como **variables independientes**, como ocurre en el **análisis de regresión**, en el **análisis discriminante** y en la **correlación canónica**. Además, las **técnicas de interdependencia** de **análisis factorial** y **clúster** normalmente requieren **variables métricas**. Con este fin, todos los debates han asumido la **medición métrica de variables**. Pero, ¿qué podemos hacer cuando las variables son **no métricas**, con dos o más categorías? ¿Se **excluyen** en muchas técnicas multivariantes las variables no métricas tales como género, situación marital u ocupación? **La respuesta es negativa**. Existen procedimientos para **incorporar las variables no métricas** a muchas de estas situaciones que requieren **variables métricas**. Usted tiene a su disposición un método para usar **variables dicotómicas**, conocidas como **variables ficticias**, que actúan como **variables de sustitución**. **Una variable ficticia es una variable dicotómica que representa una categoría de variable independiente no métrica**. Cualquier variable no métrica con k categorías puede ser representada como **variable ficticia k- l**

-Problema 33: Determine cómo el tamaño de empresa en todas sus categorías (**Firm´s size**), el Genero del Gerente (**GEN**), la variable Incentivos y recompensas del conocimiento (**X_{11}** _Knowledge_Incentives_by_Rewards_and_Recognition) influyen en la variable dependiente: Compartir conocimiento entre empleados y administradores (**X_{13}** _Knowledge_Fluence_among_Employees_and_Managers) de la base de datos **CKM_MKT_Digital_imputaciones3.sav**. Para determinarlo, se requiere precisar la **variable categórica base (o categoría de referencia y su valor siempre a 1) para que el resto de las variables se consideren en cero (0)**. Por ej. la variable **Firm _size**, tiene los valores previos: **0=Industry and services is Micro (1-10); 1= Industry and services is Little (11-50); 2= Industry is Media (51-250); 3= Services is Media (51-100); 4= Industry is Big (251 a+); 5= Services is Big (101 a+) (prueba de hipótesis por diferencia de medias**). Así se deberá escoger cuál de los valores será el de referencia a **1**; se sugiere iniciar con Industry and services is Micro (1-10) con valor de referencia a **1.**
-Teclear: Transformar->Recodificar en distintas variables cualitativas: GEN female en 1, male en 0 Firm_size; Variable de resultado: DISMicro1to10 (posteriormente: DIMedia51to250; DSM51to100; DIB251; DSB101) => Cambiar->Valores antiguos y nuevos: Seleccionar valor antiguo (1, 2, 3, 4, 5); Seleccionar valor nuevo (1, 0, 0, 0,0) ->Continuar->Aceptar. Nota: realizar con las 4 variables restantes borrando los valores anteriores. Verifique en la base de datos que todos los **1s** correspondan a **Services is Media (51-100) del campo Firm_size;** puede señalar y arrastrar el campo **DISMicro1to10** a donde mejor convenga para revisar. Con lo anterior se asegura que la variable con dato **1** se incluye en la regresión múltiple y el resto de los datos son anulados (valor **0**). Así, se construyen tantas variables **dummy** cuantas categorías tenga el campo de referencia-1 **(k-1). Las dummy son 1, el resto son cero en cada caso. Ver Figura 3.52 y Figura 3.53.**

Figura 3.52. Transformación de variables: Proceso

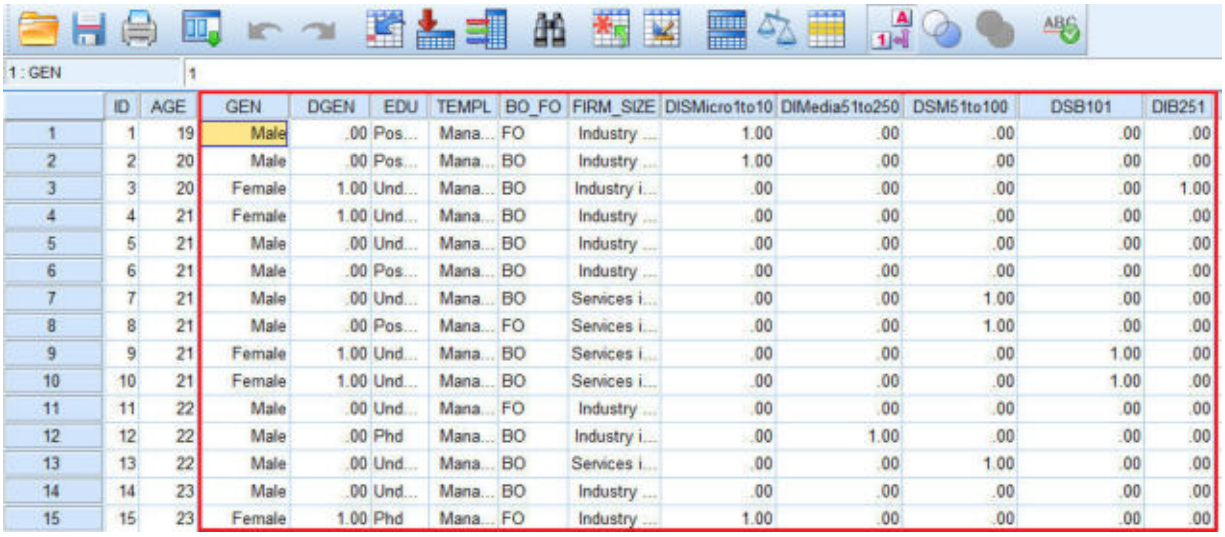


-Resultado: se obtienen todos los resultados de variables de k-1 (2 categorías-1)=1

Fuente: SPSS 20 IBM

-Resultado: se obtienen todos los resultados de variables de k-1 (6 categorías-1)=5

Figura 3.53. Transformación de variables: resultado en la base de datos.



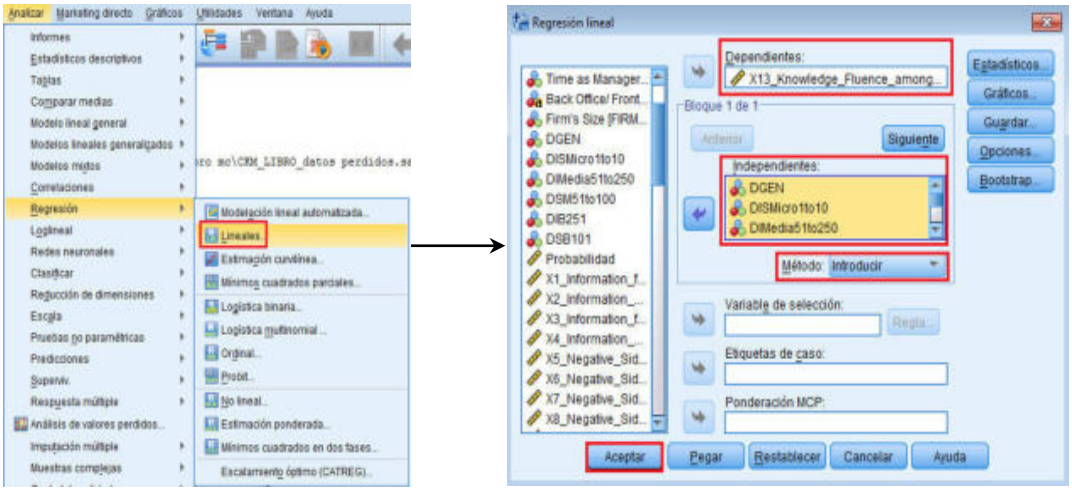
1 : GEN | 1

	ID	AGE	GEN	DGEN	EDU	TEMPL	BO_FO	FIRM_SIZE	DISMicro1to10	DIMedia51to250	DSM51to100	DSB101	DIB251
1	1	19	Male	.00	Pos...	Mana...	FO	Industry ...	1.00	.00	.00	.00	.00
2	2	20	Male	.00	Pos...	Mana...	BO	Industry ...	1.00	.00	.00	.00	.00
3	3	20	Female	1.00	Und...	Mana...	BO	Industry i...	.00	.00	.00	.00	1.00
4	4	21	Female	1.00	Und...	Mana...	BO	Industry ...	.00	.00	.00	.00	.00
5	5	21	Male	.00	Und...	Mana...	BO	Industry ...	.00	.00	.00	.00	.00
6	6	21	Male	.00	Pos...	Mana...	BO	Industry ...	.00	.00	.00	.00	.00
7	7	21	Male	.00	Und...	Mana...	BO	Services i...	.00	.00	1.00	.00	.00
8	8	21	Male	.00	Pos...	Mana...	FO	Services i...	.00	.00	1.00	.00	.00
9	9	21	Female	1.00	Und...	Mana...	BO	Services i...	.00	.00	.00	1.00	.00
10	10	21	Female	1.00	Und...	Mana...	BO	Services i...	.00	.00	.00	1.00	.00
11	11	22	Male	.00	Und...	Mana...	FO	Industry ...	.00	.00	.00	.00	.00
12	12	22	Male	.00	Phd	Mana...	BO	Industry i...	.00	1.00	.00	.00	.00
13	13	22	Male	.00	Und...	Mana...	BO	Services i...	.00	.00	1.00	.00	.00
14	14	23	Male	.00	Und...	Mana...	BO	Industry ...	.00	.00	.00	.00	.00
15	15	23	Female	1.00	Phd	Mana...	FO	Industry ...	1.00	.00	.00	.00	.00

Fuente: SPSS 20 IBM

Para comprobar, realice la regresión lineal y **teclee: Analizar->Regresión->Lineales->Seleccionar variable métrica dependiente: X_{13} ; Seleccionar variables independientes: DGEN, DISMicro1to10, DIMedia51to250; DSM51to100; DIB251; DSB101; X_{11} ->Aceptar. Ver Figura 3.54.**

Figura 3.54. Transformación de variables: Resultados



Fuente: SPSS 20 IBM

SPSS genera **tabla Resumen del modelo. Ver Figura 3.55.**

Figura 3.55. Resumen del modelo

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	.798[a]	.636	.623	.5143

a. Variables predictoras: (Constante), X11_Knowledge_Incentives_by_Rewards_and_Recognition, DSB101, DGEN, DIMedia51to250, DIB251, DISMicro1to10, DSM51to100

Fuente: SPSS 20 IBM

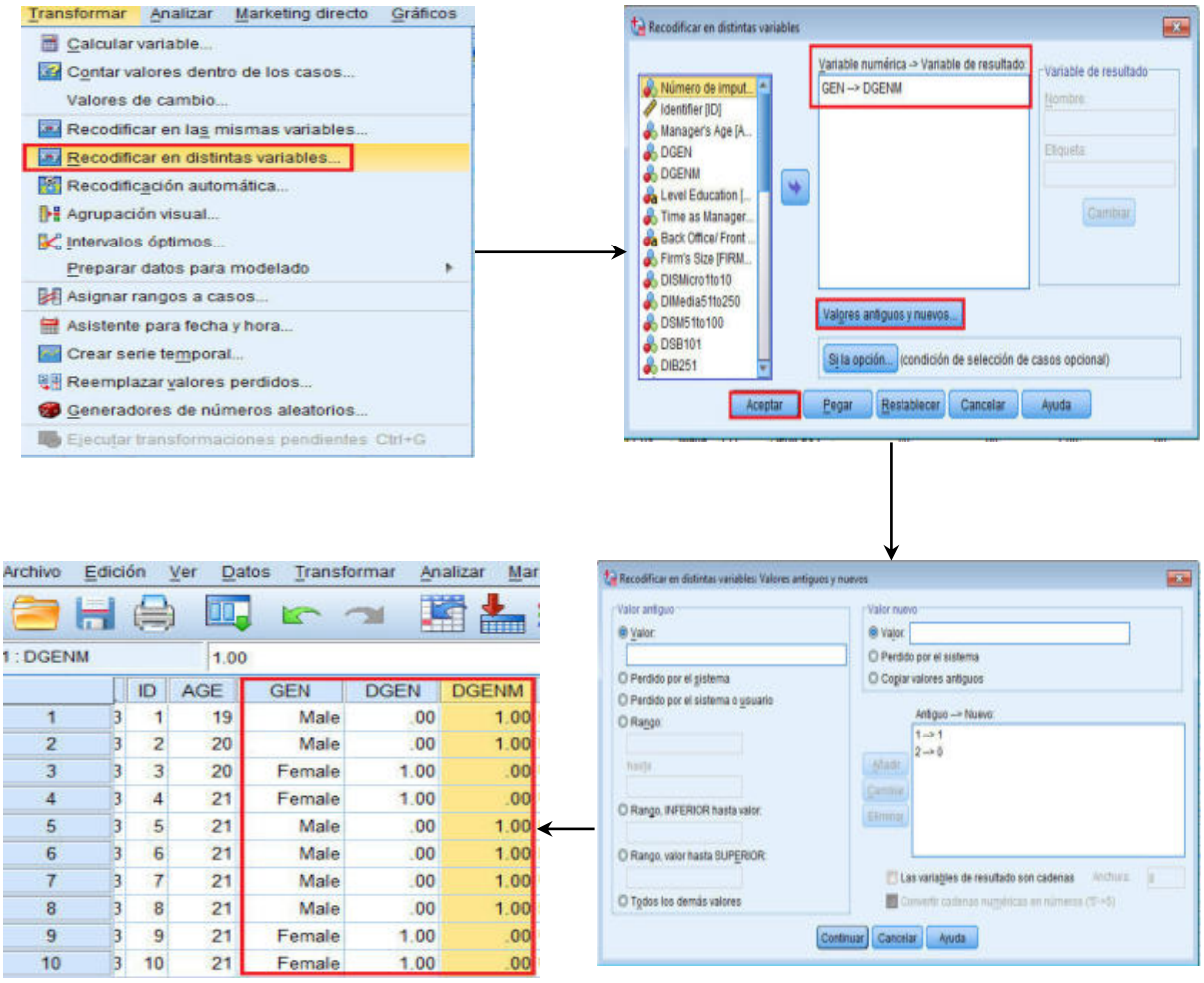
La cual, nos indica:

- Que el modelo explica el **63.6%** del fenómeno, por lo que como la variable **GEN** puso en **0** a los gerentes (male), y en **1** a las gerentes (female) y en todas la categoría de empresas hay Gran resistencia por dar incentivos para la mejora del Conocimiento

-Problema 34: Suponga ahora que desea comparar cuando los gerentes son masculinos (male); se deberá poner variable dummy DGENM en **1** y las gerentes femeninas (female) en **0**

-Teclear: Teclear: Transformar->Recodificar en distintas variables cualitativas: GEN female en 1, male en 0->Aceptar. Ver Figura 3.56.

Figura 3.56. Transformación de variables: Proceso

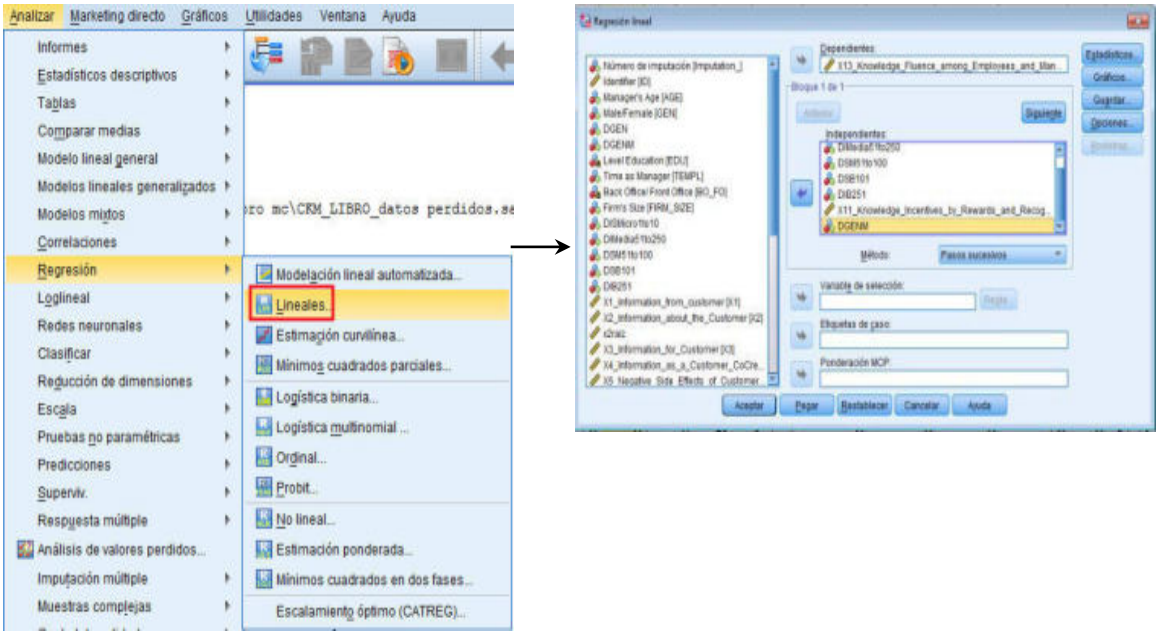


	ID	AGE	GEN	DGEN	DGENM
1	1	19	Male	.00	1.00
2	2	20	Male	.00	1.00
3	3	20	Female	1.00	.00
4	4	21	Female	1.00	.00
5	5	21	Male	.00	1.00
6	6	21	Male	.00	1.00
7	7	21	Male	.00	1.00
8	8	21	Male	.00	1.00
9	9	21	Female	1.00	.00
10	10	21	Female	1.00	.00

Fuente: SPSS 20 IBM

Realizar nueva regresión lineal.
-Teclear: Analizar->Regresión->Lineales->Seleccionar variable métrica dependiente: X_{13}; Seleccionar variables independientes: DGENM, DISMicro1to10, DIMedia51to250; DSM51to100; DIB251; DSB101; X_{11} ->Aceptar. Ver Figura 3.57.

Figura 3.57. Transformación de variables: Resultados



Fuente: SPSS 20 IBM

SPSS genera la **tabla Resumen del Modelo. Ver Figura 3.58.**

Figura 3.58. Tabla Resumen del modelo

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	.793[a]	.629	.627	.5113

a. Variables predictoras: (Constante), X11_Knowledge_Incentives_by_Rewards_and_Recognition

Fuente: SPSS 20 IBM

Donde se observa que el modelo explica el **62.9%** del fenómeno, por lo que como la variable **GEN** puso en **1** a los gerentes (male), y en **0** a las gerentes (female) y en todas la categoría de empresas.

En este caso se puede concluir que las gerentes (female) contribuyen más (**62.9%<63.6%**) a que se dé la variable $\mathbf{X_{13}}$ de Compartir conocimiento entre empleados y administradores (Knowledge_ Fluence_ among_ Employees_ and_ Managers), considerando el tamaño de la firma (Firm´s size) así como la variable $\mathbf{X_{11}}$ de Incentivos y recompensas del conocimiento (Knowledge_Incentives_by_Rewards_and_Recognition)

Referencias

Anderson, E. (1969), A Semigraphical Method for the Analysis of Complex Problems. *Technometrics* 2 (August): 387-91.

Box, G. E., y Cox D. R. (1964), An Analysis of Transformations*. Journal of the Royal Statistical Society B* (26): 211-43.

Cohen, J., y Cohen, P. (1983), Applied Multiple Regression/*Correlation Analysis for the Behavioral Sciences, 2d* ed. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Daniel, C., y Wood F. S. (1980), Fitting Equations to Data, 2d Ed. New York: Wiley-Interscience.

Dempster, A. P., y Rubín D. B. (1983), Overview, in incomplete Data in Sample Surveys: Theory and Annotated Bibliography, vol. 2. Madow, Olkin, and Rubín, cds. New York: Academic Press.

IBM (2011a). *IBM SPSS Statistics Base 20.* EUA. .Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Base.pdf

IBM (2011b).*Guía breve de IBM SPSS Statistics 20.* EUA.Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Brief_Guide.pdf

IBM (2011c). IBM SPSS Missing Values 20. EUA. .Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Missing_Values.pdf

Johnson, R. A., y Wichern D. W. (1982), Applied *Multivariate Statistical Analysis*. Upper Saddle River, N.J.: Prentice-Hall.

Little, R. J. A., y Donald B. R. (1987), *Statistical Analysis with Missing Data*. New York: Wiley.

Mosteller, F y Tukey, J.W. (1977), *Data Analysis and Regresion* . Reading , Mass: Addison-Wesley

Weisberg, S. (1985), *Applied Linear Regression*. New York: Wiley.

Capítulo 4. Confiabilidad en Cuestionarios

4.1. Análisis de validez y confiabilidad: ¿Qué es?

SPSS, tiene importantes herramientas que permiten analizar los cuestionarios diseñados por los investigadores. Para saber más, consulte: IBM, 2011a; IBM, 2011b; IBM,2011c. Así, es posible por ejemplo:

1. Partir del **cruce-tabular** que muestre las relaciones entre los diferentes indicadores o preguntas, por ejemplo, si Usted requiere conocer de los gerentes de la industria de las telecomunicaciones su género como primer pregunta y grado académico máximo como la segunda pregunta, lo podrá visualizar mejor a través de aplicar el **cruce-tabular**
2. Ahora bien, pruebas como un ***Chi-cuadrado***, también se pueden aplicar para comparar los patrones de estos datos.
3. Diferentes cuestionarios conducirán al uso de diferentes pruebas como las **pruebas *t*, o ANOVA.** Incluso la pertinencia de las variables a utilizar a través del análisis factorial, suele emplearse para asegurar la **validez** del instrumento o grado en el que un instrumento en verdad mide la variable que se busca medir. (por ejemplo, si está diseñado para medir el instrumento la innovación, no debe medir competitividad). Al respecto de la **validez**, la tenemos clasificada en :

-Validez de Contenido: grado en que un instrumento refleja un dominio específico de contenido de lo que se mide. Por ejemplo: una prueba de operaciones aritméticas no tendrá validez de contenido si incluye sólo problemas de adición y excluye problemas de sustracción, multiplicación y división (**Validez de juicio de experto**).

-Validez de Criterio: se establece al validar un instrumento de medición al compararlo con algún criterio externo que pretende medir lo mismo.

-Validez Concurrente y la Validez Predictiva. En las campañas electorales, los sondeos se comparan con los resultados finales de las elecciones. Por ejemplo: Coeficiente de Contingencias, ***Spearman – Brow, Pearson,*** **Alfa de *Cronbach*** y la Técnica ***Aiken***.

-Validez de Constructo: debe explicar el modelo teórico empírico que subyace a la variable de interés. Por ejemplo El Análisis de Factorial y Análisis de Cofactores, el Análisis de Covarianza.

Validez Total= (Validez de Contenido + Validez de Criterio + Validez de Constructo)/3

4.2. Análisis de confiabilidad: ¿Qué es?

Sin embargo, ninguno de estos diferentes análisis será significativo a menos que nuestro cuestionario sea **confiable.** La **confiabilidad es la capacidad del cuestionario para medir de forma consistente el tema bajo estudio en diferentes momentos y a través de diferentes poblaciones**. Imagine tener una cinta métrica que mide a una persona como 1 metro 70 centímetros en un día y 1 metro 50 centímetros en otro día. Esta sería una cinta métrica de muy baja confiabilidad. Existen diferentes formas de evaluar la fiabilidad como el método de las dos mitades.

Existen diversas formas de evaluar la confiabilidad, tales como:

1. Medida de estabilidad: Un mismo instrumento de medición se aplica dos más veces a un mismo grupo de personas, después de cierto periodo. Se le conoce también como **Confiabilidad por test-retest, "*r*" de *Pearson.***

2. Método de formas alternativas o paralelas. Aquí no se administra el mismo instrumento de medición, sino dos o más versiones equivalentes de este. Se le conoce también como **Coeficiente de correlación producto-momento de *Pearson*.**

3. Método de mitades partidas: Se necesita solo una aplicación, el total de los ítems se divide en dos partes y se comparan los resultados. (***Pearson y Spearman-Brown***).

4. Medidas de consistencia interna: Requiere sólo una administración. Confiabilidad del test según el método de división de las mitades por ***Rulon y Guttman***, ***Fórmula 20 de Kuder-Richardson*** y Coeficiente del **Alfa de *Cronbach***.

Por su amplia aplicabilidad, el presente documento se referirá al Alfa de ***Cronbach***

4.3. Análisis de confiabilidad: Alfa de *Cronbach*

Cuando llevamos a cabo un cuestionario que a menudo lo estamos utilizando para medir un determinado constructo, tales como la **innovación**. En el cuestionario se diseñan una serie de preguntas o indicadores acerca de este constructo, por lo que para la **"*innovación*"** se puede plantear: ¿se realiza en la organización?, ¿es a nivel de producto y/o servicio?, ¿es a nivel de mercadotecnia?, ¿se aplica a en los procesos o en el modelo de negocios? De tal manera que un gerente de una empresa de alta tecnología, con alto sentido de la estrategia

que implica la innovación al responder el cuestionario, deberá medir con **validez** el constructo. Debe ser capaz de medir a otras empresas dedicadas a la innovación apropiadamente, reportando correctamente su nivel de innovación. **El cuestionario deberá No sólo ser sólo válido sino también confiable**. Con esto, deberíamos esperar que si aplicamos el cuestionario al gerente de la empresa de alta tecnología nuevamente, **entonces mostraría el mismo resultado**. La **confiabilidad** puede evaluarse de diferentes maneras. Para lograrlo, una alternativa es la de **dividir nuestro cuestionario en dos y ver si la primera mitad de las preguntas se resuelven con el mismo resultado que la segunda mitad (confiabilidad de dos mitades).** Podemos llevar el cuestionario a niveles de negación y/o afirmación de las preguntas a fin de lograr las mismas respuestas y asegurarnos de que el instrumento es totalmente confiable, de lo contrario se deberán descartar y/o reemplazar aquellas preguntas que no logren dicho objetivo. La prueba de **Alfa de *Cronbach*** es el método más popular de examinar la **confiabilidad** y se basa en el número de **ítems, preguntas o indicadores** de un cuestionario así como el promedio de la correlación entre **ítems**. Se deberá asumir que toda respuesta a las preguntas del cuestionario por cada individuo existirá una desviación de la respuesta esperada vs. La obtenida la cual se denomina **error aleatorio**. Así:

1. Una alta correlación entre los diferentes ítems indicará que se **está midiendo lo mismo** y por lo tanto existirán valores pequeños de **error**.

2. Una baja correlación entre los diferentes ítems indicará que hay una **gran cantidad de errores y los elementos no son confiable medir la misma cuestión.**

Alfa de *Cronbach* oscila entre **0** para una prueba completamente confiable (aunque técnicamente puede bajar por debajo de **0**) a **1** para una prueba completamente fiable. ¿Qué valor de **Alfa de *Cronbach*** se debe obtener para que un cuestionario o una medida sean fiables? Hay un cierto debate alrededor de esto, con algunos estadísticos sugiriendo **0.7** o más, mientras que otros recomiendan **0.8**. Esto dependerá en cierta medida del número de elementos de la prueba y del número de participantes, pero **0.75** es un valor de compromiso razonable que tomar como punto de referencia, siendo el referente sugerido:

- **0.90 y más, excelente confiabilidad**
- **0.70 a 0.90, alta confiabilidad**
- **0.50 a 0.70 moderada confiabilidad**
- **0.50 y menor, baja confiabilidad**

Alfa de *Cronbach* se basa en los mismos supuestos que hemos considerado en nuestros capítulos sobre la correlación lineal simple y múltiple, ya que emplea un análisis correlacional. También se emplea un modelo en el que se supone que las puntuaciones observadas son las **verdadero puntaje más error**, y así, como se ha explicado, para que este modelo sea apropiado **los errores deben ser aleatorios** (y por lo tanto imparciales). **Si los supuestos NO se cumplen, el valor de Alfa de *Cronbach* puede ser una subestimación o una sobreestimación del valor correcto.**

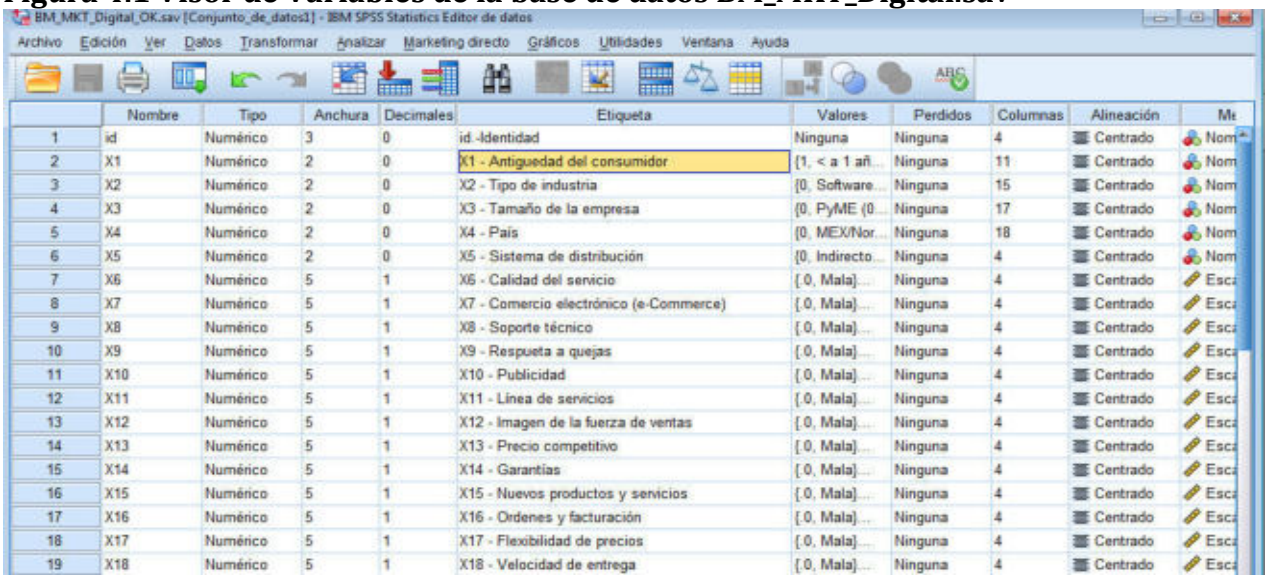
4.4. Alfa de *Cronbach*: Ejemplo

Paso 1: Objetivos

-Problema 1: La empresa **MKT Digital** ha construido un cuestionario para examinar su modelo de negocios, y desea probar su confiabilidad con una muestra de **200** participantes

y utilizando una escala de **10** puntos siendo 1 totalmente en desacuerdo y **10**, totalmente de acuerdo. Ver **Figura 4.1 y 4.2**

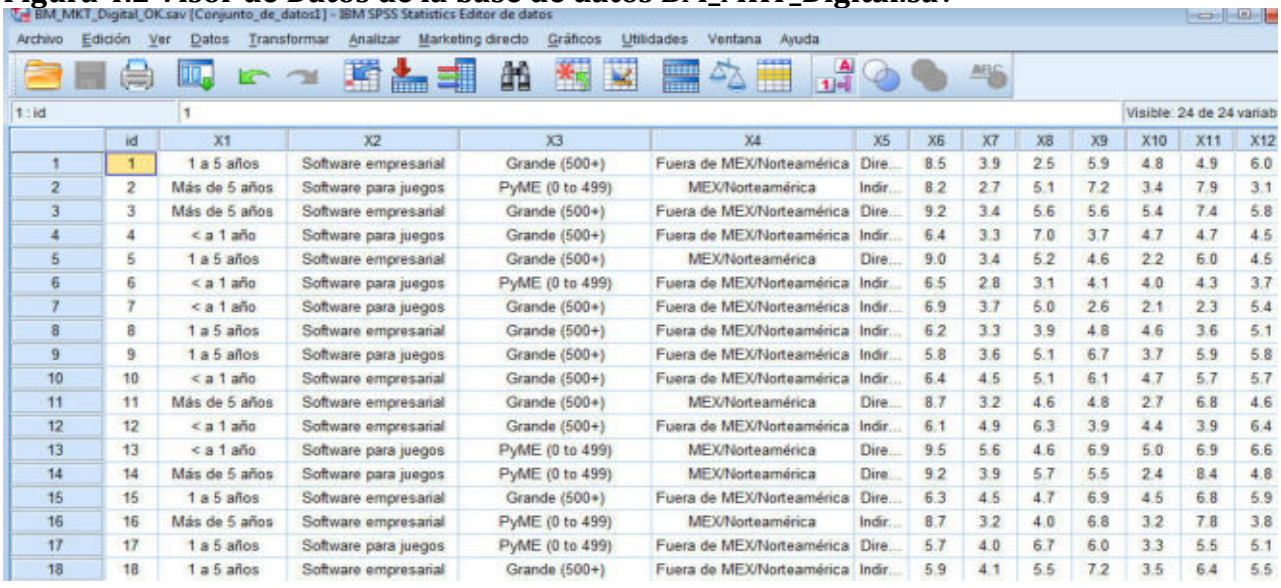
Figura 4.1 Visor de Variables de la base de datos BM_MKT_Digital.sav



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Me
1	id	Numérico	3	0	id -Identidad	Ninguna	Ninguna	4	Centrado	Nom
2	X1	Numérico	2	0	X1 - Antiguedad del consumidor	{1, < a 1 añ...	Ninguna	11	Centrado	Nom
3	X2	Numérico	2	0	X2 - Tipo de industria	{0, Software...	Ninguna	15	Centrado	Nom
4	X3	Numérico	2	0	X3 - Tamaño de la empresa	{0, PyME (0...	Ninguna	17	Centrado	Nom
5	X4	Numérico	2	0	X4 - País	{0, MEX/Nor...	Ninguna	18	Centrado	Nom
6	X5	Numérico	2	0	X5 - Sistema de distribución	{0, Indirecto...	Ninguna	4	Centrado	Nom
7	X6	Numérico	5	1	X6 - Calidad del servicio	{.0, Mala}...	Ninguna	4	Centrado	Esc
8	X7	Numérico	5	1	X7 - Comercio electrónico (e-Commerce)	{.0, Mala}...	Ninguna	4	Centrado	Esc
9	X8	Numérico	5	1	X8 - Soporte técnico	{.0, Mala}...	Ninguna	4	Centrado	Esc
10	X9	Numérico	5	1	X9 - Respuesta a quejas	{.0, Mala}...	Ninguna	4	Centrado	Esc
11	X10	Numérico	5	1	X10 - Publicidad	{.0, Mala}...	Ninguna	4	Centrado	Esc
12	X11	Numérico	5	1	X11 - Línea de servicios	{.0, Mala}...	Ninguna	4	Centrado	Esc
13	X12	Numérico	5	1	X12 - Imagen de la fuerza de ventas	{.0, Mala}...	Ninguna	4	Centrado	Esc
14	X13	Numérico	5	1	X13 - Precio competitivo	{.0, Mala}...	Ninguna	4	Centrado	Esc
15	X14	Numérico	5	1	X14 - Garantías	{.0, Mala}...	Ninguna	4	Centrado	Esc
16	X15	Numérico	5	1	X15 - Nuevos productos y servicios	{.0, Mala}...	Ninguna	4	Centrado	Esc
17	X16	Numérico	5	1	X16 - Ordenes y facturación	{.0, Mala}...	Ninguna	4	Centrado	Esc
18	X17	Numérico	5	1	X17 - Flexibilidad de precios	{.0, Mala}...	Ninguna	4	Centrado	Esc
19	X18	Numérico	5	1	X18 - Velocidad de entrega	{.0, Mala}...	Ninguna	4	Centrado	Esc

Fuente: SPSS 20 IBM

Figura 4.2 Visor de Datos de la base de datos BM_MKT_Digital.sav



	id	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
1	1	1 a 5 años	Software empresarial	Grande (500+)	Fuera de MEX/Norteamérica	Dire...	8.5	3.9	2.5	5.9	4.8	4.9	6.0
2	2	Más de 5 años	Software para juegos	PyME (0 to 499)	MEX/Norteamérica	Indir...	8.2	2.7	5.1	7.2	3.4	7.9	3.1
3	3	Más de 5 años	Software empresarial	Grande (500+)	Fuera de MEX/Norteamérica	Dire...	9.2	3.4	5.6	5.6	5.4	7.4	5.8
4	4	< a 1 año	Software para juegos	Grande (500+)	Fuera de MEX/Norteamérica	Indir...	6.4	3.3	7.0	3.7	4.7	4.7	4.5
5	5	1 a 5 años	Software empresarial	Grande (500+)	MEX/Norteamérica	Dire...	9.0	3.4	5.2	4.6	2.2	6.0	4.5
6	6	< a 1 año	Software para juegos	PyME (0 to 499)	Fuera de MEX/Norteamérica	Indir...	6.5	2.8	3.1	4.1	4.0	4.3	3.7
7	7	< a 1 año	Software para juegos	Grande (500+)	Fuera de MEX/Norteamérica	Indir...	6.9	3.7	5.0	2.6	2.1	2.3	5.4
8	8	1 a 5 años	Software empresarial	Grande (500+)	Fuera de MEX/Norteamérica	Indir...	6.2	3.3	3.9	4.8	4.6	3.6	5.1
9	9	1 a 5 años	Software para juegos	Grande (500+)	Fuera de MEX/Norteamérica	Indir...	5.8	3.6	5.1	6.7	3.7	5.9	5.8
10	10	< a 1 año	Software empresarial	Grande (500+)	Fuera de MEX/Norteamérica	Indir...	6.4	4.5	5.1	6.1	4.7	5.7	5.7
11	11	Más de 5 años	Software empresarial	Grande (500+)	MEX/Norteamérica	Dire...	8.7	3.2	4.6	4.8	2.7	6.8	4.6
12	12	< a 1 año	Software empresarial	Grande (500+)	Fuera de MEX/Norteamérica	Indir...	6.1	4.9	6.3	3.9	4.4	3.9	6.4
13	13	< a 1 año	Software para juegos	PyME (0 to 499)	MEX/Norteamérica	Dire...	9.5	5.6	4.6	6.9	5.0	6.9	6.6
14	14	Más de 5 años	Software para juegos	PyME (0 to 499)	MEX/Norteamérica	Dire...	9.2	3.9	5.7	5.5	2.4	8.4	4.8
15	15	1 a 5 años	Software empresarial	Grande (500+)	Fuera de MEX/Norteamérica	Dire...	6.3	4.5	4.7	6.9	4.5	6.8	5.9
16	16	Más de 5 años	Software empresarial	PyME (0 to 499)	MEX/Norteamérica	Indir...	8.7	3.2	4.0	6.8	3.2	7.8	3.8
17	17	1 a 5 años	Software para juegos	PyME (0 to 499)	Fuera de MEX/Norteamérica	Dire...	5.7	4.0	6.7	6.0	3.3	5.5	5.1
18	18	1 a 5 años	Software empresarial	Grande (500+)	Fuera de MEX/Norteamérica	Indir...	5.9	4.1	5.5	7.2	3.5	6.4	5.5

Fuente: SPSS 20 IBM

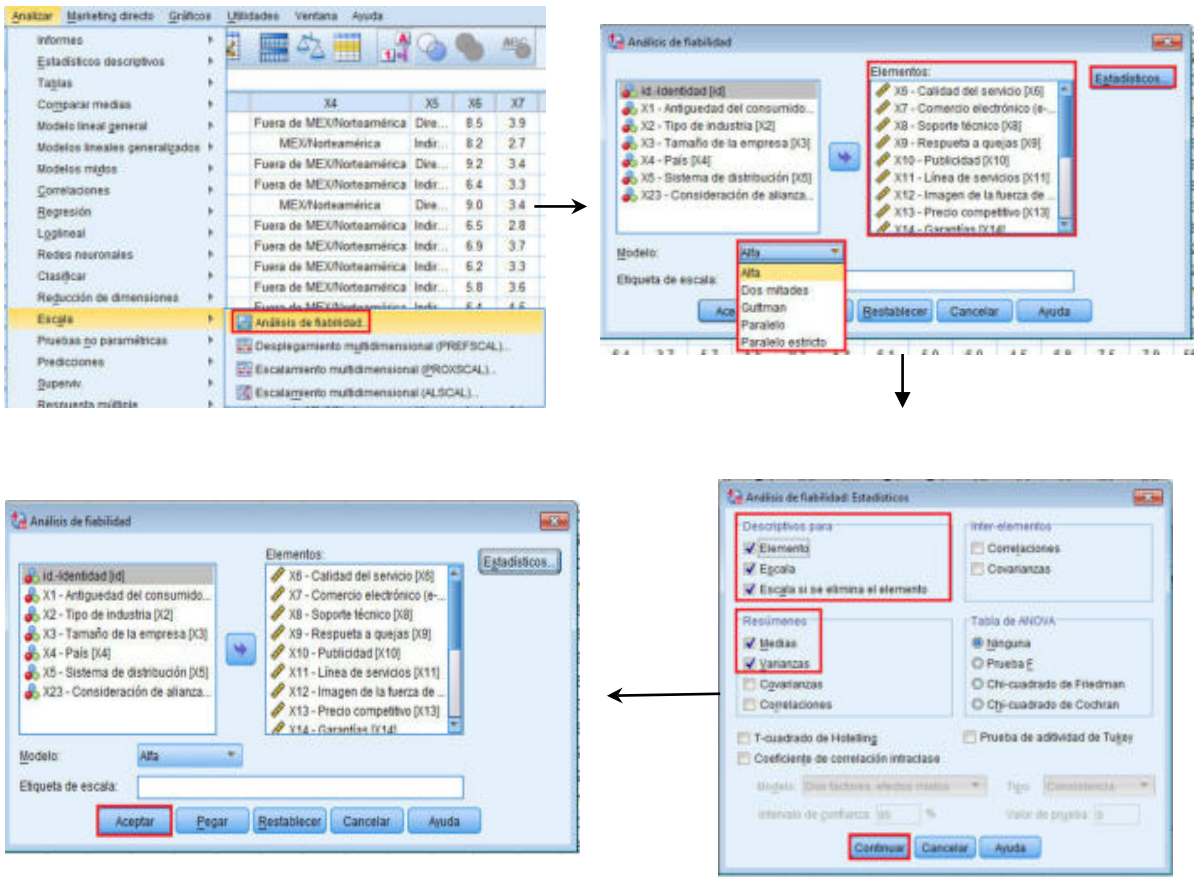
Paso 2: Diseño y Paso 3: Condiciones de aplicabilidad

- Ampliamente vistos en **Capítulo 6**.

Paso 4: Estimación y ajuste

Teclear: Analizar->Escala->Análisis de fiabilidad->Modelo: Alfa->Elementos: todas las variables métricas (X_6 a X_{22} en nuestro ejemplo) ->Estadísticos; Descriptivos para: Elementos; Escala; Escala si se elimina el elemento->Resúmenes: Medias; Varianzas->Continuar->Aceptar. Ver Figura 4.3

Figura 4.3 Proceso para calcular Alfa de *Cronbach*



Fuente: SPSS 20 IBM

- Modelos alternativos de análisis de confiabilidad se pueden encontrar en el menú desplegable **Modelo**. Sin embargo, se recomienda el uso del **Alfa de *Cronbach*** debido a su flexibilidad para aplicarlo tanto a las **respuestas dicotómicas/ binarias** como a los datos medidos en una **escala** más grande. **Alpha de *Cronbach*** nos dará un cálculo de confiabilidad basado en la **totalidad del cuestionario.**
- Otro tipo de confiabilidad, que puede que desee evaluar, es la confiabilidad de **Dos mitades (ampliamente visto en los capítulos anteriores)**. Aquí las preguntas se

están respondiendo de manera diferente en las dos partes de la cuestionario. Tenga en cuenta que la opción por defecto para la confiabilidad de **Dos mitades** en **SPS**S es agrupar por la primera mitad las primeras preguntas y luego realizar lo mismo por las preguntas finales.

Paso 5: Interpretación

- La primera tabla generada por **SPSS**, es la de **Resumen del procesamiento de los casos,** la cual confirma la cantidad de elementos que son incluidos en el análisis así como los excluidos. Ver **Figura 4.4**

Figura 4.4. Tabla Resumen del procesamiento de los casos

Resumen del procesamiento de los casos

		N	%
Casos	Válidos	200	100.0
	Excluidos[a]	0	.0
	Total	200	100.0

a. Eliminación por lista basada en todas las variables del procedimiento.

Fuente: SPSS 20 IBM

- La segunda tabla generada por **SPSS**, es la **Estadísticos de fiabilidad**. En nuestro caso el coeficiente de fiabilidad para los **17 elementos** se muestra como un **Alfa de *Cronbach***, **basada en los elementos tipificados** (estandarizada); ambos valores suelen ser muy similares **y suele el elegirse éste último (el estandarizado).**
- Una puntuación Alfa por encima de **0.75** se toma generalmente para indicar una escala de alta de confiabilidad, **generalmente se acepta que 0.5 a 0.75** que indica una escala **moderadamente confiable**, mientras que una cifra inferior indica generalmente una escala de **baja fiabilidad. En nuestro caso 0.842 es considerado de alta confiabilidad.**
- Normalmente se elige el **alfa estandarizado**. De hecho, **si bien es preferible que nuestras escalas sean similares** (por ejemplo, todas las preguntas se midan en una escala de 10 puntos), puede también que lo sea el **Alfa de *Cronbach*** simple**.** Ver **Figura 4.5.**

Figura 4.5. Estadísticos de fiabilidad

Estadísticos de fiabilidad

Alfa de Cronbach	Alfa de Cronbach basada en los elementos tipificados	N de elementos
.669	.842	17

Fuente: SPSS 20 IBM

- La siguiente tabla generada por **SPSS**, es la de **Estadísticos de los elementos.**
- La primera parte reporta un resumen de las respuestas de los participantes a las preguntas individuales, y proporciona información sobre la **Media** y la **Desviación típica** (desviación estándar) para las respuestas a cada pregunta, y un informe que indica cuántos (**N**) participantes completó la pregunta.
- De nuestro ejemplo podemos ver que todos los **200** de participantes respondieron a todas nuestras preguntas. También podemos ver qué preguntas suscitan una amplia variedad de respuestas, a través de las desviaciones estándar más grandes.
- En el ejemplo, las respuestas se midieron en una escala de **1 a 10**. Altas puntuaciones de la **Media**, por lo tanto, indican preguntas donde los participantes estuvieron finalmente de acuerdo en la escala de puntuación. Ver **Figura 4.6**

Figura 4.6. Tabla Estadísticos de los elementos

Estadísticos de los elementos

	Media	Desviación típica	N
X6 - Calidad del servicio	7.894	1.3830	200
X7 - Comercio electrónico (e-Commerce)	3.765	.7689	200
X8 - Soporte técnico	5.243	1.6552	200
X9 - Respueta a quejas	5.368	1.2100	200
X10 - Publicidad	4.061	1.1471	200
X11 - Línea de servicios	5.815	1.3174	200
X12 - Imagen de la fuerza de ventas	5.248	1.1286	200
X13 - Precio competitivo	6.971	1.5813	200
X14 - Garantías	6.048	.8753	200
X15 - Nuevos productos y servicios	5.211	1.4960	200
X16 - Ordenes y facturación	4.242	.9119	200
X17 - Flexibilidad de precios	4.464	1.1927	200
X18 - Velocidad de entrega	3.816	.7494	200
X19 - Satisfacción	6.952	1.2411	200
X20 - Probabilidad de recomendación	6.952	1.0829	200
X21 - Probabilidad de compra	7.665	.8932	200
X22 - Nivel de compra	58.200	8.9662	200

Fuente: SPSS 20 IBM

- Otra tabla generada por **SPSS**, es la de **Estadísticos de resumen de los elementos.**
- La fila **Medias de los elementos** detalla las estadísticas descriptivas de una respuesta en las preguntas individuales. Como podemos ver en el ejemplo anterior, la puntuación **Media** de los ítems es **8.701**, tal como se esperaría al promediar las puntuaciones en una gama de elementos empleando una escala de **1-10.**
- Los valores **Mínimo** y **Máximo** son las dos puntuaciones más extremas seleccionadas Por los participantes. En nuestro caso se trata de **3.765** y **58.2**, lo que indica que no hay encuestados más extremos de la escala.
- La columna **Varianza de los elementos** muestra la varianza en puntaje cuando se observan las puntuaciones en el ítem individual. Ver **Figura 4.7**

Figura 4.7. Estadísticos de resumen de los elementos

Estadísticos de resumen de los elementos

	Media	Mínimo	Máximo	Rango	Máximo/mínimo	Varianza	N de elementos
Medias de los elementos	8.701	3.765	58.200	54.435	15.458	164.445	17
Varianzas de los elementos	6.074	.562	80.392	79.830	143.160	367.187	17

Fuente: SPSS 20 IBM

- La siguiente tabla producida por **SPSS** permite examinar la confiabilidad de cada pregunta, y el efecto en el cuestionario general, **si se eliminara dicha pregunta individual**
- El resultado de **SPSS** muestra las conclusiones del análisis para cada ítem del cuestionario.
- En nuestro ejemplo, las preguntas que son potencialmente preocupantes podrían señalarse
- La columna **Correlación elemento-total corregida** muestra la relación entre las respuestas sobre preguntas individuales y la puntuación total en el cuestionario. Nosotros esperamos que una pregunta confiable tuviera una relación positiva con el total general, **idealmente por encima de 0.3.** Un elemento que muestre **una relación positiva débil o una relación negativa con el total e indica una pregunta que puede ser pobre en confiabilidad y por lo tanto afecta a las Conclusiones de toda la escala.** En el ejemplo anterior, podemos ver que la pregunta $\mathbf{X_{15}}$ es la más débil, ya que su correlación con el total global es de sólo **0.142**.
- Los efectos que pueden tener las preguntas individuales sobre la confiabilidad general del cuestionario se destacan por la **relación inversa** entre la **Correlación elemento-total corregida vs Alfa de *Cronbach* si se elimina el elemento.** La importancia de la relación débil entre la pregunta **15** y la puntuación global total en el cuestionario se refleja en el aumento de la puntuación alfa para el cuestionario **SI este elemento se omite.** Un examen de la tabla más abajo de la salida nos da un Valor alfa de **0.669 (Ver Figura 10.5)**. Si bien esta cifra es alta, la eliminación de la pregunta 1 de la El cuestionario final vería esta cifra subir a **0.649** (como puede verse en la tabla). Ver **Figura 4.8**

Figura 4.8. Estadísticos de resumen de los elementos

	Media de la escala si se elimina el elemento	Varianza de la escala si se elimina el elemento	Correlación elemento-total corregida	Correlación múltiple al cuadrado	Alfa de Cronbach si se elimina el elemento
X6 - Calidad del servicio	140.022	259.017	.395	.757	.649
X7 - Comercio electrónico (e-Commerce)	144.151	268.707	.365	.660	.659
X8 - Soporte técnico	142.673	265.796	.185	.740	.663
X9 - Respueta a quejas	142.548	250.078	.705	.795	.632
X10 - Publicidad	143.854	264.557	.339	.464	.656
X11 - Línea de servicios	142.101	251.000	.617	.974	.635
X12 - Imagen de la fuerza de ventas	142.668	263.051	.388	.815	.653
X13 - Precio competitivo	140.945	286.436	-.195	.416	.691
X14 - Garantías	141.867	269.580	.284	.748	.661
X15 - Nuevos productos y servicios	142.705	269.278	.142	.100	.667
X16 - Ordenes y facturación	143.674	259.279	.627	.648	.645
X17 - Flexibilidad de precios	143.452	271.275	.148	.970	.666
X18 - Velocidad de entrega	144.100	259.004	.785	.980	.644
X19 - Satisfacción	140.964	245.320	.814	.856	.624
X20 - Probabilidad de recomendación	140.963	254.845	.651	.639	.639
X21 - Probabilidad de compra	140.251	260.510	.597	.581	.647
X22 - Nivel de compra	89.715	83.189	.703	.780	.774

Fuente: SPSS 20 IBM

- La última tabla que el SPSS genera es la de **Estadísticos de la escala**
- **N of elementos** es el número de ítems en nuestro cuestionario En nuestro caso **17** indicadores,
- La estadística en los renglones nos reporta la parte de la estadística descriptiva para la escala como un todo. En el ejemplo, cuando el total del puntaje del cuestionario es analizado, la puntuación **Media** de los participantes es de **147.916** con una **varianza** de **278.511**, y una Desviación típica, de **16.6886** Una desviación estándar pequeña nos indica que no hay amplias variaciones en los puntajes de los participantes para todo el modelo global de puntaciones de nuestro cuestionario. Ver **Figura 4.9**

Figura 4.9. Estadísticos de la escala

Estadísticos de la escala

Media	Varianza	Desviación típica	N de elementos
147.916	278.511	16.6886	17

Fuente: SPSS 20 IBM

Referencias

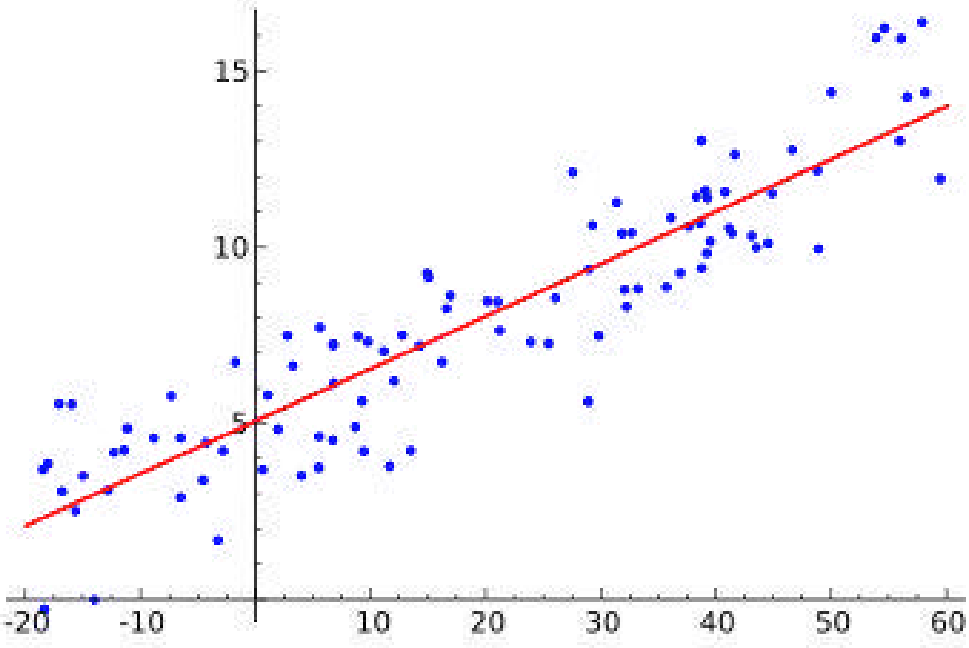
Hinton, P. R; Bfownlow, C., McMurray, I., y Cozens, B. (2004) *spss explanmed*. Routledge Taylor and Francis group. London, New York.

IBM (2011a). *IBM SPSS Statistics Base 20.* EUA. .Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Base.pdf

IBM (2011b).*Guía breve de IBM SPSS Statistics 20.* EUA.Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Brief_Guide.pdf

IBM (2011c). IBM SPSS Missing Values 20. EUA. .Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Missing_Values.pdf

Capítulo 5. Correlación y Regresión Lineal Simple y Múltiple



5.1. Correlación: ¿qué es?

Existirán ocasiones en que se requiera recoger una puntuación en dos variables de un conjunto de participantes, estaremos interesados en verificar si existe una relación entre las variables. Por ejemplo, medir la experiencia gerencial de una persona en un trabajo (número de años) y la productividad semestral para preguntarse si hay una relación entre la experiencia y la productividad. En caso de tener dos variables como éstas, generadas por los mismos participantes, es que es posible examinar la asociación entre las variables mediante una **correlación**. Se realiza una **correlación** para probar el grado en que las puntuaciones de las dos variables varían entre sí. Esto es, el grado en el cual la variación de los puntajes de una variable resulta en una correspondiente variación de los puntajes de variación de la segunda variable. La más simple, es la lineal, llamada también **regresión lineal** y que de la que es muy difícil encontrar exactamente una recta, por lo que se deberá incluir el término de **error,** expresándose: **y=a+bx**, donde **y, x**, son las variables, con **a como intercepción (constante**) y **b, la pendiente** de la recta. **Podemos calcular la línea de regresión encontrando la ecuación que nos de la menor cantidad de error, que implica:**

1. **Una correlación fuerte** indica que hay sólo una pequeña cantidad de error y los puntos se encuentran cerca de la línea de regresión;
2. **Una correlación débil** indica que hay una gran cantidad de error y los puntos están más dispersos. En el segundo caso, es probable que concluyamos que **una relación lineal no es un buen modelo para nuestros datos.**

3. Valores altos de una variable asociada con valores altos de la segunda variable indican que la **correlación es positiva.** Por ejemplo, podríamos encontrar una **correlación positiva** entre altura personal y tamaño del pie, con personas más altas que tienen pies más grandes y personas más pies más pequeños. Cuando valores altos de la primera variable se asocian con valores bajos de la segunda variable entonces nos referimos a esto como una **correlación negativa**. Por lo tanto, para un coche viajando a una velocidad constante a lo largo de una pista, encontraremos que la distancia recorrida está de forma **negativamente correlacionada** con la gasolina que queda en el tanque
4. Las estadísticas de correlación más conocidas son: ***Pearson, Spearman, Kendall tau-b*** y todos producen una estadística que va desde **-1**, lo que indica una perfecta correlación negativa, a **+1** indicando una correlación positiva perfecta. **Un valor cero no indica correlación.**

5.2. Correlación de *Pearson*: ¿qué es?

El coeficiente de correlación de *Pearson* (*r*) también se conoce, como la correlación del producto momento de ***Pearson***. Esencialmente, calcula cuánto puntaje de dos variables varían juntas (conocido como **"*producto*"**) para luego contrastarlas en cuanto varían entre sí mismas. La **variabilidad conjunta** se conoce como la **sumas de los productos** y será tanto mayor cuando altos valores de una variable corresponden a valores altos de la segunda variable. Será de un **valor negativo** cuando la correlación es **negativa**. Si la **variabilidad conjunta** coincide con la variación individual en las puntuaciones, entonces estos **valores serán iguales**, por lo que uno dividido por el otro resultará en

***r* = 1 (o -1 si las sumas de productos son negativas).**

Sí No existe variabilidad conjunta las puntuaciones no se correlacionan en absoluto y **r será cero**. Al igual que otros métodos de análisis paramétrico la **correlación de *Pearson*** se basa en una serie de supuestos:

1. La relación entre las variables es lineal.
2. Los puntos deben estar uniformemente distribuidos a lo largo de la línea recta. Esta es el supuesto de **Homocedasticidad**. Si los datos tienen los puntos desigualmente distribuidos a lo largo de la línea recta (o hay un punto o dos puntos periféricos) entonces **la correlación de *Pearson* no es una medida exacta de la asociación.**
3. Los datos se extraen de poblaciones **normalmente distribuidas**.
4. Los datos recopilados deben ser **intervalos o razones,** a partir de **distribuciones continuas.**
5. El punto importante a **recordar** es que estamos considerando una correlación lineal aquí, por lo que **nuestra suposición clave es que los puntos siguen una línea recta si están correlacionados**.
6. Si nosotros creemos que la relación entre las variables **no es lineal**, **entonces no usamos el *Pearson* sino que usan el *Spearman* o *Kendall tau-b.***
7. Se destaca que en una **predicción** se espera una correlación positiva y que esto indica una dirección. Nuestra predicción, por lo tanto, es de una cola. Si no supiéramos cual es la dirección de nuestra relación entre las dos variables, considere realizar una predicción de dos colas, y así, estaríamos buscando un positivo o un correlación negativa

5.3. Correlación de *Pearson*: Ejemplo

Paso 1: Objetivos

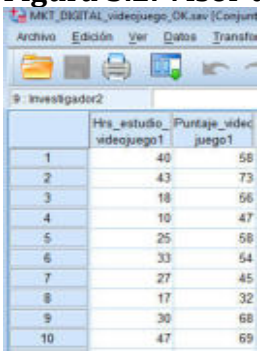
-Problema 1: La empresa **MKT Digital** postuló, mediante uno de sus investigadores que los jugadores de sus videojuegos, que mostraran mayor disposición a estudiar sus estrategias lograrían calificaciones más altas en el videojuego, que aquellos que hicieran menos estudios. El investigador anotó los resultados de diez jugadores, mostrando cuánto tiempo (en horas) pasaron estudiando (en promedio por semana durante un mes) y sus calificaciones de 0-100 obtenidas. Ver **Figura 5.1 y 5.2**

Figura 5.1. Visor de Variables base de datos MKT_DIGITAL_Videojuego.sav

	Nombre	Tipo	An...	Decimales	Etiqueta	Valores	Perdidos	Columnas
21	Errores_simulador_vuelo	Numérico	2	0	Errores por persona del simulador	Ninguna	Ninguna	8
22	AM_Diurno	Numérico	2	0	Horario Diurno	Ninguna	Ninguna	8
23	PM_Vepsertino	Numérico	2	0	Horario Vespertino	Ninguna	Ninguna	8
24	Campus	Numérico	8	0	Campus 1 o 2	{1, Campus Norte...	Ninguna	10
25	Evaluación	Numérico	2	0	Puntaje	Ninguna	Ninguna	8
26	Programa1	Numérico	2	0	Programa1	Ninguna	Ninguna	8
27	Programa2	Numérico	2	0	Programa2	Ninguna	Ninguna	8
28	Hrs_estudio_videojuego1	Numérico	2	0	Hrs estudio videojuego 1	Ninguna	Ninguna	8
29	Puntaje_videojuego1	Numérico	2	0	Puntaje Videojuego1	Ninguna	Ninguna	8

Fuente: SPSS 20 IBM

Figura 5.2. Visor de Datos base de datos MKT_DIGITAL_Videojuego.sav



	Hrs_estudio_videojuego1	Puntaje_videojuego1
1	40	58
2	43	73
3	18	56
4	10	47
5	25	58
6	33	54
7	27	45
8	17	32
9	30	68
10	47	69

Fuente: SPSS 20 IBM

Paso 2: Diseño

- En adelante para efectos de aprendizaje, sólo se tomarán de forma directa los datos con el fin de aprender a utilizar la técnica.
- Seleccione si su predicción es de una cola o de dos colas. En nuestro caso es de una cola como lo afirmamos de ser una correlación positiva.

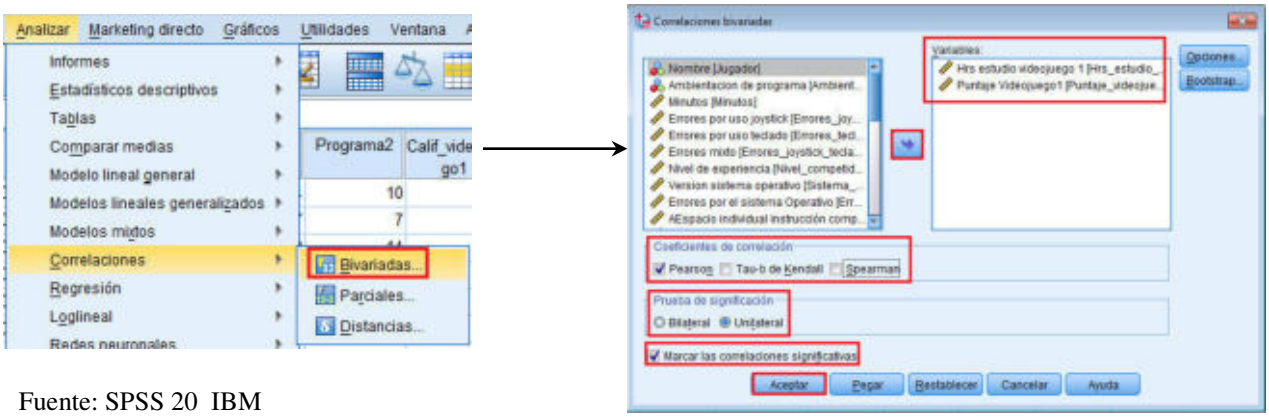
Paso 3: Condiciones de Aplicabilidad

- No olvidar realizar los análisis previos de inspección visual de la base de datos para detectar ausentes y/o atípicos (corregir en su defecto), confiabilidad, normalidad, homocedasticidad y linealidad.

Paso 4: Estimación y ajuste

-Teclear: Analizar->Correlaciones->Bivariadas->Variables->Hrs estudio videojuego1->Puntaje videojuego1->Coeficientes de correlación: *Pearson*->Prueba de significación: Unilateral->Marcar las correlaciones significativas. Ver Figura 5.3

Figura 5.3. Proceso para calcular la regresión lineal simple



Fuente: SPSS 20 IBM

- El cuadro **Prueba de significación** se señala como predeterminado: **Correlaciones significativas,** se resaltan debajo de la tabla de salida con un * para un significado de ***p* <0.05 y ** para *p* <0.01.**
- Si necesita las medias y desviaciones stándard para cada variable, haga ***click*** en **Opciones** y selecciónelas, luego Continuar y Aceptar.

Paso 5: Interpretación

- **SPSS** produce una tabla de resultados, denominada **Correlaciones**, a menos que haya seleccionado estadísticas descriptivas. Ver **Figura 5.4**

Figura 5.4.- Tabla de Correlaciones

Correlaciones

[Conjunto_de_datos1] C:\Users\Juan\Desktop\proy libro mc\MKT_DIGITAL_videojuego

Correlaciones

		Hrs estudio videojuego 1	Puntaje Videojuego1	
Hrs estudio videojuego 1	Correlación de Pearson	1	.721**	**Prueba estadística**
	Sig. (unilateral)		.009	***p* Valor**
	N	10	10	
Puntaje Videojuego1	Correlación de Pearson	.721**	1	
	Sig. (unilateral)	.009		
	N	10	10	

**. La correlación es significante al nivel 0,01 (unilateral).

Fuente: SPSS 20 IBM

- La Correlación de ***Pearson*** = **0.721**. SPSS indica con ** que es **significativa en el nivel de 0.01 para una predicción de una cola**. El valor de **p real se muestra para ser 0.009**.
- Una forma convencional de informar estas cifras sería la siguiente:

$$r = 0.721, N = 10, p < 0.01$$

- **Conclusión:** estos resultados indican que a **medida que aumenta el tiempo de estudio, el rendimiento aumenta**, lo **que es una correlación positiva.**
- Como el valor **r** reportado es positivo y $p < 0.01$, podemos afirmar que **tenemos una correlación positiva** entre nuestras dos variables y nuestra **hipótesis nula puede ser rechazada**. Si el **valor de r** fuera negativo, esto indicaría una **correlación negativa y estaría en contra de nuestra hipótesis.**
- En la tabla, se muestra la **Correlación de *Pearson*** también muestra el **valor r** cuando **Calificaciones videojuego1** es correlacionado con sí mismo, y hay un coeficiente de correlación perfecto de **1.000**. Similar, es el caso de **Calificaciones videojuego2** una correlación perfecta con sí mismo, $r = 1.000$. Estos valores son por lo **tanto no son necesarios.**
- **Se recomienda reportar gráfico de dispersión de la regresión lineal.**

5.4. Correlación de *Spearman*: ¿Qué es?

Habrá ocasiones en que desearemos correlacionar datos:

1. Cuando hay datos ordinales (una o ambas variables, no se miden como una escala de intervalo),
2. Cuando los datos no se distribuyen normalmente, o
3. Cuando Otras suposiciones de la **correlación de *Pearson*** son violadas.

Y es en estas ocasiones cuando es pertinente usar el **coeficiente de correlación de *Spearman***, que es el equivalente **no paramétrico** de la **correlación de *Pearson***. La **correlación de *Spearman*** utiliza exactamente los mismos cálculos que el de ***Pearson***, pero realiza el análisis con las **calificaciones** de las puntuaciones en lugar de en los **valores** de los datos reales. El **coeficiente de correlación de *Spearman*** se conoce como r_s. Como se usan las calificaciones en lugar de las puntuaciones reales, la **correlación de *Spearman*** se puede utilizar incluso cuando la relación entre las dos variables es **no lineal**. La **correlación de *Spearman*** puede hacer frente a unas cuantas calificaciones muy vinculadas en los datos sin necesidad preocuparse por el efecto en r_s. Sin embargo, cuando hay muchos valores vinculados, el resultado es que hará r_s más grande de lo que debería ser. En este caso, el ***Kendall tau-b*** puede utilizarse en su lugar.

5.5. Correlación de *Spearman*: Ejemplo

Paso 1: establecimiento de objetivos

-Problema 2 :A dos investigadores de la empresa **MKT Digital**, se les solicitó evaluaran en **8 jugadores** con la pregunta: "*qué tan probable es que adquirieran el videojuego probado*", en una escala de **0** (improbable) a **20** (altamente probable). Se pensó que habría cierto nivel de correlación significativa entre la evaluación delos investigadores. **La predicción no establece si esperamos un resultado de correlación positiva o negativa, por lo tanto,**

se tiene una predicción de dos colas. Si fuera el caso de predecir una correlación positiva o negativa, entonces tendríamos una predicción de una cola. Ver 5.5 y 5.6

Figura 5.5. Visor de Variables base de datos MKT_DIGITAL_Videojuego.sav

	Nombre	Tipo	An...	Decimales	Etiqueta	Valores	Perdidos	Columnas
21	Errores_simulador_vuelo	Numérico	2	0	Errores por persona del simulador	Ninguna	Ninguna	8
22	AM_Diurno	Numérico	2	0	Horario Diurno	Ninguna	Ninguna	8
23	PM_Vepsertino	Numérico	2	0	Horario Vespertino	Ninguna	Ninguna	8
24	Campus	Numérico	8	0	Campus 1 o 2	{1, Campus Norte...	Ninguna	10
25	Evaluación	Numérico	2	0	Puntaje	Ninguna	Ninguna	8
26	Programa1	Numérico	2	0	Programa1	Ninguna	Ninguna	8
27	Programa2	Numérico	2	0	Programa2	Ninguna	Ninguna	8
28	Hrs_estudio_videojuego1	Numérico	2	0	Hrs estudio videojuego 1	Ninguna	Ninguna	8
29	Puntaje_videojuego1	Numérico	2	0	Puntaje Videojuego1	Ninguna	Ninguna	8
30	Investigador1	Numérico	2	0	Calif.Investigador1	Ninguna	Ninguna	8
31	Investigador2	Numérico	2	0	Calif.Investigador2	Ninguna	Ninguna	8

Fuente: SPSS 20 IBM

Figura 5.6. Visor de Datos base de datos MKT_DIGITAL_Videojuego.sav

	Investigador1	Investigador2
1	15	8
2	13	12
3	18	4
4	11	9
5	14	16
6	16	7
7	8	16
8	12	9

Fuente: SPSS 20 IBM

Paso 2: Diseño

- En adelante para efectos de aprendizaje, sólo se tomarán de forma directa los datos con el fin de aprender a utilizar la técnica.
- A partir del comando **Analizar**, en la opción **Correlación**, **Bivariado** (relacionar las variables: investigador1, investigador2), **Coeficientes de correlación, seleccionar como opción *Spearman***, con **Prueba de significancia a Dos colas.**

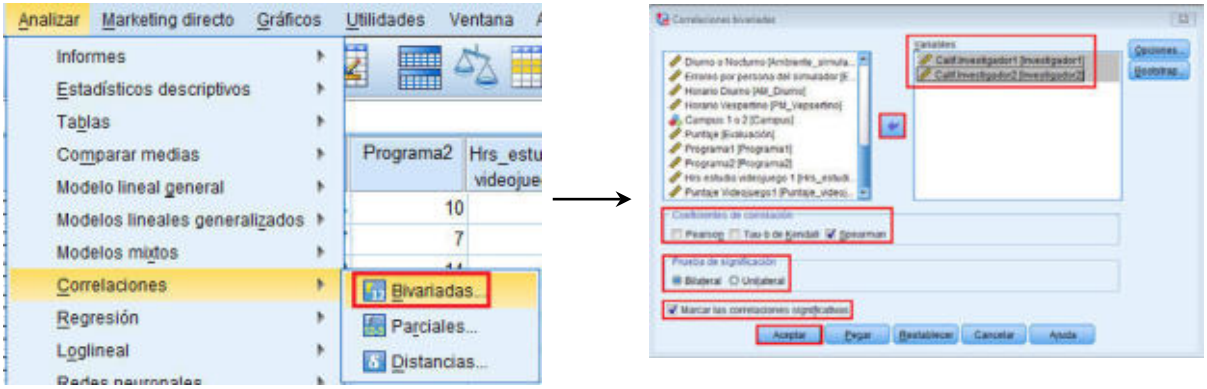
Paso 3: Condiciones de Aplicabilidad

- No olvidar realizar los análisis previos de inspección visual de la base de datos para detectar ausentes y/o atípicos (corregir en su defecto), confiabilidad, normalidad, homocedasticidad y linealidad.

Paso 4: Estimación y ajuste

- **Teclear: Analizar->Correlaciones->Bivariadas->Variables: Calif.Investigador1; Calif.Investigador2->Coeficiente de correlación: *Spearman*->Prueba de significación: Bilateral->Marcar las correlaciones significativas->Aceptar. Ver Figura 5.7.**

Figura 5.7. Proceso para calcular la regresión lineal simple con correlación de *Spearman*



Fuente: SPSS 20 IBM

El cuadro de correlaciones significativas está seleccionado como predeterminado. Las correlaciones significativas, son resaltadas con: *** para una significación de $p<0.05$ y ** para $p<0.01$**.

Paso 5: Interpretación

- A fin de comprobar si existe una **correlación significativa** entre las calificaciones de los dos investigadores, se debe observar la tabla de correlaciones. Ver **Figura 5.8**

Figura 5.8. Tabla de Correlaciones de *Spearman*

Correlaciones no paramétricas

```
[Conjunto_de_datos1] C:\Users\Juan\Desktop\proy libro mc\MKT_DIGITAL_videojuego_OK.sav
```

Correlaciones

			Calif. Investigador1	Calif. Investigador2	
Rho de Spearman	Calif.Investigador1	Coeficiente de correlación	1.000	-.735*	Prueba estadística
		Sig. (bilateral)	.	.038	*p* Valor
		N	8	8	
	Calif.Investigador2	Coeficiente de correlación	-.735*	1.000	
		Sig. (bilateral)	.038	.	
		N	8	8	

*. La correlación es significativa al nivel 0,05 (bilateral).

Fuente: SPSS 20 IBM

- **La prueba estadística correlación rho de *Spearman*** = **-0.735**. El signo denota una correlación negativa. **SPSS** también ilustra con * que es significativo en el nivel **0.05 para una predicción de dos colas.** El ***p* valor** real es **0.038**.
- Al observar la **correlación de *Spearman*** se puede ver que el **Calif. Investigador 1** está perfectamente correlacionada consigo misma, de ahí el **coeficiente de correlación rho de Spearman sea 1.000**. Es el mismo caso de **Calif. Investigador 2 con un coeficiente de correlación de rho de *Spearman* de 1.000.**
- La sintaxis de reporte, es r_s **= -0.735, N = 8, *p*<0.05**
- **Conclusión:** Estos resultados indican que a medida que las calificaciones de un investigador **aumentan** las calificaciones del otro investigador disminuyen. Por lo tanto, como la percepción de cada investigador, sobre el rendimiento de cada jugador es diferente, **estas calificaciones pueden no ser un buen indicador del desempeño real.**
- Se sugiere realizar un **diagrama de dispersión** ya que es generalmente, la estadística ilustrativa más apropiada para soportar una correlación, sin embargo, al realizar una **prueba de *Spearman*** debe utilizarse con precaución. **El coeficiente de correlación rho de *Spearman*** se produce utilizando las calificaciones de las puntuaciones en lugar de los datos reales directos, mientras que el diagrama de dispersión muestra las puntuaciones.

5.6. Correlación de *Kendall tau-b:* ¿Qué es?

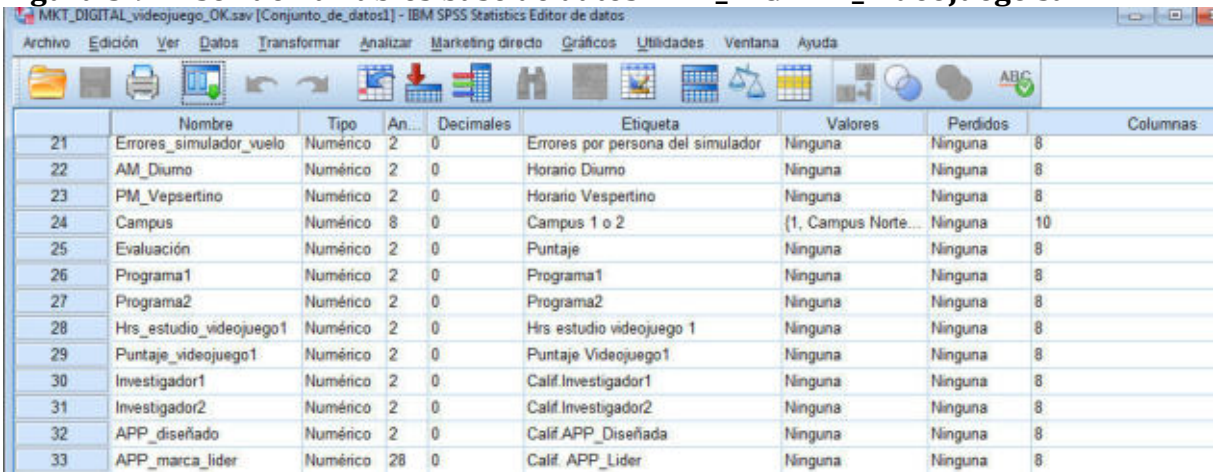
La correlación de ***Kendall tau-b*** es otro **coeficiente de correlación no paramétrico** y es un alternativa a la **correlación de *Spearman***. Es una medida de asociación entre dos variables ordinales y toma en cuenta calificaciones vinculadas, por lo que puede utilizarse para un conjunto de datos pequeños con un gran número de variables vinculadas (a diferencia de la prueba de ***Spearman***, la presencia de **calificaciones vinculadas** "***inflar***" artificialmente el valor de la estadística). Al igual que la prueba ***Spearman***, en el ***Kendall* tau-b** todas las puntuaciones se clasifican en cada variable. Sin embargo, opera en un principio diferente a las **correlaciones de *Pearson o Spearman***. La prueba de ***Kendall* tau-b** evalúa qué tan bien se encuentra el grado de clasificación de la segunda variable, respecto a la primera variable. Se ubican las calificaciones de la primera variable, a fin de colocar las calificaciones referidas a la segunda variable. Entonces, se puede observar cómo esto ordena las calificaciones de la segunda variable. En el ***Kendall* tau-b**, cada uno de los pares de calificaciones en la segunda variable es analizada. Cuando estos pares **coinciden** con el orden de las calificaciones de la primera variable, entonces el **par es concordante** y cuando el orden es invertido el **par es discordante**. **Se calcula la diferencia en el número de pares concordantes y discordantes.** Este valor se compara con el valor cuando cada par es concordante para producir ***Kendall tau-b*.** Al igual que otros coeficientes de correlación, ***Kendall tau-b*** oscila entre **-1 y +1.** Sin embargo, debido a que los métodos utilizados son diferentes, el ***Kendall* *Tau-b* producirá un valor diferente al de *Spearman,* pero las probabilidades serán muy similares**.

5.7. Correlación de *Kendall tau-b*: Ejemplo

Paso 1: Objetivo

-Problema 3: La empresa **MKT Digital** realiza una prueba de consumo piloteando la salida de una nueva aplicación de software respecto de uno previo ya existente. Decidieron preguntar si la nueva aplicación era tan satisfactoria como la aplicación de la marca líder actual. Dieron a 20 personas la aplicación actual de dicha marca líder (marca A) y también les la nueva aplicación diseñada (marca B).El orden en el que los participantes probaron cada aplicación por marca difería para contrarrestar efectos potenciales a sesgar el punto de vista de cada uno de los 20 usuarios. A cada participante se le pidió que calificara la satisfacción de uso de la aplicación en escala de 1 a 10, siendo, (1) definitivamente nada satisfactoria y (10) definitivamente muy satisfactorio. Ver **Figura 5.9 y 5.10**

Figura 5.9. Visor de Variables base de datos MKT_DIGITAL_Videojuego.sav



	Nombre	Tipo	An...	Decimales	Etiqueta	Valores	Perdidos	Columnas
21	Errores_simulador_vuelo	Numérico	2	0	Errores por persona del simulador	Ninguna	Ninguna	8
22	AM_Diumo	Numérico	2	0	Horario Diumo	Ninguna	Ninguna	8
23	PM_Vepsertino	Numérico	2	0	Horario Vespertino	Ninguna	Ninguna	8
24	Campus	Numérico	8	0	Campus 1 o 2	{1, Campus Norte...	Ninguna	10
25	Evaluación	Numérico	2	0	Puntaje	Ninguna	Ninguna	8
26	Programa1	Numérico	2	0	Programa1	Ninguna	Ninguna	8
27	Programa2	Numérico	2	0	Programa2	Ninguna	Ninguna	8
28	Hrs_estudio_videojuego1	Numérico	2	0	Hrs estudio videojuego 1	Ninguna	Ninguna	8
29	Puntaje_videojuego1	Numérico	2	0	Puntaje Videojuego1	Ninguna	Ninguna	8
30	Investigador1	Numérico	2	0	Calif.Investigador1	Ninguna	Ninguna	8
31	Investigador2	Numérico	2	0	Calif.Investigador2	Ninguna	Ninguna	8
32	APP_diseñado	Numérico	2	0	Calif.APP_Diseñada	Ninguna	Ninguna	8
33	APP_marca_lider	Numérico	28	0	Calif. APP_Lider	Ninguna	Ninguna	8

Fuente: SPSS 20 IBM

Figura 5.10. Visor de Datos base de datos MKT_DIGITAL_Videojuego.sav

	APP_diseñ...	APP_marc...
1	6	6
2	8	9
3	8	8
4	7	10
5	6	8
6	8	8
7	6	9
8	5	6
9	4	8
10	6	6
11	8	7
12	6	6
13	6	6
14	5	8
15	7	8
16	5	5
17	7	8
18	4	4
19	4	8
20	5	5

Fuente: SPSS 20 IBM

Paso 2: Diseño

- En adelante para efectos de aprendizaje, sólo se tomarán de forma directa los datos con el fin de aprender a utilizar la técnica.
- Se selecciona el comando Analizar, Correlacionar, Bivariado, ***Kendall tau-b***, una cola,
- **Como nuestra predicción indica que esperamos una correlación positiva esto indica una dirección. Nuestra predicción es, por lo tanto, de una cola. Si no supiéramos la dirección que tendría la relación entre nuestras dos variables, tendría una predicción de dos colas. En ese caso estaríamos buscando una correlación positiva o negativa.**
- Como estamos nos interesa el número de calificaciones vinculadas en nuestro conjunto de datos, vamos a llevar a cabo un ***Kendall tau-b*** en lugar de una **correlación de *Spearman.***

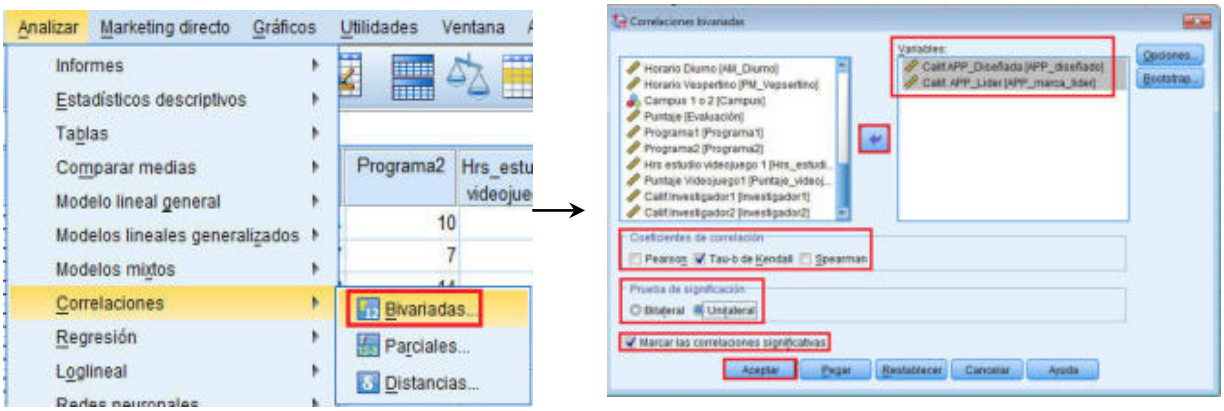
Paso 3: Condiciones de Aplicabilidad

- No olvidar realizar los análisis previos de inspección visual de la base de datos para detectar ausentes y/o atípicos (corregir en su defecto), confiabilidad, normalidad, homocedasticidad y linealidad.

Paso 4: Estimación y ajuste

Teclear: Analizar->Correlaciones->Bivariadas->Variables: Calif APP_Diseñada; Calif APP_Lider->Coeficientes de correlación: *Tau-b de Kendall*->Pruebas de significación->Unilateral->Marcar las correlaciones significativas. Ver Figura 5.11

Figura 5.11. Proceso para calcular a correlación de *Spearman*



Fuente: SPSS 20 IBM

El cuadro de correlaciones significativas está seleccionado como predeterminado. Las correlaciones significativas, son resaltadas con: * para una significación de $p<0.05$ y ** para $p<0.01$.

Paso 5: Interpretación
SPSS genera la tabla Correlaciones de *Kendall-tau-b*. Ver Figura 5.12

Figura 5.12. Tabla Correlaciones

Correlaciones no paramétricas

[Conjunto_de_datos1] C:\Users\Juan\Desktop\proy libro mc\MRT_DIGITAL_videojuego_OK.sav

Correlaciones

			Calif. APP_Diseñada	Calif. APP_Lider
Tau_b de Kendall	Calif.APP_Diseñada	Coeficiente de correlación	1.000	.397*
		Sig. (unilateral)	.	.017
		N	20	20
	Calif. APP_Lider	Coeficiente de correlación	.397*	1.000
		Sig. (unilateral)	.017	.
		N	20	20

*. La correlación es significativa al nivel 0,05 (unilateral).

Prueba estadística

***p* Valor**

Fuente: SPSS 20 IBM

De acuerdo a la tabla, el coeficiente de correlación de ***Kendall tau-b*** muestra un coeficiente de correlación de **0.397**. Como este valor es un número **positivo**, muestra que nuestros datos están correlacionados positivamente. **SPSS** También indica **con * que es significativo en el nivel de 0.05 para una predicción de una cola. El valor p real se muestra como 0.017.**

• Una forma convencional de informar estas cifras es la siguiente:

***Kendall tau - b* = 0.397, N = 20, $p<0.05$**

• **Conclusión:** Estos resultados indican que a medida que las calificaciones de la aplicación diseñada aumentan, también lo hace la de la aplicación líder. Por lo tanto, la empresa tiene confianza de anunciar el nuevo diseño de aplicación con baja afectación del competidor líder. Mientras que un diagrama de dispersión es generalmente la estadística ilustrativa más apropiada para apoyar una correlación, cuando se realiza una prueba de ***Kendall tau*-b** debe utilizarse con precaución. El **coeficiente de correlación de *Kendall tau-b*** se obtiene utilizando las calificaciones en lugar de los datos reales, mientras que el diagrama de dispersión muestra las puntuaciones. A continuación se muestra el procedimiento para producir un diagrama de dispersión.

5.8. Diagrama de dispersión: ¿Qué es?

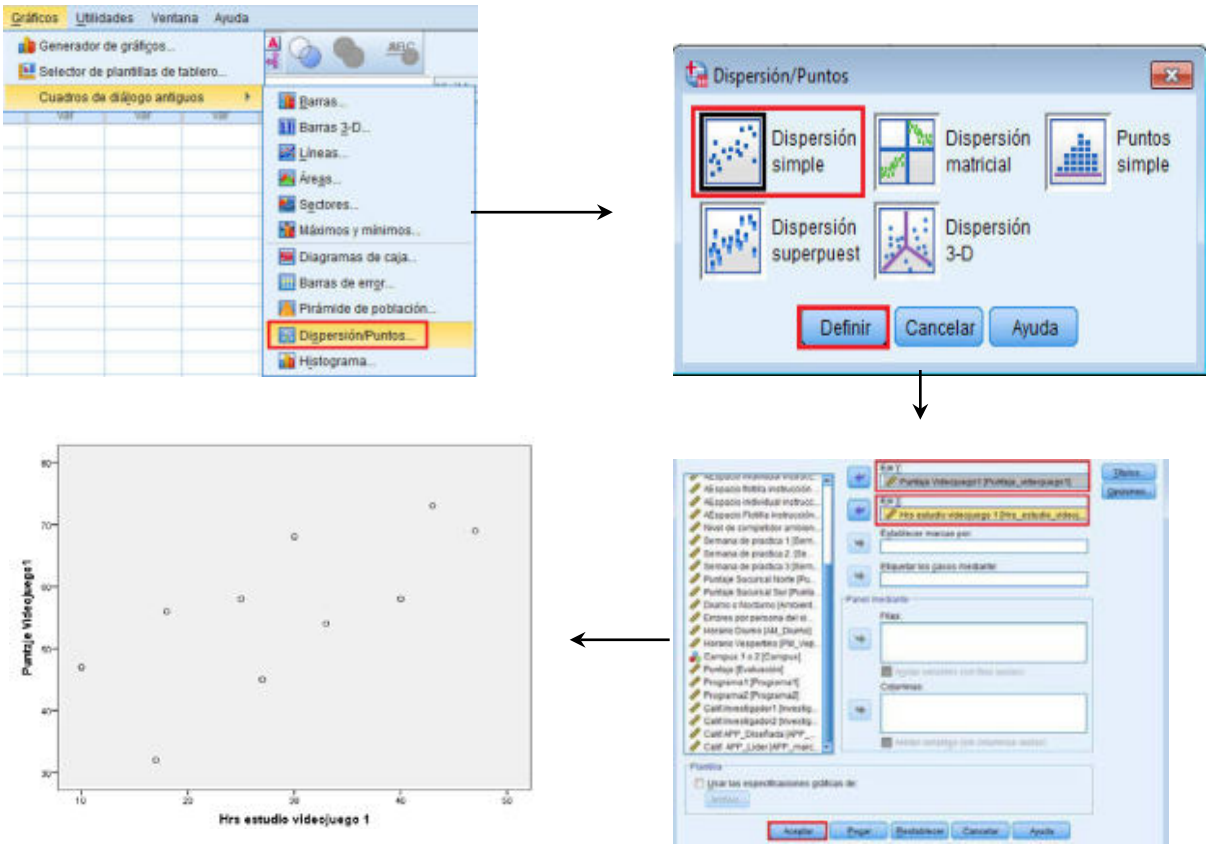
Un diagrama de dispersión ilustra las puntuaciones o datos que deseamos correlacionar, donde los ejes son representados por las dos variables. Si las puntuaciones de la primera variable aumentan y también lo hacen las puntuaciones en la segunda variable, a esto se le

conoce como una **correlación positiva**. Si las puntuaciones de la primera variable aumentan mientras que las puntuaciones en la segunda variable disminuyen a esto se le conoce como **correlación negativa. Cuando los puntos se dispersan aleatoriamente, generalmente no hay correlación entre las dos variables**. Aunque se recomienda realizar el **diagrama de dispersión para soportar ilustrativamente una correlación, debe utilizarse con precaución tomando en cuenta a las correlaciones de *Spearman* y *Kendall tau-b* ya que estas utilizan como análisis no paramétrico, puntajes de calificación en lugar de los datos reales directos y el diagrama de dispersión muestra las puntuaciones reales directas**.
Al producir un diagrama de dispersión puede pedir a **SPSS** que genere **la línea de regresión (línea de mejor ajuste)**. Esta línea en particular, minimiza la distancia de los puntos a la línea recta.

5.9. Diagrama de dispersión: Ejemplo

-Problema 4. Del problema de Correlación Bivariada, realice el correspondiente **diagrama de dispersión.**
-Teclear: Gráficos->Cuadro de diálogo antiguos->Dispersión/puntos->Dispersión simple->Definir->Eje Y: Puntaje Videojuego 1; Eje X: Hrs estudio videojuego 1 (aquí es donde se ubica la variable predictor) ->Aceptar. Ver Figura 5.13

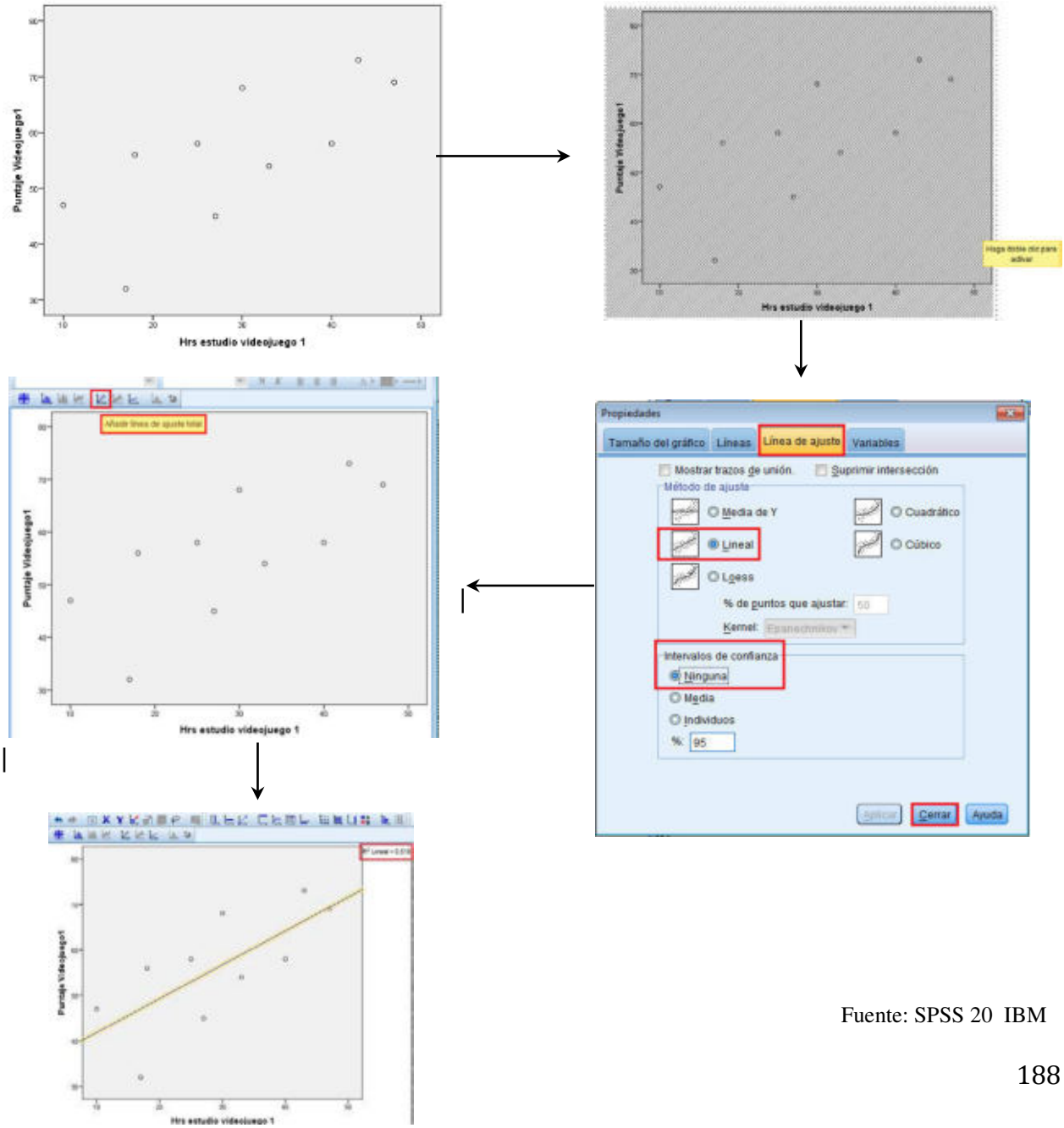
Figura 5.13. Proceso de generación diagrama de dispersión



Fuente: SPSS 20 IBM

-Problema 5. Del problema anterior, insertar **recta de regresión**. Aunque la relación lineal es positiva, como puede verse en el gráfico anterior, la línea de regresión permitirá emitir un juicio más preciso. Para insertar la línea de regresión, haga doble ***click*** dentro del diagrama de dispersión y el gráfico SPSS aparecerá la ventana del editor. Así:
- **Teclear: Gráficos->Cuadro de diálogo antiguos->Dispersión/puntos->Dispersión simple->Definir->Eje Y: Puntaje Videojuego 1; Eje X: Hrs estudio videojuego 1 (aquí es donde se ubica la variable predictor) ->En gráfico doble click->Añadir línea de ajuste total->Línea de ajuste->Lineal->Intervalos de confianza->Ninguna->Cerrar** Ver **Figura 5.14.**

Figura 5.14. Proceso de generación recta de regresión



Fuente: SPSS 20 IBM

5.10. Correlación parcial: ¿Qué es?

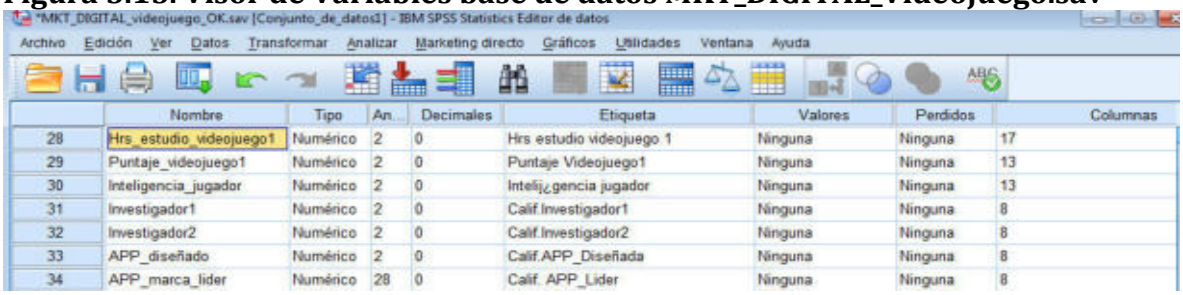
Hasta ahora, en el ejemplo hemos analizado mostrar una correlación significativa entre Hrs de estudio del videojuego1 vs Puntaje videojuego1. Sin embargo, podríamos decidir una tercera variable: inteligencia, que podría estar influyendo en la correlación. Esto es, si la inteligencia tiene una correlación positiva con Hrs de estudio de videojuego1, o sea, que los jugadores más inteligentes invierten más tiempo en estudiar el videojuego1 y si también se correlaciona positivamente con la obtención de mayor puntaje al jugarlo, o sea, que los jugadores más inteligentes obtienen mayor puntaje, significa que **la correlación de Hrs de estudio del videojuego1 y obtención de mayor puntaje al jugarlo, son debidos simplemente a la inteligencia del jugador. Si retiráramos el efecto de la inteligencia, la relación del tiempo de estudio con el rendimiento podría desaparecer.**

5.11. Correlación parcial: Ejemplo

Paso 1: Objetivos

-Problema 6: Para responder a la pregunta de la influencia de la inteligencia en el tiempo de estudio vs. Puntaje en el videojuego1, necesitaremos analizar la correlación de tiempo de estudio y puntaje para después de eliminar los efectos de la inteligencia. Si la correlación desaparece, entonces sabremos que se debió al tercer factor. Para ello, calculamos una **correlación parcial. Ver Figura: 5.15 y 5.16**

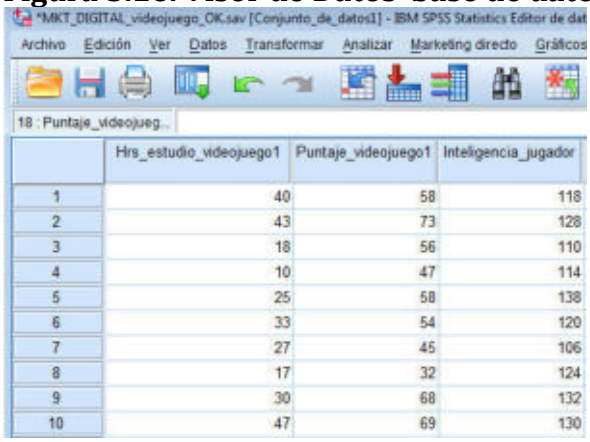
Figura 5.15. Visor de Variables base de datos MKT_DIGITAL_Videojuego.sav



	Nombre	Tipo	An...	Decimales	Etiqueta	Valores	Perdidos	Columnas
28	Hrs_estudio_videojuego1	Numérico	2	0	Hrs estudio videojuego 1	Ninguna	Ninguna	17
29	Puntaje_videojuego1	Numérico	2	0	Puntaje Videojuego1	Ninguna	Ninguna	13
30	Inteligencia_jugador	Numérico	2	0	Intelij¿gencia jugador	Ninguna	Ninguna	13
31	Investigador1	Numérico	2	0	Calif.Investigador1	Ninguna	Ninguna	8
32	Investigador2	Numérico	2	0	Calif.Investigador2	Ninguna	Ninguna	8
33	APP_diseñado	Numérico	2	0	Calif.APP_Diseñada	Ninguna	Ninguna	8
34	APP_marca_lider	Numérico	28	0	Calif. APP_Lider	Ninguna	Ninguna	8

Fuente: SPSS 20 IBM

Figura 5.16. Visor de Datos base de datos MKT_DIGITAL_Videojuego.sav



	Hrs_estudio_videojuego1	Puntaje_videojuego1	Inteligencia_jugador
1	40	58	118
2	43	73	128
3	18	56	110
4	10	47	114
5	25	58	138
6	33	54	120
7	27	45	106
8	17	32	124
9	30	68	132
10	47	69	130

Fuente: SPSS 20 IBM

Paso 2: Diseño

- En adelante para efectos de aprendizaje, sólo se tomarán de forma directa los datos con el fin de aprender a utilizar la técnica.
- A partir del comando **Analizar**, en la opción **Correlación**, **Parcial** (relacionar las variables: Hrs de estudio del videojuego1, Puntaje del videojuego1 vs la variable Inteligencia del jugador), **Coeficientes de correlación, seleccionar como opción *Spearman*,** con **Prueba de significancia a Una cola.**

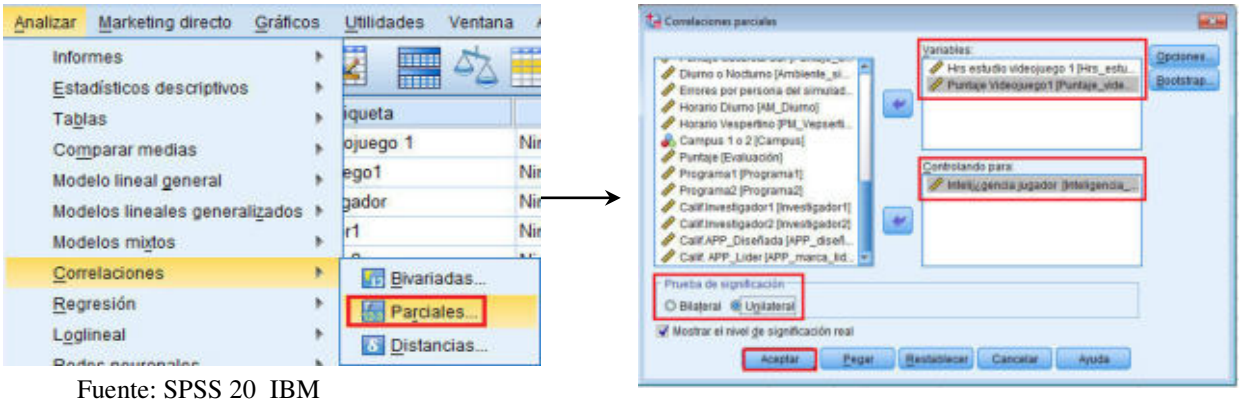
Paso 3: Condiciones de Aplicabilidad

- No olvidar realizar los análisis previos de inspección visual de la base de datos para detectar ausentes y/o atípicos (corregir en su defecto), confiabilidad, normalidad, homocedasticidad y linealidad.

Paso 4: Estimación y ajuste

-Teclear: Analizar->Correlaciones->Parciales->Variables: Hrs estudio videojuego1; Puntaje Videojuego1->Controlando para: Inteligencia jugador->Prueba de significación: Unilateral->Aceptar. Ver Figura 5.17.

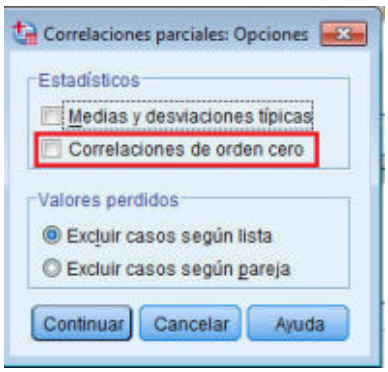
Figura 5.17. Proceso de cálculo de una correlación parcial



Fuente: SPSS 20 IBM

- Si se requirieran los coeficientes de correlación para las tres variables sin control de los puntajes de la variable inteligencia, haga ***click*** en el botón **Opciones** y **marque la Correlaciones de orden cero. Ver Figura 5.18**

Figura 5.18. Opción correlaciones de orden cero



Fuente: SPSS 20 IBM

Paso 5: Interpretación

- **SPSS** genera la tabla de **Correlaciones. Ver Figura 5.19**

Figura 5.19. Tabla Correlaciones

Correlaciones

Variables de control			Hrs estudio videojuego 1	Puntaje Videojuego1	
Intelij¿gencia jugador	Hrs estudio videojuego 1	Correlación	1.000	.665	**Prueba estadística**
		Significación (unilateral)	.	.025	***p* Valor**
		gl	0	7	
	Puntaje Videojuego1	Correlación	.665	1.000	
		Significación (unilateral)	.025	.	
		gl	7	0	

Fuente: SPSS 20 IBM

- La correlación de la prueba estadística es de **0.665**. El ***p* Valor** se muestra como **0.025**. Como este valor es inferior a **0.05**, **existe una correlación significativa**.
- Una forma convencional de informar estas cifras es:

r = 0.665, gl = 7, p <0.05.

- **Conclusión:** estos resultados indican que **a medida que aumenta Hrs de estudio videojuego1, el Puntaje Videojuego 1 aumenta, y esto se logra cuando se han controlado los efectos de la inteligencia del jugador. Esto resulta en una correlación positiva. Por lo tanto, la relación entre Hrs de estudio videojuego1 y el Puntaje Videojuego 1 NO es resultado de la inteligencia del jugador.**
- La **tabla de Correlaciones** también muestra que las **Hrs de estudio videojuego1** está perfectamente correlacionado consigo mismo, ***r* = 1.000**. Del mismo modo, los resultados del **Puntaje del Videojuego 1** tienen una perfecta correlación consigo misma, ***r* = 1.000. Por lo tanto, estos valores no son necesarios.**

5.12. Regresión lineal múltiple: ¿qué es?

Es con mucho, la **técnica de dependencia** más versátil y ampliamente utilizada, aplicable en cualquier ámbito de las **ciencias de la administración**, como lo es por ejemplo, la toma de decisiones en los negocios. Sus usos, van de los problemas más generales a los más específicos, relacionando en cada caso un factor (o factores) con un resultado específico. Por ejemplo, es el fundamento de los modelos de liderazgo, de modelos organizacionales, modelos de gestión de conocimiento, modelos de innovación, etc. que permiten realizar la predicción de comportamiento, de estructuras, de cómo transferir óptimamente el conocimiento o de qué proceso de innovación es pertinente desarrollar en ciertos sectores de la industria de seguir ciertas tácticas estratégicas. En la mercadotecnia por ejemplo, los modelos de regresión se utilizan también para estudiar cómo los consumidores toman decisiones o forman impresiones o actitudes. Otras aplicaciones incluyen la evaluación de los determinantes de la efectividad de un programa (por ejemplo, qué factores ayudan a mantener la satisfacción de servicio, calidad o lograr una mejor penetración de mercado) y de la viabilidad de un nuevo producto o el rendimiento esperado de una nueva acción que cotiza en bolsa. Incluso, aunque estos ejemplos ilustren sólo un pequeño subconjunto de

todas las aplicaciones, demuestran que es una poderosa herramienta analítica diseñada para explorar **todos los tipos de relaciones de dependencia.** El análisis de regresión lineal múltiple **es una técnica estadística general utilizada para analizar las relaciones entre una única variable, criterio y varias variables independientes**. Formulación básica es:

$$Y_1 = X_1 + X_2 + X_3 + \ldots\ldots X_n$$

(Métrica) (Métrica)

Existen líneas generales para juzgar la conveniencia de la regresión lineal múltiple respecto a varios tipos de problemas, tales como el interpretar los resultados de su aplicación tanto desde un punto de vista estadístico como de toma de decisiones, o las posibles transformaciones de los datos para remediar las violaciones de los diversos supuestos, junto con una serie de procedimientos de diagnóstico que identifican observaciones con una influencia particular sobre los resultados.

Esta técnica estadística puede usarse para analizar la relación entre una única **variable dependiente (criterio)** y varias **variables independientes (llamadas predictores o valor teórico)** que en conjunto se le conoce como **ecuación de regresión**, es el ejemplo de **valor predictivo** más ampliamente reconocido entre todas las técnicas multivariantes**.** El objetivo de la técnica es usar las **variables independientes** cuyos valores son **conocidos** para **predecir la única variable criterio** que seleccione. Cada **variable predictor** es ponderada, de forma que **las ponderaciones indican su contribución relativa** asegurando la **máxima predicción conjunta**. Estas ponderaciones facilitan también la interpretación de la influencia de cada variable en la realización de la predicción, **aunque la correlación entre las variables independientes complica el proceso de interpretación.**

Como técnica de dependencia, al utilizarla, deberá de ser capaz de distinguir y dividir las variables entre: **independientes y dependientes** y su uso se condiciona a que éstas sean **todas métricas**. Sin embargo, bajo ciertas circunstancias, **es posible incluir datos no métricos** para las **variables independientes** (**transformando** los datos **ordinales** o los **nominales** en **variables ficticias**) o la **variable criterio** (mediante el uso de **una medida binaria** en la técnica especial de la **regresión logística**). En resumen, al aplicar el análisis de regresión múltiple:

1. Los datos deben ser **métricos** o **apropiadamente transformados** y
2. Antes de derivar la ecuación de regresión, debe decidir qué variable va a ser **dependiente** y cuál de las restantes variables será **independiente.**

El objetivo del análisis de regresión lineal **es predecir una única variable criterio a partir del conocimiento de una o más variables independientes**. Cuando el problema implica:

1. Una **única variable independiente**, la técnica estadística se denomina **regresión simple**.
2. **Dos o más variables independientes**, se denomina **regresión múltiple**.

Para dar mayor referencia a los principios básicos relacionados con la utilización de esta técnica, se aborda el caso presentado por Hair (et al. 1999) con fines ilustrativos del que proporcionan los resultados de un pequeño estudio de ocho familias y su uso de tarjetas de crédito. Se identificaron tres factores potenciales (tamaño familiar, ingresos familiares y el número de automóviles poseídos), y se recogieron datos de cada una de las ocho familias

(**ver Figura 5.20**.). En la terminología del análisis de regresión, la **variable criterio (*Y*)** es el número de tarjetas de crédito utilizado y las tres variables (***X₁,X₂,X₃....***) representan **el tamaño de familia, los ingresos familiares y el número de automóviles poseídos,** respectivamente. La exposición de este ejemplo se divide **en 3 partes** para ayudar a entender cómo la regresión estima, la relación entre la variable independiente y la variable criterio.

Los tres temas que se van a tratar son:

1. Predicción sin una variable independiente, utilizando sólo una medida única **la media**,
2. Predicción usando una única **variable independiente** (**regresión simple**), y
3. Predicción usando varias **variables independientes** (**regresión múltiple**).

Figura 5.20. Resultados de la encuesta sobre el uso de tarjetas de crédito.

Familia ID	Número de tarjetas de crédito que tiene	Tamaño de familia	Ingreso familiar ($)	Número de posesión de automóviles
1	4	2	14	1
2	6	2	16	2
3	6	4	14	2
4	7	4	17	1
5	8	5	18	3
6	7	5	21	2
7	8	6	17	1
8	10	6	25	2

Fuente: Hair et al. (1999)

5.12.1. Predicción sin variable independiente

Para tener la capacidad de contrastar la estimación con la primera ecuación de regresión, empecemos con el cálculo de una línea básica con la que compararemos la capacidad de predicción de nuestros modelos de regresión. La línea básica debería representar nuestra mejor **predicción sin el uso de variables independientes**, entre las que se encuentran opciones como la **predicción perfecta**, un **valor especificado previamente** o una de las **medidas de tendencia central**, como **la media**, **la mediana** o **la moda.** Sin embargo, la **línea predictor** utilizada en la regresión es la **media simple de la variable dependiente**, lo cual tiene varias propiedades deseables. En nuestro ejemplo, la **media aritmética** del número de tarjetas utilizadas es **7**. Nuestra predicción podría ser ***"el número medio de tarjetas de crédito mantenidas por una familia es siete"***. También podríamos poner esto como la siguiente ecuación de regresión:

Predicción del número de tarjetas de crédito (***Y***)= Número medio de tarjetas de crédito (***y***)**.**
Ver Figura 5.21.

Figura 5.21. Predicción de la línea de base con el uso de la media de la variable criterio.

Familia ID	Número de tarjetas de crédito que tiene	Predicción de la línea base(a)	Error de la predicción (b)	Error de la predicción elevado al cuadrado
1	4	7	**-3**	**9**
2	6	7	**-1**	**1**
3	6	7	**-1**	**1**
4	7	7	**0**	**0**
5	8	7	**+1**	**1**
6	7	7	**0**	**0**
7	8	7	**+1**	**1**
8	10	7	**+3**	**9**
Total	56		**0**	**22**

Valor teórico de la regresión: ***Y=y*** Ecuación de predicción: ***y=7***

a).-Número de tarjetas de crédito utilizadas= 56/8=7
b).-Error de predicción referido al valor real de la variable dependiente menos el valor de predicción

Sin embargo, aún debe responder todavía a una cuestión: **¿qué precisión tiene la predicción?** Dado que la **media no dará una predicción perfecta de cada valor de la variable criterio**, tiene que crear alguna manera de valorar la exactitud de predicción, que se podría usar tanto con **la predicción de la línea de base como con los modelos de regresión** que creó. El modo habitual de evaluar la adecuación de una variable predictor es **examinar los errores en la predicción de la variable criterio** cuando se usa para la predicción. Por ejemplo, con la predicción planteada, se dice que cada familia usa siete tarjetas de crédito, de forma que se está **sobrestimando** el número de tarjetas de crédito utilizado por la **familia 1 en 3** (ver **Figura 5.21.**). Por tanto, el **error es +3.** Si este procedimiento fuese seguido para cada familia, algunas estimaciones **serían demasiado altas**, otras serían **demasiado bajas** y a la vez **otras podrían ser correctas**. Aunque se podría esperar la obtención de **una medida útil de exactitud de predicción** con una simple **suma de los errores**, esto **no sería de utilidad porque los errores que proceden del uso del valor medio siempre sumarían cero.** Por tanto, **la suma simple de errores nunca cambiaría**, independiente del grado de éxito que tuvimos con la predicción de la variable criterio con el uso de la media. Para solucionar este problema, **elevamos al cuadrado el error y sumamos los resultados.** El total, denominado como la **suma de los errores al cuadrado** (**SSE. *Square sum error***), proporciona una medida de la precisión predictiva que **varía según la cantidad de errores de predicción**. El objetivo es obtener la **suma de los errores al cuadrado más pequeña posible**, dado que esto indicaría que **nuestras predicciones serían las más precisas.** Se elige **la media aritmética** porque **siempre producirá una suma de errores al cuadrado más pequeña que cualquier otra medida de tendencia central incluida la mediana, la moda, cualquier otro valor único o cualquier otra medida estadística más sofisticada**.
Por tanto, para nuestra encuesta de **ocho familias**, la utilización de la media como nuestra línea básica de predicción nos proporciona el **mejor predictor único del número de tarjetas de crédito con una suma de errores al cuadrado de 22** (vea **Figura 5.21**). En nuestra discusión de la **regresión simple o múltiple**, se usará esta predicción a partir de la

media como argumento para la comparación, dado que **representa la mejor predicción sin utilizar variables independientes.**

5.12.2. Predicción mediante una única variable independiente.

Siempre deberá estamos interesados en mejorar sus predicciones. En la sección precedente, se expuso que **la media es el mejor predictor si no utilizamos otras variables independientes**. Pero en nuestra encuesta de ocho familias también **recogimos información sobre otras medidas que podrían actuar como variables independientes.** Determinemos si el conocimiento de una de estas variables independientes nos ayudará en nuestras predicciones por lo que se refiere a la **regresión simple.**

La regresión simple es otro procedimiento para predecir datos (al igual que la media predice datos), y utiliza la misma regla: **minimizar la suma de los errores cuadrados de la predicción**. Se sabe, que **sin utilizar el tamaño de la familia podemos describir mejor el número de tarjetas de crédito mantenidas como el valor de la media, siete**. El **objetivo** para la **regresión simple es encontrar una variable independiente que mejore la predicción de la línea de base.**

5.12.3. El coeficiente de correlación (*r*)

Basados en el dato del tamaño de la familia, es posible mejorar las predicciones **reduciendo los errores de predicción**. Para hacerlo, los errores de predicción en el **número de tarjetas de crédito** mantenidas deben estar asociadas (**correlacionado**) con el **tamaño de la familia.** El concepto de **correlación**, representado por el **coeficiente de correlación (*r*)**, es fundamental para el análisis de regresión y **describe la relación entre dos variables**. Se dice **que dos variables están correlacionadas si los cambios en una variable están asociados con los cambios en la otra variable**. De esta forma, en la medida que una variable cambia, se sabe cómo está cambiando la otra. Si el tamaño de la familia está correlacionada con el uso de tarjetas de crédito, se escribe entonces la relación como sigue:

Número previsto de tarjetas de crédito = **Cambio en el número de tarjetas mantenidas Con el cambio unitario de X_1 ó $Y= b_1, X_1$** * **Valor de X_1**

Ver **Figura 5.22**

Figura 5.22. Mejorando la exactitud de predicción con la adición de una constante en una ecuación de regresión

Valor de X_1	Variable dependiente	Predicción de la línea	Error de la predicción
1	4	2	2
2	6	4	2
3	8	6	2
4	10	8	2
5	12	10	2

Fuente: Hair et al. (1999)

Parte A: Predicción **sin la constante**
Ecuación de predicción: ***Y=2 X_1***

Valor de X_1	Variable dependiente	Predicción de la línea	Error de la predicción
1	4	2	**2**
2	6	4	**2**
3	8	6	**2**
4	10	8	**2**
5	12	10	**2**

Fuente: Hair et al. (1999)

Parte B: Predicción **con una constante de 2.0**
Ecuación de predicción:
Y= 2.0+ 2 X_1

Valor de X_1	Variable dependiente	Predicción de la línea	Error de la predicción
1	4	2	**0**
2	6	6	**0**
3	8	8	**0**
4	10	10	**0**
5	12	12	**0**

Fuente: Hair et al. (1999)

En la **Figura 5.22** se muestra una ilustración del procedimiento para algunos datos hipotéticos con una única variable independiente $\mathbf{X_1}$. Si encontramos que conforme aumenta $\mathbf{X_1}$ en una unidad, aumenta la **variable criterio (sobre la media) por dos**, entonces es posible **hacer predicciones para cada valor de la variable independiente.** Por ejemplo, cuando X, tiene un valor de 4, la predicción tendría un valor de 8 (véase **Parte A**). Por tanto, el valor de predicción siempre es dos veces el valor de X, (2 $\mathbf{X_1}$). Sin embargo, muchas veces, nos encontramos que **la predicción esta mejorada por la adición de un valor constante**. En la Parte A de la **Figura 5.22.** se puede observar que la simple predicción de **dos veces $\mathbf{X_1}$ es errónea** en cada caso. Por tanto, si cambiamos nuestra descripción para **añadir una constante de dos a cada predicción**, nos proporciona **predicciones perfectas en todos los casos** (véase **Parte B**). Observaremos que cuando se estima una ecuación de regresión, normalmente merece la **pena incluir una constante**.

5.12.4. Especificación de la ecuación de regresión simple

Podemos seleccionar la "***mejor***" variable independiente en nuestro estudio del uso de tarjetas de crédito en base a los coeficientes de correlación dado que cuanto más alto es el coeficiente de correlación, más fuerte es la relación y por tanto más grande es la exactitud de predicción. La **Figura 5.23.** Contiene una matriz de correlaciones entre la variable criterio (***Y***) y las variables independientes ($\mathbf{X_1}, \mathbf{X_2}$ o $\mathbf{X_3}$).

Figura 5.23. Matriz de correlación por el estudio de uso de tarjetas de crédito

Variable	Y	X_1	X_2	X_3
Y. Número de tarjetas de crédito usadas	1.000			
X_1. Tamaño de familia	0.866	1.000		
X_2. Ingreso familiar	0.829	0.673	1.000	
X_3. Número de automóviles	0.342	0.192	0.301	1.000

Fuente: Hair et al. (1999)

Observando la primera columna, podemos ver que $\mathbf{X_1}$, el tamaño de familia, tiene la correlación más alta con la variable criterio y por tanto es la mejor candidata para nuestra primera regresión simple. La matriz de correlación también contiene las correlaciones entre las variables independientes, aspecto muy importante en la regresión múltiple (dos o más variables in dependientes). Ahora podemos estimar nuestro primer modelo de regresión simple para la muestra de ocho familias y ver cómo se ajusta la descripción a nuestros datos. El procedimiento es como sigue:

Número previsto de tarjetas de crédito mantenidas =	Constante + Cambio en el número de tarjetas de crédito con diferentes tamaños de familia **ó** $Y= b_0+ b_1X_1$	*	Tamaño de familia

En la ecuación de regresión, representamos la **constante como *b*** también llamada **coeficiente de regresión**, denotando el cambio estimado en la **variable criterio** por un cambio unitario de la **variable independiente**. El **error de predicción**, la **diferencia entre los valores reales y de predicción de la variable criterio** se denomina **residuo** (**e**). El **análisis de regresión** también permite que las pruebas estadísticas de la **constante y los coeficientes de regresión** pueden determinar si son sustancialmente **diferentes de cero** (es decir, que tienen un impacto diferente al cero). Utilizando el procedimiento matemático conocido como **mínimos cuadrados [**Johnson et al.1982 Neter, 1989**],** po-demos estimar los valores de tal forma que la **suma de los errores cuadrados de la predicción se minimiza**. Para este ejemplo, los valores apropiados son una **constante (b_0) de 2.87** y un **coeficiente de regresión** (b_1) **de 0.97** por tamaño de familia. La ecuación indica que por cada miembro adicional de familia, la posesión de tarjetas de crédito es más alta **como media un 0.97.** La ecuación indica que por cada miembro adicional de familia, la posesión de tarjetas de crédito es más alta como media un **0.97**. Sólo se puede interpretar la **constante 2,87** dentro de la gama de valores para la variable independiente. En este caso, un tamaño de familia de **cero** no es posible por lo que la **constante por sí sola no tiene un sentido práctico**. Sin embargo, esto no anula su uso, dado que ayuda en la predicción de uso de tarjetas de crédito para cada tamaño de familia posible (en nuestro ejemplo: **de 1 a 5**). En los casos en los que las variables independientes pueden adquirir **valores de cero**, la constante tiene una interpretación directa. Para algunas situaciones especiales **donde se conoce que la relación específica pasa por el origen,** la denominación de constante podría ser eliminada (denominado **"*regresión en el origen*"**). En estos casos, la interpretación de los residuos y los coeficientes de regresión cambia ligeramente.

Se muestra la ecuación de regresión simple y las predicciones y residuos para cada una de las ocho familias en la **Figura 5.24.**

Figura 5.24. Resultados de la regresión simple con el uso del tamaño de familia como la variable independiente

Familia ID	Número de tarjetas de crédito que tiene	Tamaño de familia (X_1)	Predicción de regresión simple	Error de la predicción	Error de la predicción elevado al cuadrado
1	4	2	4.81	-0.81	0.66
2	6	2	4.81	1.19	1.42
3	6	4	6.75	-0.75	0.56
4	7	4	6.75	0.25	0.06
5	8	5	7.72	0.28	0.08
6	7	5	7.72	-0.72	0.52
7	8	6	8.69	-0.69	0.48
8	10	6	8.69	1.31	1.72
Total	56				5.50

Nota: Valor teórico de regresión: $Y= b_0 + b_1X_1$; Ecuación de predicción: $Y=2.87+0.97X_1$
Fuente: Hair et al. (1999)

Dado que hemos utilizado el mismo criterio (minimizar la suma de los errores al cuadrado o mínimos cuadrados), podemos determinar si nuestro conocimiento del tamaño familiar nos ha ayudado a predecir mejor la posesión de tarjetas de crédito cuando se compara la predicción de regresión simple con la predicción de la línea básica. La **suma de los errores al cuadrado utilizando la media era 22**. Ahora, la **suma de los errores al cuadrado es 5.50** (véase la **Figura 5.24.**). Utilizando el procedimiento de los mínimos cuadrados y una única variable independiente, vemos que nuestra nueva aproximación, la regresión simple, es mejor que usar sólo la media

5.12.5. La creación de un intervalo de confianza para la predicción

Dado que no podemos conseguir **predicciones perfectas** de la **variable dependiente**, podríamos desear estimar el **rango de valores** que la variable a predecir puede tomar, en lugar de basamos exclusivamente en una estimación simple (**puntual**). La estimación puntual es nuestra mejor estimación de la **variable dependiente** y puede demostrarse que va a ser la mejor predicción para cualquier valor dado de la variable independiente. Utilizando esta **estimación puntual**, podemos calcular el rango de los valores a predecir basándose en una medida de los errores de predicción que esperamos realizar. Conocido como **el error estándar de la estimación** (**SEE**), esta medida es, sencillamente, la **desviación estándar de los errores de predicción**. Recordemos de la estadística elemental que podemos construir un intervalo de confianza para una variable sobre su valor medio añadiendo (más o menos) un cierto número de desviaciones estándar. Por ejemplo, añadiendo (más o menos) **1.96 desviaciones estándar de la media**, se define un rango que incluye el **95 %** de los valores de una variable. Podemos seguir un método similar para las predicciones que realizamos de un modelo de regresión. Utilizando la estimación puntual, podemos añadir (más o menos) un cierto número de erro res estándar de la estimación (dependiendo del nivel de confianza deseado y del tamaño muestral) para

establecer los límites superiores e inferiores de nuestras predicciones hechas con cualquier variable(s) independiente(s). El error estándar de la estimación (SEE) se calcula mediante:

Error estándar de estimación (**SEE**)= $\sqrt{\frac{\text{Suma de errores al cuadrado}}{\text{Tamaño muestral} - 2}}$

El número de **SEE** utilizados para derivar el intervalo de confianza se determina por el nivel de significación (**alfa**) y el tamaño muestral (**N**), que da un valor ***t***. El intervalo de confianza se calcula entonces con el límite inferior siendo igual al valor previsto menos (**SEE** X valor ***t***) y se calcula el límite superior como el valor previsto más (**SEE** X valor ***t)***. Para nuestro ejemplo de la regresión simple, **SEE** = **0.957** (la raíz cuadrada del valor **5.50** dividido por **6**). Se construye el intervalo de confianza para las predicciones seleccionando el número de errores estándar a añadir (más/menos) mediante la búsqueda en una tabla para la distribución **t** y la selección del valor para una nivel de confianza dado con **6 grados de libertad** (**tamaño muestral menos el número de coeficientes**, (**8 - 2 = 6**) es **2.447**. La cantidad añadida (más/menos) al valor previsto es entonces (**0.957 X 2.447**), **o 2.34**. Si sustituimos el tamaño medio de las familias (**4.25**) en la ecuación de regresión, entonces el valor previsto es **6.99** (**difiere de la media de siete sólo en una centésima**). El rango esperado va entonces de **4.65** (**6.99 - 2.34**) a **9.33** (**6.99 + 2.34**). Para una discusión más detallada de estos intervalos de confianza, ver **[**Neter et al. 1989**].**

5.12.6. Valoración de la exactitud de predicción

Si la **suma de los errores al cuadrado** (***SSE***) representa una medida de nuestros errores de predicción, deberíamos ser capaces de determinar una medida de nuestro éxito predictivo, que llamaremos la **suma de los cuadrados de la regresión** (***SSR***). Conjuntamente, estas dos medidas deberían igual a la **suma total de los cuadrados** (***TSS***), el mismo valor que nuestra predicción de la línea de base. En la medida en que el investigador añade variables **independientes,** el total de la suma de los cuadrados puede ahora dividirse en:

1. La **suma de los cuadrados** prevista por la variable independiente, también conocida como **la suma de los cuadrados de la regresión**, y
2. La **suma de los errores al cuadrado:**

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \% \quad \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad ! \quad \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$TSS \quad \% \quad SSE \quad ! \quad SSR$$

Suma total de los cuadrados % Suma de los errores al cuadrado ! Suma de los cuadrados de la regresión

Donde:

$\bar{y}$ = media de todas las observaciones

y_i = valor de las observaciones individuales

$\hat{y}$ = valor previsto para la observación

Podemos utilizar esta división de la suma total de cuadrados para aproximar hasta qué punto el valor teórico de la regresión describe la posesión familiar de tarjetas de crédito. El número medio de tarjetas de crédito mantenidas por nuestra muestra de familias es nuestro mejor estimador del número mantenido por cualquier familia. Sabemos que esto no es una estimación extremadamente precisa, pero es la mejor predicción disponible sin usar otras variables. La predicción de la línea básica con la utilización de la media fue calculada mediante el cálculo de la suma de los errores al cuadrado en la predicción (suma de los cuadrados= **22**). Ahora que hemos ajustado un modelo de regresión utilizando el tamaño de la familia, ¿se explica así la variación mejor que con la media? Sabemos que es algo mejor porque la suma de los errores al cuadrado es ahora **5.50**. Podemos observar así el alcance de nuestro modelo con la investigación de esta mejora.

Donde:

Suma de los errores al cuadrado (predicción de la línea básica)	(SS_{Total} o SST)	22,0
– Suma de los errores al cuadrado (regresión simple)	(SS_{Error} o SSE)	–5,5
Suma de los errores al cuadrado explicados (regresión simple)	($SS_{Regresión}$ o SSR)	16,5

Podemos utilizar esta división de la suma total de cuadrados para aproximar hasta qué punto el por tanto, por lo tanto explicamos **16.5 errores al cuadrado** cambiando de la media a un modelo de regresión utilizando el tamaño de la familia. Esto supone una mejora del **75% (16.5/22 = 0.75)** sobre el uso de la línea básica. Otra forma de expresar este nivel de precisión predictiva es el coeficiente de determinación $\boldsymbol{R^2}$**,** el ratio de la suma al cuadrado de la regresión sobre el total de la suma de los cuadrados como muestra la siguiente ecuación:

Coeficiente de determinación $\boldsymbol{R^2}$ = (suma de los cuadrados de la regresión/suma del total de los cuadrados).

Teniendo en cuenta que el modelo de regresión que utiliza el tamaño de la familia predice perfectamente todas las tarjetas de crédito mantenidas por las familias, $\boldsymbol{R^2 = 1.0}$ y que la utilización del tamaño de la familia no ofreció mejores predicciones que utilizando la media, $\boldsymbol{R^2 = 0}$**,** cuando la ecuación de la regresión contiene más de una variable independiente, el valor del $\boldsymbol{R^2}$ representa el efecto combinado del valor teórico en el conjunto en la predicción. El valor del $\boldsymbol{R^2}$ es simplemente la correlación al cuadrado de los reales y los valores previstos. Cuando se utiliza el coeficiente de correlación (***r***) para evaluar la relación entre las variables dependientes e independientes, el signo del coeficiente de correlación (***+r, -r*)** denota la pendiente de la línea de regresión. Sin embargo, la **"*fuerza*"** de la relación se representa mejor por el $\boldsymbol{R^2}$**,** que es, por supuesto, siempre positiva. En nuestro

ejemplo el R^2 =0.75, que indica que **el 75% de la variación en la variable dependiente se explica por la variable independiente.** Cuando las discusiones mencionan la variación de la variable dependiente, se refieren a esta sima total de cuadrados que el análisis de regresión intenta predecir con una o más variables independientes.

5.12.7 Predicción utilizando varias variables independientes: Análisis de regresión múltiple

Hemos demostrado previamente cómo una regresión simple mejora nuestra predicción de lapo sesión de tarjetas de crédito. Usando los datos del tamaño de la familia, predecimos el número de tarjetas de crédito que poseería una familia mejor de lo que podríamos realizar utilizando simple mente la media aritmética. Este resultado pone de manifiesto la cuestión de si pudiésemos mejorar nuestra predicción utilizando datos adicionales obtenidos de las mismas familias. ¿Mejoraría nuestra predicción del número de tarjetas de crédito si utilizáramos datos del tamaño de la familia además de datos de otra variable, quizá la renta familiar?

5.12.8 La multicolinealidad

La capacidad de una **variable independiente adicional** de mejorar la predicción de una **variable criterio** tiene relación no sólo con la **correlación con la variable dependiente**, sino también respecto de las **correlaciones de las variables independientes** adicionales en función de las **variables independientes** ya presentes en la ecuación de regresión.
La **Colinealidad** es la asociación, medida como correlación, entre dos variables independientes. La **Multicolinealidad** se refiere a la correlación entre tres o más variables independientes (evidenciada cuando se hace la regresión de una respecto de las otras). Aunque existe una distinción precisa en términos estadísticos, es muy común en la práctica utilizar estos términos indistintamente. **El impacto de la multicolinealidad consiste en reducir el poder predictivo de cualquier variable independiente individual en la medida en que está asociado con las otras variables independientes. Al incrementarse la colinealidad, la varianza única explicada por cada variable independiente se reduce y el porcentaje de predicción compartida aumenta**. Dado que sólo se puede calcular una vez se ha realizado la predicción compartida, la predicción global aumenta mucho más lentamente conforme se añaden variables independientes con un nivel de **multicolinealidad alta.** Para maximizar la predicción de un número específico de variables independientes, deberá **buscar otras variables independientes que tienen una multicolinealidad baja** con las otras variables independientes pero que también tienen correlaciones altas con la **variable dependiente.** Estos conceptos se abordarán más adelante con sus implicaciones para la selección de **variables independientes** y la interpretación del valor teórico de regresión.

5.12.9. La ecuación del análisis de regresión múltiple

Para mejorar aún más nuestra predicción de la posesión de tarjetas de crédito, utilicemos unos da tos adicionales obtenidos de nuestras ocho familias. La segunda variable independiente para ser incluida en el modelo de regresión es el ingreso familiar ($\mathbf{X_2}$), que tiene la siguiente correlación más alta con la variable dependiente. Aunque V tiene bastante correlación con $\mathbf{X_1}$ (que ya se contempla en la ecuación), es todavía la siguiente mejor variable para entrar porque $\mathbf{X_3}$, tiene una correlación mucho más baja con la variable

dependiente. Simplemente, ampliamos nuestro modelo de regresión simple para incluir dos variables independientes como sigue:
Predicción del número de tarjetas de crédito utilizadas= $b_0 + b_1X_1 + b_2X_2 + e$
Donde:
b_0 =Número constante de tarjetas de crédito independientemente del tamaño y renta familiar
b_1 = Cambio en la posesión de tarjetas de crédito asociado a un cambio unitario en el tamaño familiar
b_2 = Cambio en la posesión de tarjetas de crédito asociado con un cambio unitario en la renta familiar
X_1= Tamaño de la familia
X_2= Ingreso de la familia

El modelo de regresión múltiple con dos variables independientes, cuando se estima con el procedimiento de mínimos cuadrados, tiene un constante de **0.482** con unos coeficientes de regresión de **0.63** y de **0.216** para X_1 y X_2 respectivamente. De nuevo podemos hallar nuestro residuo en la predicción de **Y,** restando de nuestra predicción el valor efectivo. Elevamos entonces al cuadrado el **error de predicción** resultante, como en la **Figura 5.25.** La suma de los errores al cuadrado es de **3.04** para nuestra predicción utilizando tanto la renta familiar como el tamaño de la familia. Se puede comparar con el valor del modelo de regresión simple de **5.50** (ver **Figura 5.24**), utilizando sólo el tamaño familiar para la predicción. Usando se añade al análisis de la regresión la renta de la familia, aumenta a **0.86:** R^2 (el **tamaño de la familia**+ ingreso familiar)=(22-3.4)/(22.0)= **18.96/22.0= 0.86**
Esto significa que la inclusión de la renta familiar en el análisis de regresión aumenta la predicción en un **11 %** (**0.86-0.75**), debido al incremento de la potencia predictiva del ingreso familiar.

Figura 5.25. Resultados de la regresión múltiple utilizando el tamaño familiar y la renta familiar como variables independientes.

Valor teórico de la regresión: $Y_0=b_0+b_1X_1+b_2X_2$

Ecuación de predicción: $Y=0.482+0.63\ X_1+0.216X_2$

Familia ID	Número de tarjetas de crédito que tiene	Tamaño de familia (X_1)	Ingreso familiar (X_2)	Predicción de regresión simple	Error de la predicción	or de la pre
1	4	2	14	4.76	-0.76	0.58
2	6	2	16	5.20	0.80	0.64
3	6	4	14	6.03	-0.03	0.00
4	7	4	17	6.68	0.32	0.10
5	8	5	18	7.53	0.47	0.22
6	7	5	21	8.18	-1.18	1.39
7	8	6	17	7.95	0.05	0.00
8	10	6	25	0.67	0.33	0.11
Total	56					3.04

Fuente: Hair et al. (1999)

5.12.10. La adición de una tercera variable independiente

Finalmente, se observa un **incremento en la exactitud de predicción** que se gana **con el cambio de la ecuación de regresión simple a la ecuación de regresión múltiple,** pero también tenemos que tener en cuenta que en algún momento la **adición de variables independientes** también será **menos ventajosa** e incluso en algunos casos es contraproducente. En nuestra encuesta de la utilización de tarjetas de crédito, tenemos otra adición posible de la ecuación de regresión múltiple, el número de posesión de automóviles ($\mathbf{X_3}$). Si ahora especificamos la ecuación de regresión para incluir todas las tres variables independientes, podemos observar alguna mejora en la ecuación de regresión, pero no de la misma envergadura vista anteriormente. El valor $\boldsymbol{R^2}$ aumenta a **0.87**, lo que representa sólo un incremento del **0.01** sobre el modelo anterior de regresión múltiple. Además, tal y como se aborda en una sección posterior, el coeficiente de regresión para el $\mathbf{X_3}$ no es estadísticamente significativo. Por tanto, en este caso, el investigador hará mejor en emplear el modelo de regresión múltiple con dos variables independientes (tamaño de familia e ingreso) y no utilizar la tercera variable independiente (número de posesión de automóviles) para hacer predicciones.

El análisis de regresión es una técnica de dependencia simple y sencilla que puede proporcionar al investigador tanto predicción como explicación. El ejemplo previo ha ilustrado los conceptos y procedimientos básicos del análisis de regresión con el fin de desarrollar un conocimiento de la racional y las cuestiones de este procedimiento en su forma más básica. Las siguientes secciones tratan estas cuestiones en más detalle y proporcionan un proceso de decisión para aplicar el análisis de regresión a cualquier problema de investigación.

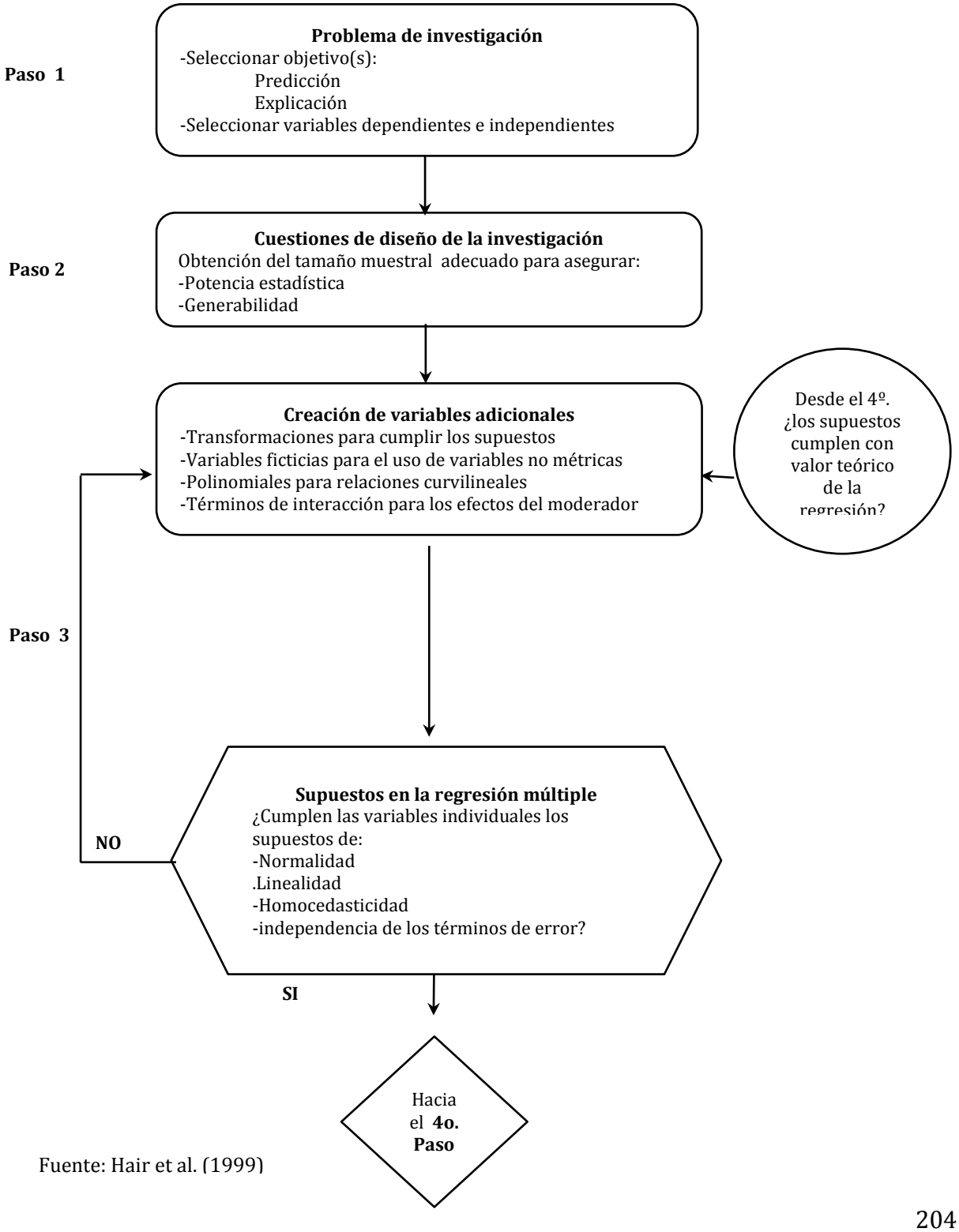
5.13. Regresión lineal múltiple: Proceso de decisión

Hasta aquí, la discusión con ejemplos de qué son las regresiones simples y múltiples. Donde diversos factores afectaban a nuestra capacidad para encontrar el mejor modelo de regresión. En las siguientes secciones, utilizaremos el proceso de modelización en **seis pasos** introducido en el **Capítulo 2** como esquema para discutir los factores que afectan a la creación, estimación, interpretación y validación de un análisis de regresión. El proceso, en términos generales consta de:

1. Comienza con la especificación de los objetivos del análisis de regresión, incluyendo la selección de las variables **dependientes e independientes**. Entonces deberá procede a diseñar el análisis de regresión, con la consideración de factores tales como el **tamaño muestral** y la necesidad de **transformaciones de variables.**
2. Una vez **formulado el modelo de regresión**, **se contrastan** en primer lugar los **supuestos subyacentes** al **análisis de regresión** para las **variables** individualmente.
3. **Si se cumplen todos los supuestos**, entonces es cuando se estima el modelo.
4. Una vez que se obtienen los **resultados**, se lleva a cabo **análisis de diagnóstico** para asegurar que el modelo global cumple los **supuestos de regresión** y que **ninguna observación** tiene una influencia indebida sobre los resultados.
5. El siguiente paso es **la interpretación del valor teórico** de la regresión, donde se **examina el papel jugado por cada variable independiente en la predicción de la medida dependiente**.
6. Finalmente, **los resultados se validan** para asegurar la **generalidad a su población**.

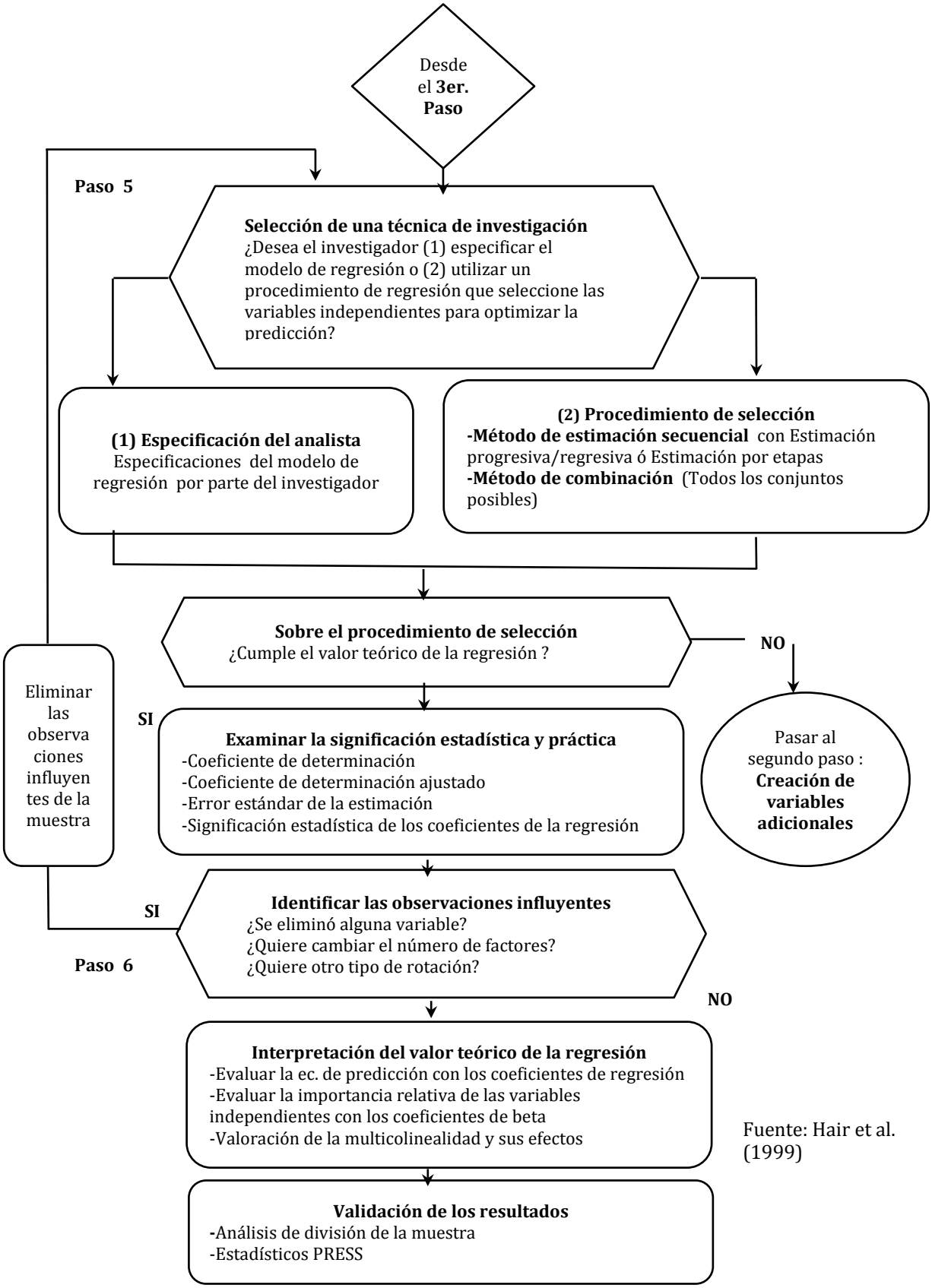
Las **Figuras 5.26 y 5.27** proporcionan los pasos 1-3 y 4-6 respectivamente de una representación gráfica de proceso de construcción del modelo de regresión múltiple para su discusión a detalle

Figura 5.26. Diagrama de flujo pasos 1-3 del análisis de regresión lineal múltiple



Fuente: Hair et al. (1999)

Figura 5.27 Diagrama de flujo pasos 4-7 del análisis de regresión lineal múltiple



Fuente: Hair et al. (1999)

5.14. Regresión lineal: Objetivos

Paso 1: Establecimiento de objetivos

El análisis de regresión lineal múltiple, una forma de modelo lineal general, es una técnica estadística de análisis multivariante utilizada para examinar las relaciones entre una **única variable criterio** y **un conjunto de variables independientes.** El punto de partida, como en todas las técnicas estadísticas multivariantes, **es el problema a investigar.** La flexibilidad y la capacidad de adaptación de la regresión múltiple permite utilizarlos **con casi cualquier relación de dependencia**. En la selección de la aplicación adecuada del análisis de regresión lineal, deberá considerar **3 puntos** fundamentales:

1. La conveniencia del programa de investigación,
2. La especificación de una relación estadística y
3. La selección de las variables dependientes e independientes**.**

5.14.1. Problemas de investigación adecuados para la regresión múltiple

La regresión lineal múltiple es con mucho **la técnica multivariante** más utilizada. Con sus amplias posibilidades de aplicación, el **análisis de regresión lineal múltiple** se utiliza para diversos propósitos.

Las crecientes aplicaciones de la regresión múltiple, sin embargo, se agrupan en 2 amplias clases de problemas de investigación:

1. **Predicción** y
2. **Explicación.**

Estos problemas de investigación **no son mutuamente excluyentes**, y se puede llevar a cabo una aplicación del análisis de regresión múltiple para cualquiera de los dos tipos de problemas de investigación.

5.14.2. Predicción con regresión lineal múltiple

Uno de los objetivos fundamentales de la regresión lineal múltiple es la **predicción de la variable criterio con un conjunto de variables independientes.** De hecho, **el primer objetivo es maximizar la potencia conjunta de predicción de las variables independientes tal y como se representan en el valor teórico.** Esta combinación lineal de variables independientes se construye de tal forma que se convierta en un **predictor óptimo de la variable criterio**. La regresión lineal múltiple proporciona un medio objetivo de evaluar el poder predictivo de un conjunto de variables independientes. Al centrarse es este objetivo, el estará principalmente interesado en **conseguir la máxima predicción**. La regresión lineal múltiple proporciona muchas opciones tanto en la forma como en la especificación de las variables independientes que **puedan modificar el valor teórico para aumentar su poder predictivo**. **Muchas veces la predicción se maximiza a expensas de la interpretación**.

Un ejemplo es una **variante** del análisis de regresión, el **análisis de series temporales**, en el cual el **único propósito es predecir** y la interpretación de los resultados es útil sólo como un medio de incrementar la precisión predictiva.

En otras situaciones, la precisión predictiva es crucial para **asegurar la validez del conjunto de variables independientes**, teniendo en cuenta la **ulterior interpretación del valor teórico.**

Las medidas de precisión predictiva y los test estadísticos se forman en relación con la significación del poder predictivo que pueda obtenerse.
En todos los casos, tanto **si la predicción es o no es el objetivo principal**, el **análisis de regresión lineal** debe conseguir niveles aceptables de precisión predictiva para justificar su aplicación. Así, deberá asegurarse de tener en cuenta tanto la significación práctica como la estadística (vea el paso cuarto, a continuación de esta discusión).
La **regresión lineal múltiple** puede también conseguir un **segundo objetivo** de comparación de dos o más conjuntos de variables independientes para **averiguar el poder predictivo de cada valor teórico**. Ilustrativo de una **aproximación de modelización confirmatoria**, este uso de la regresión múltiple se centra en la **comparación de resultados entre dos o más alternativas o modelos en competencia**. El objetivo principal de este tipo de análisis es el **poder predictivo relativo entre modelos**, aunque en cualquier situación la predicción del modelo elegido debe demostrar tanto **significación práctica como estadística.**

5.14.3. Explicación con regresión múltiple

La **regresión lineal múltiple** proporciona también un medio de evaluar objetivamente el **grado y carácter de la relación entre las variables dependientes e independientes** al formar el valor teórico. Las **variables independientes**, además de su predicción conjunta de la **variable dependiente**, pueden considerarse también por su **contribución individual al valor teórico y a sus predicciones**. La interpretación del valor teórico puede tomarse desde alguna de estas **3 perspectivas:**

1. **La importancia de las variables independientes,**
2. Los **tipos de relaciones** encontradas o
3. Las **interrelaciones entre las variables independientes.**

La interpretación más directa del **valor teórico** de la regresión es una determinación de la importancia relativa de cada variable independiente en la predicción de la medida independiente. En todas las aplicaciones, la selección de variables independientes se basaría en sus relaciones teóricas con la variable dependiente. El análisis de regresión proporciona un medio de evaluar objetivamente la **magnitud y dirección (positiva o negativa) de cada relación con la variable independiente**. El carácter de la regresión múltiple, que la diferencia de sus contrapartidas univariantes, es la **evaluación simultánea de relaciones entre cada variable independiente y las medidas de la dependiente**. Al realizar esta evaluación simultánea, se determina la importancia relativa de cada predictor.
Además de evaluar la importancia de cada variable, la regresión **le permite también la evaluación de la naturaleza de las correlaciones entre las variables independientes y la variable dependiente**. Así también se disponen de **transformaciones** para **evaluar si existen otros tipos de relación**, particularmente las **relaciones curvilineales.** Esta flexibilidad le asegura que pueda examinar la verdadera naturaleza de las relaciones **más allá de la supuesta relación lineal**. Finalmente, la regresión múltiple proporciona también **una idea** de las relaciones entre las **variables independientes** en sus **predicciones** de la **variable dependiente**. Estas interpretaciones son importantes por **2 razones:**

1. La **correlación** entre las **variables independientes** puede hacer que algunas variables sean **redundantes** en su esfuerzo predictivo. Como tal, **no son necesarias para**

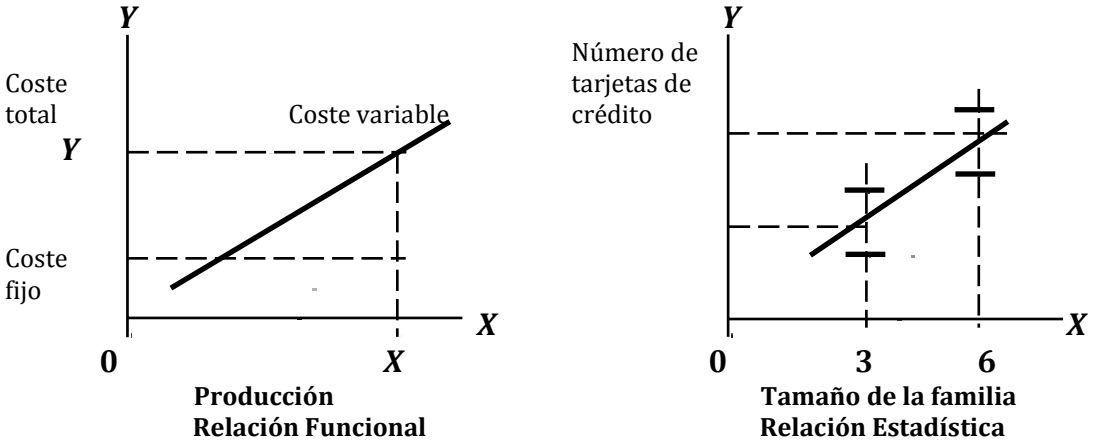
producir una predicción óptima. No se trata de reflejar sus relaciones individuales con la variable criterio sino que indica **que en un contexto multivariante, no son necesarias** si se emplea otro conjunto de variables independientes para explicar esta varianza.

2. Deberá mostrarse precavido con la **determinación de la importancia** de las variables independientes basadas solamente en el **valor teórico** derivado, porque las relaciones entre las variables independientes pueden **"*enmascara*r"** relaciones que no se necesitan para propósitos predictivos pero que representan sin embargo hallazgos sustantivos. Las interrelaciones entre las variables pueden extenderse no sólo a su poder predictivo sino también a las interrelaciones entre sus efectos estimados. Esto se ve mejor cuando el efecto de una variable independiente es contingente con otra variable independiente. La regresión múltiple proporciona un diagnóstico que puede determinar si existen tales efectos basados en razones empíricas o teóricas. Las indicaciones de un alto grado de interrelaciones (**multicolinealidad**) entre las variables independientes pueden sugerir **el uso de las escalas sumadas**.

5.14.4. Especificación de la relación estadística para la regresión múltiple

Una regresión múltiple es apropiada cuando está interesado en una relación estadística, **no funcional**, Por ejemplo, la siguiente relación**: Coste total = Coste variable + Coste fijo**
Si el coste variable es de **$2** por unidad, el coste fijo es de **$500** y producimos **100** unidades suponemos que el coste total será exactamente **$700** y que cualquier desviación de los **$700** causada por nuestra incapacidad para medir el coste dado que la relación entre costes es fija. A esto se le denomina **relación funcional** porque esperamos que no existirá un error en nuestra predicción. Pero en el ejemplo anterior sobre la muestra de datos que representa el comportamiento humano, estábamos suponiendo que nuestra descripción del uso de las tarjetas de crédito era sólo aproximada y **no una predicción perfecta.** Se pensaba que era una relación estadística porque siempre existiría un **componente aleatorio** en la relación examinada. Encontramos dos familias con dos miembros, dos con cuatro miembros, etc., que tenían distinto número de tarjetas de crédito. En una **relación estadística** se observará más de un valor de la variable dependiente para cualquier valor de una variable predictor. La **variable criterio** se supone que es una **variable aleatoria**, y para un **predictor** dado sólo podemos esperar estimar el **valor medio de la variable criterio** asociado con él. En el ejemplo de la **regresión simple**, las dos familias con cuatro miembros mantienen una media de **6.5** tarjetas de crédito, y nuestra predicción era de **6.75**. **Nuestra predicción no es tan precisa como desearíamos, pero es mejor que usar nada más que la media de 7 tarjetas de crédito.** Se supone que **el error** es el resultado de un **comportamiento aleatorio** entre los poseedores de las tarjetas. Así, **una relación funcional calcula un valor exacto, mientras que una relación estadística estima un valor medio. Aquí nos centraremos en las relaciones estadísticas. Ver Figura 5.28.**

Figura 5.28. Comparación de las relaciones funcionales y estadísticas.



5.14.5. Selección de variables dependientes e independientes para la regresión múltiple

El logro final de cualquier técnica multivariante, incluyendo las regresiones múltiples, comienza en este punto. Dado que la **regresión múltiple** muestra una relación de dependencia, **deberá especificar qué variable es la que se usa como criterio y qué variables se usan como predictor**. La selección de ambos tipos de variables debería basarse principalmente en fundamentos conceptuales o teóricos. Deberá por tanto, tomar las decisiones fundamentales de la **selección de variables**, incluso aunque tenga opciones y comandos de programas de software para ayudarle en la estimación del modelo, si no emite juicios durante la selección de la variable y en su lugar:

1. **Selecciona las variables indiscriminadamente** o
2. **Permite que la selección de una variable independiente se base exclusivamente en bases empíricas**, se incumplirán varios de los principios básicos del desarrollo del modelo.

La selección de una **variable criterio** está muchas veces dictada por el problema de la investigación. Pero en muchos casos, **deberá ser consciente del error de medida**, especialmente en la **variable independiente**. El **error de medida** se refiere al grado en que la variable es **una medida precisa y consistente** del concepto que está siendo estudiado. Si la variable que se utiliza **como dependiente tiene un error sustancial de medida,** entonces **incluso las mejores variables independientes pueden ser incapaces de conseguir niveles aceptables de precisión predictiva**.

El error de medida puede venir de diversas fuentes (Vea **Capítulo 2**). El error de medida que es problemático puede ser abordado mediante el uso de las **escalas aditivas**. Debe siempre interesarse por la **obtención de la mejor medida de las variables dependientes e independientes**, basadas ambas en factores empíricos y conceptuales. El supuesto más problemático en la selección de la variable independiente es el **error de especificación,** que hace referencia a la inclusión de **variables irrelevantes o a la omisión de variables relevantes** del conjunto de variables independientes. Aunque la inclusión de una variable irrelevante no sesgue los resultados de las otras variables independientes, tiene cierto impacto sobre ellos:

1. **Reduce la parsimonia del modelo**, que puede ser crítica en la interpretación de los resultados.
2. Las variables adicionales pueden **enmascarar** o desplazar los efectos de variables más útiles, especialmente si se utiliza alguna forma jerárquica de estimación del modelo (véase paso 4 para más detalle).
3. Finalmente, las **variables adicionales pueden hacer que las contrastaciones de la significación estadística de las variables independientes sean menos precisas** y reduzcan la significación estadística y práctica del análisis.

Dados los problemas asociados con la adición de variables irrelevantes, ¿debe fijarse en las **variables relevantes excluidas**? La respuesta es **definitivamente sí**, porque la exclusión de las **variables relevantes** puede **sesgar seriamente los resultados y afectar negativamente cualquier interpretación de ellos**. En el caso más simple, **las variables omitidas no están correlacionadas con las variables incluidas**, y el único efecto es **reducir la precisión predictiva conjunta del análisis.** Pero cuando existe correlación entre las variables incluidas y las omitidas, los efectos de las variables incluidas pueden verse segados en la medida en que están correlacionadas con las variables omitidas. **Cuanto mayor sea la correlación, mayor será el sesgo.** Los efectos estima dos para las variables incluidas representan ahora no sólo sus efectos reales sino también los efectos que las variables incluidas comparten con las variables omitidas. Esto nos puede llevar a **serios problemas en la interpretación de los modelos y en la evaluación de la significación estadística y práctica.**

Deberá ser cuidadoso en la selección de las variables para **evitar ambos tipos de errores de especificación**. Quizá los mayores problemas consistan en la **omisión de las variables relevantes**, dado que los efectos de las variables no pueden evaluarse sin su inclusión. Esto intensifica la necesidad de un soporte práctico y teórico de todas las variables incluidas o excluidas en un análisis de regresión múltiple.

Los errores de medida afectan también a las variables independientes reduciendo su poder predictivo en la medida en que aumenta el error de medida. **La regresión múltiple no tiene medios directos para corregir los niveles conocidos de error de medida** para las variables independientes.

Si el investigador sospecha que el error de medida es problemático en las variables independientes **deberá utilizar los modelos de ecuaciones estructurales como un medio de tratar los errores de medida en la estimación de los efectos de las variables independientes.**

5.15. Regresión lineal múltiple: Diseño

Paso 2: Diseño

En el diseño de un análisis de regresión múltiple, el investigador debe considerar asuntos tales el tamaño muestral, la naturaleza de las variables independientes y la posible creación de variables para representar las especiales relaciones entre las variables dependientes e independientes. Al hacerlo, debe mantener siempre el criterio de significación práctica y estadística. La capacidad de las de las regresiones múltiples para realizar muchos tipos de investigaciones se ve enormemente influenciada por los supuestos del diseño de la investigación que se discutirán a continuación.

5.15.1. Tamaño muestral

El tamaño muestral utilizado en la regresión múltiple es quizá el elemento aislado más influyente bajo control del investigador en el diseño del análisis. Los efectos del tamaño muestral se ven más directamente en la potencia estadística del test de significación y la generalización del resultado. Trataremos ambos asuntos en las secciones siguientes.

5.15.2. Potencia estadística y tamaño muestral

El tamaño muestral tiene un impacto directo en la conveniencia y la potencia estadística de la regresión múltiple. Muestras pequeñas, habitualmente caracterizadas por tener menos de **20 observaciones**, **son apropiadas sólo para análisis de regresión simple con una única variable independiente.** Incluso en estas situaciones, sólo se pueden detectar relaciones muy fuertes con cierto grado de certidumbre. De la misma forma, las muestras muy grandes, de **1000 observaciones o más**, hacen **los test de significación estadística demasiado sensibles, indicando que casi cualquier relación es estadísticamente significativa**. Con muestras muy grandes deberá asegurarse que el **criterio de significación práctica se cumpla a la vez que la significación estadística**.

La **potencia de la regresión múltiple** se refiere a la probabilidad de detectar como **estadísticamente significativo** un nivel específico de o un **coeficiente de regresión** para un **nivel de significación** especificado y un **tamaño de muestra específico** (vea **Capítulo 2** para una discusión más detallada). El **tamaño muestral** tiene un impacto directo y cuantificable **sobre la potencia.** Vea la **Figura 5.29.**

Figura 5.29. Mínimo que se puede encontrar estadísticamente significativo con una potencia de 0.80 para diferentes variables independientes y tamaños muestrales

	Nivel de significación (alfa)=0.01 Número de variables Independientes				Nivel de significación (alfa)=0.05 Número de variables Independientes			
Tamaño muestral	2	5	10	20	2	5	10	20
20	45	56	71	NA	39	48	64	NA
50	23	29	36	49	19	23	29	42
100	13	16	20	26	10	12	15	21
250	5	7	8	11	4	5	6	8
500	3	3	4	6	3	4	5	9
1000	1	2	2	3	1	1	2	2

Fuente: Hair et al. (1999)

La **Figura 5.29** ilustra la interacción entre el tamaño muestral, el nivel de significación (alfa) elegido y el número de variables independientes para detectar un R^2 significativo. Los valores de la tabla son el mínimo R^2 que el tamaño muestral especificado detectará como **estadísticamente significativo** y el nivel alfa especificado con una probabilidad (**potencia**) de **0.80.** Por ejemplo.

1. Si empleara **5 variables independientes,** especifica el nivel de significación de **0.05** y está satisfecho al detectar el R^2 del **80%** de las veces que ocurre (**potencia de 0.80**), una muestra de **50** encuestados detectará valores de R^2 del **23%** y superior.
2. **Si la muestra aumenta en 100** detectarán valores del R^2 del **12%** o superiores.
3. pero si los **50** encuestados es todo lo que tiene el investigador y quiere un nivel de significación del **0.01,** el investigador detectará valores de R^2 sólo por encima del **29%.**
4. As , debera considerar el papel del taman o muestral en la contrastacio n de la significacio n antes de la recogida de datos. Si se esperan relaciones de biles, puede hacer juicios contrastados en la medida en que el taman o de muestra necesario detecte razonablemente las relaciones, si existen. Por ejemplo, de la misma **Figura 5.29** demuestra que los **tamaños de muestra 100** detectara n claramente valores de R^2 bajos(**del 10 al 15%**) por encima de **10 variables independientes** y un nivel de significacio n de **0.05.** sin embargo, si el taman o muestral **es menor de 50 observaciones** en estas situaciones, puede detectarse el **doble de R^2 .** Debera ser consciente de la potencia anticipada de cualquier ana lisis de regresio n propuesto y entender los elementos del disen o de la investigacio n que pueden cambiarse para cumplir los requerimientos de un **análisis aceptable [**Masan, et al. 1991**].**
5. El investigador puede determinar también el tamaño muestral necesario para detectar los efectos de las variables independientes individuales dado el efecto del tamaño esperado (**correlación**), el nivel alfa y la potencia deseada.

Para saber ma s, vea **análisis de potencia (**Cohen, 1983) o un programa de computador para calcular el **tamaño muestral para una situación dada [**BMDP Statistical Software 1991**].**

5.15.3. Generalización y tamaño muestral

Además de tener un papel importante en la determinación de la **potencia estadística**, el **tamaño muestral** también **afecta a la generalización de los resultados**, en función del ratio de observaciones sobre las variables independientes. Una norma general es que **este ratio nunca debería caer por debajo de cinco (5), lo que significa que existirán cinco (5) observaciones para cada variable independiente presente en el valor teórico.** Conforme cae este ratio por debajo de cinco, el investigador puede incurrir en el riesgo de "***sobre ajustar***" el valor teórico respecto de la muestra, haciendo que los **resultados sean muy específicos para esa muestra** y por tanto con **falta de generalización**. Mientras que **el ratio mínimo es 5 a 1,** el nivel deseado está entre **15 y 20 observaciones para cada variable independiente.** Cuando este nivel se alcance, **los resultados deberían ser generalizables si el tamaño muestral es representativo**. Sin embargo, si se emplea un proceso por pasos (**cuarto paso bajo la aproximación a la estimación del modelo**), el nivel recomendado aumenta de **50 a l.** En casos donde la muestra disponible no cumpla estos criterios, deberá asegurarse la validación de la generalización de los resultados.

5.15.4. Predictores de efectos fijos frente a predictores de efectos aleatorios

Los ejemplos de los **modelos de regresión** vistos hasta este punto han sido formulados bajo el supuesto de que **los niveles de las variables predictor son fijos**. **Por ejemplo**, si queremos saber el **impacto sobre los tres niveles de preferencia de una nueva esencia para perfume,** presentaremos **tres tandas** de este tipo de **esencias** a un grupo de personas. Entonces **predecimos el ratio** de preferencia sobre **cada esencia**, utilizando el **nivel de percepción como predictor.** Hemos fijado el nivel de **esencia** y estamos interesados en su efecto en estos niveles. No suponemos que los tres niveles sean una muestra aleatoria de un gran número de posibles niveles de esencia. Una **variable predictor aleatoria** es aquella en que los niveles de la variable predictor se seleccionan aleatoriamente. Cuando se usa, el interés no está sólo en los ni veles examinados sino en **la mayor cantidad de posibles niveles del predictor de los que se seleccionan en una muestra.** La mayor parte de **los modelos de regresión basados en una encuesta de datos son modelos de efectos aleatorios.** A modo de ilustración, se dirigió una encuesta para ayudar a evaluar la relación entre la **edad** del encuestado y la frecuencia de sus **visitas al médico**. La variable predictor **"*edad del encuestado*"** se seleccionó aleatoriamente de la población, y **la inferencia en relación con la población es el objetivo, no sólo el conocimiento de los individuos de la muestra.** Los procedimientos de estimación de modelos que utilizan ambos tipos de variables son las mismas excepto para los **términos de error.** En los **modelos de efectos aleatorios**, una parte del **error aleatorio procede del muestreo de los predictores.** Sin embargo, los procedimientos estadísticos basados en el modelo fijo son bastante robustos, así que utilizar el análisis estadístico como si se es tuviese tratando con un modelo fijo (como suponen la mayor parte de los programas de análisis) puede considerarse como una aproximación razonable.

5.15.5. Creación de variables adicionales

La relación básica representada en la **regresión múltiple** es la **asociación lineal** entre variables de pendientes e independientes métricas basada en **la correlación** momento-producto. Muchas veces se enfrentará al **problema de incorporar datos no métricos**, tales como género u ocupación, en una ecuación de regresión. Sin embargo, **la regresión se limita a los datos métricos**. Además, tiene **la incapacidad de modelizar directamente las relaciones no lineales,** lo que puede suponer una restricción para Usted cuando se enfrenta con situaciones en las que una **relación no lineal** (**por ejemplo, en forma de U**) es **sugerida por la teoría o es detectada cuando examinan los datos**. En estas situaciones, **deben crearse nuevas variables mediante transformaciones**, en la medida en que la regresión múltiple descansa completamente en los tipos de variables del modelo **que sólo representan relaciones lineales**. **La transformación** de los datos ofrece un medio de **modificar tanto las variables dependientes como las independientes** por una de estas dos razones:

1. **Mejorar o modificar la relación entre las variables** dependientes o independientes o
2. **Permitir el uso de variables no métricas** en el valor teórico de la regresión.

Las transformaciones de los datos pueden basarse en **razones tanto "*teóricas*"** (transformaciones cuya conveniencia es sugerida estrictamente por la naturaleza de los datos) **o "*derivadas de los datos*"** (transformaciones sugeridas estrictamente por el examen de los datos). En cualquier caso debe proceder muchas veces por **ensayo y error**,

evaluando constantemente las mejoras frente a la necesidad de transformaciones adicionales. Las transformaciones se pueden realizar mediante software estad stico, aunque existen otros me todos ma s sofisticados y complicados .Ver **[**Box & Cox, 1964**]**

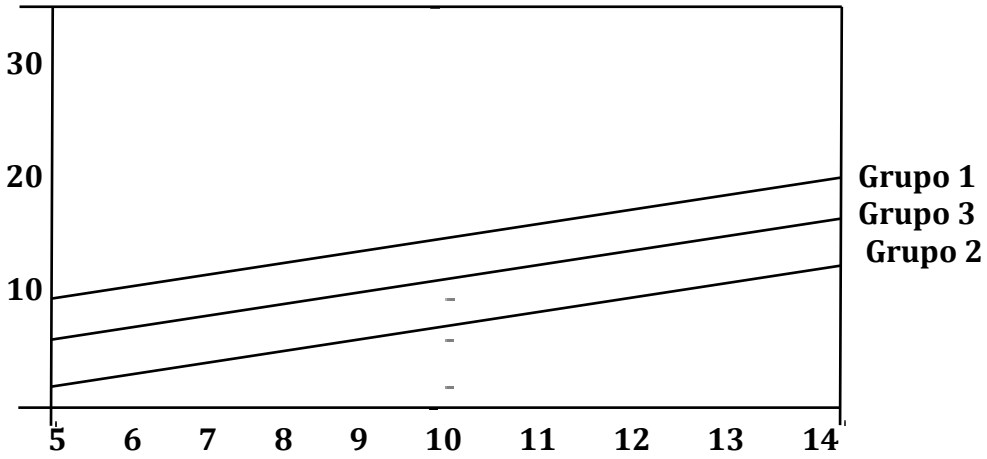
5.15.6. Incorporación de datos no métricos con variables ficticias

Una situación común a la que se enfrentará es la presencia de **variables independientes no métricas**. No obstante, **hasta ahora todas nuestras explicaciones** han supuesto **medidas métricas** tanto para las variables **independientes** como las variables **dependientes**. Cuando la **variable criterio** se mide como una **variable dicotómica (0, 1),** lo más apropiado es el **análisis discriminante** o una forma especializada de regresión **(regresión logística),** |pero, ¿qué hacer cuando las **variables independientes son no métricas**, con **dos o más categorías**? Como se vio en **Capítulo 3** anterior, se deberá considerar el concepto de **variables dicotómicas**, conocidas **como variables ficticias**, que actúan en lugar de **las variables independientes**. Cada variable representa una categoría de **variable independiente no métrica** y **cualquier variable no-métrica con k categorías puede representarse como una variable ficticia k- l.**

Existen **2 formas de codificación de variable ficticia:**

1. La más común es la **codificación de indicador** en la que se representa la categoría por **1 o 0**. Los coeficientes de regresión para la variable ficticia representan desviaciones para cada grupo de encuestados formado por una variable ficticia de la categoría de referencia (es decir, el grupo omitido que recibe todos los ceros) res pecto la variable dependiente. Estas diferencias de grupo pueden ser valoradas directamente, dado que los coeficientes están en las mismas unidades que la variable dependiente. Esta forma de codificación de la variable ficticia puede ser mostrada como atajos diferentes de los grupos (véase **Figura 5.30**). En este ejemplo, se representa una variable no métrica de tres categorías por dos variables ficticas (**D1y D2**) dando valores a los **grupos 1 y 2**, respecto del **grupo 3**, **la categoría de referencia**. Los coeficientes de regresión son **2.0** para **D1** y **-3.0** para **D2.** Estos coeficientes se traducen en tres líneas paralelas. Ver **Figura 5.30.**

Figura 5.30. La incorporación de variables no métricas mediante variables ficticias (D_1=1, D_2 =0)



Ecuaciones de regresión con variables ficticias (D_1 y D_2)	
Especificadas	$Y=a+b_1X+b_2D_1+b_3D_2$
Estimadas	
Globales	$Y=2+1.2X+ 2D_1-3D_2$
Específicas de grupo	
Grupo 1(D_1=1, D_2 =0)	$Y=2+1.2X+2(1)$
Grupo 2 (D_1=0, D_2 =1)	$Y=2+1.2X-3(1)$
Grupo 3 (D_1=0, D_2 =0)	$Y=2+1.2X$

Fuente: Hair et al. (1999)

El grupo de referencia (**grupo 3**) se define por la ecuación de regresión con las dos **variables ficticias igual a cero**. La línea del **grupo 1** está **dos unidades** por encima de la línea para el grupo de referencia. La línea del **grupo 2** está **tres unidades** por de bajo de la línea del **grupo de referencia 3**. Las líneas paralelas indican que las variables ficticias No cambia la naturaleza de la relación, pero solamente estipula los atajos diferentes entre los grupos. Esta forma es más apropiada cuando existe un grupo de comparación lógica, **tal como ocurre en un experimento.** Siempre que se utilice la codificación mediante variable ficticia, debemos ser conscientes del grupo de comparación y recordar que los coeficientes presentan las diferencias entre la media del grupo y la del grupo de referencia.

2. **Por determinación de efectos.** Es exactamente igual que la codificación de indicador excepto que el grupo omitido o de comparación (**el grupo donde todos son ceros**) se le da ahora el valor **de -1 en lugar de 0** para las variables ficticias. Así, **los coeficientes representan diferencias para cualquier grupo respecto de la media de todos los grupos (en vez de la media del grupo omitido).**

Ambas formas de variable ficticia darán exactamente **los mismos resultados predictivos, coeficientes de determinación y coeficientes de regresión de las variables continuas.** Las únicas diferencias estarán en la interpretación de los coeficientes de la variable ficticia.

5.15.7. Representación de efectos curvilíneos con polinomios

Existen varios tipos de **transformaciones de datos apropiados para convertir en lineal una relación curvilínea.** Los métodos directos, discutidos en el **Capítulo 2**, implican modificaciones de los valores a través de **ciertas transformaciones aritméticas** (por ejemplo, **calculando la raíz cuadrada o el logaritmo de las variables**). Sin embargo, tales transformaciones tienen varias limitaciones:

1. Son **útiles sólo en relaciones curvilineales simples** (una relación con sólo un punto de giro o inflexión).
2. **No ofrecen medios estadísticos para evaluar si el modelo lineal o curvilineal es el más apropiado.**
3. Finalmente, **sólo se pueden utilizar para relaciones univariantes y no para la interacción entre variables cuando nos encontramos con más de una variable independiente**.

Discutiremos a continuación un medio de **crear variables que modelizan explícitamente los componentes curvilineales de la relación y ponen de manifiesto las limitaciones inherentes a las transformaciones de los datos.** Los **polinomios** son transformaciones potenciales de una **variable independiente** que an ade **un componente no lineal para cada potencia adicional de la variable independiente.** La potencia de 1 ($\mathbf{X^1}$) representa el componente lineal y es la forma que vamos a discutir a lao largo de este cap tulo. La potencia segunda, la variable al cuadrado ($\mathbf{X^2}$), representa el componente cuadra tico. **En términos gráficos, $\mathbf{X^2}$, representa el primer punto de inflexión.** Un componente cu bico, representado por la variable elevada el cubo ($\mathbf{X^3}$), an ade un **segundo punto de inflexión.** Con estas variables e incluso con **potencias superiores,** pueden incluirse relaciones ma s complejas de las que son posibles explicar so lo con transformaciones. Por ejemplo, en un modelo de regresio n simple, un modelo curvilineal con un punto de giro puede modelizarse con la ecuacio n:

$$Y=b_0+b_1X_1+b_2X_1^2$$

Donde:

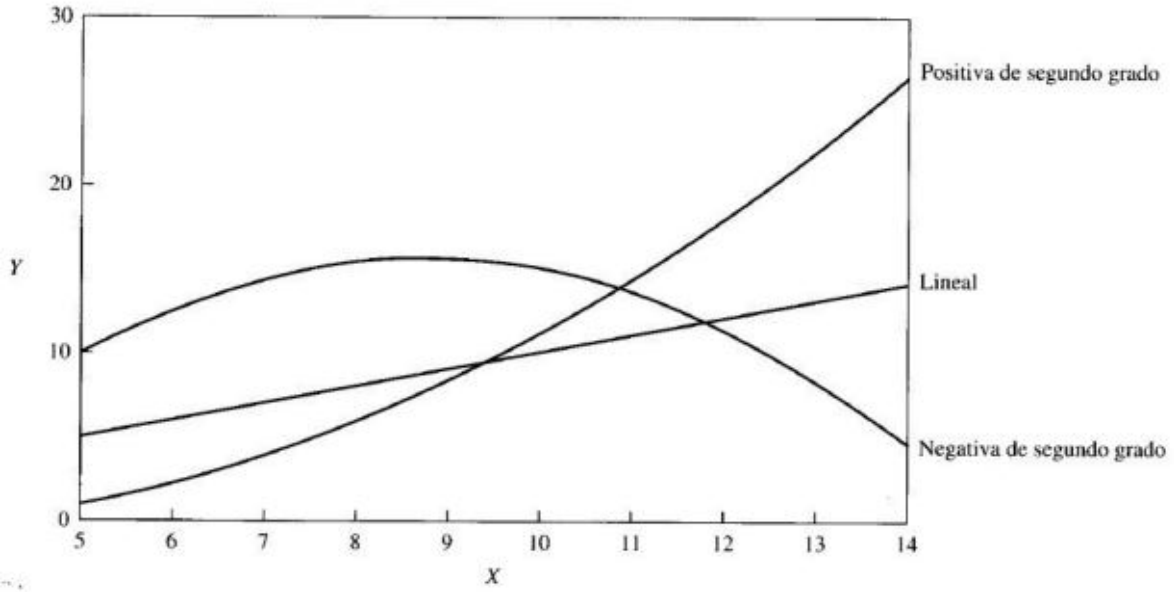
b_0**=constante**

b_1X_1**=efecto lineal de** X_1

$b_2X_1^2$**=efecto curvilineal de** X_1

Aunque puede añadirse cualquier número de componentes no lineales, el término cúbico es la mayor potencia utilizada habitualmente. A medida que cada nueva variable entra en la regresión, podemos realizar un test estadístico directo de los componentes no lineales que no podemos hacer con las transformaciones de los datos. En la Figura 5.31 se muestran tres relaciones (dos no lineales y una lineal).

Figura 5.31. Representación de las relaciones no lineales con polinomios.



A efectos interpretativos, **el término cuadrático positivo indica una curva en forma de U hacia arriba,** mientras que un **coeficiente negativo indica una relación con la n hacia abajo.** Los **polinomios multivariantes** se crean cuando la ecuación de la regresión **contiene dos o más variables independientes.** Seguimos el mismo procedimiento para crear los términos del polinomio al igual que antes, pero ahora debemos crear además un término adicional, el término de interacción $(\mathbf{X_1}, \mathbf{X_2})$, que se necesita para cada combinación de variables para representar completamente los efectos multivariantes. En **términos gráficos, un polinomio multivariante de dos variables se representa como una superficie con un pico o un valle.** Para polinomios de orden superior, es mejor hacer la interpretación dibujando la superficie desde los valores previstos. **¿Cuántos términos deberían añadirse?** Una práctica común es **empezar con el componente lineal** y entonces **añadir secuencialmente polinomios de orden superior** hasta que se llegue a la **no significación.** El uso de los polinomios no está exento de problemas potenciales:

1. Cada término adicional requiere **un mayor grado de libertad**, que puede ser particularmente restrictivo con tamaños muestrales pequeños. Esto no ocurre con las trasformaciones de datos.
2. Se introduce la **multicolinealidad** en los términos adicionales y hace que la comprobación de la **significación** estadística de los términos de los polinomios sea **inapropiada.**
3. En su lugar el investigador tiene que comparar los valores $\boldsymbol{R^2}$ del modelo de ecuación con los términos lineales con la $\boldsymbol{R^2}$ para la ecuación con los términos de los polinomios.
4. Las pruebas para la significación estadística del $\boldsymbol{R^2}$ en aumento representan la manera apropiada de valorar el impacto de los polinomios.

5.15.8. Representación de la interacción o efectos moderadores

Las **relaciones no lineales** arriba mencionadas requieren la creación de una variable adicional (por ejemplo, el término al cuadrado) para representar un **cambio de pendiente en la relación sobre el rango de la variable independiente**. Ésta se centra en la relación entre una única **variable independiente y la variable dependiente**. Pero ¿**qué ocurre si una relación de variable dependiente/independiente se ve afectada por otra variable independiente?** A esto se le llama el **efecto moderador**, que ocurre cuando la variable moderador, una segunda variable independiente. Cambia la forma de la relación entre otra variable independiente y la variable criterio. También se conoce como un **efecto interacción** y es similar al término interacción que se encuentra en el **análisis de la varianza y el análisis multivariante de la varianza**.

El efecto moderador más comúnmente empleado en la regresión múltiple es el **moderador cuasi o bilineal**, donde la pendiente de la relación de una variable independiente ($\mathbf{X_1}$) cambia junto con los valores de la variable moderador ($\mathbf{X_2}$) **[**Jaccard, J. et al1990**].**

En nuestro ejemplo anterior del uso de tarjetas de crédito, supongamos que la renta familiar ($\mathbf{X_2}$) se encontró que era un **moderador positivo** de la relación entre el tamaño de la familia ($\mathbf{X_1}$) y el uso de tarjetas de crédito (**Y**). Esto significaría que el cambio esperado en el uso de tarjetas de crédito basado en el tamaño de la familia ($\boldsymbol{b_1}$, el coeficiente de regresión de $\mathbf{X_1}$) podría ser menor para familias con rentas bajas y mayor para familias con rentas altas. **Sin el efecto moderador,** suponemos que el tamaño de la familia tiene un efecto "***constante***" sobre el número de tarjetas de crédito utilizadas. Pero **los términos de interacción** nos dicen que esta relación cambia, según el nivel de renta de la familia. Nótese que esto no significa necesariamente que los efectos del tamaño o la renta familiar por sí mismos no sean importantes, sino que **el término de interacción complementa su explicación del uso de tarjetas de crédito**. El **efecto moderador** se representa en la regresión múltiple por un término bastante similar a los polinomios descritos previamente para representar **efectos no lineales**. El término moderador es una variable compuesta formada por la multiplicación de $\mathbf{X_1}$ por el moderador $\mathbf{X_2}$, que entra en la ecuación de regresión. De hecho, **el término no lineal** puede ser **considerado como una forma de interacción**, donde la variable independiente se "***modera***" a sí misma, elevándose al cuadrado ($\boldsymbol{X_1 X_1}$). La relación moderadora se representa como:

$$\boldsymbol{Y=b_0X_1+b_1X_1+b_2X_2+b_3X_1X_2}$$

Donde:

$\boldsymbol{b_0}$**=** constante

$\boldsymbol{b_1X_1}$**=** efecto lineal de $\boldsymbol{X_1}$

$\boldsymbol{b_2X_2}$**=** efecto lineal de $\boldsymbol{X_2}$

$\boldsymbol{b_3X_1X_2}$**=** efecto moderador de $\boldsymbol{X_2}$ *sobre* $\boldsymbol{X_1}$**.**

Dada la multicolinealidad entre las variables antiguas y nuevas, se emplea un enfoque parecido a la comprobación para la significación de efectos polinomiales (no lineales). Para determinar si el **efecto moderador es significativo,** deberá estimar:

1. La ecuación original (**sin moderar**) y a continuación estima la relación moderada. Si el cambio en el $\boldsymbol{R^2}$ es **estadísticamente significativo**, entonces nos hallamos en presencia de un efecto **moderador significativo.** Por lo que solamente se valora el efecto de incremento, no las variables individuales.

2. La interpretación de los coeficientes de regresión cambia ligeramente en las relaciones moderadas. El coeficiente b_3 el **efecto moderador**, indica que el cambio unitario en el efecto de $\mathbf{X_1}$ cuando $\mathbf{X_2}$ cambia.
3. Los coeficientes ***b1*** y ***b2*** representan los efectos de $\mathbf{X_1}$ y $\mathbf{X_2}$, respectivamente, cuando el resto de las **variables independientes es cero**.
4. En la relación **sin moderar**, el coeficiente b_1 presenta el efecto de $\mathbf{X_1}$ para todos los niveles de $\mathbf{X_2}$ y viceversa para b_2. Por tanto, en la regresión sin moderar, los coeficientes de regresión b_1 y b_2 se "***promedian***" respecto de los niveles del resto de las otras variables independientes
5. Una **relación moderada** se **separan del resto de las variables independientes.**
6. Para determinar el efecto total de una variable independiente, **se deben combinar los efectos separados y moderados**.
7. El efecto total conjunto de $\mathbf{X_1}$ para cualquier valor de $\mathbf{X_2}$ se puede calcular sustituyendo el valor de $\mathbf{X_2}$ en lo siguiente:

$$b\ (Total)=b_1+b_3X_2$$

Por ejemplo, supongamos una regresión moderada que se resuelve en los siguientes coeficientes: b_1**= 2,0** y b_3**= 0,5**. Si el valor de X_2 puede tomar valores entre uno (**1**) y siete (**7**), puede **calcular el efecto total de** $\mathbf{X_1}$ para cualquier valor de $\mathbf{X_2}$. Cuando $\mathbf{X_2}$ **=3**, el efecto total de $\mathbf{X_1}$ es **3.5 [2.0 + 0.5(3)]**. Cuando **X=7**, el **efecto total de** $\mathbf{X_1}$ es ahora **5.5 [2.0 + 0.5 (7)]**. Se puede apreciar **al efecto moderador**, haciendo que la relación de $\mathbf{X_1}$ y la variable criterio cambie, dado el nivel de $\mathbf{X_2}$.

Para ver más sobre relaciones moderadas consulte **[**Cohen, 1983**,** Jaccard et al1990**].**

La creación de nuevas variables le proporcionará una flexibilidad enorme en la representación de relaciones dentro de los modelos de regresión. No obstante, **demasiadas** veces el deseo de obtener un ajuste mejor de un modelo **lleva a la inclusión de estas relaciones especiales sin apoyo teórico.** En estos casos, el investigador corre un riesgo mayor de encontrar resultados con poca o ninguna generalización. En su lugar, con el empleo de estas **variables adicionales**, el investigador tiene que **estar guiado por una teoría respaldada por el análisis empírico.** De esta manera, se puede alcanzar tanto la **significación práctica como estadística.**

5.16. Regresión lineal múltiple: Supuestos

Paso 3: supuestos de aplicabilidad

Se Ha mostrado cómo son posibles mejoras en la predicción de la variable criterio añadiendo variables independientes e incluso transformándolas para representar aspectos de las relación que no son lineales.

Pero para hacerlo debemos hacer varios s**upuestos sobre las relaciones entre las variables dependientes e independientes** que afectan al procedimiento estadístico (mínimos cuadrados) utilizado para la regresión múltiple.

En las siguientes secciones se discutirá la contrastación de los supuestos y las acciones correctivas que se deben tomar si se incumplen los resultados

5.16.1. Valoración de las variables individuales frente al valor teórico

Los supuestos subyacentes de la regresión múltiple se aplican tanto a las variables individuales (dependientes e independientes) como a la relación global. En el **Capítulo 3** se examinaron los métodos disponibles para evaluar los supuestos de las variables individuales. Pero en la **regresión múltiple**, una vez que se ha **calculado el valor teórico**, **actúan colectivamente en la predicción de la variable dependiente**. Este hecho implica que se deben evaluar los supuestos no sólo de las variables individuales sino del valor teórico en sí mismo. Esta sección se **centra en el examen del valor teórico y de su relación con la variable dependiente** en el cumplimiento de los su puestos de la regresión múltiple. Estos análisis deben realizarse después de que se haya estimado el modelo de regresión en el **paso cuarto**. Por tanto, la **contrastación** de los supuestos debe tener lugar no sólo en las fases iniciales de la regresión sino también después de que el modelo se ha estimado. La cuestión básica es si, en el proceso de cálculo de los coeficientes de regresión y predicción del variable criterio, se cumplen los supuestos del análisis de la regresión. **¿Son los errores en la predicción un resultado de una falta efectiva de relación entre variables o son provocados por ciertas características de los datos no contemplados por el modelo de regresión?**

Los supuestos que se van a examinar son los siguientes:

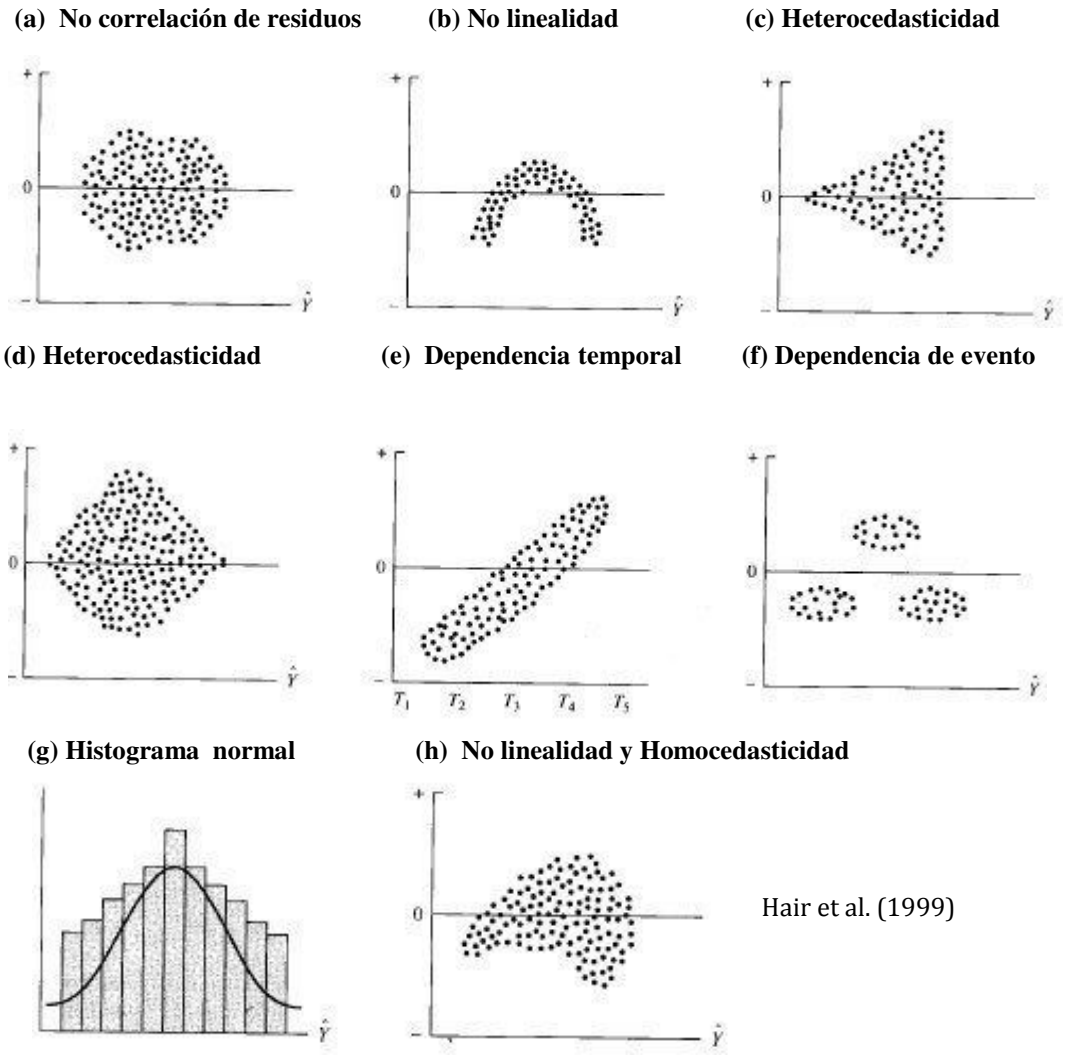
- La linealidad del fenómeno medido
- La varianza constante del término de error
- La independencia de los términos de error
- La normalidad de la distribución del término de error

La medida principal del error de predicción del valor teórico es el residuo (la diferencia entre los valores observados y las predicciones de la variable criterio). **Los gráficos de residuos y de las variables independientes** o de las predicciones constituyen el método básico de **identificación de los incumplimientos de los supuestos** para el conjunto de la relación. Cuando se examinan los **residuos**, se recomienda cierta forma de **estandarización**, con el fin de hacer los residuos directamente comparables. (En su forma original, los valores con sobre predicción tienen mayores residuos.)

El método más ampliamente utilizado es el residuo basado en la *t de Student*. Su valor corresponde a valores *t*. Esta correspondencia facilita la evaluación de la significación estadística de residuos particularmente elevados. Existen también unas series de test estadísticos que pueden **complementar** el examen de los gráficos residuos.

El gráfico de residuos más habitual se forma con los **residuos (r_1) frente a los valores de la predicción de la dependiente (Y_1).** Para un modelo de **regresión simple**, se pueden trazar los **residuos** respecto de las **variables dependientes o independientes**, dado que están **directamente relacionados**. En la **regresión múltiple**, sin embargo, **sólo los valores dependientes representan el efecto total del valor teórico de regresión**. Por tanto, a no ser que el análisis dual pretenda concentrarse en una sola variable, se usan las **variables dependientes pronosticadas.** Los incumplimientos de cada supuesto pueden identificarse por **específicas de los residuos. La Figura 5.32** contiene unos cuantos gráficos de residuos que muestran los puestos básicos discutidos en las secciones siguientes. Un gráfico de especial interés es el de **no-correlación de residuos** (**Figura 5.32a**), el gráfico de los residuos **cuando se cumplen los supuestos**. Los gráficos de no-correlación de los residuos **se distribuyen aleatoriamente**, una **dispersión relativamente igual a cero** y una tendencia no muy fuerte a que sea **mayor o menor que cero.** Asimismo, **no se encuentra ninguna pauta o regularidad para valores elevados o reducidos de la variable independiente**. Los gráficos residuales restantes serán utilizados ilustrar los métodos de examen de incumplimientos de los supuestos que subyacen el análisis regresión.

Figura 5.32. Análisis gráfico de los residuos



Hair et al. (1999)

5.16.2. Linealidad del fenómeno

Es la relación entre variables dependientes o independientes que representa **el grado de cambio en la variable dependiente asociado con la variable independiente**. El **coeficiente de regresión es constante** a lo largo del rango de valores de la variable independiente. **El concepto de correlación está basado en una relación linea**l, siendo por tanto un supuesto cr tico del ana lisis de regresio n. La linealidad se puede examinar fa cilmente en los **gráficos de residuos**. La **Figura 5.32b** muestra una forma típica de residuos que indican la existencia de una **relación no lineal** no representada en el modelo habitual. **Cualquier modelo curvilineal de los residuos indica que la acción correctiva aumentará tanto la precisión predictiva del modelo como la validez de los coeficientes estimados.** Las soluciones mediante transformaciones de los datos fueron expuestas anteriormente (**Capítulo 3 y ver** Mosteller& Tukey 1977**).** Así, Usted puede tambie n querer incluir relaciones no lineales en los modelos de regresión. En el **segundo paso** se presentan las transformaciones de datos, tales como la creacio n de **términos polinomiales**, o bien me todos espec ficos como la **regresión no lineal** que pueden tratar los **efectos curvilineales** de las variables independientes o relaciones no lineales ma s complejas.

En la **regresión múltiple con más de una variable independiente**, el **examen de los residuos** mostraría los efectos combinados de todas las **variables independientes**, pero **no podemos examinar el efecto de cualquier variable independiente separadamente en un gráfico de residuos**. Para hacerlo utilizamos lo que se denomina **gráfico de regresión parcial**, que muestra la relación de una **única variable independiente con relación a la variable dependiente**. **Difiere del gráfico de residuos** que acabamos de discutir en que **la línea** que atraviesa la nube de puntos, que era horizontal en la **Figura 5.32.** Análisis gráfico de los residuos, **tendrá ahora pendiente positiva o negativa** dependiendo de si el coeficiente de regresión para esa variable independiente es positivo o negativo. El examen de los residuos alrededor de esta línea se hace exactamente igual que antes. En los **gráficos de regresión parciales**, las pautas **curvilineales** de los residuos indican **una relación no lineal** entre una **variable independiente y la variable dependiente**.

Este es el método más útil **cuando tenemos varias variables independientes**, en la medida en que podemos decir qué variables específicas **incumplen el supuesto de linealidad** y aplicar los remedios necesarios. También se facilita la identificación de **atípicos** o de **observaciones influyentes** sobre la base del uso de una sola variable independiente a la vez.

5.16.3. Varianza constante del término de error

La presencia de varianzas desiguales (**heterocedasticidad**) es uno de los supuestos que se **incumple** más habitualmente. El **diagnóstico** se realiza mediante **gráficos de residuos** o test estadísticos simples. El **gráfico de los residuos** (**basados en la *t de Student***) frente a los **valores de la variable dependiente** se compone con el gráfico de **no-correlación de residuos** (vea **Figura 5.32a**) y muestra una **forma consistente si la varianza no es constante**.

Quizá **la forma más común es la triangular**, en cualquier dirección (**Figura 5.32c**). Puede esperarse una **forma de diamante** (**Figura 5.32c**) en el caso donde se espera más variación en los valores intermedios que en los extremos. En la mayoría de las ocasiones, **muchos incumplimientos ocurren simultáneamente**, tal como **la no-linealidad y la heterocedasticidad** mostrada en la **Figura 5.32h.** Las soluciones para cada uno de los problemas de incumplimiento a menudo corrigen otros problemas.
Cada software estadístico de datos, tiene test estadísticos de heterocedasticidad, por ejemplo, el **SPSS** ofrece el **test de *Levene* de homogeneidad de varianza, que mide la igualdad de varianzas para un único par de varianzas.** Su uso es particularmente recomendable porque es el que **menos queda afectado por desviaciones de la normalidad,** otro de los problemas que ocurren con frecuencia en la regresión. Si existe **heterocedasticidad**, podemos aplicar **2** soluciones:

1. Si el incumplimiento puede atribuirse a una **única variable criterio**, puede utilizarse el procedimiento de **mínimos cuadrados ponderados**.
2. Más directas y fáciles son, sin embargo, diversas transformaciones de estabilización de varianza discutidas en el **Capítulo 3**, que permiten usar las variables transformadas en el modelo de regresión.

5.16.4. Independencia de los términos de error

En la **regresión** suponemos que cada **variable predictor es independiente**. Con esto queremos decir que el valor de **la predicción no está relacionado con cualquier otra predicción**; esto es, no están ordenadas por otra variable. Para identificar este hecho, se utiliza **el gráfico de residuos** respecto a cualquier posible **varianza secuencial**. Si los **residuos son independientes**, la forma puede **parecer aleatoria y similar al gráfico de no-correlación de los residuos**.
Los **incumplimientos** quedarán identificados por una forma consistente de los **residuos**:

1. La **Figura 5.32e** representa **un gráfico de residuos** que muestra una asociación entre los **residuos y tiempo**, una variable de secuencia habitual.
2. Otra forma habitual se muestra en la **Figura 5.32f**. Esta forma ocurre cuando las condiciones básicas del modelo cambian pero no se incluyen en el modelo. **Por ejemplo**, las ventas de trajes de baño se miden mensualmente en 12 meses, con 2 estaciones invernales frente a una única estación de verano, sin estimarse un indicador estacional. La forma de los residuos mostrará **residuos negativos** para los meses de **invierno** frente a **residuos positivos** para los meses de **verano.**
3. Las transformaciones de los datos, tales como primeras diferencias en un modelo de series temporales, inclusión de variables indicador o modelos de regresión especialmente formulados pueden solucionar este problema.

5.16.5. Normalidad de la distribución del término de error

Quizá **el incumplimiento de supuestos más frecuente** es **la no-normalidad de las variables independientes y dependientes, o ambos [**Seer, 1984**]**. El **diagnóstico** se puede realizar:

1. El más simple para el conjunto de variables predictor en la ecuación es un **histograma de residuos**, donde se puede **comprobar visualmente si la distribución se aproxima a la normal** (ver **Figura 5.32g**). Aunque **atractivo por su simplicidad**, este método es

particularmente **difícil en muestras pequeñas**, donde la **distribución** puede estar **deformada.**

2. Un método mejor es utilizar los **gráficos de probabilidad normal**. Difieren de los **gráficos de residuos** en que los **residuos estandarizados** se comparan con la **distribución normal.** La **distribución normal** traza una **línea recta diagonal y los gráficos de residuos se comparan con la diagonal**. Si una distribución es **normal**, **la línea de residuos seguirá de cerca la diagonal**. Con el mismo procedimiento se pueden comparar las variables dependientes e independientes separada mente respecto de la **distribución normal [**Daniel % Wood 1980**].** El **Capítulo 3** da una mayor discusión de la interpretación de gráficos de **probabilidad normal.**

El **análisis de residuos,** bien con los **gráficos de residuos** o bien con **test estadísticos**, proporciona un **conjunto simple pero potente** de instrumentos analíticos para examinar la conveniencia del **modelo de regresión**. Frecuentemente, sin embargo, **estos análisis no se realizan y se dejan intactos los incumplimientos de los supuestos.** Lo anterior, provoca:

1. Los usuarios **no son conscientes de las potenciales imprecisiones que puedan presentar.**
2. Éstas van desde **test de significación de coeficientes inapropiados** (**por mostrar significación cuando no está presente o viceversa**)
3. **Predicciones sesgadas e imprecisas de la variable dependiente**.

Es altamente recomendable fehacientemente que estos métodos se apliquen para conjunto de datos del modelo de regresión. La aplicación de los remedios, particularmente la transformación de los datos, aumentará la confianza en las interpretaciones y las predicciones de la regresión múltiple.

5.17. Regresión lineal múltiple: Estimación y valoración

Paso 4: estimación y ajuste

Hasta aquí, se ha realizado:

1. Especificación los **objetivos del análisis de regresión**
2. **Selección de las variables dependientes e independientes**,
3. Confrontación de los **resultados del diseño de la investigación** y
4. Evaluación de las variables a la hora de **cumplir los supuestos** de la regresión,

Por lo que se encuentra preparado para **estimar el modelo de regresión y evaluar la precisión predictiva conjunta de las variables independientes** (vea **Figura 5.8**). A este nivel, se deben lograr **3** tareas básicas:

1. **Seleccionar un método** para especificar el modelo de regresión a estimar,
2. **Evaluar la significación estadística** del modelo con junto en la predicción de la variable criterio, y
3. Determinar si cualquiera de las observaciones ejerce una **indebida influencia sobre los resultados.**

5.17.1. Aproximaciones generales a la selección de variables

En la mayoría de los casos de regresión múltiple, se tiene un número posible de **variables independientes** entre las **cuales elegir** para incluirlas en la ecuación de regresión. A veces el conjunto de **variables independientes** puede estar muy definido y el modelo de

regresión se usa esencialmente en una **aproximación confirmatoria**. **En otros casos**, puede **desear elegir entre el conjunto de variables independientes.** Existen varias aproximaciones (**métodos de búsqueda secuencial y procesos combinatorios**) para ayudarlo en la búsqueda del **"*mejor*"** modelo de regresión. A continuación se discutirá cada uno de estos modelos de especificación del modelo de regresión.

5.17.2. Especificación confirmatoria

La más simple aunque quizá más exigente aproximación de especificación del modelo de regresión es emplear ésta perspectiva, en la cual deberá **especificar por completo el conjunto de variables independientes a incluir**. En comparación con otras aproximaciones, **se tiene control total sobre la selección de la variable**. Aunque la **especificación confirmatoria** es simple en su concepción, **debe estar seguros de que el conjunto de variables consigue la máxima predicción mientras mantiene un modelo de parsimonia.** Las pautas generales para el desarrollo del modelo se incluyen en todas las técnicas de análisis multivariante.

5.17.3. Métodos de búsqueda secuencial

Estos métodos tienen en común la aproximación general de estimación de las ecuaciones de regresión con un conjunto de variables y a continuación añadir o eliminar se lectivamente variables hasta que se consiga alguna medida criterio conjunta.

Esta aproximación proporciona un método objetivo de selección de variables que **maximizan la predicción con el número más pequeño de variables empleadas.**

Existen **2** tipos de aproximaciones de búsqueda secuencial:

1. **La estimación por etapas, y**
2. **La eliminación progresiva y regresiva.**

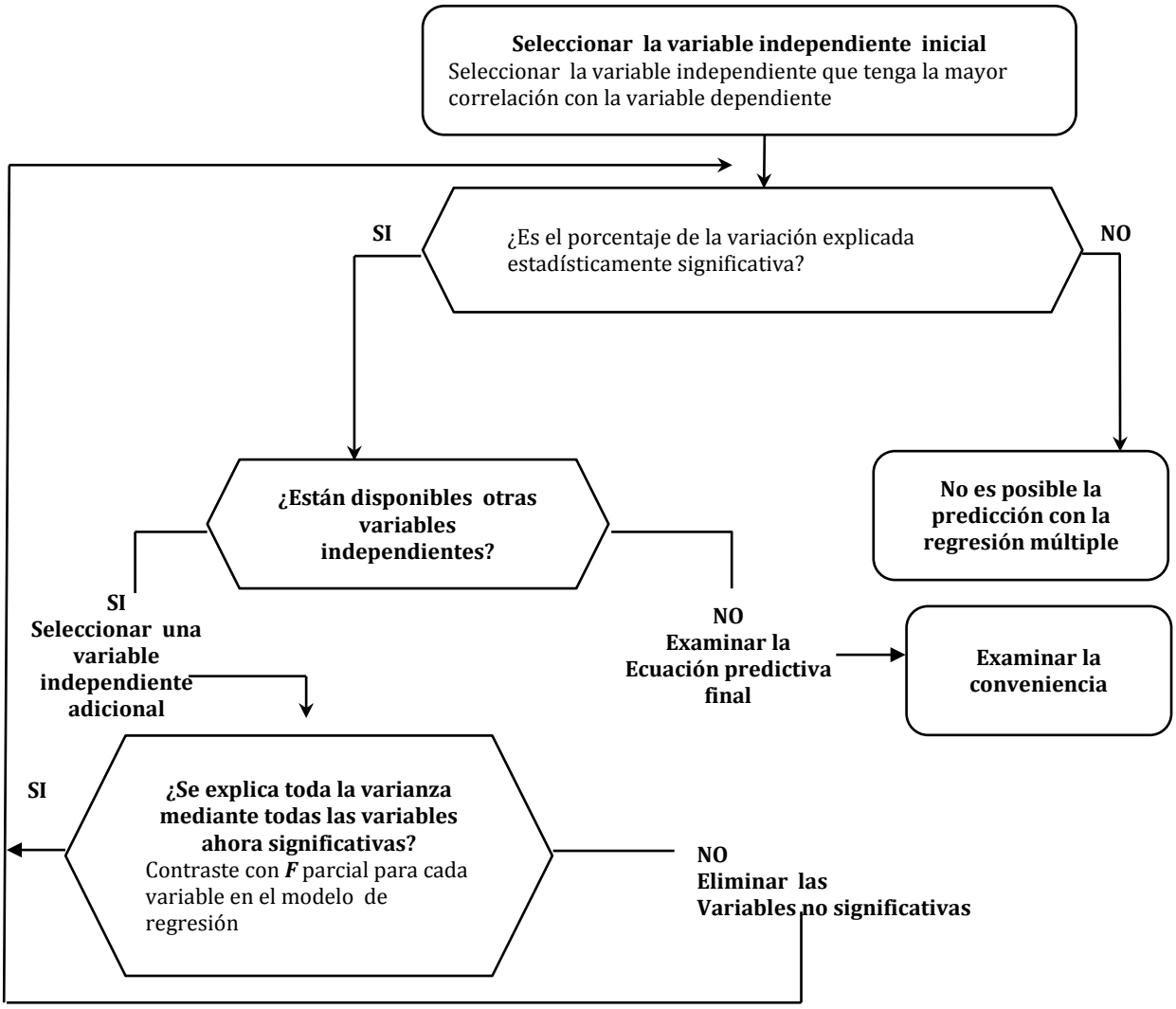
En cada aproximación, se valoran las variables individualmente en función de su contribución a la predicción de la variable dependiente y se añaden o eliminan según su contribución relativa.

5.17.4. Estimación por etapas (paso a paso o *Stepwise*)

Es quizá la aproximación más popular para seleccionar variables. Esta aproximación **permite examinar la contribución de cada variable predictor** al modelo de regresión. Se considera **la inclusión de cada variable antes de desarrollar la ecuación.** Se añade primero la variable independiente con la **contribución más grande.** Las variables independientes se seleccionan entonces para la inclusión basada en su **contribución incremental sobre la(s) variable(s) ya existente(s) en la ecuación.** Se ilustra el procedimiento por etapas en la **Figura 5.33**. Las cuestiones específicas en cada etapa son:

1. Empezar con el modelo de **regresión simple** en el cual sólo se utiliza la única variable predictor que es la que se muestra más altamente correlacionada con la variable criterio. La ecuación sería: $Y= b_0+ b_1X_1$.
2. Examinar los coeficientes de correlación parcial para **encontrar una variable predictor adicional que explique además de una parte significativa,** la mayor parte del error que queda de la primera ecuación de regresión.

Figura 5.33 Diagrama de flujos del procedimiento de estimación por etapas.



Fuente: Hair et al. (1999)

3. **Recalcular** la ecuación de regresión utilizando las **dos variables predictor**, y **examinar el valor parcial *F*** de la variable original del modelo para ver si todavía realiza una **contribución significativa**, dada la presencia de la **nueva variable predictor. Si no lo hace, eliminamos la variable**. Esta capacidad de eliminar variables presentes en el modelo **distingue el modelo por etapas de los modelos de adición progresiva/eliminación regresiva**. Si las variables originales todavía representan una contribución significativa, la ecuación sería: $Y= b_0 + b_1X_1+bx_2$
4. Continúe este procedimiento **examinando todas las variables independientes no presentes en el modelo** para determinar si deberían incluirse en la ecuación. **Si se incluye una nueva variable independiente**, hay que examinar todos los predictores previamente incluidos en el modelo para juzgar si se deben mantener. **Existe un sesgo**

potencial en el procedimiento por etapas que resulta de considerar sólo una variable a seleccionar cada vez. Supongamos que las variables $\mathbf{X_3}$ y $\mathbf{X_4}$ explicaran conjuntamente una parte significativa de la varianza (**cada una considerando la presencia de la otra**), **pero no son significativas por sí solas**. **Nota: En esta situación, ninguna debería ser considerada para el modelo final.**

5.17.5. La adición progresiva (*Forward*) y la eliminación regresiva (*Backward*)

La adición progresiva y la eliminación regresiva son fundamentalmente procesos de **ensayo y error** para buscar los mejores estimadores de la regresión. **El modelo de adición progresiva** es **similar** al procedimiento por etapas arriba explicada, mientras que el **procedimiento de eliminación regresiva** implica calcular una ecuación de regresión con todas las variables independientes, para a continuación **ir eliminando** las variables independientes que no contribuyan significativamente.

La distinción principal de la **aproximación por etapas** respecto de los procedimientos de **adición progresiva y eliminación regresiva** es su **capacidad de añadir o eliminar las variables en cada etapa.** Una vez que se añade o elimina una variable en los esquemas de adición progresiva o eliminación **regresiva**, **no existe posibilidad de revertir la acción posteriormente.**

5.17.6. Aproximaciones generales a la selección de variables

Advertencias sobre los métodos de búsqueda secuenciales

Deberá estar consciente **de 2 advertencias** cuando se usa cualquier procedimiento de búsqueda secuencial:

1. La **multicolinealidad** entre **variables independientes** puede tener un impacto sustancial sobre la especificación final del modelo. Examinemos esta situación con dos variables altamente correlacionadas que tienen similares correlaciones con la variable independiente. El criterio de **inclusión o eliminación** en estas aproximaciones es **maximizar el incremento de potencia predictiva de la variable adicional.** Si una de estas variables entra en el modelo de regresión, es muy probable que la otra variable también entre, dado que estas variables están altamente correlacionadas **y existe poca varianza singular** para cada variable por separado (véase más sobre la **multicolinealidad**).Por esta razón, se deben evaluar los efectos de la **multicolinealidad** en la interpretación del modelo y **examinar las correlaciones directas de todas las variables independientes potenciales.** Esto ayudará a **evitar concluir que las variables independientes que no entren en el modelo no sean trascendentes cuando en realidad están altamente relacionadas con la variable dependiente, pero también correlacionadas con las variables ya existentes en el modelo.** Aunque las aproximaciones de búsqueda secuencial maximizarán la capacidad predictiva del modelo de regresión, debe ser cuidadoso en la interpretación del modelo.
2. Principalmente al procedimiento por etapas. En esta aproximación, **los test de significación múltiple se realizan en el proceso de estimación del modelo**. Para asegurar que la tasa de error conjunto a lo largo de todos los test de significación es razonable, por lo que **deberá emplear umbrales muy conservadores (por ejemplo, 0.01) al añadir o destruir las variables.**

5.17.7. Métodos combinatorios

Son fundamentalmente un **proceso de búsqueda generalizada** a lo largo de todas las **combinaciones posibles de variables independientes.** El procedimiento más conocido es la **regresión parcial combinando variables**. Se examinan todas las **combinaciones posibles** de las **variables independientes para identificar el conjunto de variables que mejor se ajusta**. **Por ejemplo**, en un modelo con **diez variables** independientes, existen **1.204 regresiones** posibles (**1 ecuación** con una única constante, **10 ecuaciones** con una única variable independiente, **45 ecuaciones** con todas las combinaciones posibles de dos variables, etc.). Con algoritmos especializados, este proceso se puede gestionar incluso para problemas muy grandes, **identificando la mejor ecuación de regresión conjunta para cualquier número de medidas de ajuste predictivo**. Deberá recordar que supuestos tales como la **multicolinealidad**, la identificación de **atípicos** y **observaciones influyentes** así como la **interpretación de los resultados no están orientadas a la selección del modelo final**.

Cuando se han considerado estos supuestos, la **"*mejor*"** ecuación puede tener **problemas serios que afecten a su conveniencia**, pudiendo ser elegido en última instancia otro modelo.

5.17.8. Perspectiva de las aproximaciones de la selección de modelos.

Independientemente de que se elija un **método combinacional**, de **búsqueda secuencial o confirmatorio,** el criterio más importante **es el conocimiento sustantivo del investigador de la situación**, que es lo que determina las variables que se van incluir así como los signos esperados y la magnitud de sus coeficientes. Sin este conocimiento, la regresión resultante **puede tener una elevada precisión predictiva sin relevancia teórica o gerencial.**

El investigador **no debería guiarse completamente por estos métodos sino que en su lugar debería utilizarlos después de una cuidadosa consideración** de las aproximaciones alternativas para a continuación aceptar los resultados sólo después de un cuidadoso escrutinio.

5.17.9. Contrastación del cumplimiento de los supuestos de regresión.

Con las variables independientes seleccionadas y los coeficientes de regresión estimados, deberá ahora evaluar el modelo estimado a la hora de cumplir los **supuestos** subyacentes en la regresión múltiple.

Como se discutió en el **paso tercero**, las variables individuales deben cumplir los **supuestos de linealidad, varianza constante, independencia y normalidad**.

Además de las variables individuales, el valor teórico de la regresión debe también cumplir estos supuestos. Los **test de diagnósticos** expuestos en el **paso tercero** pueden aplicarse a la evaluación del **efecto colectivo del valor teórico** a través del **examen de los residuos**.

Si se encuentran incumplimientos sustanciales, el investigador debe tomar medidas correctivas para posteriormente volver a estimar el modelo de regresión.

5.17.10. Examen de la significación estadística del modelo

Si fuéramos a tomar muestras repetidas de ocho familias y preguntáramos cuántos miembros de la familia y tarjetas de crédito tienen, **rara vez** obtendríamos exactamente los mismos valores para $Y = b_0 + b_1 X_1$ para todas las muestras. Esperaríamos **variaciones al azar causadas por las diferencias entre las muestras**. **Generalmente tomaríamos sólo una muestra y basaríamos sobre ella nuestro modelo predictivo.**

Con una sola muestra y modelo de regresión, necesitamos **comprobar la hipótesis con relación a nuestro modelo predictivo** para asegurar que representa la población de todas las familias que tienen tarjetas de crédito en lugar de representar sólo a nuestra muestra de ocho familias.

Estos test pueden tomar una o dos formas básicas:

1. **Un test de varianza explicada (coeficiente de determinación), y**
2. **Los test de coeficientes**

5.17.11. Significación del modelo en su conjunto: el coeficiente de determinación.

Para contrastar la hipótesis de que la cantidad de variación explicada por el modelo de regresión es más que la variación explicada por la media (es decir, que $R^2 > 0$), se utiliza el **ratio F.** La prueba del **estadístico de la F** se define como:

$$\text{Ratio } F = \frac{\dfrac{\text{Suma de los errores al cuadrado}_{\text{regresión}}}{\text{Grados de libertad}_{\text{regresión}}}}{\dfrac{\text{Suma de los errores al cuadrado}_{\text{total}}}{\text{Grados de libertad}_{\text{residuos}}}} = \frac{SSE_{\text{regresión}}/df_{\text{regresión}}}{SSE_{\text{total}}/df_{\text{residual}}}$$

grados de libertad$_{\text{regresión}}$= Número de los coeficientes estimados (incluida la constante)-1.

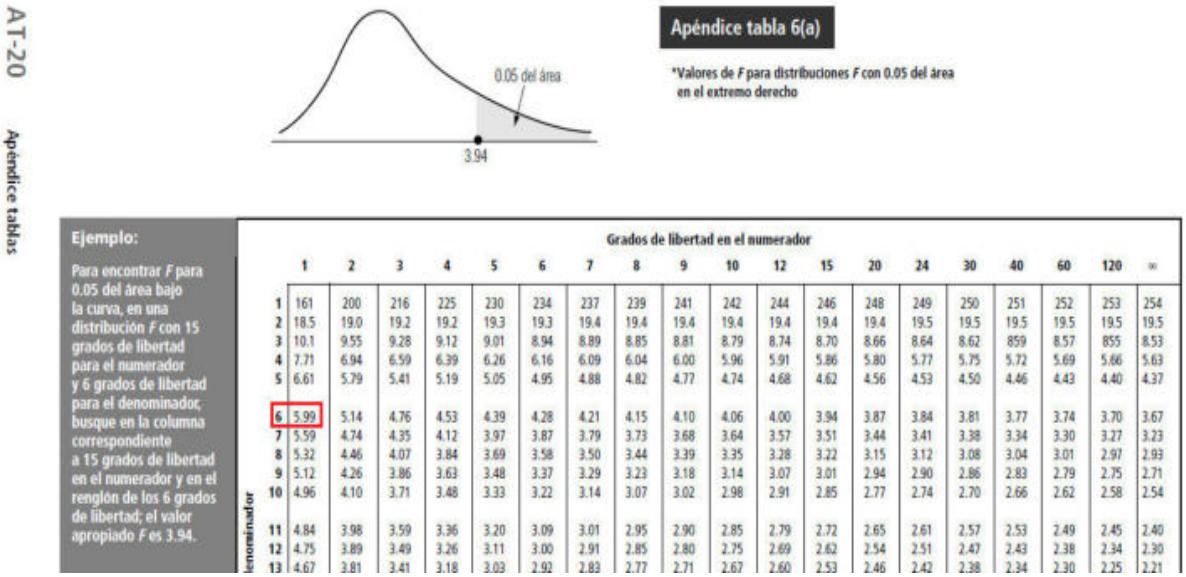
Grados de libertad$_{\text{residuos}}$=tamaño muestral – el número de los coeficientes estimados (incluida la constante).

Deben destacarse dos importantes características de este **ratio:**

1. **Cada suma de los cuadrados se divide por sus grados de libertad**. Es simplemente la varianza de los errores de predicción.
2. **Si el ratio de la varianza** explicada respecto a la varianza sobre la básica (alrededor de la media) **es alto**, el valor teórico de regresión tiene que tener un **valor significativo** en la explicación de la variable dependiente.

En el ejemplo, **el ratio *F*** del modelo de **regresión simple** discutido previamente en el capítulo es: **(16.5 /1) / (5.50 /6) = 18.0**. La tabla del **estadístico de la F de 1 con seis grados de libertad para un nivel de significación de 0.05 proporciona 5.99. Ver Figura 5.34.**

Dado que el **ratio *F* (es mayor que el valor de tabla (18.0>5.99), rechazamos** la hipótesis de que la reducción en el error que obtenemos utilizando el tamaño de la familia para predecir la posesión de tarjetas de crédito era un producto del azar. Este resultado significa que, considerando la muestra utilizada para la estimación, **podemos explicar 18 veces más variación que cuando utilizamos la media**, y que esto no es muy probable que ocurra **por azar** (**menos del 5 % de las veces**).



Apéndice tabla 6(a)

*Valores de *F* para distribuciones *F* con 0.05 del área en el extremo derecho

0.05 del área

3.94

Ejemplo:

Para encontrar *F* para 0.05 del área bajo la curva, en una distribución *F* con 15 grados de libertad para el numerador y 6 grados de libertad para el denominador, busque en la columna correspondiente a 15 grados de libertad en el numerador y en el renglón de los 6 grados de libertad; el valor apropiado *F* es 3.94.

Grados de libertad en el numerador

Denominador	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	859	8.57	855	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21

Figura 5.34 Valores *F* al 0.05

Fuente: Levin, R.,I.; Rubin, D.S. (2004). *Estadística para Administración y Economía*.7ª. Edición. México: Prentice-Hall

De la misma forma, el **ratio *F*** del modelo de **regresión múltiple con dos variables independientes (página 2)** es **(18.96/2)/(3.04/5)= 15.59.** El modelo de **regresión múltiple** es también estadísticamente significativo, indicando que la variable independiente adicional era sustancial al añadirse a la capacidad predictiva del modelo. Sabemos que R^2 está influencido por el número de variables predictor relativas al tamaño muestral. Se han propuesto varias reglas, que van desde **10 a 15 observaciones** por predictor a un mínimo absoluto de **4 observaciones por variable independiente.** A medida que se llega a estos límites, necesitamos ajustar la inflación del R^2 del "***sobreajuste***" de los datos. Como parte de todos los programas de regresión, se da un coeficiente de regresión ajustado **(R^2 ajustado)** junto con los coeficientes de determinación. Interpretado igualmente que el coeficiente de regresión sin ajustar, el **R^2 ajustado se hace más pequeño a medida que tenemos menos observaciones por variable independiente.** El **R^2 ajustado** es particularmente útil para **comparar las diferentes ecuaciones de regresión estimadas con variables independientes o diferentes tamaños muestrales,** dado que marca límites para el número específico de variables independientes y para el tamaño muestral sobre el que se basa cada modelo. En nuestro ejemplo del uso de tarjetas de crédito, el R^2 para el modelo de regresión simple es de **0.751** y el R^2 ajustado es **0.709.** Conforme añadimos la segunda variable independiente, el R^2 aumenta a **0.861,** pero el **R^2 ajustado sólo aumenta a 0.806.** En los dos casos, el **R^2 ajustado** refleja el ratio descendiente de los coeficientes estimados al tamaño muestral y **compensa un "*sobre ajuste*" de los datos.**

5.17.12. Test de significación de los coeficientes de regresión

La prueba de significación estadística de los coeficientes estimados del análisis de regresión es apropiada y necesaria **cuando el análisis se basa en una muestra de la población y no es un censo**. Cuando utilizamos una muestra para estimar el modelo de regresión, el investigador no está interesado, en la regresión estimada sólo para la muestra, **sino en la generalización de los resultados** para la población. **Para cada muestra extraída de la población, se obtendrá un valor diferente**. Para **muestras pequeñas**, los coeficientes estimados variarán ampliamente de muestra a muestra. Pero a medida que el tamaño **muestral aumenta**, las muestras se hacen más representativas de la población y **la variación en los coeficientes estimados para estas muestras mayores se espera que sean más pequeñas.** Esto es verdad hasta que se estima el análisis utilizando la **población.** En este caso, **no hay necesidad para la significación estadística** porque la **"*muestra*"** es igual a, y por tanto perfectamente representativa de la población. La variación esperada de los coeficientes estimados (tanto los coeficientes constantes como de regresión) se denomina **el error estándar de los coeficientes**. La significación estadística de los coeficientes de regresión proporciona una estimación probabilística de fundamento estadístico sobre si los coeficientes estimados a lo largo de un gran número de muestras de un cierto tamaño serán diferentes de cero. **Si el tamaño muestral es pequeño**, la variación puede ser muy grande como para decir con el necesario grado de certeza (**nivel de significación**) **que el coeficiente no es igual a cero**. Sin embargo**, si el tamaño muestra es grande**, el test tiene una mayor precisión porque la variación en los coeficientes es menor. **Muestras más grandes no garantizan que los coeficientes no sean iguales a cero, pero hacen los test más precisos.**

Como ejemplo suponga que se obtuvieron **20** muestras aleatorias de cada uno de los cuatro tamaños muestrales (**10, 25,50 y 100 encuestados**) de una gran base de datos. Se realizó una regresión simple para cada muestra y los coeficientes de regresión estimados se registran en la **Figura 5.35**.

Como se puede observar, la variación en los coeficientes estimados es mayor para muestras de l0 encuestados, que varían desde un coeficiente bajo **2.20 al más elevado de 6.06.** A medida que el **tamaño muestral aumenta de 25 a 50 encuestados, la variación de los coeficientes baja considerablemente**.

Finalmente, las muestras de **100 encuestados** tienen un rango de casi la mitad que las muestras de 10 encuestados (**2.10 vs. 3.86**). Con esto podemos ver que la capacidad de los **test de significación para de terminar si el coeficiente es realmente mayor que cero es más precisa con tamaños muestrales superiores.**

Figura 5.35. Variación muestral para los coeficientes de regresión estimada

	Tamaño muestral			
Muestra	**10**	**25**	**50**	**100**
1	2.58	2.52	2.97	3.60
2	2.45	2.81	2.91	3-70
3	2.20	3.73	3.58	3.88
4	6.06	5.64	5.00	4,20
5	2.59	4.00	4.08	3.16
6	5.06	3.08	3.89	3.68
7	4.68	2.66	3.07	2.80

8	6.00	4.12	3.65	4.58
9	3.91	4.05	4.62	3.34
10	3.04	3.04	3.68	3.32
11	3.74	3.45	4.04	3.48
12	5.20	4.19	4.43	3.23
13	5.82	4.68	5.20	3.68
14	2.23	3.77	3.99	4.30
15	5.17	4.88	4.76	4.90
16	3.69	3.09	4.02	3.75
17	3.17	3.14	2.91	3.17
18	2.63	3.55	3.72	3.44
19	3.49	5.02	5.85	4.31
20	4.57	3.61	5.12	4.21
Mínimo	2.20	2.52	2.91	2.80
Máximo	6.06	5.64	5.85	4.90
Alcance	3.86	3.12	2.94	2.10
Desviación STD	1.28	0.85	0.83	0.54

Fuente: Hair et al. (1999)

5.17.13. Contrastes de significación en el ejemplo de regresión simple

Cuando abordamos el modelo de regresión simple para el ejemplo de uso de tarjetas de crédito, dijimos que la ecuación de regresión para el número de tarjetas de crédito es $\boldsymbol{Y= b_0+ b_1X_1 = 2.87 + 0.971}$ (tamaño de la familia). Contrastaremos dos hipótesis para este modelo de regresión:

-Hipótesis l: **El valor de la constante 2,87 se debe al error muestral, siendo cero el término constante real apropiado para la población**.

Con esta hipótesis, estaríamos comprobando simplemente si el **término constante debería considerarse apropiado para nuestro modelo predictivo**. Si se encuentra que **no es significativamente distinto de cero**, supondríamos que el **término constante no se utilizaría para propósitos predictivos**.

El contraste apropiado es **el test de la *t***, que se encuentra habitualmente en la mayoría de los programas informáticos de regresión. **El valor *t* de un coeficiente es el coeficiente dividido por el error estándar**. **Por ejemplo**, un coeficiente de **2.5 con un error estándar de 0.5 tendría un valor *t* de 5.0**. Para determinar si el **coeficiente es diferente de cero** de forma significativa**, se compara el valor *t* calculado con el valor de la tabla para el tamaño muestral** y **el nivel de confianza seleccionado**.

Si nuestro valor es mayor que el valor de la tabla, podemos estar seguros (en nuestro nivel de confianza) que el coeficiente tiene un efecto estadístico significativo en el valor teórico de regresión.

Desde un punto de vista **práctico**, **este test rara vez es necesario**. Si los datos utilizados para desarrollar el modelo no incluyeron ciertas observaciones con todos los medios para cero, **el término constante está "*fuera*" de los datos** y actúa sólo para posicionar al modelo. **No será necesario entonces contrastar el modelo.**

-Hipótesis 2. El coeficiente **0.971** indica que **un aumento de una unidad en el tamaño familiar se asocia con un aumento en el número medio de tarjetas de crédito**

mantenidas por **0.971** y que este coeficiente **también difiere significativamente de cero.**
Si así ocurriera por un **error muestral**, **concluiríamos que el tamaño de la familia no tiene impacto sobre el número de tarjetas de crédito mantenidas**. Téngase en cuenta que **este no es un test de un cierto valor exacto del coeficiente sino más bien de si debiera utilizarse**. **Otra vez, el test apropiado es el test de la *t.***
Deberá recordar que el **test estadístico de los coeficientes de regresión** sirve para **asegurar que para todas las posibles muestras que pudiesen extraerse**, **e**l **coeficiente de regresión debería ser diferente de cero**.
En el ejemplo, el **error estándar del tamaño de familia** en el modelo de regresión simple es **0.229**. El valor calculado ***t*** es **4.25 (0.971/ 0.229**), lo cual tiene una probabilidad de **0.005**. Esto significa que podemos estar seguros con un alta nivel de exactitud (**99.5 por ciento**) **que el coeficiente debería ser incluido en la ecuación de regresión.**
Los **contrastes de significación de los coeficientes de regresión** proporcionan una valoración empírica de su **"*verdadero*"** impacto. Aunque **no constituye una prueba de validez,** determina si **los impactos** representados por los **coeficiente**s son generalizables para **otras muestras de esta población**. Muchas veces los investigadores **olvidan que los coeficientes estimados en el análisis de regresión son específicos para la muestra** utilizada en la estimación. Representan las mejores estimaciones para aquella muestra de observaciones, pero como muestran los citados resultados, **los coeficientes pueden variar notablemente de una muestra a otra.** Esto indica la necesidad de una actuación coordinada para validar cualquier análisis de regresión **sobre muestra(s) diferente(s).** De esta manera, el investigador tiene **que esperar que cambien los coeficientes**, pero el intento es para demostrar que **generalmente se mantiene la relación en otras muestras para que se pueda suponer que los resultados sean generalizables para cualquier muestra obtenida de la población.**

5.17.14. Identificación de observaciones influyentes

Hasta el momento, nos hemos centrado en la identificación de pautas generales en el conjunto de observaciones. Aquí desviaremos nuestra atención a las **observaciones individuales**, con el objetivo de **encontrar las observaciones que caen fuera de las pautas generales del conjunto de datos** o que **ejercen una fuerte influencia en los resultados de la regresión.** Recordemos que estas observaciones **no son necesariamente malas** en el sentido de que deban ser **omitidas.** En muchos casos representan los **elementos diferenciadores del conjunto de datos.** Sin embargo, **debemos identificarlas y evaluar su impacto antes de empezar.** La siguiente sección introduce el concepto de **observaciones influyentes y su impacto potencial** sobre los resultados de la regresión.
Las **observaciones influyentes** se clasifican en **3 casos:**

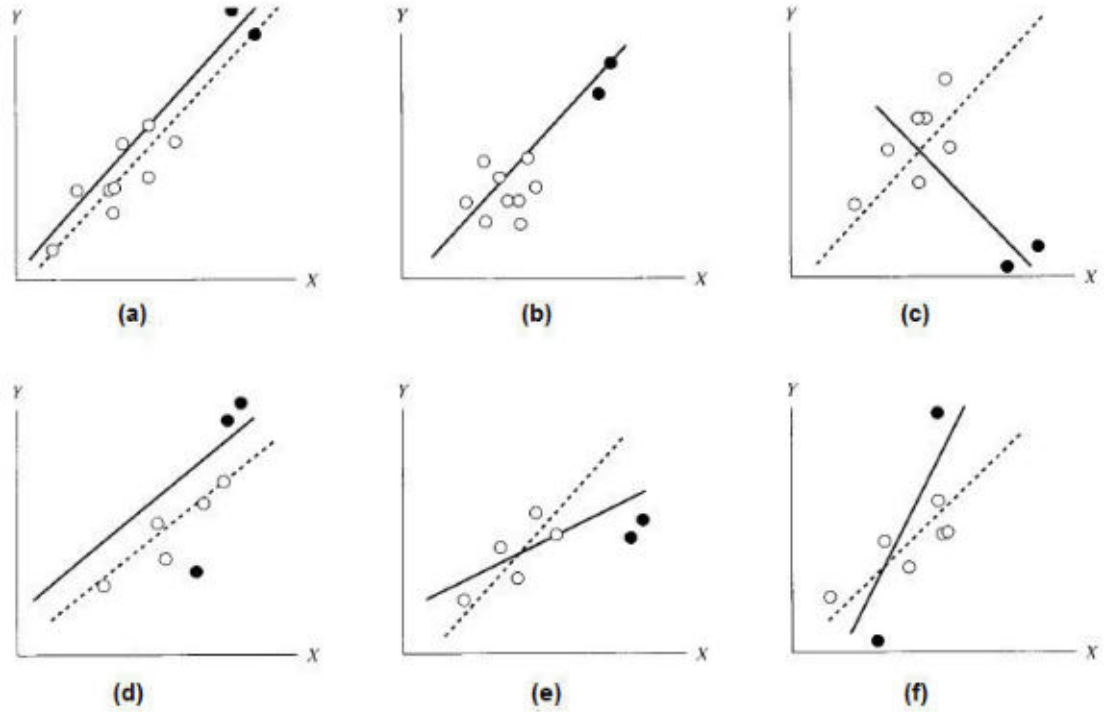
1. Atípicos,
2. Puntos de apalancamiento e
3. influyentes.

Los **atípicos** son observaciones que tienen **grandes valores residuales y pueden identificarse sólo con respecto a un modelo de regresión específico,** tradicionalmente la única forma de observación influyente considerada en los modelos de regresión, y se han desarrollado métodos de **regresión específicos** (**regresión robusta**) para tratar

específicamente con el impacto de atípicos sobre los resultados de la regresión. Los puntos de apalancamiento son observaciones diferentes del resto de las observaciones de los valores de las variables independientes. Su impacto es particularmente destacado en los coeficientes estimados de una o más variables predictor. Finalmente, las observaciones influyentes en sentido amplio, incluyen todas las observaciones que tienen un efecto desproporcionado sobre los resultados de la regresión. Las observaciones influyentes no sólo incluyen potencialmente atípicos y puntos de apalancamiento, sino que pueden incluir otras observaciones. Además, no todos los atípicos y puntos de apalancamiento son necesariamente observaciones influyentes.

Las observaciones influyentes pueden mostrar muchas formas. La **Figura 5.36** muestra varias formas de observaciones influyentes y su correspondencia con los residuos. En cada caso, el residuo de los puntos influyentes (la distancia perpendicular desde el punto a la línea de la regresión estimada) no se espera que sea muy grande como para clasificarse como atípico. Por tanto, centrándose sólo en grandes residuos, generalmente ignoraríamos estas observaciones influyentes adicionales. En la **Figura 5.36a**, el punto de influencia es **"*bueno*",** reforzando la pauta general de los datos y reduciendo el error estándar de la predicción y los coeficientes. Es un punto de apalancamiento pero tiene un valor residual pequeño o casi cero, en la medida en que es bien predicho por el modelo de regresión.

Figura 5.36. Formas de las observaciones influyentes



Notas:

Pendiente de la regresión sin influyentes

Observación típica

Pendiente de la regresión con influyentes

Observación influyente

Fuente: Hair et al. (1999)

Sin embargo, los puntos influyentes pueden tener también un efecto que es contrario a la pauta general de los datos restantes y sin embargo tener residuos pequeños (vea **Figura 5.36b y 5.36c**). En la **Figura 5.36b**, dos observaciones influyentes son las que cuentan casi por completo para la relación observada, dado que sin ellas no surge una pauta real del resto de los datos. No se identificarían tampoco si sólo se considerasen los residuos grandes, dado que su valor residual sería pequeño. En la **Figura 5.36c**, se ve un efecto incluso más profundo donde las observaciones influyentes contrarrestan la pauta general del resto de los datos. En este caso, los datos **"*reales*"** tendrían mayores residuos que los puntos influyentes **"*malos*".** Las observaciones influyentes pueden afectar sólo a una parte de los resultados, como en la **Figura 5.36d**, donde la pendiente permanece constante pero se desplaza la constante. Finalmente, los puntos influyentes múltiples pueden reforzar el mismo resultado. En la **Figura 5.36e**, dos puntos influyentes pueden tener la misma relación relativa, haciendo la detección más difícil. Y en la **5.36f**, los puntos influyentes tienen posiciones muy diferentes pero un efecto similar los resultados. Esto ejemplos ilustran que debemos desarrollar un instrumental superior de métodos para identificar estos casos influyentes

Los procedimientos para identificar todo tipo de observaciones influyentes son muy numerosos aunque están peor definidos que muchos otros aspectos del análisis de regresión. Todos los programas informáticos ofrecen un análisis de residuos de los que aquellos con mayores valores (particularmente los residuos estandarizados mayores que **2.0**) pueden identificarse fácilmente. Más aún, la mayoría de los programas informáticos ofrecen hoy en día al menos alguna medida de diagnóstico para la identificación de los puntos de apalancamiento y otras observaciones influyentes.

La necesidad de un estudio adicional de los puntos de apalancamiento y los influyentes se pone de manifiesto cuando vemos la sustancial medida en la que la generalización de los resultados y las conclusiones sustantivas (la importancia de las variables, nivel de ajuste, etc.) puede modificarse por un número relativamente pequeño de observaciones. Sean "***buenas***" (acentuando los resultados) o **"*malas*"** (cambiando sustancialmente los resultados), estas observaciones deben identificarse para evaluar su impacto. Influyentes, atípicos y puntos de apalancamiento se basan en alguna de las **4** condiciones siguientes:

1. Un error en la entrada de observaciones o datos.
2. Una observación válida aunque excepcional que es explicable por una situación extraordinaria.
3. Una observación excepcional sin una explicación plausible.
4. Una observación ordinaria en sus características individuales pero excepcionales en su combinación de características.

Pueden recomendarse varios cursos de acción para tratar con las observaciones influyentes de los diferentes tipos. Para un **error** en la observación se puede corregir el dato o eliminar la observación. Con observaciones válidas pero excepcionales (**condición 2**), está autorizada la eliminación del caso a menos que las variables que reflejan la situación extraordinaria se incluyan en la ecuación de regresión. La observación inexplicable (**condición 3**) presenta en un problema especial, dado que no existe razón para eliminar el caso, aunque su inclusión no puede justificarse. Finalmente, la observación que es ordinaria en una variable aunque excepcional en su combinación de características

(**condición 4**) indica modificaciones para la base conceptual del modelo de regresión y debería retenerse.
En todas las situaciones, se recomienda al investigador que elimine las observaciones verdaderamente excepcionales pero que esté en guardia contra la destrucción de observaciones que, aun que diferentes, sean representativas de la población. Recordemos que el objetivo es asegurar el modelo más representativo para la muestra de datos de tal forma que refleje de la mejor forma posible la población de la que se ha extraído. Esto se extiende incluso al mejor ajuste predictivo, dado que ciertos atípicos pueden ser casos válidos que el modelo intentaría predecir, aunque sea pobre mente. El investigador debería ser consciente de los casos donde los resultados cambiarían sustancialmente destruyendo sólo una observación aislada o un número muy pequeño de observaciones.

5.18. Regresión lineal múltiple: Interpretación

Paso 5: interpretación

La siguiente por realizar es **interpretar el valor teórico de la regresión** evaluando **los coeficientes de regresión** estimados para la explicación de la variable dependiente. Como veremos a lo largo de nuestra exposición, Usted deberá **evaluar no sólo el modelo de regresión que se estimó sino también el potencial de variables independientes que se omitieron si se empleó una aproximación combinatoria o de búsqueda secuencial.** En esas aproximaciones, la **multicolinealidad** puede afectar sustancialmente a las variables incluidas en última instancia en el valor teórico de la regresión. Por tanto, además de evaluar los coeficientes estimados, debe evaluar también **el impacto potencial de las variables omitidas** para asegurar que la **significación práctica se evalúa a la vez que la significación estadística.**

5.18.1. Utilización de los coeficientes de regresión

Los coeficientes de regresión estimados se usan para calcular los valores de la predicción para cada observación y para expresar el **cambio esperado de la variable dependiente** para cada unidad de cambio en las **variables independientes.** Además de hacer la **predicción**, nos gustaría saber qué **variable independiente es la más útil** en la predicción de la **variable dependiente**. En el ejemplo de la regresión múltiple discutido anteriormente, nos gustaría saber **qué variable** (el tamaño de la familia o el ingreso familiar) **es la más útil en la predicción** del número de tarjetas de crédito mantenidas por una familia. Desafortunadamente, los coeficientes de regresión (b_0, b_1, b_2) **no nos ofrecen esta información.** Para ilustrar por qué, podemos utilizar el siguiente caso. Supongamos que deseamos predecir los **gastos mensuales de los jóvenes en** (***Y***), utilizando **dos variables independientes**:
X_1 es la renta de los padres en miles de dólares y
X_2 es la asignación mensual de los adolescentes medida en usd.
Supongamos que encontramos el siguiente modelo por un procedimiento de los mínimos cuadrados ordinarios: $Y= -0.01+X_1+0.001X_2$

Puede asumirse que X_1 es más importante porque su coeficiente es **1000 veces mayor** que el coeficiente de X_2**.** Este supuesto es cierto, desde luego ya que un incremento de **$10** en la renta de los padres produce un cambio en **1*$10/$1.000** en las compras medias de

(dividimos **$10** por **1.000** porque el valor $\mathbf{X_1}$ se mide en miles de usd). Un cambio en **$10** en la asignación mensual de los jóvenes produce un cambio de (**0.001*$10**) en los gastos medios en (**Y**) o un cambio de **0.01** en el número medio de (**Y**) (porque la asignación de los jóvenes se mide en dólares).

Un cambio de $10 en la renta de los padres produce el mismo efecto de un cambio de $10 en la asignación de los jóvenes. Ambas variables son igualmente importantes, pero los coeficientes de regresión no revelan directamente este hecho. Podemos resolver este problema mediante el uso de un coeficiente de regresión modificado **llamado el coeficiente beta.**

5.18.2. Estandarización de los coeficientes de regresión: Los coeficientes beta

Si cada una de nuestras **variables predictor ha sido estandarizada** antes de estimar la ecuación de regresión, **nos encontraríamos con diferentes coeficientes de regresión**. Los coeficientes resultantes de los datos estandarizados se denominan **coeficientes beta**. Su valor reside en que **eliminan el problema de tratar con diferentes unidades de medida** (como se ha ilustrado previamente), y **reflejan el impacto relativo sobre la variable criterio de un cambio en una desviación estándar de cada variable.** Ahora que tenemos una unidad común de medida, **podemos determinar qué variable es la más influyente.** Se deben tener en cuenta **3** precauciones cuando se utilizan los coeficientes beta:

1. Deben utilizarse como guía de la importancia relativa de las variables individuales independientes únicamente cuando la **colinealidad** es mínima.
2. Los valores beta pueden interpretarse **sólo en el contexto de las otras variables de la ecuación**. Por ejemplo, un valor beta para el **tamaño familiar** refleja su importancia sólo con relación al **ingreso familiar**, **no en ningún sentido absoluto**. Si se añade otra variable independiente a la ecuación, **el coeficiente beta para el tamaño familiar probablemente cambiaría**, dado que podría existir una cierta relación entre el tamaño de la familia y la nueva variable independiente.
3. **Los niveles** (es decir, **familias de 5, 6 y 7**) **afectan al valor beta**. Si encontramos familias de tamaño **8, 9 o 10**, **el valor beta probablemente cambiará.**

En resumen, utilice los **coeficientes beta sólo como guía de la importancia relativa de las variables predictor** incluidas en la ecuación, y sólo en el rango de valores para el que realmente existe una muestra de datos.

5.18.3. Evaluación de la multicolinealidad

Un supuesto clave en la interpretación del valor teórico de la regresión es la **correlación entre las variables predictor**. Se trata de un **problema de datos**, **no un problema de especificación del modelo**. La situación **ideal** para sería tener una cantidad de **variables independientes que estuvieran altamente correlacionadas con la variable dependiente**, **pero con poca correlación entre sí.**

Sin embargo, en la mayoría de las situaciones, especialmente las situaciones que incluyen datos de respuesta de consumidores, habrá algo de **multicolinealidad.** En otras ocasiones, como las de la utilización de **variables ficticias** para representar **variables no métricas** o **términos polinomiales para efectos no lineales**, el investigador **crea situaciones de alta multicolinealidad**.

Así, su tarea es

1. **Valorar el grado de multicolinealidad y**
2. **Determinar su impacto en los resulta dos y las soluciones pertinentes**.

Con lo anterior, veremos **los efectos de la multicolinealidad,** procedimientos de diagnóstico útiles y los posibles remedios

5.18.4. Los efectos de la multicolinealidad

Se pueden clasificar en términos de **explicación y estimación**:

1. **Los efectos sobre la explicación** conciernen principalmente a la capacidad del procedimiento de regresión y a la capacidad del investigador para representar y comprender los efectos de cada variable independiente en el valor teórico de regresión. Conforme ocurre la **multicolinealidad** (incluso para niveles relativamente bajos de aproximadamente **0.30**), el proceso para la separación de los **efectos de los individuos es cada vez más difícil, debido a** :

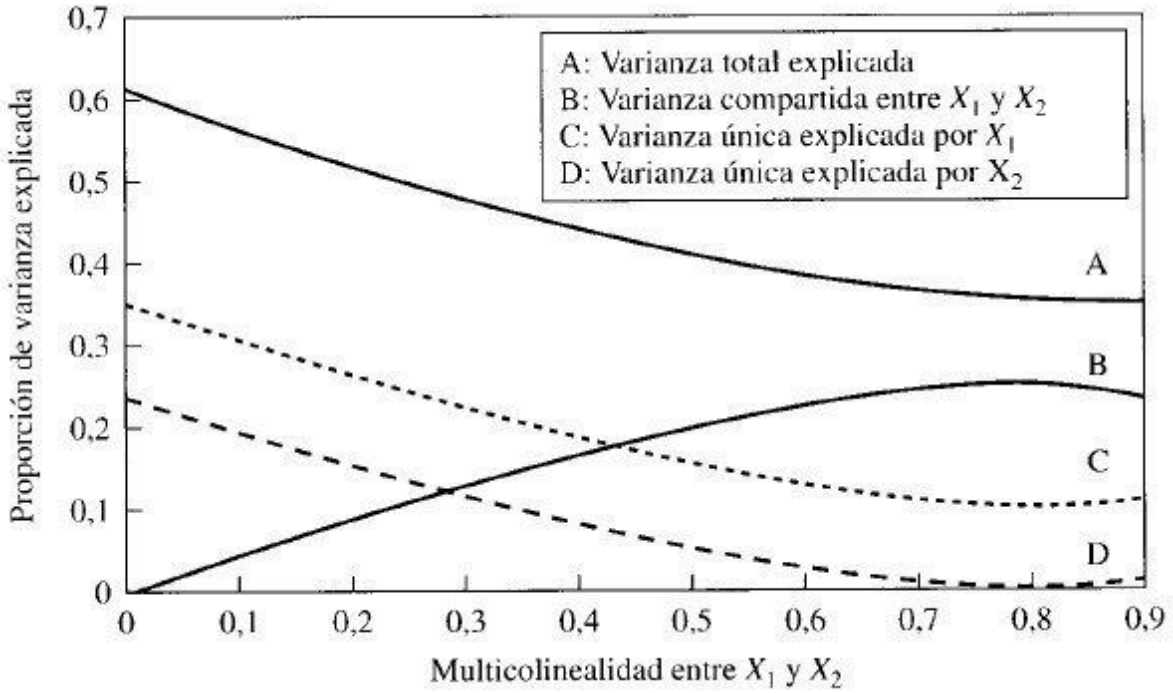
   -Limita el tamaño del coeficiente de determinación y hace más difícil añadir una predicción explicatoria extra con variables adicionales.

   -Se hace difícil determinar la contribución de cada variable debido a que los efectos de las variables independientes son **"*mixtos*"** o se confunden.

   La **multicolinealidad** tiene como resultado porciones más grandes de **varianza** compartida y niveles más bajos de varianza única a partir de los cuales se pueden determinar los efectos de las variables independientes individuales.

   Por ejemplo, supongamos que una sola variable independiente ($\mathbf{X_1}$) tiene una correlación de **0.60** con la **variable dependiente** y una segunda variable in dependiente ($\boldsymbol{X_2}$) tiene una correlación de **0.50**. Entonces $\mathbf{X_1}$ explicaría un **36%** (**obtenido con la elevación al cuadrado de la correlación de 0.60**) de la varianza de la variable dependiente y $\mathbf{X_2}$, explicaría un **25%** por ciento (la **correlación de 0.50 elevado al cuadrado**). Si las dos variables independientes no están correlacionadas la una con la otra en absoluto, **no existe un "*solapamiento*" o contribución, de su poder de predicción.** La explicación total sería la suma de las dos o el **61%**. Pero, **conforme aumenta la colinealidad,** alguna contribución del poder de predicción tiene lugar, y **el poder de predicción colectivo de las variables independientes disminuye.** El **Apartado 5.20** proporciona más detalle sobre el cálculo de las predicciones de varianza únicas y compartidas entre las variables independientes correlacionadas. La **Figura 5.37** muestra las proporciones de la **varianza compartida y única** para nuestro ejemplo de **dos variables independientes en los casos diversos de colinealidad.** Si la colinealidad de estas variables **es cero,** entonces **las variables individuales predicen un 36% y 25% de la varianza en la variable dependiente,** para una predicción global ($\boldsymbol{R^2}$) de **61%.** Pero conforme **aumenta la multicolinealidad, la varianza total explicada disminuye.** Además, la cantidad de varianza única para las variables independientes se reduce a niveles que contribuyen al hecho de que los efectos individuales sean bastante problemáticos.

Figura 5.37. Proporciones de varianza únicas y compartidas por niveles de la multicolinealidad.



Correlación entre las variables dependientes e independientes:
$\mathbf{X_1}$ y dependiente (0.60), y $\mathbf{X_2}$ dependiente (0.50)
Fuente: Hair et al. (1999)

2. Además de los efectos sobre la explicación, la multicolinealidad puede tener efectos sustantivos sobre la **estimación** de las coeficientes de regresión y sus pruebas de significación estadísticas, por ejemplo:
   -El caso extremo de la **multicolinealidad** en la cual dos o más variables están perfectamente correlacionadas, que se domina **singularidad**, **impide la estimación de cualquier coeficiente**. En este caso, **la singularidad tiene que ser eliminada para que la estimación de las coeficientes pueda proceder**. -Incluso si la **multicolinealidad no es perfecta**, altos grados de multicolinealidad pueden tener como resultado **la incorrecta estimación de los coeficientes de regresión.**

El siguiente ejemplo (vea **Figura 5.38**) ilustra el caso, con el examen de la **matriz de correlación y las regresiones simples** es evidente que la relación entre **Y y $\mathbf{X_1}$** es **positiva**, mientras que la relación entre **Y y $\mathbf{X_2}$ es negativa**. La ecuación de regresión múltiple, sin embargo, **no mantiene las relaciones de las regresiones simples**.

Figura 5.38. Estimaciones de regresión con datos multicolineales

	Variables en la Regresión		
A. Datos	**Variables Dependientes**	**Variables Independientes**	
Encuestado	**Y**	**X**	
1	5	6	13
2	3	8	13
3	9	8	11
4	9	10	11
5	13	10	9
6	11	12	9
7	17	12	7
8	15	14	7
B. Matriz de Correlación	**Y**	**X_1**	**X_2**
Y	1.000		
X_1	0.823	1.000	
X_2	-0.977	-0.913	1.000
C. Estimaciones de regresión			
Regresión simple (X_1)	**Y**=-4.75 + 1.5 **X_1**		
Regresión simple (X_2)	**Y**= -9.75 + 1.95 **X_2**		
Regresión múltiple (X_1 Y X_2)	**Y**= 44.75 + -0.75 **X_1**+-2.72 **X_2**		

Fuente: Hair et al. (1999)

El observador casual que examina solamente las coeficientes de regresión múltiple pensaría que ambas relaciones (**Y y X_1**, **Y y X_2**) son negativas, mientras que sabemos que esto no es el caso para **Y** y **X_1.** La señal del coeficiente de regresión de **X_1** es equivocada en un sentido intuitivo, pero la fuerte correlación negativa entre **X_2** y **X_2** tiene como resultado el reverso de las señales para **X_2.** A pesar de que estos efectos sobre el procedimiento de estimación ocurren principalmente a niveles altos de multicolinealidad (**por encima de 0.80**) la posibilidad de resultados engañosos exige un escrutinio de cada valor teórico de regresión para una posible multicolinealidad.

5.18.5. La identificación de la multicolinealidad

Al momento, se ha visto que los efectos de la **multicolinealidad** pueden ser sustanciales. En múltiples análisis de regresión, la evaluación de la **multicolinealidad** se debe realizar en **2 pasos**:

1. **Identificación de la magnitud de la colinealidad** y
2. **La evaluación del grado en que los coeficientes estimados se ven afectados.**

Si se dicta una acción correctiva, existen varias alternativas.

Aquí algunas de ellas:

- El medio más simple y obvio de identificar la colinealidad es un **examen de la matriz de correlación de las variables independientes**.
- La presencia de una correlaciones generalmente de **0.90** en adelante, **es la primera indicación de una elevada colinealidad.**
- La **ausencia** de elevados valores de correlación **no asegura una falta de colinealidad.** La colinealidad puede deberse a los efectos combinados de dos o más variables independientes.

- Dos de las medidas más comunes para evaluar la colinealidad de parejas o de múltiples variables son:
  -El **valor de tolerancia** y
  -Su inverso (**el factor de inflación de la varianza VIF**).
- Estas medidas nos dan el grado en el que cada variable independiente se explica por otras variables independientes. En términos simples, cada variable independiente se convierte en una **variable criterio y se realiza la regresión con el resto de las variables independientes.**
- **La tolerancia** es la cantidad de variabilidad de las variables independientes seleccionadas no explicadas por el resto de las variables independientes.
- Por tanto un **valor de tolerancia reducido (y elevados valores VIF)** denotan una **elevada colinealidad**.
- Puede ponerse un **umbral de tolerancia en un valor de 0.10** que corresponde a valores **VIF por encima de 10**.
- **Cada investigador debe determinar el grado de colinealidad que aceptará**, en la medida en que los límites por defecto o los recomendados pueden aceptar todavía una colinealidad sustancial.
- Por ejemplo, el límite sugerido para el **valor de tolerancia de 0.10** corresponde a una **correlación múltiple de 0.95**. Más aún, una correlación múltiple de **0.9** entre una variable independiente y el resto (similar a la regla que aplicamos en la matriz de correlación pareada) resultaría en un valor de tolerancia de **0.19**.
- Por tanto cualquier variable con un **valor de tolerancia por debajo de 0.19** (o por encima de un **VIF de 5.3**) estaría correlacionada en más de **0.90**.
- Se sugiere que se especifiquen siempre los **valores de tolerancia** en los programas de regresión, en la medida en que los valores por defecto que excluyen las variables colineales pueden mantener todavía **grados muy elevados de colinealidad**
- Por ejemplo, **el valor de tolerancia por defecto en el SPSS para excluir una variables es 0.0001,** lo que significa que aunque más de un **99.99%** de la varianza es prevista por el resto de las variables independientes**, la variable podría incluirse en la ecuación de regresión**. **Pueden realizarse estimaciones de los efectos reales de una elevada colinealidad sobre los coeficientes estimados, pero queda fuera del alcance de este texto (véase Neter et al. 1989).**
- Incluso con los diagnósticos que utilizan **VIF o valores de tolerancia**, no necesitamos saber qué variables están correlacionadas. Un procedimiento desarrollado por **Belsley et. al. 1980** permite la identificación de variables intercorrelaciondas, incluso si tenemos correlaciones entre varias variables, proporciona al investigador un diagnóstico de mayor potencia en la evaluación de la medida e impacto de la multicolinealidad.

5.18.6. Remedios para la multicolinealidad

Se extienden **desde una modificación del valor teórico de regresión hasta el uso de los procedimientos de estimación especializada**. Una vez que se ha determinado el grado de colinealidad, Usted tiene varias opciones:

• **Omitir una o varias variables independientes correlacionadas e identificar otras variables in dependientes para ayudar con la predicción.** Sin embargo, debe tener

cuidado cuando sigue esta opción, para evitar la creación de un **error** de especificación cuando se eliminan una o más variables independientes.

- **Utilizar el modelo con las variables correlacionados sólo para predecir** (es decir, no intentando interpretar los coeficientes de regresión).
- **Utilizar las correlaciones simples entre cada variable independiente y cada variable dependiente** para entender la relación entre la variable independiente-dependiente.
- **Utilizar un método más sofisticado de análisis, como una regresión bayesiana (o un caso especial-regresión de tipo cresta)** o una **regresión de componentes principales** para obtener un modelo que refleje más claramente los efectos simples de las variables independientes. Estos procedimientos se discuten con más detalle en varios textos **[Belsley et. al. 1980, Neter et al. 1989].**

Cada una de estas opciones exige al investigador pronunciarse sobre las variables incluidas en **el valor teórico de regresión**, guiada siempre por el marco **teórico del estudio.**

5.19. Regresión lineal: Validación de resultados

Paso 6: validación

Después de **identificar el mejor modelo de regresión**, el paso final consiste en asegurarse de que represente a la **población general (generalización)** y que sea apropiada para situaciones en las cuales será utilizada **(transferibilidad).**

La mejor guía es ver en qué medida se ajusta a un modelo teórico o a un conjunto de resultados validados previamente sobre el mismo asunto. En muchos casos, sin embargo, anteriores resultados o la teoría no están disponibles. Por tanto, discutiremos también las aproximaciones empíricas a la validación del modelo.**5.7.1. Muestras adicionales o muestras divididas**

5.19.1. Muestras adicionales o muestras divididas

La aproximación más apropiada para la validación empírica es **contrastar el modelo de regresión mediante la extracción de una nueva muestra de la población general**. Una nueva muestra asegurará la **representatividad** y puede utilizarse en varias formas:

1. El modelo original **puede predecir valores** con la nueva muestra, además del ajuste predictivo.
2. Se **puede estimar un modelo separado** con la nueva muestra para a continuación compararla con la ecuación original sobre las características de las variables incluidas: signo, tamaño e importancia relativa de las variables y precisión predictiva. En ambos casos se determina la validez del modelo original comparándolo con los modelos de regresión estimados con la nueva muestra
3. Muchas veces la capacidad de recoger nuevos datos está limitada o impedida por factores tales como los costes, la falta de tiempo o la disponibilidad de los encuestados. Cuando este es el caso, se **puede entonces dividir la muestra en dos partes: una submuestra de estimación para crear el modelo de regresión y una submuestra de validación/duración utilizada para "*contrastan*" la ecuación. Existen muchos procedimientos, tanto aleatorios como sistemáticos, para dividir los datos mediante la extracción de dos muestras independientes del conjunto de datos.**

Todos los programas informáticos de amplia difusión tienen opciones específicas que permiten la estimación y la validación de las submuestras por separado.

4. Tanto si se extrae una nueva muestra como si no, es probable que **existan diferencias entre el modelo original y otros esfuerzos de validación.** Usted pasa ahora a ser **un mediador** entre los distintos resultados, buscando el mejor modelo en las diferentes muestras. La necesidad de continuos esfuerzos de validación y de refinamiento del modelo nos recuerda **que ningún modelo de regresión, a menos que se estime del conjunto de la población, es el modelo final y absoluto.**

5.19.2. Cálculo de PRESS

Una aproximación alternativa a la obtención de muestras adicionales con el fin de validar el modelo es emplear la muestra original de forma especializada mediante el cáculo del **estadístico PRESS,** una medida similar a $\boldsymbol{R^2}$ utilizada para evaluar la precisión predictiva del modelo de regresión estimado. **Difiere de las principales aproximaciones en que se estima no uno, sino n-1 modelos de regresión.**

El procedimiento, similar a las técnicas de **"*bootstrapping*", tratadas en nuevas técnicas multivariantes** omite una observación en la estimación del modelo de regresión y predice a continuación las observaciones omitidas con el modelo estimado.

De esta forma, **la observación no puede afectar a los coeficientes del modelo utilizado para calcular su valor predictivo.** El procedimiento se aplica otra vez, omitiendo otra observación, estimando un nuevo modelo y realizando la predicción. **Pueden sumarse los residuos de las observaciones para obtener una medida conjunta del ajuste predictivo.**

5.19.3. Comparación de los modelos de regresión

Cuando comparamos modelos de regresión, el término de comparación más común suele ser el **ajuste predictivo conjunto.** Discutimos previamente que el coeficiente de determinación **($\boldsymbol{R^2}$)** nos ofrece esta información, pero tiene inconveniente: **por muchas variables que se incluyan, nunca puede disminuir.** Por tanto, aunque incluyéramos todas las variables independientes, nunca encontraríamos un **($\boldsymbol{R^2}$)** más elevado, aunque podamos conseguir el mismo $\boldsymbol{R^2}$ con un número más pequeño de variables o encontrar que un número más pequeño de predictores arroja casi el mismo valor. Por tanto, para comparar entre modelos con diferentes números de predictores, utilizaremos el $^{\mathbf{2}}$ **ajustado,** que es útil en la comparación de modelos entre diferentes conjuntos de datos, en la medida en que se compensarán los diferentes tamaños muestrales.

5.19.4. Predicción del modelo

Las predicciones del modelo siempre pueden realizarse aplicando el modelo estimado para un **nuevo conjunto de valores de las variables independientes y calculando los valores de las variables criterio.** Sin embargo, al hacerlo debemos considerar varios factores que pueden tener un serio impacto en la calidad de las nuevas predicciones:

1. Cuando aplicamos el modelo a una **nueva muestra**, debemos recordar que las **predicciones contienen ahora no sólo las variaciones muestrales respecto de la muestra original sino también la muestra nuevamente extraída**. Por tanto

deberíamos calcular siempre los intervalos de confianza de nuestras predicciones además de los puntos estimados para ver el rango esperado de los valores criterio.

2. **Asegúrese de que las condiciones y relaciones medidas en el momento en que la muestra original fue tomada no han cambiado. Por ejemplo,** en nuestro ejemplo de las tarjetas de crédito, si la mayor parte de las compañías comenzaran a cargar mayores comisiones por las tarjetas de crédito, la posesión efectiva de tarjetas de crédito puede cambiar sustancialmente, aunque esta información no esté incluida en el modelo.
3. **No se debe utilizar el modelo para estimar más allá del rango de las variables independientes que se encuentran en la muestra. Por ejemplo**, si las familias más grandes tienen seis miembros, puede ser imprudente predecir la posesión de tarjetas de crédito para familias de diez miembros. Uno no puede suponer que las relaciones son las mis mas para valores de las variables independientes sustancialmente superiores que aquellos de la muestra original.

5.20. Regresión lineal: Apartado cálculo de varianza

La base para la estimación de todas las relaciones de regresión es la **correlación**, que **mide la asociación entre dos variables**. En el análisis de la regresión, las **correlaciones** entre **las variables independientes y las variables dependientes** proporcionan la base para conformar la **regresión del valor teórico** mediante la estimación de los **coeficientes** de regresión (**ponderaciones**) para cada **variable independiente** que maximice la predicción (**varianza explicada**) de la variable de pendiente. Cuando el **valor teórico** contiene una única variable independiente, el cálculo de los coeficientes de regresión es directo y se basa en la **correlación directa o univariante entre la variable independiente y dependiente. El porcentaje de la varianza explicada de la variable dependiente es simplemente el cuadrado de la correlación directa.**

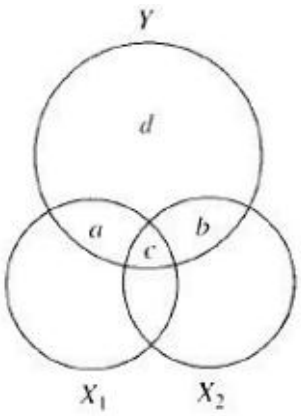
Pero a medida que se añaden variables independientes al valor teórico, los cálculos deben considerar también las **intercorrelaciones entre las variables independientes.** Si las variables independientes están correlacionadas, entonces **"*comparten*"** algo de su poder predictivo. Dado que usamos sólo la **predicción del conjunto del valor teórico, la varianza compartida no debe ser "*contabilizada dos veces*"** mediante el uso de las correlaciones directas.

Pero a medida que se añaden variables independientes al valor teórico, los cálculos deben considerar también las intercorrelaciones entre las variables independientes. Si las variables independientes están correlacionadas, entonces "***comparten***" algo de su poder predictivo. Dado que usamos sólo la predicción del conjunto del valor teórico, la varianza compartida no debe ser "***contabilizada dos veces***" mediante el uso de las correlaciones directas. Por tanto, calculamos correlaciones adicionales para representar estos efectos compartidos:

1. La primera es el **coeficiente de correlación parcial,** que es la correlación de una variable independiente ($\boldsymbol{X_i}$) y una variable dependiente $\boldsymbol{Y}$ cuando se han suprimido los efectos de las otras variables independientes tanto de $\boldsymbol{X_i}$ e $\boldsymbol{Y}$.
2. Una segunda forma de correlación es la **correlación semiparcial**, que refleja la correlación entre una variable dependiente e independiente mientras se controlan los efectos predictivos de todas las otras variables independientes sobre **X**.

Las dos formas de correlación difieren en que **la correlación parcial elimina los efectos de las otras variables independientes sobre X_i** e ***Y***., mientras que la **correlación semiparcial elimina sólo los efectos sobre X_i .**
La **correlación parcial** representa el aumento en el efecto predictivo de una variable independiente respecto al efecto colectivo del resto y se utiliza para identificar variables independientes que tienen el mayor aumento de poder predictivo dado un conjunto de variables independientes ya incluidos en el valor teórico. La correlación semiparcial representa la relación única prevista por una variable independiente una vez que se ha obtenido la predicción compartida con el resto de las variables independientes. Por tanto, la correlación semiparcial se utiliza para el "***reparto***" de la varianza entre las variables independientes. Elevando al cuadrado la correlación semiparcial obtenemos la varianza única explicada por la variable independiente. La figura **5.39** representa la varianza única y compartida entre las variables independientes correlacionadas.

Figura 5.39 varianza única y compartida entre las variables independientes correlacionadas



a= **varianza de *Y*** explicada únicamente por X_1
b= varianza de *Y* explicada únicamente porX_2
c= varianza de *Y* explicada conjuntamente por X_1 **y** X_2
d= varianza de *Y* NO explicada únicamente por X_1 **o** X_2
Fuente: Hair et al. (1999)

La varianza asociada con la correlación parcial de X_2 controlada por X_1 puede representarse como: **b/ (d+b)** donde **(d+b)** representa la varianza sin explicar después de tener en cuenta X_1**.** La parte de correlación de X_2 controlada por X_1 es **b(a+b+c+d),** donde **(a+b+c+d),** representa el total de la varianza de **Y** y **b** en la cantidad únicamente explicada por X_2**.** El investigador puede también determinar la varianza única y compartida para las variables independientes a través de cálculos simples. La **correlación semiparcial** entre la variable dependiente **(*Y*)** y una variable independiente **(X_1)** mientras se controla una segunda variable independiente **(X_2)** se calcula mediante la siguiente ecuación:
Correlación semiparcial de ***Y*, X_1 ,** dado X_2 **=** (correlación ***Y*, X_1 -** (correlación ***Y*, X_2***Correlación **Y,** X_1X_2))/√1.0- (correlacion X_1 X_2)^2 .

Un ejemplo simple con dos variables independientes (X_1 y X_2) ilustrará el calculo de ambos tipos de varianzas de la variable dependiente (Y). Ver **Figura 5.40**

Figura 5.40 Las correlaciones directas y la correlación entre X_1 y X_2

	Y	X_1	X_2
Y	1.0		
X_1	0.6	1.0	
X_2	0.5	0.7	1.0

Fuente: propia

Las correlaciones directas de **0.60** y **0.50** representan relaciones claramente fuertes con **Y,** mientras que la correlación **0.7 entre X_1 y X_2** significa que una **parte sustancial de esta potencia predictiva puede estar compartida.** La correlación semiparcial de X_1 e **Y** controlado X_2 **($r_{y,X1(X2)}$)** y la varianza única prevista por X_1**,** se calcula como:

X_2rY, X_1 (X_2)) = 0.6- (0.5*0.7) / ($\sqrt{1.0- 0.7^2}$); varianza única prevista por X_1 = **0.35^2 =0.1225**

Dado que la correlación directa de X_1 e Y es **0.6** sabemos también que la **varianza** total prevista por X_1 es 1, **0.6^2** o **0.36.** Si la varianza única es **0.1225,** entonces **la varianza compartida debe ser 0.2375 (0.36- 0.1225).**

Podemos calcular la varianza única explicada por X_2**,** y confirmar la cantidad de varianza compartida mediante lo siguiente:

X_2 ($r_{Y,X1(X2)}$)= 0.5- (0.6*0.7) / ($\sqrt{1.0- 0.7^2}$ =0.11) Varianza única prevista por X_1 = 0.11^2= 0.0125

Con la varianza total explicada por **X_2** siendo **0.5^2o 0.25,** la varianza compartida calculada siendo **0.2375 (0.25- 0.0125).** Esto confirma la cantidad encontrada en los cálculos de **X_1.**

Varianza de Y explicada únicamente por X_1 =0.1225
Varianza de Y explicada únicamente por X_2 =0.0125
Varianza de Y explicada conjuntamente por X_1 y X_2 =0.2375
Varianza de Y NO explicada únicamente por X_1 y X_2 =0.3725

Estos cálculos pueden extenderse a más de dos variables, pero en la medida en que el número de variables independientes aumenta, es fácil dejar a los programas estadísticos realizar los cálculos. El cálculo de la **varianza única y compartida** ilustra el efecto de la multicolinealidad sobre la capacidad de las variables independientes para predecir la variable dependiente. La **Figura 5.40** muestra estos efectos cuando nos enfrentamos con niveles de **multicolinealidad altos o bajos.**

5.21. Regresión lineal múltiple: Resumen para aplicar

- La **Regresión Lineal** es la técnica estadística multivariable más conocida y aplicada en todas partes del mundo, ya que constituye el medio a partir del cual se ha desarrollado la econometría. La regresión lineal se aplica tanto a datos de corte transversal, es decir, a observaciones referidas a un mismo momento de tiempo como pueden ser los datos de encuestas, familias, empresas, etc., como a datos de series temporales.
- El modelo de regresión lineal más conocido y utilizado es el que considera que el **regresor** es una función lineal de ***k – 1*** **regresores** y de una perturbación aleatoria,

existiendo además un **regresor ficticio correspondiente al término independiente**. Designado por $\boldsymbol{Y}$, al regresando, por $\boldsymbol{X_{2i}, X_{3i},..., X_{ki}}$, a los regresores y por $\boldsymbol{Y_i}$, a la perturbación aleatoria, el modelo teórico de regresión lineal viene dado, para la observación genérica **tésima**, por la siguiente expresión:

$$\boldsymbol{Y_t = b_1, + b_2 bY_{2t} + b_3 bY_{3t} + \ldots\ldots + b_k bY_{kt} + Y_t}$$

- El término de regresión fue introducido por **Francis Galton (1886)** y corroborada su ley por ***Karl Pearson*** **(1903).** En términos generales se puede decir que el análisis de regresión lineal trata del estudio de la dependencia de una variable a explicar con respecto a una o más variables explicativas.
- Los **objetivos** que se pretenden conseguir con este tipo de análisis son varios, entre los más importantes son:
  1. **Determinar la estructura o forma de la relación**, es decir, la ecuación matemática que relaciona las variables independientes con las dependientes.
  2. **Verificar hipótesis** deducidas de la teoría analizada
  3. **Predecir** los valores de la variable dependiente y realizar simulaciones.
- La variable dependiente puede expresarse con diversos términos: variable explicada, predicha, regresada y respuesta y la terminología empleada para la variable independiente es como variable explicativa, predictor, regresor y variable de control estímulo. Matemáticamente la relación entre la variable explicada y las variables explicativas se puede expresar como:

$$\boldsymbol{Y = f(X)}$$

- La letra $\boldsymbol{Y}$ representa la variable dependiente y la $\boldsymbol{X}$ $\boldsymbol{(X_1, X_2, ..., X_K)}$ representan las variables explicativas. Si el número de variables independientes es una nos encontramos ante un modelo de regresión simple, por el contrario, si son más de una se trata de un modelo de regresión lineal múltiple
- El análisis de regresión es, con mucho, la técnica multivariable más utilizada y versátil, aplicable en muchísimos campos de la toma de decisiones en las ciencias de la administración.
- El análisis de regresión es una técnica estadística utilizada para analizar la relación entre una sola variable dependiente y varias independientes, siendo su formulación básica la siguiente:
- El objetivo de esta técnica es usar las variables independientes, cuyos valores se conocen, para predecir el de la variable dependiente. Cada variable independiente está ponderada por unos coeficientes que indican la contribución relativa de cada una de las variables para explicar la dependiente.
- El análisis de regresión es, con mucho, la técnica multivariable más utilizada y versátil, aplicable en muchísimos campos de la toma de decisiones en las ciencias de la administración.
- El análisis de regresión es una técnica estadística utilizada para analizar la relación entre una sola variable dependiente y varias independientes, siendo su formulación básica la siguiente:

$$\boldsymbol{Y_1 = X_1 + X_2 + \ldots\ldots + X_n}$$

***(métrica) (métricas*)**

- El objetivo de esta técnica es usar las variables independientes, cuyos valores se conocen, para predecir el de la variable dependiente. Cada variable independiente está

ponderada por unos coeficientes que indican la contribución relativa de cada una de las variables para explicar la dependiente.

- El proceso se basa en los siguientes pasos:

Primer Paso: Para utilizar esta técnica los investigadores deben considerar 3 elementos:

1. Definición del problema de investigación;
2. Especificación clara de la relación estadística y;
3. Seleccionar la variable dependiente e independiente.

Segundo Paso: Esta etapa involucra 3 decisiones básicas:

1. El tamaño de la muestra;
2. Medición de los elementos de la relación de dependencia y;
3. La naturaleza de las variables independientes.

Tercera Paso: Esta etapa se centra en el supuesto de 4 características:

1. La linealidad de la medición del fenómeno;
2. La varianza del término del error;
3. La independencia del término del error y;
4. La normalidad de los errores.

Cuarto Paso: En esta etapa se deben considerar 3 elementos:

1. Seleccionar el método para estimar la regresión lineal
2. Determinar el nivel de significancia y;
3. Determinar el grado de influencia en los resultados.

Quinto Paso: El investigador tendrá que interpretar los resultados de la regresión en términos de una Predicción o una Explicación, en función de la estimación de los coeficientes.

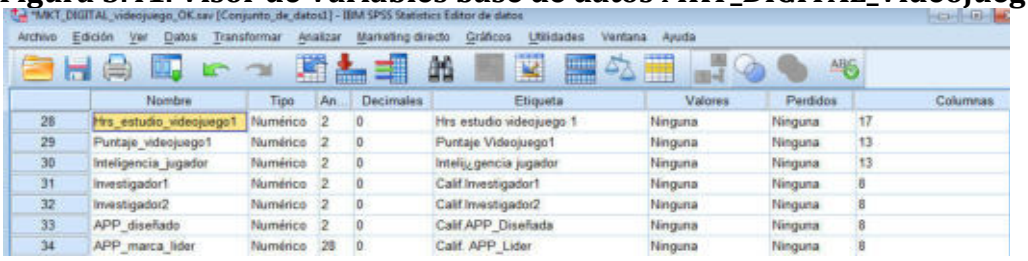
Sexta Paso: El investigador validará los resultados obtenidos de la regresión para que puedan ser generalizables a la población objeto de estudio.

5.22. Regresión lineal simple: Ejemplos.

Paso 1: Objetivos

-Problema 7 :Retomando el ejemplo al inicio del capítulo donde se relacionaban las **Hrs estudio videojuego1, Puntaje del videojuego 1 y la Inteligencia del jugador,** del que habíamos identificado con una correlación positiva entre las 2 primeras variables: **Hrs estudio videojuego1, Puntaje del videojuego 1.** Es posible que deseemos investigar esta relación al analizar si las **Hrs estudio videojuego1 predice** con seguridad **Puntaje del videojuego 1**. Para ello usamos una **regresión lineal simple. Ver Figura 5.41 y 5.42**

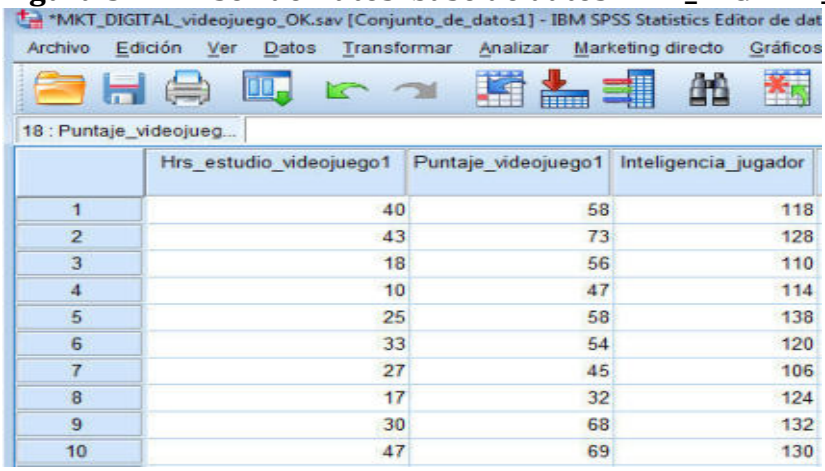
Figura 5.41. Visor de Variables base de datos MKT_DIGITAL_Videojuego.sav



	Nombre	Tipo	An...	Decimales	Etiqueta	Valores	Perdidos	Columnas
28	Hrs_estudio_videojuego1	Numérico	2	0	Hrs estudio videojuego 1	Ninguna	Ninguna	17
29	Puntaje_videojuego1	Numérico	2	0	Puntaje Videojuego1	Ninguna	Ninguna	13
30	Inteligencia_jugador	Numérico	2	0	Inteligencia jugador	Ninguna	Ninguna	13
31	Investigador1	Numérico	2	0	Calif Investigador1	Ninguna	Ninguna	8
32	Investigador2	Numérico	2	0	Calif Investigador2	Ninguna	Ninguna	8
33	APP_diseñado	Numérico	2	0	Calif APP_Diseñada	Ninguna	Ninguna	8
34	APP_marca_lider	Numérico	28	0	Calif. APP_Lider	Ninguna	Ninguna	8

Fuente: SPSS 20 IBM

Figura 5.42. Visor de Datos base de datos MKT_DIGITAL_Videojuego.sav



	Hrs_estudio_videojuego1	Puntaje_videojuego1	Inteligencia_jugador
1	40	58	118
2	43	73	128
3	18	56	110
4	10	47	114
5	25	58	138
6	33	54	120
7	27	45	106
8	17	32	124
9	30	68	132
10	47	69	130

Fuente: SPSS 20 IBM

Paso 2: Diseño

- En adelante para efectos de aprendizaje, sólo se tomarán de forma directa los datos con el fin de aprender a utilizar la técnica.
- En **SPSS** se utilizarán los comandos **Analizar, Regresión, Lineal, selección de variable dependiente: Puntaje del videojuego1; variable independiente: Hrs estudio videojuego 1**

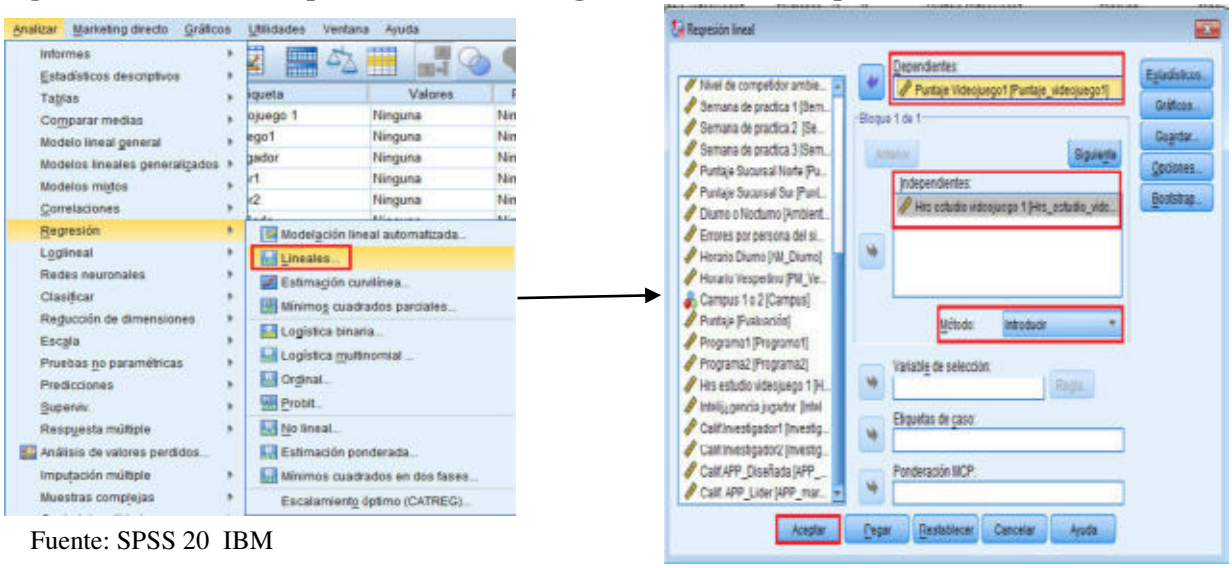
Paso 3: Condiciones de Aplicabilidad

- No olvidar realizar los análisis previos de inspección visual de la base de datos para detectar ausentes y/o atípicos (corregir en su defecto), confiabilidad, normalidad, homocedasticidad y linealidad.

Paso 4: Estimación y ajuste

-Teclear: Analizar->Regresión->Lineales->Dependientes: Puntaje Videojuego1->Independientes: Hrs estudio Videojuego1 ->Método: Introducir. Ver Figura 5.43

Figura 5.43. Proceso para calcular la regresión lineal simple



Fuente: SPSS 20 IBM

Paso 5: Interpretación

2. **SPSS** genera la primera **tabla Variables introducidas/eliminadas** y nos recuerda que estamos prediciendo los valores de **Puntaje del videojuego 1** (la **variable dependiente**) **Hrs estudio Videojuego1** (la **variable independiente**). Ver **Figura 5.44**

Figura 5.44. Tabla Variables introducidas/eliminadas

Variables introducidas/eliminadas[a]

Modelo	Variables introducidas	Variables eliminadas	Método
1	Hrs estudio videojuego 1[b]	.	Introducir

a. Variable dependiente: Puntaje Videojuego1

b. Todas las variables solicitadas introducidas.

Fuente: SPSS 20 IBM

3. La siguiente tabla que **SPSS** genera, es el **Resumen del Modelo**, que nos proporciona el **coeficiente de correlación de *Pearson*. Ver Figura 5.45**

Figura 5.45. Tabla Resumen del Modelo

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	.721[a]	.519	.459	9.144

a. Variables predictoras: (Constante), Hrs estudio videojuego 1

Fuente: SPSS 20 IBM

4. El valor ***R* cuadrado** en la **tabla resumen del modelo** muestra la **cantidad de varianza en la variable dependiente que puede explicarse por la variable independiente.**

- En nuestro ejemplo, la variable independiente de **Hrs estudio videojuego1** representa el **51.9** de la varianza en el **puntuación videojuego1.**
- El valor ***R* (0.721)** indica que a medida que **Hrs estudio videojuego1** aumenta, las **puntuaciones videojuego1,** aumentan, y esta es una correlación positiva, con ***r*=0.712.** sabemos que esto es estadísticamente significativo como resultado de la **correlación de *Pearson*.**
- El **R cuadrado corregida** se ajusta para un sesgo en ***R* cuadrado.** es sensible al número de variables y las puntuaciones que hay, y la R^2**corregida lo arregla.**
- **El error típ.de la estimación,** es una medida de la variabilidad del múltiplo de correlación.

- La siguiente tabla que reporta **SPSS** es la de **ANOVA. Ver Figura 5.46**

Figura 5.46. Tabla ANOVA

ANOVA[a]

Prueba estadística

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	723.038	1	723.038	8.647	.019[b]
	Residual	668.962	8	83.620		
	Total	1392.000	9			

p Valor

a. Variable dependiente: Puntaje Videojuego1

b. Variables predictoras: (Constante), Hrs estudio videojuego 1

Fuente: SPSS 20 IBM

- La **ANOVA** prueba la significatividad del modelo de regresión. En nuestro ejemplo, la variable independiente, **Hrs estudio videojuego1**, explica una cantidad significativa de la varianza en la variable dependiente: **Puntuaciones Videojuego1.**
- Al igual que con cualquier **ANOVA**, las piezas esenciales de información necesarias son **gl**, el valor de la **prueba estadística** F y el **valor de probabilidad**. Podemos ver que:

$$F\,(1,8) = 8.647,\ p < 0.05$$

5. Conclusión: La regresión es estadísticamente significativa.

- La tabla final que reporta **SPSS** es la de **Coeficientes y esta es la que permite generar la ecuación de regresión. Ver Figura 5.47**

Figura 5.47. Tabla Coeficientes

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	34.406	7.893		4.359	.002
	Hrs estudio videojuego 1	.745	.253	.721	2.941	.019

a. Variable dependiente: Puntaje Videojuego1

Fuente: SPSS 20 IBM

- **La columna B de Coeficientes no estandarizados** nos da el valor de la **intercepción** (fila **(Constante)** y la pendiente de la **línea de regresión** (de la fila **Hrs estudio videojuego1**). Esto, nos da la siguiente ecuación de regresión:

Puntaje Videojuego 1= 34.406 + 0.745 Hrs estudio videojuego1

- La columna de **Coeficientes tipificados Beta** nos informa de la contribución que un individuo aporta al modelo. De la tabla anterior podemos ver que la variable **Hrs estudio videojuego1 "*contribuye*"** un **0.721** al rendimiento del **Puntaje Videojuego1**, que es el **valor *r* de *Pearson***.
- El **valor de *t*** (**t = 4.359, *p* <0.01**) para **Constante** nos dice que **la intercepción es significativamente diferente de cero.**

- El valor de t para **Hrs estudio videojuego1** (t **= 2.941**, p **<0.05**) muestra que **la regresión es significativa.**

Como se ha visto, en la regresión múltiple debemos decidir qué variable va a ser la **dependiente** (**variable de criterio**) y qué variables deben ser **independientes** (**variables predictoras**). **SPSS NO puede elegir la variable dependiente** por Usted, ya que esta decisión necesita una justificación conceptual, por lo que se requiere de una razón académica para elegir esta variable. Como estamos probando una relación lineal, debemos ser capaces de elaborar el modelo lineal que brinde el mejor ajuste de los datos, llamado **regresión múltiple**., con la ecuación:

$$Y= a+b_1X_1+b_2X_2+b_3X_3+...+error$$

Donde Y es la variable dependiente, **a** es la intersección, X_1, X_2, X_3, etc. son las **variables independientes** b_1, b_2, b_3, etc. son los **coeficientes de las variables independientes.** La regresión múltiple puede ser vista como un modelo más complejo ya que emplea **más de Una variable independiente como predictor de la variable dependiente**, pudiendo examinar la contribución de cada variable independiente a la **predicción.** Si bien sigue siendo una relación lineal ya no podemos representarlo como una línea recta simple.
En el caso previo, determinamos que la variable el **Puntaje Videojuego1** es la variable **dependiente** ya que se tiene el interés en analizar la **predicción** de esta variable, y se desea ver en qué medida la variación en las puntuaciones con las **Hrs de estudio videojuego1** y la **inteligencia del jugador** son capaces de predecir la variación en el **Puntaje Videojuego1**. **El coeficiente de correlación múltiple** (R) nos reporta un valor de la **fuerza de la relación**. Pero con correlación múltiple **estamos menos interesados** en R **que en** R^2, ya que este valor nos dice **cuánto es la variación en la variable dependiente** que puede ser contabilizada por **las variables predictoras**. Por ejemplo, si R **= 0.60,** entonces R^2 **= 0.36**, lo que indica que e**l 36 % de la variación en la variable dependiente puede ser explicado por la variación en las variables independientes.** Observe que usamos R para distinguirla de **r**, que se utiliza cuando sólo se correlacionan dos variables.
Lo interesante de la **regresión múltiple** es que al tener una serie de variables, algunas serán más importantes que otras para **predecir la variación en la variable dependiente** y algunos no tendrán casi ninguna influencia en absoluto. En **SPSS** podemos elegir:

1. Incluir **todas las variables independientes** en nuestro cálculo de regresión (**método enter**).
2. Podemos usar una variedad de otros métodos para la ecuación de regresión, con el número apropiado de variables. Podemos decidir **qué variables incluir** (**método *forward*) o excluir** (***backward***) en los cálculos de regresión o
3. Ambos (**por el método escalonado**).

Así, nosotros podemos terminar con un modelo que incluya las variables que consideramos importantes en la **predicción** y **excluye aquellas que sólo tienen un efecto trivial sobre la variable dependiente.**
A medida que estamos produciendo este modelo, tenemos que ser conscientes de ciertas características del análisis, como la **Multicolinealidad** donde **dos o más predictores independientes** están altamente correlacionados entre sí, pero al serlo, es muy probable también que ambos estén midiendo lo mismo, suponiendo que miden cuestiones diferentes.

Una solución posible es eliminar una de las variables, otra solución, es combinarlas entre sí. Como regla general, las variables **predictoras** pueden correlacionarse entre sí como máximo **0.8 antes de que haya motivo para la preocupación sobre multicolinealidad.** Por último, cabe decir que las condiciones a buscar es que la relación de las variables sea lineal

Paso 1: Objetivos

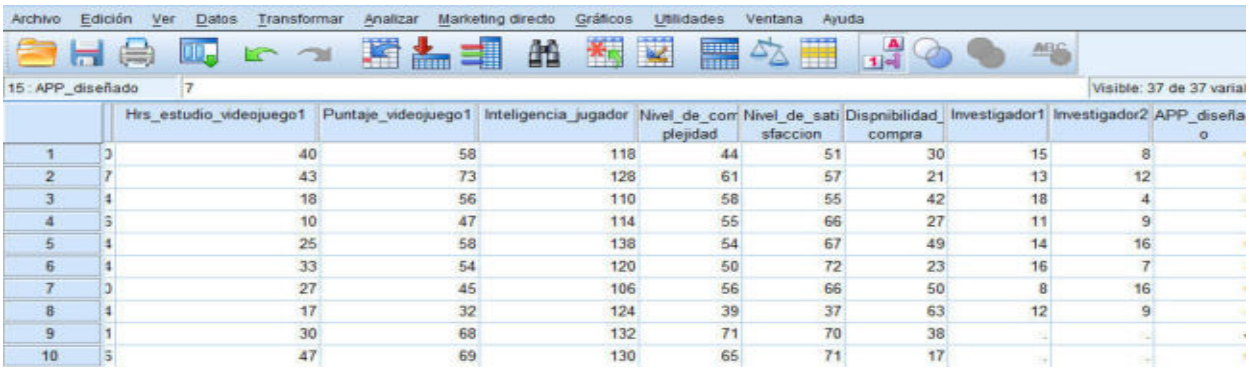
-Problema 8: La empresa **MKT Digital,** requiere investigar cuál es el mejor predictor de los jugadores de un videojuego de reciente diseño. Los datos se recogen de 10 participantes, cual incluye: Horas de estudio por semana durante 1 mes (Hrs estudio videojuego1), Puntaje obtenido al jugar el videojuego (Puntaje Videojuego1), Inteligencia del jugador (uso de prueba estándar), Nivel de complejidad percibido, Nivel de satisfacción, Disponibilidad para comprarlo. Ver **Figura 5.48.**

Figura 5.48. Visor de Variables de la base de datos MKS: Digital.SAV

	Nombre	Tipo	An...	Decimales	Etiqueta	Valores	Perdidos	Colum
28	Hrs_estudio_videojuego1	Numérico	2	0	Hrs estudio videojuego 1	Ninguna	Ninguna	17
29	Puntaje_videojuego1	Numérico	2	0	Puntaje Videojuego1	Ninguna	Ninguna	13
30	Inteligencia_jugador	Numérico	2	0	Intelij¿gencia jugador	Ninguna	Ninguna	13
31	Nivel_de_complejidad	Numérico	2	0	Nivel percibido de complejidad	Ninguna	Ninguna	8
32	Nivel_de_satisfaccion	Numérico	2	0	Nivel de satisfacción del videojuego	Ninguna	Ninguna	8
33	Dispnibilidad_compra	Numérico	8	0	Disponibilidad de adquirirlo	Ninguna	Ninguna	8
34	Investigador1	Numérico	2	0	Calif.Investigador1	Ninguna	Ninguna	8
35	Investigador2	Numérico	2	0	Calif.Investigador2	Ninguna	Ninguna	8
36	APP_diseñado	Numérico	2	0	Calif.APP_Diseñada	Ninguna	Ninguna	8
37	APP_marca_lider	Numérico	28	0	Calif. APP_Lider	Ninguna	Ninguna	8

Fuente: SPSS 20 IBM

Figura 5.49. Visor de Datos de la base de datos MKS: Digital.SAV



Archivo Edición Ver Datos Transformar Analizar Marketing directo Gráficos Utilidades Ventana Ayuda

15 : APP_diseñado 7 Visible: 37 de 37 varia

	Hrs_estudio_videojuego1	Puntaje_videojuego1	Inteligencia_jugador	Nivel_de_complejidad	Nivel_de_satisfaccion	Dispnibilidad_compra	Investigador1	Investigador2	APP_diseñado
1	40	58	118	44	51	30	15	8	
2	43	73	128	61	57	21	13	12	
3	18	56	110	58	55	42	18	4	
4	10	47	114	55	66	27	11	9	
5	25	58	138	54	67	49	14	16	
6	33	54	120	50	72	23	16	7	
7	27	45	106	56	66	50	8	16	
8	17	32	124	39	37	63	12	9	
9	30	68	132	71	70	38	.	.	
10	47	69	130	65	71	17	.	.	

Fuente: SPSS 20 IBM

Paso 2: Diseño

- En adelante para efectos de aprendizaje, sólo se tomarán de forma directa los datos con el fin de aprender a utilizar la técnica.
- Se aplica el comando Analizar, Regresión, Lineal, se elige el método de cómo las variables independientes serán estudiadas (**introducir, pasos sucesivos, eliminar, atrás, adelante**), declaración de las variables dependiente e independientes, estadísticos, ajustes.

Paso 3: Condiciones de Aplicabilidad

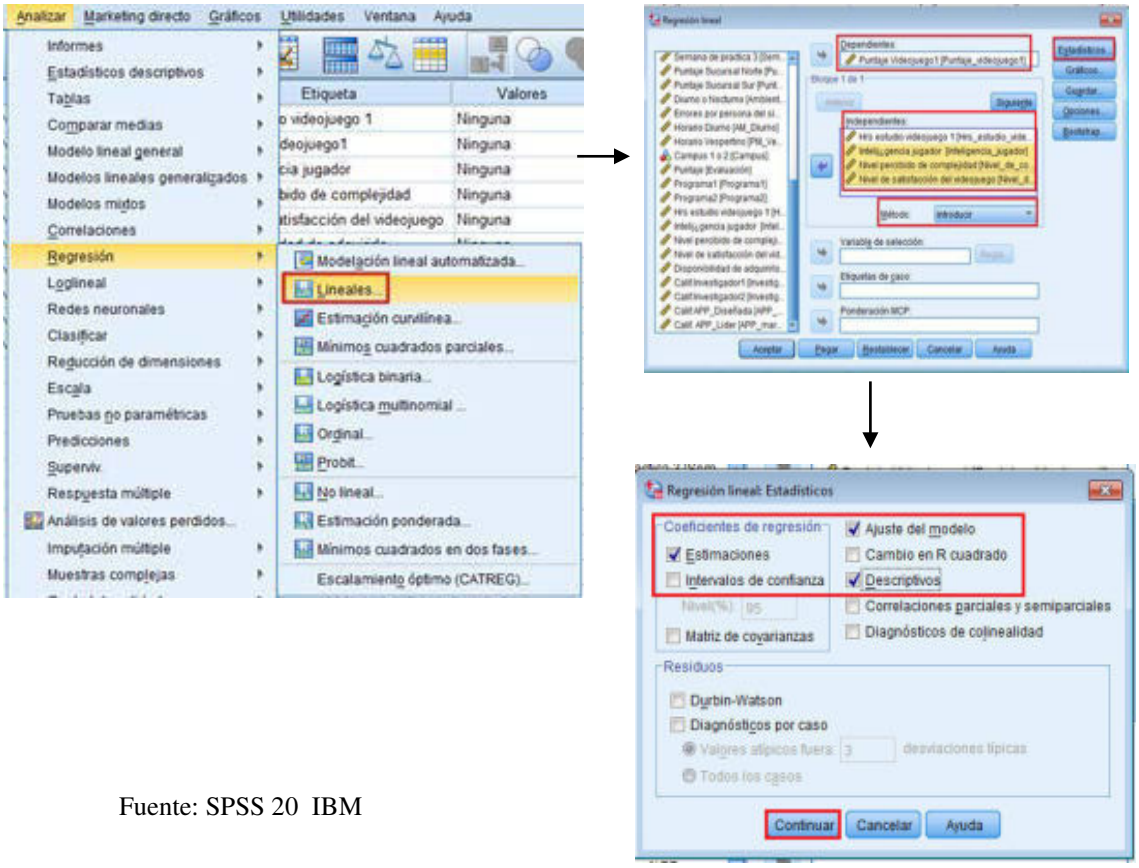
- No olvidar realizar los análisis previos de inspección visual de la base de datos para detectar ausentes y/o atípicos (corregir en su defecto), confiabilidad, normalidad, homocedasticidad y linealidad.

5.22.1. Regresión lineal múltiple. Método: Introducir.

Paso 4: Estimación y ajuste

-Teclear: Analizar ->Regresión->Lineales->Dependientes: Puntaje Videojuego1->Independientes: Hrs estudio videojuego 1; inteligencia jugador; Nivel percibido de complejidad; Nivel de satisfacción videojuego->Método: Introducir->Estadísticos->Coeficientes de regresión: Estimaciones; ajuste del modelo, descriptivos->Continuar->OK

Figura 5.50. Proceso estimación regresión lineal múltiple



Fuente: SPSS 20 IBM

Paso 5: Interpretación

La primera tabla producida por **SPSS** es s **Estadístico Descriptivos, Ver Figura 5.51**

Figura 5.51. Proceso estimación regresión lineal múltiple

Estadísticos descriptivos

	Media	Desviación típica	N
Puntaje Videojuego1	56.00	12.437	10
Hrs estudio videojuego 1	29.00	12.037	10
Inteligencia jugador	122.00	10.242	10
Nivel percibido de complejidad	55.30	9.452	10
Nivel de satisfacción del videojuego	63.80	7.193	10
Disponibilidad de adquirirlo	31.80	12.136	10

Fuente: SPSS 20 IBM

La tabla de **Estadísticos descriptivos** nos permite observar las variaciones en las puntuaciones.

- Se muestran las puntuaciones medias de cada variable.
- La desviación típica (desviación estándar) muestra la difusión de las puntuaciones para cada variable.
- **N** representa el número de participantes.

La segunda tabla producida por **SPSS** es la **tabla de Correlaciones** de todas las variables. Cada par de variables es correlacionada y los resultados se colocan en la tabla, presentando detalles del **valor de correlación *r* de *Pearson*, valor de probabilidad y número de participantes. Ver Figura 5.52.**

Figura 5.52. Tabla de correlaciones

		Puntaje Videojuego1	Hrs estudio videojuego 1	Inteligencia jugador	Nivel percibido de complejidad	Nivel de satisfacción del videojuego	Disponibilidad de adquirirlo
Correlación de Pearson	Puntaje Videojuego1	1.000	.721	.483	.740	-.010	-.111
	Hrs estudio videojuego 1	.721	1.000	.373	.293	-.026	-.352
	Inteligencia jugador	.483	.373	1.000	.264	.326	-.195
	Nivel percibido de complejidad	.740	.293	.264	1.000	.352	.150
	Nivel de satisfacción del videojuego	-.010	-.026	.326	.352	1.000	-.051
	Disponibilidad de adquirirlo	-.111	-.352	-.195	.150	-.051	1.000
Sig. (unilateral)	Puntaje Videojuego1	.	.009	.079	.007	.489	.380
	Hrs estudio videojuego 1	.009	.	.144	.206	.472	.159
	Inteligencia jugador	.079	.144	.	.231	.179	.295
	Nivel percibido de complejidad	.007	.206	.231	.	.159	.340
	Nivel de satisfacción del videojuego	.489	.472	.179	.159	.	.444
	Disponibilidad de adquirirlo	.380	.159	.295	.340	.444	.
N	Puntaje Videojuego1	10	10	10	10	10	10
	Hrs estudio videojuego 1	10	10	10	10	10	10
	Inteligencia jugador	10	10	10	10	10	10
	Nivel percibido de complejidad	10	10	10	10	10	10
	Nivel de satisfacción del videojuego	10	10	10	10	10	10
	Disponibilidad de adquirirlo	10	10	10	10	10	10

Fuente: SPSS 20 IBM

- A partir de esta tabla, podemos obtener una idea de las variables que muestran una correlación significativa (encuadrados). Se observa, que la variable **Puntaje Videojuejo1** están positivamente correlacionado con las puntuaciones tanto de **Hrs estudio videojuego1** (**p=0.009 <0.01**) y de **Inteligencia jugador** (**p=0.007 <0.01**). **Ver Figura 5.52**
- Esta tabla es muy útil, dado que permite ver en un sólo vistazo el modelo y su multicolinealidad.

Aplicando el método **Introducir**, **SPSS** ingresa todas las variables que hemos elegido para entrar en la ecuación de regresión múltiple. Ver **Figura 5.53**

Figura 5.53. Tabla variables introducidas/eliminadas

Variables introducidas/eliminadas[a]

Modelo	Variables introducidas	Variables eliminadas	Método
1	Disponibilidad de adquirirlo, Nivel de satisfacción del videojuego, Hrs estudio videojuego 1, Inteligencia jugador, Nivel percibido de complejidad[b]	.	Introducir

a. Variable dependiente: Puntaje Videojuego1

b. Todas las variables solicitadas introducidas.

Fuente: SPSS 20 IBM

- La tabla muestra la confirmación del método **Introducir** de regresión así como las variables que se han introducido en la ecuación de regresión.
- Como se observa, **todas las variables** se han introducido como **variables predictoras** para determinar la **variable criterio Puntaje Videojuego1**.

La siguiente tabla que SPSS genera para analizar es la de **Resumen del modelo**. Como sólo se ha seleccionado uno bloque del método **Introducir**, sólo se ha producido un modelo. Ver **Figura 5.54.**

Figura 5.54. Tabla Resumen del modelo

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	.959[a]	.921	.821	5.256

a. Variables predictoras: (Constante), Disponibilidad de adquirirlo, Nivel de satisfacción del videojuego, Hrs estudio videojuego 1, Inteligencia jugador, Nivel percibido de complejidad

Fuente: SPSS 20 IBM

- El valor ***R* cuadrado** en la tabla **Resumen del modelo** muestra la cantidad **de varianza en la variable dependiente que puede ser explicada por las variables independientes.**
- En nuestro ejemplo, las variables independientes representan conjuntamente el **92.1%** de la varianza en la variable **Puntaje Videojuego1.**
- El valor ***R*** (**0.959[a]**) indica el **coeficiente de correlación múltiple entre todas las variables independientes introducidas y la variable dependiente.**
- El valor **R Cuadrado ajustado** se ajusta a un sesgo en **R cuadrada** a medida que aumenta el número de variables. Con sólo unas **pocas variables predictoras**, el **R cuadrado ajustado debe ser similar al valor R cuadrado**. Normalmente tomaremos el valor **R cuadrado**, con **máximo 2 variables independientes**, pero se recomienda

seleccionar el valor ***R* cuadrado corregida,** cuando tenga varias variables independientes (**más de 2**).

- El **Error típ. de la estimación** es una medida de la variabilidad de la correlación múltiple.

SPSS genera también una tabla **ANOVA**, **que prueba la significación de la regresión. Ver Figura 5.55.**

Figura 5.55. Tabla ANOVA

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	1281.511	5	256.302	9.279	.025[b]
	Residual	110.489	4	27.622		
	Total	1392.000	9			

Prueba estadística

p Valor

a. Variable dependiente: Puntaje Videojuego1

b. Variables predictoras: (Constante), Disponibilidad de adquirirlo, Nivel de satisfacción del videojuego, Hrs estudio videojuego 1, Inteligencia jugador, Nivel percibido de complejidad

Fuente: SPSS 20 IBM

- Podemos ver en nuestra tabla que **Sig.** (***p* Valor**)=**0.025<0.05** nuestros predictores son Significativamente mejor de lo que se esperaría por casualidad. La línea de regresión predicha Por las variables independientes explica una cantidad significativa de la varianza en la variable dependiente. Normalmente se informará de manera similar a otras **ANOVA:**

$$F(5,4) = 9.279; p < 0.05$$

- La división de la **Suma de cuadrados** por los **grados de libertad (gl)** nos da la **Media Cuadrática o varianza**. Podemos ver que la regresión explica mucho más varianza que el **error o Residual.**
- Calculamos **R cuadrada** dividiendo la **Suma de Cuadrados de regresión** por la **Suma Total de Cuadrados**.

$$(1281.511/1392.000)=0.921=R\ cuadrada$$

Otra tabla que genera **SPSS**, es la **tabla Coeficientes**, la cual muestra qué variables son individualmente predictoras significativas de nuestra variable dependiente. Los valores predictivos significativos se muestran encuadrados. **Ver Figura 5.56**

Figura 5.56. Tabla Coeficiente

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados		
		B	Error típ.	Beta	t	Sig.
1	(Constante)	-5.849	24.491		-.239	.823
	Hrs estudio videojuego 1	.424	.181	.410	2.345	.079
	Inteligencia jugador	.304	.198	.250	1.531	.201
	Nivel percibido de complejidad	.884	.222	.672	3.975	.016
	Nivel de satisfacción del videojuego	-.552	.282	-.319	-1.961	.121
	Disponibilidad de adquirirlo	-.036	.164	-.035	-.218	.838

a. Variable dependiente: Puntaje Videojuego1

Fuente: SPSS 20 IBM

- La columna **B de coeficientes no estandarizados**, nos reporta los **coeficientes de las variables en la ecuación de regresión incluyendo todas las variables predictoras:**

Puntaje Videojuego1= -5.849+0.424 Hrs estudio videojuego1 + 0.304 Inteligencia jugador + 0.884 Nivel percibido de complejidad – 0.552 Nivel de satisfacción del videojuego- 0.036 Disponibilidad de adquirirlo.

- Recuerde que el método **Introducir** ha incluido todas las variables en la ecuación de regresión aunque sólo uno de ellos es un **predictor significativo**. La columna de **Coeficiente tipificado Beta** muestra la contribución que una variable hace al modelo. La ponderación que se hace de beta es la cantidad promedio en que la **variable dependiente aumenta cuando la variable independiente aumenta en una desviación estándar (todas las demás variables independientes se mantienen constantes)**. **Como están estandarizados, podemos compararlos**. Observe que la influencia más grande sobre la variable **Puntaje Videojuego1** proviene de la variable **Nivel percibido de complejidad** (**0.672**) y la siguiente **variable Horas estudio videojuego1** (**0.410**).
- Se realizan **pruebas *t*** para probar la **hipótesis de dos colas** que el **valor Beta es significativamente mayor o menor que cero**. Esto también nos permite ver cuales predictores son significativos.
- De nuestro ejemplo, observamos que la puntuación del **Nivel percibido de complejidad** es significativa (***p*=0.016<0.05**), por lo tanto, con el método **Introducir**, la puntuación de la variable **Nivel percibido de complejidad** es el único predictor significativo (se muestra encuadrado en la tabla de Coeficientes). El siguiente valor de t más grande **Hrs estudio videojuego1**, pero aquí su ***p*=0.079> 0.05.**
- Los Coeficientes no estandarizados de error tip. proporciona una estimación de la variabilidad de los coeficientes.

Como se ha visto, en la regresión múltiple debemos decidir qué variable va a ser la **dependiente** (**variable de criterio**) y qué variables deben ser **independientes** (**variables predictoras**). **SPSS NO puede elegir la variable dependiente** por Usted, ya que esta decisión necesita una justificación conceptual, por lo que se requiere de una razón

académica para elegir esta variable. Como estamos probando una relación lineal, debemos ser capaces de elaborar el modelo lineal que brinde el mejor ajuste de los datos, llamado **regresión múltiple**., con la ecuación:

$$Y= a+b_1X_1+b_2X_2+b_3X_3+...+error$$

Donde Y es la variable dependiente, **a** es la intersección, X_1, X_2, X_3, etc. son las **variables independientes** b_1, b_2, b_3, etc. son los **coeficientes de las variables independientes.** La regresión múltiple puede ser vista como un modelo más complejo ya que emplea **más de Una variable independiente como predictor de la variable dependiente**, pudiendo examinar la contribución de cada variable independiente a la **predicción.** Si bien sigue siendo una relación lineal ya no podemos representarlo como una línea recta simple.
En el caso previo, determinamos que la variable el **Puntaje Videojuego1** es la variable **dependiente** ya que se tiene el interés en analizar la **predicción** de esta variable, y se desea ver en qué medida la variación en las puntuaciones con las **Hrs de estudio videojuego1** y la **inteligencia del jugador** son capaces de predecir la variación en el **Puntaje Videojuego1**. **El coeficiente de correlación múltiple** (***R***) nos reporta un valor de la **fuerza de la relación**. Pero con correlación múltiple **estamos menos interesados** en ***R*** **que en** R^2, ya que este valor nos dice **cuánto es la variación en la variable dependiente** que puede ser contabilizada por **las variables predictoras**. Por ejemplo, si ***R*** **= 0.60,** entonces R^2 **= 0.36**, lo que indica que el **36 % de la variación en la variable dependiente puede ser explicado por la variación en las variables independientes.** Observe que usamos ***R*** para distinguirla de ***r***, que se utiliza cuando sólo se correlacionan dos variables.
Lo interesante de la **regresión múltiple** es que al tener una serie de variables, algunas serán más importantes que otras para **predecir la variación en la variable dependiente** y algunos no tendrán casi ninguna influencia en absoluto. En **SPSS** podemos elegir:

1. Incluir **todas las variables independientes** en nuestro cálculo de regresión (**método enter**).
2. Podemos usar una variedad de otros métodos para la ecuación de regresión, con el número apropiado de variables. Podemos decidir **qué variables incluir** (**método forward**) **o excluir** (***backward***) en los cálculos de regresión o
3. Ambos (**por el método escalonado**).

Así, nosotros podemos terminar con un modelo que incluya las variables que consideramos importantes en la **predicción** y **excluye aquellas que sólo tienen un efecto trivial sobre la variable dependiente.**
A medida que estamos produciendo este modelo, tenemos que ser conscientes de ciertas características del análisis, como la **Multicolinealidad** donde **dos o más predictores independientes** están altamente correlacionados entre sí, pero al serlo, es muy probable también que ambos estén midiendo lo mismo, suponiendo que miden cuestiones diferentes. Una solución posible es eliminar una de las variables, otra solución, es combinarlas entre sí. Como regla general, las variables **predictoras** pueden correlacionarse entre sí como máximo **0.8 antes de que haya motivo para la preocupación sobre multicolinealidad.**
Por último, cabe decir que las condiciones a buscar es que la relación de las variables sea lineal.

Paso 1: Objetivos

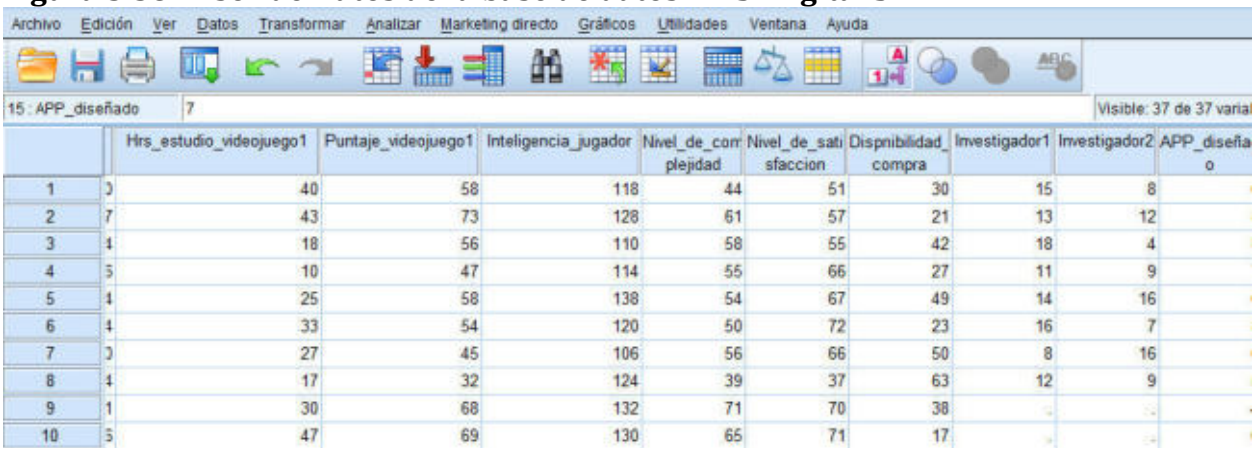
-Problema 9. La empresa **MKT Digital,** requiere investigar cuál es el mejor predictor de los jugadores de un videojuego de reciente diseño. Los datos se recogen de 10 participantes, cual incluye: Horas de estudio por semana durante 1 mes (Hrs estudio videojuego1), Puntaje obtenido al jugar el videojuego (Puntaje Videojuego1), Inteligencia del jugador (uso de prueba estándar), Nivel de complejidad percibido, Nivel de satisfacción, Disponibilidad para comprarlo. Ver **Figura 5.57.**

Figura 5.57. Visor de Variables de la base de datos MKS: Digital.SAV

	Nombre	Tipo	An...	Decimales	Etiqueta	Valores	Perdidos	Colum
28	Hrs_estudio_videojuego1	Numérico	2	0	Hrs estudio videojuego 1	Ninguna	Ninguna	17
29	Puntaje_videojuego1	Numérico	2	0	Puntaje Videojuego1	Ninguna	Ninguna	13
30	Inteligencia_jugador	Numérico	2	0	Intelij¿gencia jugador	Ninguna	Ninguna	13
31	Nivel_de_complejidad	Numérico	2	0	Nivel percibido de complejidad	Ninguna	Ninguna	8
32	Nivel_de_satisfaccion	Numérico	2	0	Nivel de satisfacción del videojuego	Ninguna	Ninguna	8
33	Dispnibilidad_compra	Numérico	8	0	Disponibilidad de adquirirlo	Ninguna	Ninguna	8
34	Investigador1	Numérico	2	0	Calif.Investigador1	Ninguna	Ninguna	8
35	Investigador2	Numérico	2	0	Calif.Investigador2	Ninguna	Ninguna	8
36	APP_diseñado	Numérico	2	0	Calif.APP_Diseñada	Ninguna	Ninguna	8
37	APP_marca_lider	Numérico	28	0	Calif. APP_Lider	Ninguna	Ninguna	8

Fuente: SPSS 20 IBM

Figura 5.58. Visor de Datos de la base de datos MKS: Digital.SAV



Archivo Edición Ver Datos Transformar Analizar Marketing directo Gráficos Utilidades Ventana Ayuda

15 : APP_diseñado 7 Visible: 37 de 37 varia

	Hrs_estudio_videojuego1	Puntaje_videojuego1	Inteligencia_jugador	Nivel_de_complejidad	Nivel_de_satisfaccion	Dispnibilidad_compra	Investigador1	Investigador2	APP_diseñado
1	40	58	118	44	51	30	15	8	
2	43	73	128	61	57	21	13	12	
3	18	56	110	58	55	42	18	4	
4	10	47	114	55	66	27	11	9	
5	25	58	138	54	67	49	14	16	
6	33	54	120	50	72	23	16	7	
7	27	45	106	56	66	50	8	16	
8	17	32	124	39	37	63	12	9	
9	30	68	132	71	70	38	.	.	
10	47	69	130	65	71	17	.	.	

Fuente: SPSS 20 IBM

Paso 2: Diseño

- En adelante para efectos de aprendizaje, sólo se tomarán de forma directa los datos con el fin de aprender a utilizar la técnica.
- Se aplica el comando Analizar, Regresión, Lineal, se elige el método de cómo las variables independientes serán estudiadas (**introducir, pasos sucesivos, eliminar, atrás, adelante**), declaración de las variables dependiente e independientes, estadísticos, ajustes.

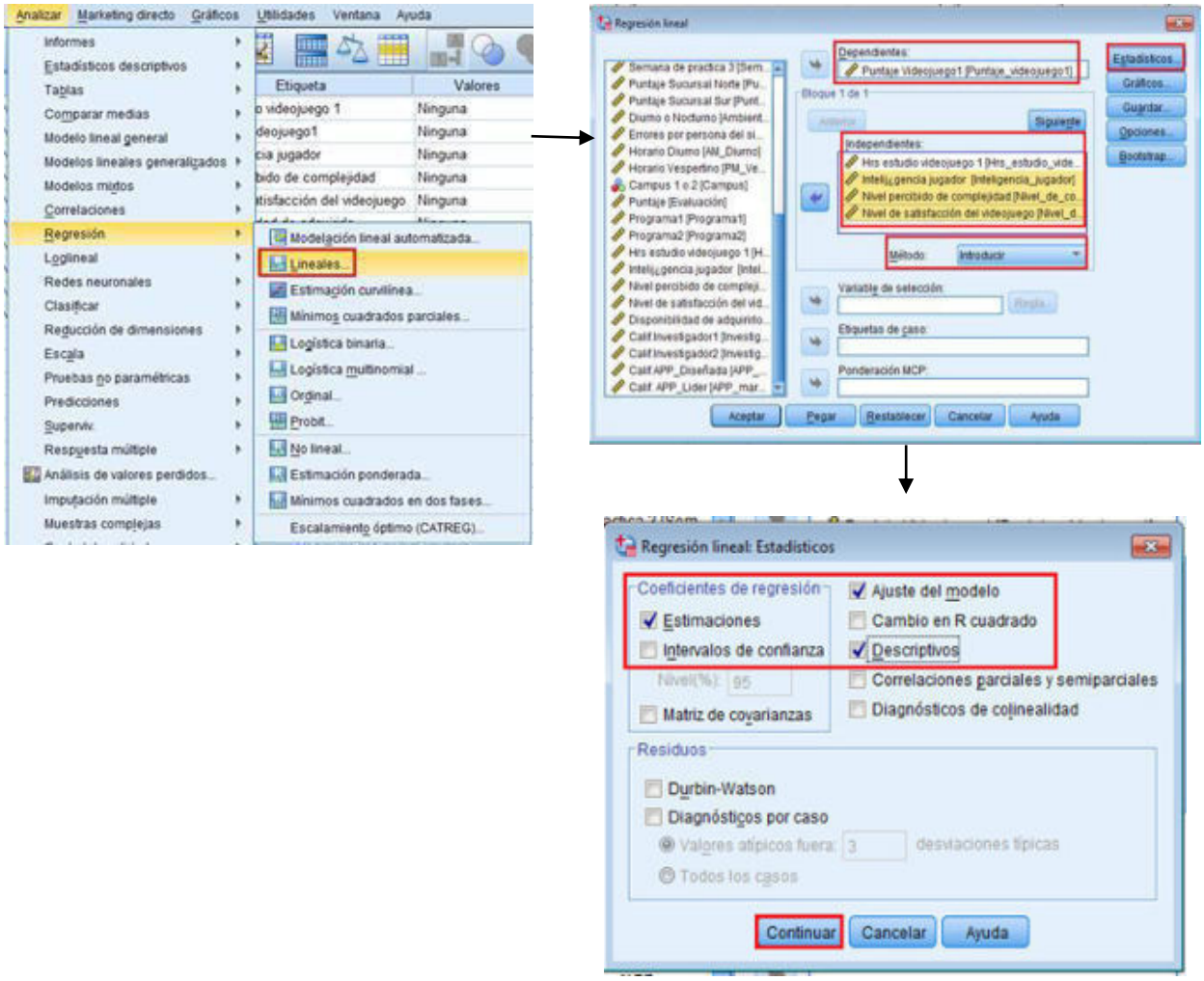
Paso 3: Condiciones de Aplicabilidad

- No olvidar realizar los análisis previos de inspección visual de la base de datos para detectar ausentes y/o atípicos (corregir en su defecto), confiabilidad, normalidad, homocedasticidad y linealidad.

Paso 4: Estimación y ajuste

-Teclear: Analizar ->Regresión->Lineales->Dependientes: Puntaje Videojuego1->Independientes: Hrs estudio videojuego 1; inteligencia jugador; Nivel percibido de complejidad; Nivel de satisfacción videojuego->Método: Introducir->Estadísticos->Coeficientes de regresión: Estimaciones; ajuste del modelo, descriptivos->Continuar->OK. Ver Fogura 5.59

Figura 5.59. Proceso estimación regresión lineal múltiple



Fuente: SPSS 20 IBM

Paso 5: Interpretación

- La primera tabla producida por **SPSS** es s **Estadístico Descriptivos, Ver Figura 5.60**

Figura 5.60. Proceso estimación regresión lineal múltiple

Estadísticos descriptivos

	Media	Desviación típica	N
Puntaje Videojuego1	56.00	12.437	10
Hrs estudio videojuego 1	29.00	12.037	10
Inteligencia jugador	122.00	10.242	10
Nivel percibido de complejidad	55.30	9.452	10
Nivel de satisfacción del videojuego	63.80	7.193	10
Disponibilidad de adquirirlo	31.80	12.136	10

Fuente: SPSS 20 IBM

La tabla de **Estadísticos descriptivos** nos permite observar las variaciones en las puntuaciones.

- Se muestran las puntuaciones medias de cada variable.
- La desviación típica (desviación estándar) muestra la difusión de las puntuaciones para cada variable.
- **N** representa el número de participantes.

- La segunda tabla producida por **SPSS** es la **tabla de Correlaciones** de todas las variables. Cada par de variables es correlacionada y los resultados se colocan en la tabla, presentando detalles del **valor de correlación *r* de *Pearson*, valor de probabilidad y número de participantes. Ver Figura 5.61**

Figura 5.61. Tabla de correlaciones

		Puntaje Videojuego1	Hrs estudio videojuego 1	Inteligencia jugador	Nivel percibido de complejidad	Nivel de satisfacción del videojuego	Disponibilidad de adquirirlo
Correlación de Pearson	Puntaje Videojuego1	1.000	.721	.483	.740	-.010	-.111
	Hrs estudio videojuego 1	.721	1.000	.373	.293	-.026	-.352
	Inteligencia jugador	.483	.373	1.000	.264	.326	-.195
	Nivel percibido de complejidad	.740	.293	.264	1.000	.352	.150
	Nivel de satisfacción del videojuego	-.010	-.026	.326	.352	1.000	-.051
	Disponibilidad de adquirirlo	-.111	-.352	-.195	.150	-.051	1.000
Sig. (unilateral)	Puntaje Videojuego1	.	.009	.079	.007	.489	.380
	Hrs estudio videojuego 1	.009	.	.144	.206	.472	.159
	Inteligencia jugador	.079	.144	.	.231	.179	.295
	Nivel percibido de complejidad	.007	.206	.231	.	.159	.340
	Nivel de satisfacción del videojuego	.489	.472	.179	.159	.	.444
	Disponibilidad de adquirirlo	.380	.159	.295	.340	.444	.
N	Puntaje Videojuego1	10	10	10	10	10	10
	Hrs estudio videojuego 1	10	10	10	10	10	10
	Inteligencia jugador	10	10	10	10	10	10
	Nivel percibido de complejidad	10	10	10	10	10	10
	Nivel de satisfacción del videojuego	10	10	10	10	10	10
	Disponibilidad de adquirirlo	10	10	10	10	10	10

Fuente: SPSS 20 IBM

A partir de esta tabla, podemos obtener una idea de las variables que muestran una correlación significativa (encuadrados). Se observa, que la variable **Puntaje Videojuejo1** están positivamente correlacionado con las puntuaciones tanto de **Hrs estudio videojuego1** (***p*=0.009 <0.01**) y de **Inteligencia jugador** (***p*=0.007 <0.01**). **Ver Figura 5.61**

Esta tabla es muy útil, dado que permite ver en un sólo vistazo el modelo y su multicolinealidad.

- Aplicando el método **Introducir**, **SPSS** ingresa todas las variables que hemos elegido para entrar en la ecuación de regresión múltiple. Ver **Figura 5.62**

Figura 5.62. Tabla variables introducidas/eliminadas

Variables introducidas/eliminadas[a]

Modelo	Variables introducidas	Variables eliminadas	Método
1	Disponibilidad de adquirirlo, Nivel de satisfacción del videojuego, Hrs estudio videojuego 1, Inteligencia jugador, Nivel percibido de complejidad[b]	.	Introducir

a. Variable dependiente: Puntaje Videojuego1

b. Todas las variables solicitadas introducidas.

Fuente: SPSS 20 IBM

- La tabla muestra la confirmación del método **Introducir** de regresión así como las variables que se han introducido en la ecuación de regresión.
- Como se observa, **todas las variables** se han introducido como **variables predictoras** para determinar la **variable criterio Puntaje Videojuego1**.

- La siguiente tabla que SPSS genera para analizar es la de **Resumen del modelo**. Como sólo se ha seleccionado uno bloque del método **Introducir**, sólo se ha producido un modelo. Ver **Figura 5.63**

Figura 5.63. Tabla Resumen del modelo

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	.959[a]	.921	.821	5.256

a. Variables predictoras: (Constante), Disponibilidad de adquirirlo, Nivel de satisfacción del videojuego, Hrs estudio videojuego 1, Inteligencia jugador, Nivel percibido de complejidad

Fuente: SPSS 20 IBM

- El valor ***R* cuadrado** en la tabla **Resumen del modelo** muestra la cantidad **de varianza en la variable dependiente que puede ser explicada por las variables independientes.**
- En nuestro ejemplo, las variables independientes representan conjuntamente el **92.1%** de la varianza en la variable **Puntaje Videojuego1.**
- El valor **R** (**0.959[a]**) indica el **coeficiente de correlación múltiple entre todas las variables independientes introducidas y la variable dependiente.**

- El valor **R Cuadrado ajustado** se ajusta a un sesgo en **R cuadrada** a medida que aumenta el número de variables. Con sólo unas **pocas variables predictoras**, el **R cuadrado ajustado debe ser similar al valor R cuadrado**. Normalmente tomaremos el valor **R cuadrado**, con **máximo 2 variables independientes**, pero se recomienda seleccionar el valor **R cuadrado corregida,** cuando tenga varias variables independientes (**más de 2**).
- El **Error típ. de la estimación** es una medida de la variabilidad de la correlación múltiple.
- **SPSS** genera también una tabla **ANOVA**, **que prueba la significación de la regresión. Ver Figura 5.64**

Figura 5.64. Tabla ANOVA

ANOVA[a]

Prueba estadística

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	1281.511	5	256.302	9.279	.025[b]
	Residual	110.489	4	27.622		
	Total	1392.000	9			

***p* Valor**

a. Variable dependiente: Puntaje Videojuego1

b. Variables predictoras: (Constante), Disponibilidad de adquirirlo, Nivel de satisfacción del videojuego, Hrs estudio videojuego 1, Inteligencia jugador, Nivel percibido de complejidad

Fuente: SPSS 20 IBM

- Podemos ver en nuestra tabla que **Sig.** (***p* Valor**)=**0.025<0.05** nuestros predictores son Significativamente mejor de lo que se esperaría por casualidad. La línea de regresión predicha Por las variables independientes explica una cantidad significativa de la varianza en la variable dependiente. Normalmente se informará de manera similar a otras **ANOVA:** ***F (5,4) = 9.279; p< 0.05***
- La división de la **Suma de cuadrados** por los **grados de libertad (gl)** nos da la **Media Cuadrática o varianza**. Podemos ver que la regresión explica mucho más varianza que el **error o Residual.**
- Calculamos **R cuadrada** dividiendo la **Suma de Cuadrados de regresión** por la **Suma Total de Cuadrados**: ***(1281.511/1392.000)=0.921=R cuadrada***
- Otra tabla que genera **SPSS**, es la **tabla Coeficientes**, la cual muestra qué variables son individualmente predictoras significativas de nuestra variable dependiente. Los valores predictivos significativos se muestran encuadrados. **Ver Figura 5.65**

Figura 5.65. Tabla Coeficientes

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados		
		B	Error típ.	Beta	t	Sig.
1	(Constante)	-5.849	24.491		-.239	.823
	Hrs estudio videojuego 1	.424	.181	.410	2.345	.079
	Inteligencia jugador	.304	.198	.250	1.531	.201
	Nivel percibido de complejidad	.884	.222	.672	3.975	.016
	Nivel de satisfacción del videojuego	-.552	.282	-.319	-1.961	.121
	Disponibilidad de adquirirlo	-.036	.164	-.035	-.218	.838

a. Variable dependiente: Puntaje Videojuego1

Fuente: SPSS 20 IBM

- La columna **B de coeficientes no estandarizados**, nos reporta los **coeficientes de las variables en la ecuación de regresión incluyendo todas las variables predictoras:**

Puntaje Videojuego1= -5.849+0.424 Hrs estudio videojuego1 + 0.304 Inteligencia jugador + 0.884 Nivel percibido de complejidad – 0.552 Nivel de satisfacción del videojuego- 0.036 Disponibilidad de adquirirlo

- Recuerde que el método **Introducir** ha incluido todas las variables en la ecuación de regresión aunque sólo uno de ellos es un **predictor significativo**.
- La columna de **Coeficiente tipificado Beta** muestra la contribución que una variable hace al modelo. La ponderación que se hace de beta es la cantidad promedio en que la **variable dependiente aumenta cuando la variable independiente aumenta en una desviación estándar** (**todas las demás variables independientes se mantienen constantes**). **Como están estandarizados, podemos compararlos**. Observe que la influencia más grande sobre la variable **Puntaje Videojuego1** proviene de la variable **Nivel percibido de complejidad** (**0.672**) y la siguiente **variable Horas estudio videojuego1** (**0.410**).
- Se realizan **pruebas *t*** para probar la **hipótesis de dos colas** que el **valor Beta es significativamente mayor o menor que cero**. Esto también nos permite ver **qué predictores son significativos.**
- De nuestro ejemplo, observamos que la puntuación del **Nivel percibido de complejidad** es significativa (***p*=0.016<0.05**), por lo tanto, con el método **Introducir**, la puntuación de la variable **Nivel percibido de complejidad** es el único predictor significativo (se muestra encuadrado en la tabla de Coeficientes). El siguiente valor de t más grande **Hrs estudio videojuego1**, pero aquí su ***p*=0.079> 0.05.**
- La columna de **Coeficientes no estandarizados de error tip.** proporciona una estimación de la variabilidad de los coeficientes.

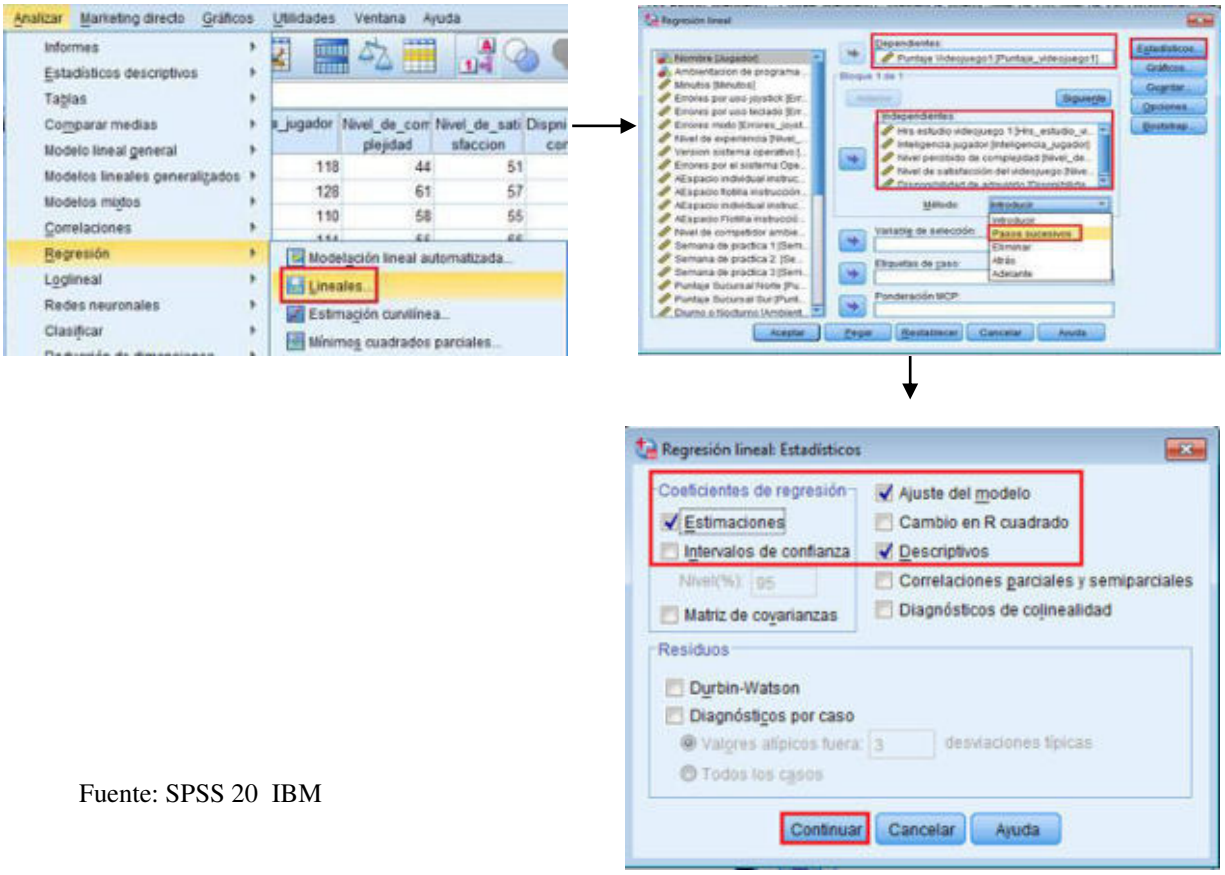
5.22.2. Regresión lineal múltiple. Método: Pasos sucesivos

En contraste con el método **Introducir**, el método Paso a Paso añade variables predictoras a la regresión que mejor se correlacionan con la variable dependiente, y resta las variables predictoras que menos se correlacionan. De esta manera se genera una ecuación de regresión utilizando variables predictor que hacen una contribución significativa a la predicción. Las tablas de resultados son muy similares a las producidas por el método Introducir pero pueden tener más filas, dependiendo de cuántos modelos la regresión ha producido.

Paso 4: Estimación y Ajuste

-Teclear: Analizar->Regresión->Lineales->Dependientes: Puntaje Videojuego1->Independientes: Hrs estudio videojuego 1; inteligencia jugador; Nivel percibido de complejidad; Nivel de satisfacción videojuego->Método: Pasos sucesivos->Estadísticos->Coeficientes de regresión: Estimaciones; ajuste del modelo, descriptivos->Continuar->OK. Ver Figura 5.66

Figura 5.66. Proceso estimación regresión lineal múltiple método: paso sucesivos



Fuente: SPSS 20 IBM

SPSS genera la tabla **Estadísticos descriptivos** y la **tabla de Correlación** de manera similar al método **Introducir.** Otra tabla a analizar es la de **Variables introducidas/eliminadas. Ver Figura 5.67**

Figura 5.67. Tabla variables introducidas/eliminadas

Variables introducidas/eliminadas[a]

Modelo	Variables introducidas	Variables eliminadas	Método
1	Nivel percibido de complejidad	.	Por pasos (criterio: Prob. de F para entrar <= .050, Prob. de F para salir >= .100).
2	Hrs estudio videojuego 1	.	Por pasos (criterio: Prob. de F para entrar <= .050, Prob. de F para salir >= .100).

a. Variable dependiente: Puntaje Videojuego1

Fuente: SPSS 20 IBM

- La tabla muestra que el método **Pasos sucesivos** de regresión ha sido usado.

Obsérvese que **SPSS** ha introducido en la ecuación de regresión las dos variables independientes (**Nivel percibido de complejidad y Hrs de estudio videojuego1**) que están significativamente correlacionados con la variable dependiente **Puntaje Videojuego1**.

Al observar la tabla **Resumen del Modelo**, utilizando el método **Pasos sucesivos** se han producido dos modelos. El **Modelo 1** incluye los resultados del **Nivel percibido de complejidad**, mientras que el Modelo **2** incluye los resultados del **Nivel percibido de complejidad y Hrs de estudio videojuego1**.**Ver Figura 5.68**

Figura 5.68. Tabla Resumen del modelo

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	.740[a]	.548	.491	8.871
2	.909[b]	.825	.776	5.891

a. Variables predictoras: (Constante), Nivel percibido de complejidad

b. Variables predictoras: (Constante), Nivel percibido de complejidad, Hrs estudio videojuego 1

Fuente: SPSS 20 IBM

- El valor de **R cuadrado** muestra la cantidad de varianza en la **variable dependiente** que puede ser explicada por las variables independientes.
- En nuestro ejemplo anterior:

Modelo 1. La variable independiente **Nivel percibido de complejidad** representa el **54.8 %** de la varianza en las puntuaciones del examen de ciencias. **Modelo 2.** Las variables independientes **Nivel percibido de complejidad** y **Hrs estudio videojuego1** tienen el **82.5 %** la varianza en las calificaciones del **Puntaje videojuego1**

- El valor de **R (0.740)** en el **Modelo 1** es el **coeficiente de correlación múltiple** entre las variables predictoras y la variable dependiente. Como en la variable **Puntaje Videojuego1** es el único en este modelo en el que podemos ver que el **valor de R es el mismo valor que el Coeficiente de correlación de *Pearson* en nuestra matriz de correlación por pares.**

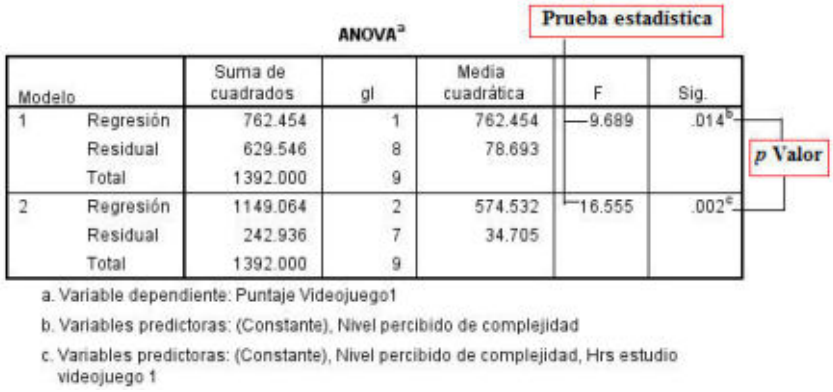
• En el **Modelo 2** las variables independientes **Nivel percibido de complejidad** y **Hrs de estudio videojuego1**, generan un coeficiente de correlación múltiple, ***R* = 0.909**.

• El ***R* Cuadrado corregida** se ajusta para un sesgo en **R cuadrado**. Con sólo unas pocas variables predictoras (a menos 2 variables independientes), El **R Cuadrado corregida** ajustado debe ser similar al valor de **R Cuadrado**. Normalmente tomaríamos el valor de **R Cuadrado** pero se recomienda seleccionar el valo**r R Cuadrado corregida** al tener más de **2 variables independientes.**

• **El error típ. de la estimación** es una medida de la variabilidad de la correlación múltiple.

Al usar el método de pasos sucesivos, **SPSS** genera una tabla **ANOVA. Ver Figura 5.69.**

Figura 5.69. Tabla ANOVA



ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	762.454	1	762.454	9.689	.014[b]
	Residual	629.546	8	78.693		
	Total	1392.000	9			
2	Regresión	1149.064	2	574.532	16.555	.002[c]
	Residual	242.936	7	34.705		
	Total	1392.000	9			

a. Variable dependiente: Puntaje Videojuego1

b. Variables predictoras: (Constante), Nivel percibido de complejidad

c. Variables predictoras: (Constante), Nivel percibido de complejidad, Hrs estudio videojuego 1

Fuente: SPSS 20 IBM

- El **ANOVA** prueba la significatividad de cada modelo de regresión para ver si la regresión predicha por las variables independientes, explican una cantidad significativa de la varianza en a variable dependiente.
- Al igual que con cualquier **ANOVA**, los elementos esenciales de información necesarios son el **gl,** el valor ***F*** el **valor de la probabilidad.** Ambos modelos de regresión explican una cantidad significativa de la variación en la variable dependiente.

Modelo 1: F (1,8) = 9.689; p < 0.05

Modelo 2: F (2,7) = 16.555; p <0.01

- La división de la **Suma de cuadrados** por los **grados de libertad (gl)** nos da la **Media Cuadrática o varianza**. Podemos ver que la regresión explica mucho más varianza que el **error o Residual.**
- Calculamos **R cuadrada** dividiendo la **Suma de Cuadrados de regresión** por la **Suma Total de Cuadrados**.

(762.454/1392.000)=0.548=R cuadrada

- Al igual que con el método **Introducir**, **SPSS** produce la **tabla Coeficientes**. Al aplicar el método de **Pasos sucesivos,** sólo se incluyen las variables seleccionadas para el modelo final. **Ver Figura 5.70**

Figura 5.70. Tabla Coeficientes

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	2.151	17.526		.123	.905
	Nivel percibido de complejidad	.974	.313	.740	3.113	.014
2	(Constante)	-2.615	11.726		-.223	.830
	Nivel percibido de complejidad	.761	.217	.579	3.504	.010
	Hrs estudio videojuego 1	.569	.171	.551	3.338	.012

a. Variable dependiente: Puntaje Videojuego1

Fuente: SPSS 20 IBM

- La columna **B Coeficientes no estandarizados** nos reporta los coeficientes de las variables en la ecuación de la regresión para cada modelo. Así:

Modelo 1: Puntaje Videojuego1= 2.151 + 0.974 Nivel percibido de complejidad

Modelo 2: Puntaje Videojuego1= −2.615 + 0.761 Nivel percibido de complejidad + 0.569 Hrs estudio Videojuego1

- La columna de **Coeficiente tipificado Beta** muestra la contribución que una variable hace al modelo. La ponderación que se hace de beta es la cantidad promedio en que la **variable dependiente aumenta cuando la variable independiente aumenta en una desviación estándar** (**todas las demás variables independientes se mantienen constantes**).**Como están estandarizados, podemos compararlos**.
- Se realizan **pruebas *t*** para probar la hipótesis de **dos colas** de que **el valor beta es significativamente mayor o menor que cero**. Esto también nos permite ver **qué predictores son significativos**.
- Observando los valores de **Sig.** en nuestro ejemplo, podemos ver que para el **Modelo 1** que la puntuación de la variable **Nivel percibido de complejidad** es significativo (***p*=0.014 < 0.05**). Sin embargo, con el **Modelo 2** tanto las puntuaciones del **Nivel percibido de complejidad** (***p*=0.01 <0.05**) y de **Hrs estudio videojuego 1** (***p*= 0.012 <0.05**) son predictores significativos (valores encuadrados en la tabla de coeficientes).
- Se recomienda utilizar el **Modelo 2 porque aporta más a la varianza.**
- La columna de **Coeficientes no estandarizados de error tip.** proporciona una estimación de la variabilidad de los coeficientes.
- Cuando se excluyen las variables del modelo sus **valores beta**, los **valores t** y los **valores significativos** son mostrados en la tabla **Variables Excluidas**. **Ver Figura 5.71.**

Figura 5.71. Tabla Variables Excluidas

Variables excluidas[a]

Modelo		Beta dentro	t	Sig.	Correlación parcial	Estadísticos de colinealidad
						Tolerancia
1	Hrs estudio videojuego 1	.551[b]	3.338	.012	.784	.914
	Inteligencia jugador	.309[b]	1.311	.231	.444	.930
	Nivel de satisfacción del videojuego	-.309[b]	-1.261	.248	-.430	.876
	Disponibilidad de adquirirlo	-.227[b]	-.937	.380	-.334	.978
2	Inteligencia jugador	.150[c]	.848	.429	.327	.835
	Nivel de satisfacción del videojuego	-.233[c]	-1.476	.190	-.516	.858
	Disponibilidad de adquirirlo	-.005[c]	-.024	.982	-.010	.806

a. Variable dependiente: Puntaje Videojuego1

b. Variables predictoras en el modelo: (Constante), Nivel percibido de complejidad

c. Variables predictoras en el modelo: (Constante), Nivel percibido de complejidad, Hrs estudio videojuego 1

Fuente: SPSS 20 IBM

El valor **Beta dentro** proporciona una estimación del peso de beta, si es que se incluyó en el modelo.

- Los resultados de las **pruebas *t*** para cada variable independiente se detallan con su valor de probabilidad.
- Del **Modelo 1** podemos ver que el **valor de *t*** para **Hrs estudio videojuego1** es significativo (***p=0.012<0.05***). Sin embargo, como hemos utilizado el método de **Pasos sucesivos ésta variable ha sido excluida del modelo.**
- Como la variable **Hrs estudio videojuego1** se ha incluido en el **Modelo 2**, se ha eliminado de esta tabla.
- Como la variable **Nivel de complejidad** está presente en ambos modelos no se menciona en la **tabla Variables excluidas.**
- El valor de **Correlación Parcial** indica la contribución que el **predictor excluido haría si decidimos incluirlo en nuestro modelo.**
- Las **Estadísticas de colinealidad con los valores de tolerancia** como regla general, **un valor de tolerancia por debajo de 0.1 indica un problema grave.**

5.22.3. Regresión lineal múltiple. Método: Pasos suscesivos con prueba de datos.

Paso 1: Establecimiento de los Objetivos

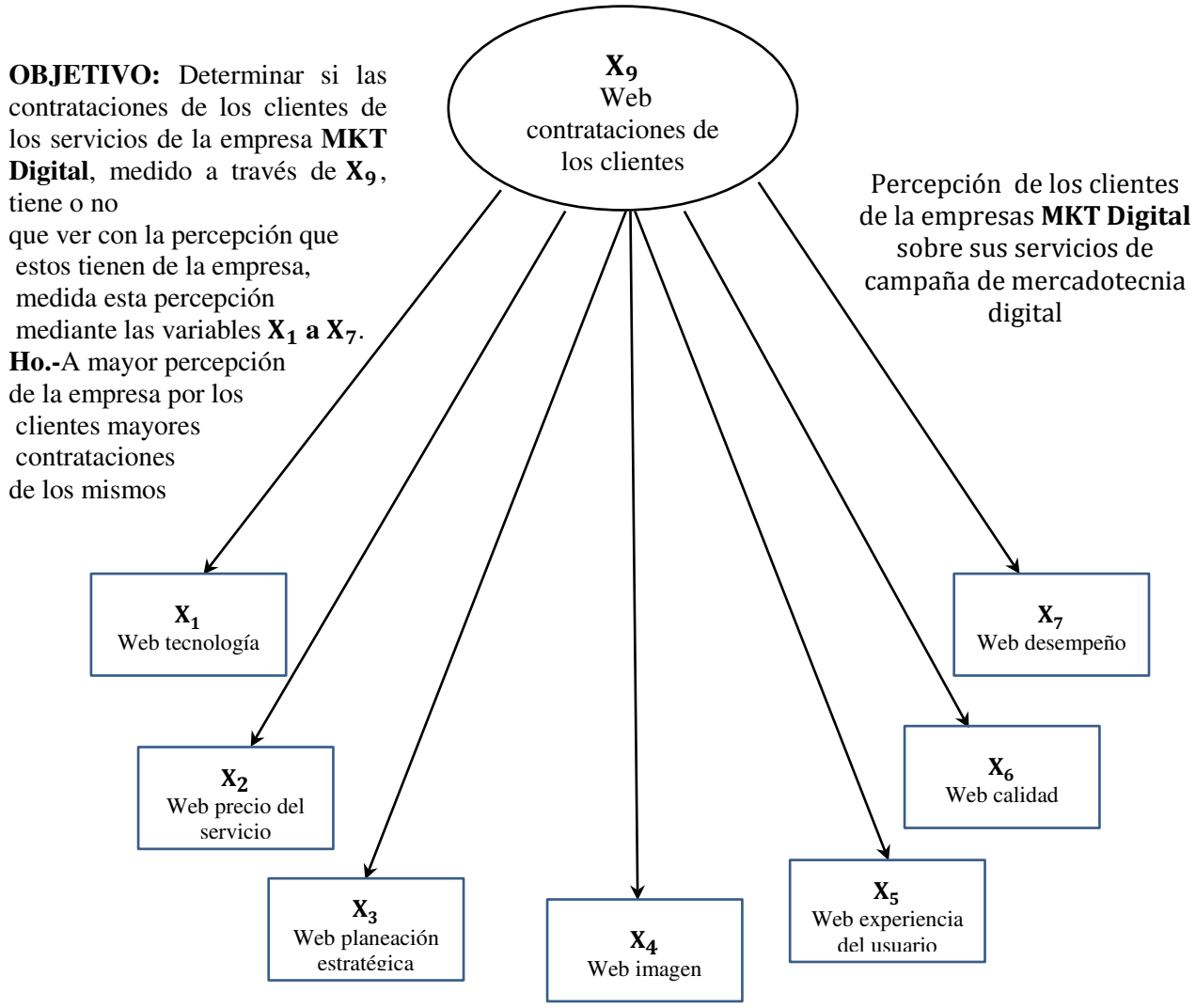
Problema 5: Determinar si el grado de relación con los clientes de la empresa **MKT Digital** (medido a través de **X_9.-Web contrataciones de clientes**) tiene o no que ver con la percepción que tienen éstos de **MKT Digital**, basada dicha percepción en las variables **X_1 a $_7$.** El **CEO**, considera lógico que el grado de relación de los clientes este influenciado por esas variables (así, se establecen las variables dependiente e independientes) que al estar basadas en percepciones de los clientes, orilla a aplicar relaciones estadísticas con error.

Para poder aplicar el modelo de regresión lineal, los investigadores seleccionaron la variable: contrataciones de los clientes (**X_9**) como una variable dependiente (**Y)**, para que pueda ser predecida por un grupo de **7** variables independientes que representan la Percepción de los clientes de la empresas **MKT Digital** sobre sus servicios de campaña de mercadotecnia digital (**X_1-X_7**). Se cumple la identificación de variables dependientes e independientes.

Adicionalmente a los resultados que se obtengan de la predicción del nivel de la relación con los clientes, los investigadores también están muy interesados en identificar las variables que mayores efectos, para implementar campañas de mercadotecnia digital más efectivas.

Ver Figura 5.72

Figura 5.72. Modelo problema para análisis de regresión lineal



Fuente: propia

Paso 2: Desarrollo del plan de análisis

- Para el caso de nuestro ejemplo se cuenta con **100 observaciones y 7** variables (relación de **14 a1**). En el diseño de un plan de análisis basado en la regresión lineal, el investigador debe tomar en cuenta un tema fundamental: **el tamaño de la muestra.**
- El tamaño de la muestra es el factor más importante para la fiabilidad de los resultados que puede controlar el investigador. Con muestras pequeñas **(<20 observaciones)** el análisis de regresión lineal sólo será adecuado cuando exista una única variable independiente y, aun así, solo las relaciones fuertes podrán detectarse con cierta certeza.
- Por el contrario, con **tamaños muestrales superiores a las 1000 encuestas u observaciones, los test de significatividad se vuelven demasiado sensibles**

haciendo que casi todas las relaciones o hipótesis sean estadísticamente significativa.

- El poder de una regresión lineal hace referencia a la probabilidad de que un R^2 sea significativo, dado un nivel de significatividad, un tamaño muestral y un número de variables independientes predeterminadas.

Tamaño de la muestra:

- Gran influencia sobre la significatividad de la relación (ver **Figura 5.73**)
- **Nunca menos de 5 observaciones por cada variable independiente, lo adecuado son ratios de 15 a 20.**
- Para el caso de nuestro ejemplo se cuenta con **100 observaciones y 7 variables** (relación de **14 a 1**).

Figura 5.73 Mínimo R^2 que se puede encontrar estadísticamente significativo con una potencia de 0.80 para diferentes variables independientes y tamaños muestrales

	Nivel de significación (alfa)=0.01 Número de variables Independientes				Nivel de significación (alfa)=0.05 Número de variables Independientes			
Tamaño muestral	2	5	10	20	2	5	10	20
	2				2			
20	45	56	71	NA	39	48	64	NA
50	23	29	36	49	19	23	29	42
100	13	16	20	26	10	12	15	21
250	5	7	8	11	4	5	6	8
500	3	3	4	6	3	4	5	9
1000	1	2	2	3	1	1	2	2

Fuente: Cohen y Cohen 1983

Paso 3: Condiciones de Aplicabilidad del Análisis de Regresión

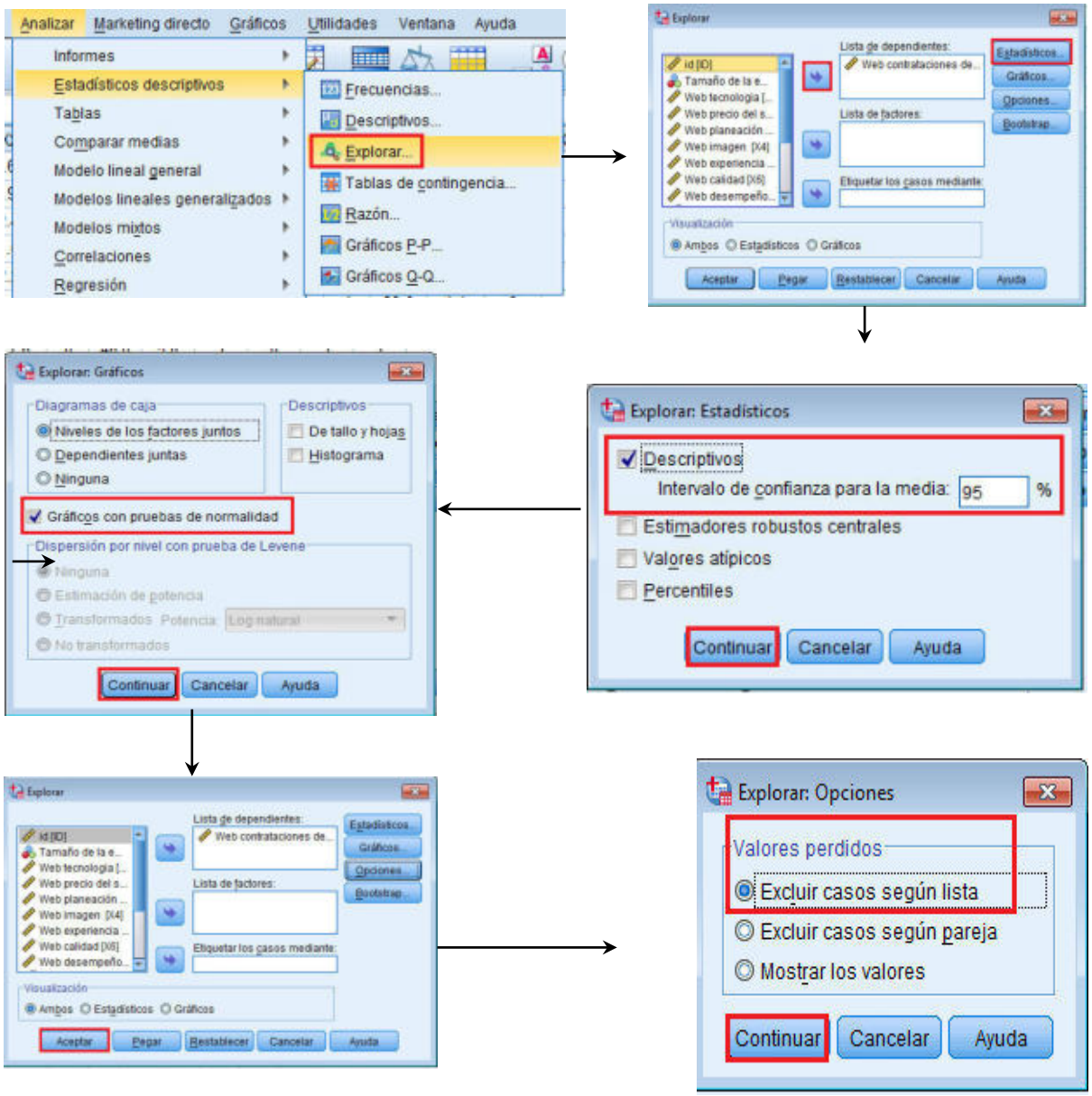
Las condiciones de aplicabilidad del análisis de regresión, deben considerarse en **2** etapas:

1. **Antes de estimar el modelo:** Sólo pueden comprobarse en las variables independientes y la dependiente de manera individual.
2. **Después de estimar el modelo:** Podrá evaluarse si se cumplen las condiciones de manera conjunta y, por ello, los resultados son fiables. Recordemos que las **3 condiciones** que debían cumplir las variables dependientes e independientes eran las de **normalidad, homoscedasticidad y linealidad.**

- **Normalidad:** Dado que la **p>0.05** **Se Acepta** H_0.-La variable X_9 **SI** tienen una población con distribución (**No Esencial**). Ver **Figura 5.74**
- **Homoscedasticidad: La base de datos se considera cumple con homocedasticidad** (**No Esencial**)
- **Linealidad:** Los gráficos de dispersión no parecer indicar la existencia de relaciones no lineales entre la variable dependiente y las independientes. **(Esencial por las correlaciones entre variables**)

Para el caso de la **normalidad:**
-Teclear: Analizar->Estadísticos descriptivos->Explorar->Seleccionar lista de variables dependientes: X_9 Web contrataciones de los clientes->Estadísticos->Seleccionar: Descriptivos a intervalo de confianza para la media: 95% ->Continuar->Seleccionar: Gráficos con prueba de normalidad->Continuar->Opciones->Valores perdidos: Excluir casos según lista->Continuar ->Aceptar. Ver **Figura 5.74.**

Figura 5.74.- Proceso para verificar distribución normal de la variable X_9



Fuente: SPSS 20 IBM

SPSS genera tablas y gráficos que reportan que la variable **X_9**. Contrataciones de los clientes **SÍ** responde a una distribución normal.
Dado que la ***p*>0.05 , Se Acepta H_0.-**Las variables **X_9 SI** tienen una población con distribución normal. Ver **Figura 5.75.**

Figura 5.75. Resultados Prueba de normalidad variable X_9

Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Web contrataciones de clientes	.079	100	.131	.985	100	.320

a. Corrección de la significación de Lilliefors

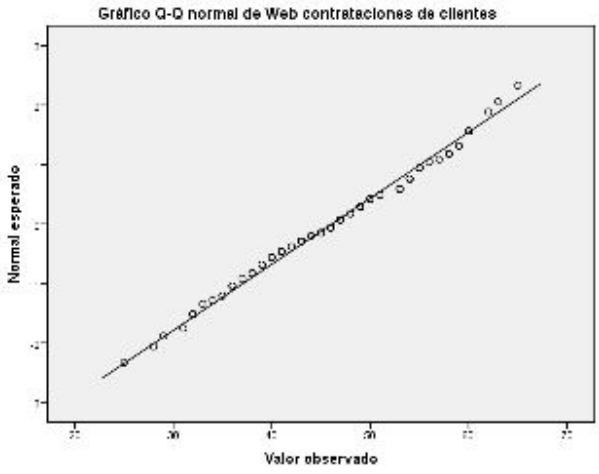
Fuente: SPSS 20 IBM

Resumen del procesamiento de los casos

	Casos					
	Válidos		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Web contrataciones de clientes	100	100.0%	0	0.0%	100	100.0%

Descriptivos

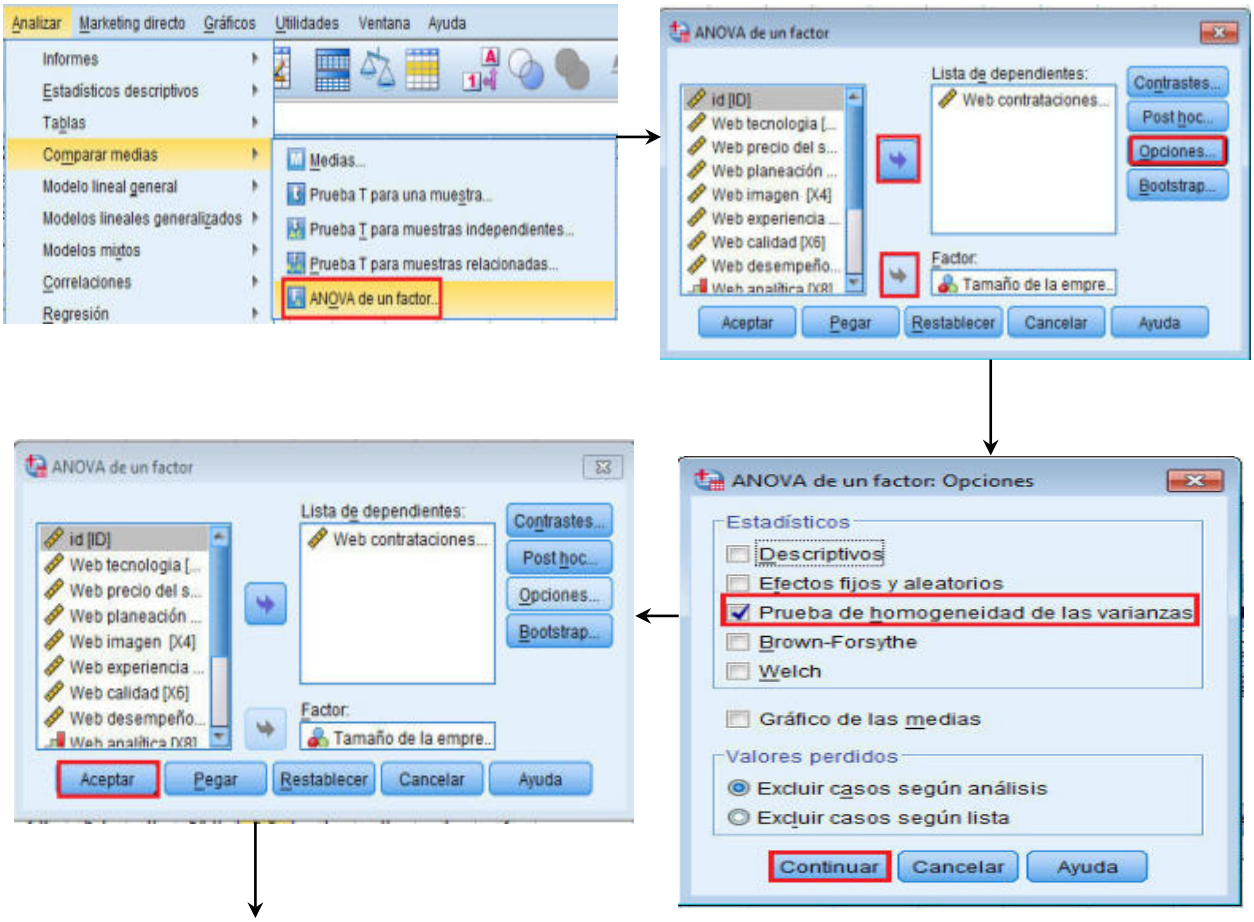
			Estadístico	Error típ.
Web contrataciones de clientes	Media		46.100	.8889
	Intervalo de confianza para la media al 95%	Límite inferior	44.316	
		Límite superior	47.884	
	Media recortada al 5%		46.167	
	Mediana		46.500	
	Varianza		80.798	
	Desv. típ.		8.9888	
	Mínimo		25.0	
	Máximo		65.0	
	Rango		40.0	
	Amplitud intercuartil		14.8	
	Asimetría		-.063	.241
	Curtosis		-.725	.478



Fuente: SPSS 20 IBM

Para el caso de **homocedasticidad.**
-Teclear: Analizar-> Comparar medias->ANOVA de un factor>Selección Variables métricas (X_9) en Lista de dependientes; Selección Variable de agrupación nominal: Tamaño de la empresa (V_3) en Factor->Opciones->Estadísticos: prueba de homogeneidad de varianzas->Continuar->Aceptar. Ver Figura 5.76.

Figura 5.76.- Proceso para verificar la homocedasticidad de la variable X_9



ANOVA de un factor

[Conjunto_de_datos1] C:\Users\Juan\Desktop\proy libro mc\WEB

Prueba de homogeneidad de varianzas

Web contrataciones de clientes

Estadístico de Levene	gl1	gl2	Sig.
.077	2	97	.926

ANOVA de un factor

Web contrataciones de clientes

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	157.377	2	78.689	.973	.381
Intra-grupos	7841.623	97	80.841		
Total	7999.000	99			

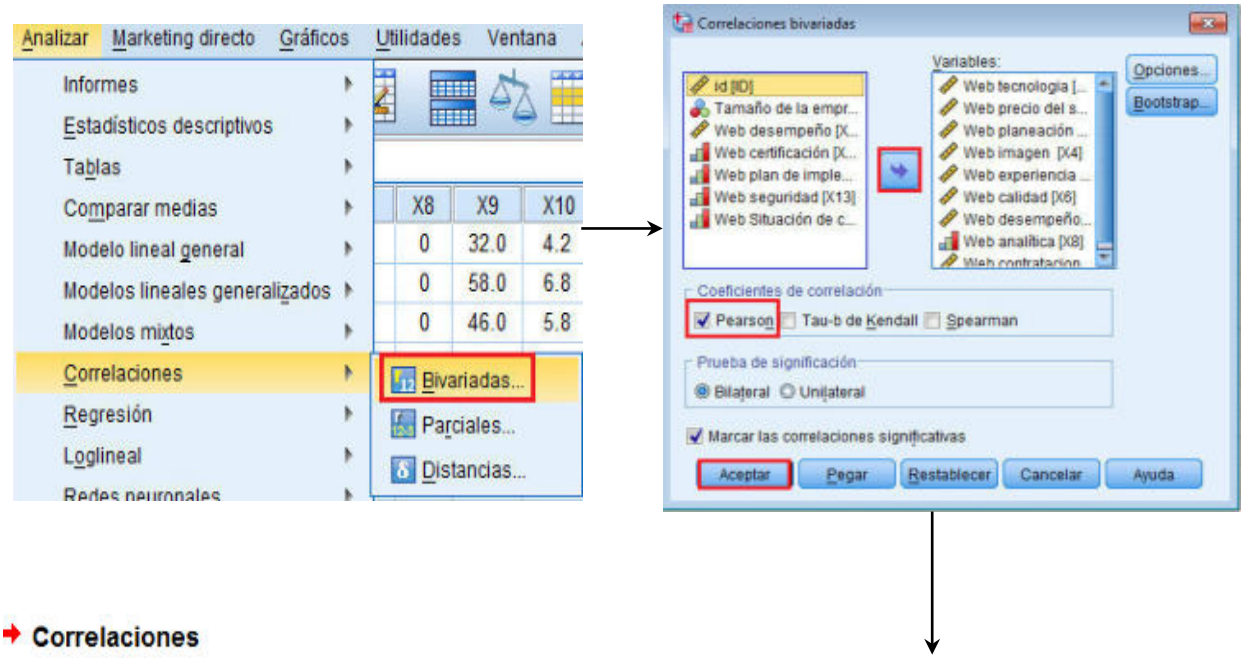
Fuente: SPSS 20 IBM

Dado que la $p>0.05$ **Se Acepta** H_0.-La variable X_9 SI tienen una población con homocedasticidad respecto al tamaño de la empresa V_3

Para el caso de **linealidad**

-Teclear: Analizar->Selección de variables: X_1-X_7 y X_9->Coeficientes de correlación: *Pearson*->Aceptar. Ver Figura 5.77

Figura 5.27.- Proceso para verificar la linealidad de las variables X_1-X_7 y X_9



Correlaciones

[Conjunto_de_datos1] C:\Users\Juan\Desktop\proy libro mc\WEB diseño.sav

Correlaciones

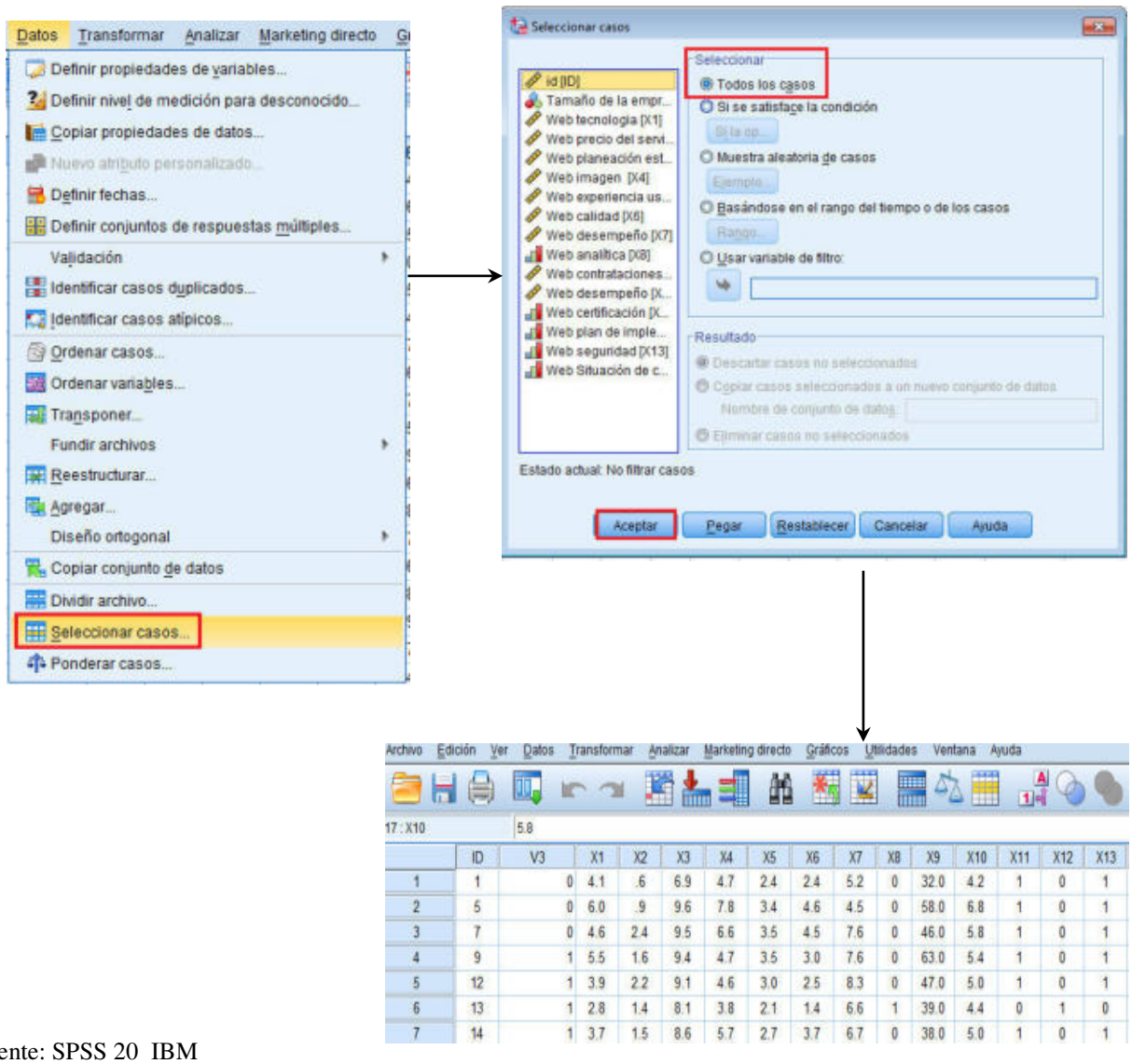
		Web tecnologia	Web precio del servicio	Web planeación estrategica	Web imagen	Web experiencia usuario	Web calidad
Web tecnologia	Correlación de Pearson	1	-.349**	.509**	.050	.612**	.077
	Sig. (bilateral)		.000	.000	.618	.000	.444
	N	100	100	100	100	100	100
Web precio del servicio	Correlación de Pearson	-.349**	1	-.487**	.272**	.513**	.185
	Sig. (bilateral)	.000		.000	.006	.000	.065
	N	100	100	100	100	100	100
Web planeación estrategica	Correlación de Pearson	.509**	-.487**	1	-.116	.067	-.035
	Sig. (bilateral)	.000	.000		.250	.510	.731
	N	100	100	100	100	100	100
Web imagen	Correlación de Pearson	.050	.272**	-.116	1	.299**	.788**
	Sig. (bilateral)	.618	.006	.250		.003	.000
	N	100	100	100	100	100	100
Web experiencia usuario	Correlación de Pearson	.612**	.513**	.067	.299**	1	.240*
	Sig. (bilateral)	.000	.000	.510	.003		.016
	N	100	100	100	100	100	100
Web calidad	Correlación de Pearson	.077	.185	-.035	.788**	.240*	1
	Sig. (bilateral)	.444	.065	.731	.000	.016	
	N	100	100	100	100	100	100

Fuente: SPSS 20 IBM

Dada la inserción de la nueva variable $\mathbf{X_9}$, se sugiere realizar todo el proceso mostrado en **AFE** de análisis de matriz de correlación. Dado que cumple con los previos de normalidad y homocedasticidad, continuaremos con el análisis de regresión lineal

-Problema 6: Para quitar filtros base de datos **WEB_MKT_Digital. sav. Teclear: Datos->Seleccionar casos->Seleccionar: Todos los casos. Ver Figura 5.78.**

Figura 5.78.- Proceso para quitar filtros a la base de datos



	ID	V3	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
1	1	0	4.1	.6	6.9	4.7	2.4	2.4	5.2	0	32.0	4.2	1	0	1
2	5	0	6.0	.9	9.6	7.8	3.4	4.6	4.5	0	58.0	6.8	1	0	1
3	7	0	4.6	2.4	9.5	6.6	3.5	4.5	7.6	0	46.0	5.8	1	0	1
4	9	1	5.5	1.6	9.4	4.7	3.5	3.0	7.6	0	63.0	5.4	1	0	1
5	12	1	3.9	2.2	9.1	4.6	3.0	2.5	8.3	0	47.0	5.0	1	0	1
6	13	1	2.8	1.4	8.1	3.8	2.1	1.4	6.6	1	39.0	4.4	0	1	0
7	14	1	3.7	1.5	8.6	5.7	2.7	3.7	6.7	0	38.0	5.0	1	0	1

Fuente: SPSS 20 IBM

- La **ausencia de normalidad** puede corregirse, como se indicó al inicio de esta materia, transformando las variables originales mediante logaritmos neperianos. Sin embargo, varios autores establecen que la no normalidad de las variables no tiene efectos significativos en el **estadístico *F*** (Hair et al., 2010).
- Varios investigadores consideran que la no existencia de **homoscedasticidad** de las variables, no tiene un efecto significativo en el **estadístico *F*** (Hair et al., 1999).

- **La linealidad** se deberá soportar con lo encontrado en el análisis factorial en el estudio de las matrices de correlación El investigador debería estimar el modelo considerando las variables transformadas y sin transformar, para después, cuando se compruebe si, de manera global, se violan las hipótesis señaladas, mantener las variables de la manera que menos distorsión provoquen respecto al cumplimiento de estas hipótesis.

Paso 4: Estimación del Modelo y Ajuste del mismo

- Habiendo sido especificados los objetivos del análisis, seleccionado las variables dependientes e independientes y comprobadas las condiciones de aplicabilidad del modelo de regresión lineal, el investigador está preparado para estimar el modelo y establecer **la bondad del mismo (su ajuste)**. Esta tarea se desdobla en tres decisiones elementales:

1. **Seleccionar un método para estimar el modelo**
2. Establecer la **significatividad (0.01 o 0.05)** global del modelo estimado y de los coeficientes de cada una de las variables independientes
3. Determinar si hay observaciones **que ejercen una influencia** no deseable sobre los resultados.

En esta primera decisión, el investigador debe optar entre dos alternativas:

1. Decidir aquellas **variables independientes** que, según su conocimiento sobre el tema, pueden ejercer algún tipo de influencia sobre la variable dependiente e incluirlas
2. O bien recurrir a **procedimientos secuenciales**, en los cuales es el propio programa quien va introduciendo y eliminando del análisis aquellas variables que aseguren la mejor especificación del modelo.

En el primer tipo de aproximación, el investigador debe estar muy
seguro de que no está dejando fuera variables relevantes, ni introduciendo variables irrelevantes, que puedan distorsionar el modelo de regresión lineal.

En el **segundo enfoque**, el **proceso iterativo** asegura que se acaban considerando las variables que mejor pueden explicar el comportamiento de la variable dependiente, por este motivo desarrollaremos en este tema este último enfoque.

- **Los métodos secuenciales estiman la ecuación de regresión añadiendo o eliminando** (según los dos enfoques que veremos) aquellas variables que cumplen determinados criterios.
- **Esta aproximación ofrece un procedimiento objetivo para seleccionar las variables,** que maximiza la capacidad predictiva del modelo con el menor número posible de variables independientes.
- Aunque este enfoque parece ideal, hay que tener en cuenta que es muy sensible al efecto de la **multicolinealidad** y, por ello, su determinación y corrección es crítica en estos modelos.

Eliminación hacia atrás: Es básicamente un procedimiento de **prueba y error**, y comienza estimando una recta de regresión con todas las variables independientes posibles **y luego va eliminando aquellas que no contribuyen significativamente**. Los pasos son los siguientes:

1. Cálculo de una recta de regresión con todas las variables independientes posibles.
2. Cálculo de un **estadístico *F*** parcial para cada variable que computa la varianza que explicaría el modelo si se eliminasen todas las variables menos esa.

3. Se eliminan las variables con ***F*** parciales que indican que no realizan una contribución estadísticamente significativa.
4. Después de eliminar esas variables se vuelve a estimar la recta de regresión con las que quedan.
5. Se vuelve al **paso 2** hasta que sólo quedan las variables significativas.

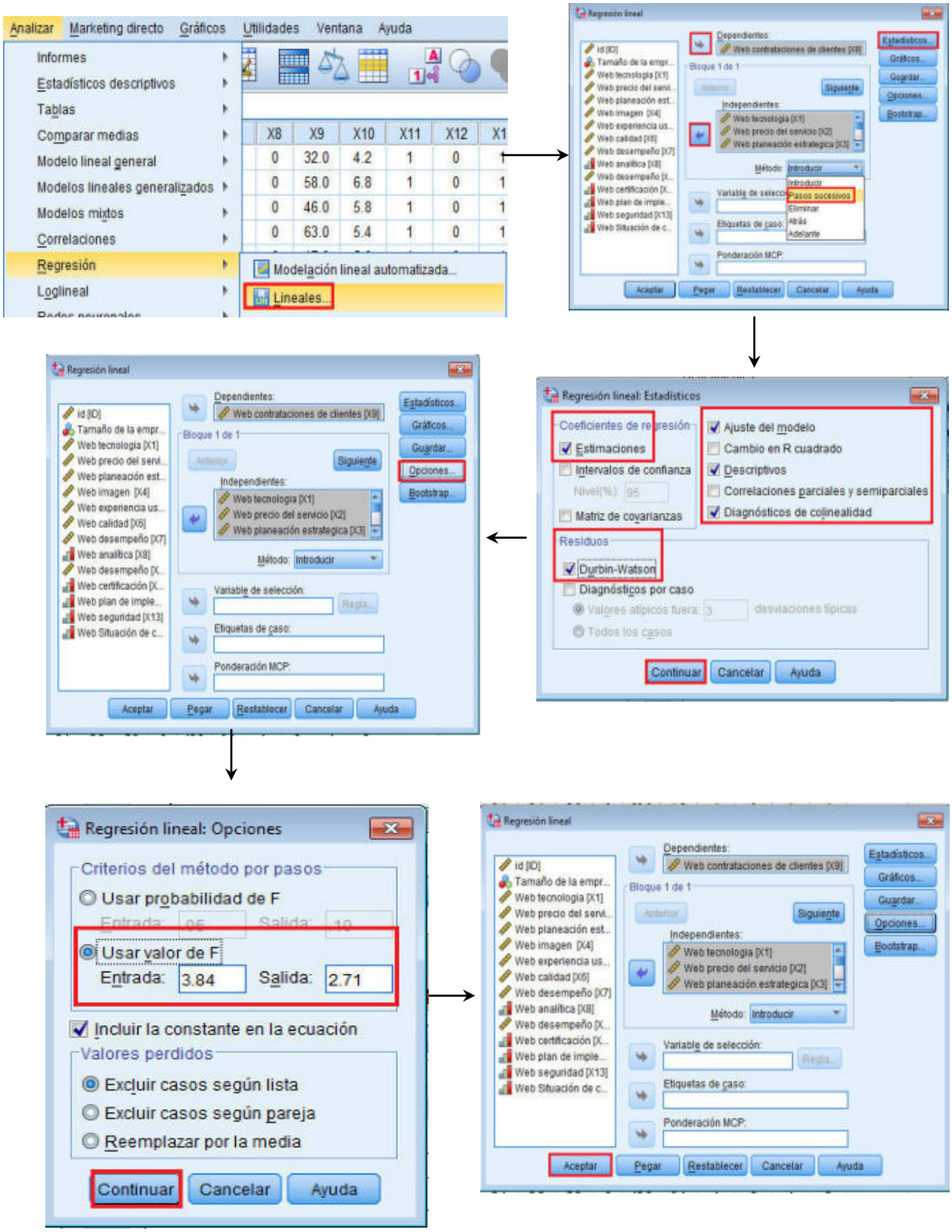
Estimación paso a paso: Es el procedimiento secuencial más utilizado dado que permite analizar la contribución de cada variable independiente por separado al modelo de regresión lineal. **Se diferencia del anterior en que evalúa una a una las variables antes de incorporarlas al modelo de regresión lineal** y, además, puede eliminar una variable después de haberla introducido en una etapa anterior. Los pasos que sigue son los siguientes:

1. Comienza con el modelo de regresión más simple, que es el formado por la constante y aquella variable que está más correlacionada con la variable dependiente.
2. Examina los coeficientes de correlación parcial para encontrar la variable independiente la mayor proporción del error que se comete con la recta de regresión anterior.
3. Vuelve a calcular la ecuación de regresión utilizando ahora las dos variables independientes seleccionadas y analiza el valor de la F parcial de la primera variable para ver si todavía lleva a cabo una contribución significativa dado que hemos incluido una variable adicional. Si no lo hace la elimina y en caso contrario la mantiene.

El proceso continúa examinando todas las variables independientes para ver cuál debe ser introducida en la ecuación. Cuando se incluye una nueva se examinan las ya introducidas para determinar cuál debe permanecer y así hasta que ninguna variable cumple el criterio de entrada.

-Problema 7: realice la regresión lineal de su modelo: ¿cuál es el nivel de relación de las contrataciones de sus clientes (X_9) respecto a la percepción que tienen de la empresa MKT Digital sobre sus servicios de campaña de mercadotecnia digital? Teclear: Analizar-Regresión->Lineales->Selección variable dependiente (X_9); Selección de variables independientes (X_1 - X_7) ->Método: Pasos sucesivos->Estadísticos->Coeficientes de regresión: Estimaciones; Ajuste del modelo; Descriptivos; Diagnósticos de colinealidad; Residuos: Durbin-Watson->Continuar->Opciones->Usar valor de F-> Continuar->Aceptar. Ver Figura 5.79.

Figura 5.79.-Proceso de regresión lineal



Fuente: SPSS 20 IBM

SPSS produce la tabla de **Estadísiticos descriptivos. Ver Figura 5.80**

Figura 5.80. Tabla Estadísticos descriptivos

Regresión

[Conjunto_de_datos1] C:\Users\Juan\Desktop\proy lit

Estadísticos descriptivos

	Media	Desviación típica	N
Web contrataciones de clientes	46.100	8.9888	100
Web tecnologia	3.515	1.3207	100
Web precio del servicio	2.364	1.1957	100
Web planeación estrategica	7.894	1.3865	100
Web imagen	5.248	1.1314	100
Web experiencia usuario	2.916	.7513	100
Web calidad	2.665	.7706	100
Web desempeño	6.971	1.5852	100

Fuente: SPSS 20 IBM

Sus clientes realizan en promedio el **46% compras Correlaciones 50%+1.**

SPSS produce la **tabla de Correlaciones**. Ver **Figura 5.81.**

Figura 5.81.-Tabla de Correlaciones

Correlaciones

		Web contrataciones de clientes	Web tecnologia	Web precio del servicio	Web planeación estrategica	Web imagen	Web experiencia usuario	Web calidad	Web desempeño
Correlación de Pearson	Web contrataciones de clientes	1.000	.676	.082	.559	.224	.701	.255	-.192
	Web tecnologia	.676	1.000	-.349	.509	.050	.612	.077	-.483
	Web precio del servicio	.082	-.349	1.000	-.487	.272	.513	.185	.470
	Web planeación estrategica	.559	.509	-.487	1.000	-.116	.067	-.035	-.448
	Web imagen	.224	.050	.272	-.116	1.000	.299	.788	.200
	Web experiencia usuario	.701	.612	.513	.067	.299	1.000	.240	-.055
	Web calidad	.255	.077	.185	-.035	.788	.240	1.000	.177
	Web desempeño	-.192	-.483	.470	-.448	.200	-.055	.177	1.000
Sig. (unilateral)	Web contrataciones de clientes	.	.000	.209	.000	.012	.000	.005	.028
	Web tecnologia	.000	.	.000	.000	.309	.000	.222	.000
	Web precio del servicio	.209	.000	.	.000	.003	.000	.032	.000
	Web planeación estrategica	.000	.000	.000	.	.125	.255	.366	.000
	Web imagen	.012	.309	.003	.125	.	.001	.000	.023
	Web experiencia usuario	.000	.000	.000	.255	.001	.	.008	.293
	Web calidad	.005	.222	.032	.366	.000	.008	.	.039
	Web desempeño	.028	.000	.000	.000	.023	.293	.039	.

Fuente: SPSS 20 IBM
Nota: Datos en encuadre rojo con valor <=0.05

- Se observan **28** datos por lado de la matriz bajo/sobre la diagonal
- La mayoría deberá ser 50%+1=**28/2+1=15** correlaciones deben ser significativas para cumplir
- Existen **21** al **0.05** con este también cumple, son significativas. Se recomienda usar. **SI CUMPLE CONDICION**

SPSS, genera la **tabla Variables introducidas/eliminadas. Ver Figura 5.82**

Figura 5.82. Tabla Variables introducidas/eliminadas.

Variables introducidas/eliminadas[a]

Modelo	Variables introducidas	Variables eliminadas	Método
1	Web experiencia usuario	.	Por pasos (criterio: F para entrar >= 3.840, F para salir <= 2.710).
2	Web planeación estrategica	.	Por pasos (criterio: F para entrar >= 3.840, F para salir <= 2.710).
3	Web calidad	.	Por pasos (criterio: F para entrar >= 3.840, F para salir <= 2.710).

a. Variable dependiente: Web contrataciones de clientes

Fuente: SPSS 20 IBM

- Sólo estas **3/7** variables explican más (**no se necesitan 7**) la variable x_9 **(contratación de servicios de los clientes)** a partir de cómo perciben a la empresa **MKT Digital.** El resto de las variables no aportan lo suficiente.

SPSS genera la tabla **Resumen del modelo con 3 modelos** resultantes. Ver **Figura 5.83**

Figura 5.83.Tabla Resumen del modelo

Resumen del modelo[d]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	.701[a]	.491	.486	6.4458	
2	.869[b]	.755	.750	4.4980	
3	.876[c]	.768	.761	4.3948	1.963

a. Variables predictoras: (Constante), Web experiencia usuario

b. Variables predictoras: (Constante), Web experiencia usuario , Web planeación estrategica

c. Variables predictoras: (Constante), Web experiencia usuario , Web planeación estrategica, Web calidad

d. Variable dependiente: Web contrataciones de clientes

Fuente: SPSS 20 IBM

Modelo 1

- La variable $\mathbf{X_5}$ (Web experiencia del usuario) se relaciona en un **70%** con un nivel de contratación de servicios a la empresa **MKT Digital.**
- Recordar que $\mathbf{X_5}$ Web experiencia, en el proceso de análisis factorial fue la variable común o excluida de los grupos formados dada su alta importancia en la correlación, así, se concluye que 7/10 clientes compran a **MKT Digital** por la experiencia que provoca al usuario sus servicios.
- Así la variable explica en un **R|2** aprox. **49%** a $\mathbf{X_9}$**.** Se prefiere usar el $\boldsymbol{R^2}$corregido.

Modelo 2

- Las variables $\mathbf{X_5}$ (Web experiencia del usuario) y $\mathbf{X_3}$ (Web planeación estratégica,) juntas, explican como $\boldsymbol{R^2}$corregido, el **75% de** $\mathbf{X_9}$. La diferencia entre modelo 1 y 2 es notoria (0.486 a 0.75)

Modelo 3

- Las variables $\mathbf{X_5}$ (Web experiencia del usuario), $\mathbf{X_3}$ (Web planeacio n estrate gica), y $\mathbf{X_6}$ (Web calidad) juntas, explican como $\boldsymbol{R^2}$ corregido, el **76% de** $\mathbf{X_9}$. La diferencia entre modelo 1 y 2 es muy baja (0.75 a 0.761). Es decisio n del investigador el elegir si se incluyen 1-2 o 3 variables para explicar el feno meno. En este caso es posible so lo adoptar los 2 primeras variables: $\mathbf{X_5}$**y** $\mathbf{X_3}$
- Nota: $\boldsymbol{R^2}$ corregido, hasta valores de 20% son todavía útiles en ciencias de la administración (mínimo 70% ciencias duras)
- **Durbin-Watson <2 (1.963) indica que el modelo de regresión es el adecuado**

SPSS produce la tabla **ANOVA. Ver Figura 5.84.**

Tabla 5.84. ANOVA

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	3927.309	1	3927.309	94.525	.000[b]
	Residual	4071.691	98	41.548		
	Total	7999.000	99			
2	Regresión	6036.513	2	3018.256	149.184	.000[c]
	Residual	1962.487	97	20.232		
	Total	7999.000	99			
3	Regresión	6144.812	3	2048.271	106.049	.000[d]
	Residual	1854.188	96	19.314		
	Total	7999.000	99			

a. Variable dependiente: Web contrataciones de clientes

b. Variables predictoras: (Constante), Web experiencia usuario

c. Variables predictoras: (Constante), Web experiencia usuario , Web planeación estrategica

d. Variables predictoras: (Constante), Web experiencia usuario , Web planeación estrategica, Web calidad

Fuente: SPSS 20 IBM

Modelo 1

- La variable $\mathbf{X_5}$ (Web experiencia del usuario) tiene una ***F*= 94.525 y es significativa**

Modelo 2

- Las variables $\mathbf{X_5}$ (Web experiencia del usuario) y $\mathbf{X_3}$ (Web planeación estratégica) tienen una ***F*= 149.84 y es significativa**

Modelo 3

- Las variables $\mathbf{X_5}$ (Web experiencia del usuario), $\mathbf{X_3}$ (Web planeación estratégica) y $\mathbf{X_6}$ (Web calidad) tienen una ***F*= 106.049 y es significativa. Disminuye respecto al Modelo 2. Así, con 2 variables se maximiza la recta, en lugar de insertar la 3ª variable.** A decisión del investigador si se incluye o no la 3er. Variable

SPSS genera la **tabla Coeficientes. Ver Figura 5.85.**

Figura 5.85.Tabla Coeficientes

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Estadísticos de colinealidad	
		B	Error típ.	Beta			Tolerancia	FIV
1	(Constante)	21.653	2.596		8.341	.000		
	Web experiencia usuario	8.384	.862	.701	9.722	.000	1.000	1.000
2	(Constante)	-3.489	3.057		-1.141	.257		
	Web experiencia usuario	7.974	.603	.666	13.221	.000	.996	1.004
	Web planeación estrategica	3.336	.327	.515	10.210	.000	.996	1.004
3	(Constante)	-6.514	3.248		-2.005	.048		
	Web experiencia usuario	7.623	.608	.637	12.548	.000	.937	1.068
	Web planeación estrategica	3.376	.320	.521	10.560	.000	.993	1.007
	Web calidad	1.400	.591	.120	2.368	.020	.940	1.064

a. Variable dependiente: Web contrataciones de clientes

Fuente: SPSS 20 IBM

Que complementa su información con la **tabla Resumen del modelo. Ver Figura 5.86.**

Figura 5.86. Resumen del modelo

Resumen del modelo[d]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	.701[a]	491	.486	6.4458	
2	.869[b]	.755	.750	4.4980	
3	.876[c]	.768	.761	4.3948	1.963

a. Variables predictoras: (Constante), Web experiencia usuario

b. Variables predictoras: (Constante), Web experiencia usuario , Web planeación estrategica

c. Variables predictoras: (Constante), Web experiencia usuario , Web planeación estrategica, Web calidad

d. Variable dependiente: Web contrataciones de clientes

Fuente: SPSS 20 IBM

Modelo 1

- **B**=0.701= **R** (Ver ambas Tablas)

Recordando que la ecuación de la Recta, es: $y=a+bx+e$

$$X_9=21.653+8.384X$$

Modelo 2

- La constante **disminuyó** (**-3.489**) la recta cae. **FIV**, se toma el mayor (**0, sin problemas; 0.3-2 con probable problema y >2 con problemas de multicolinealidad**). No hay problemas de multicolinealidad

$$X_9=-3.489+7.974X_5+3.336X$$

Modelo 3

- La constante disminuyó (**-6.514**) la recta cae. **FIV**, se toma el mayor. No hay problemas de multicolinealidad.

$$X_9= -6.514+7.623X_5+3.376X_3+1.4X_6$$

SPSS genera la **tabla Variables excluídas. Ver Figura 5.87.**

Figura 5.87. Tabla Variables excluídas.

Variables excluidas[a]

Modelo		Beta dentro	t	Sig.	Correlación parcial	Estadísticos de colinealidad		
						Tolerancia	FIV	Tolerancia mínima
1	Web tecnologia	.396[b]	4.812	.000	.439	.626	1.599	.626
	Web precio del servicio	-.377[b]	-5.007	.000	-.453	.737	1.357	.737
	Web planeación estrategica	.515[b]	10.210	.000	.720	.996	1.004	.996
	Web imagen	.016[b]	.216	.830	.022	.911	1.098	.911
	Web calidad	.092[b]	1.242	.217	.125	.942	1.061	.942
	Web desempeño	-.154[b]	-2.178	.032	-.216	.997	1.003	.997
2	Web tecnologia	.016[c]	.205	.838	.021	.405	2.469	.405
	Web precio del servicio	-.020[c]	-.267	.790	-.027	.464	2.156	.464
	Web imagen	.095[c]	1.808	.074	.181	.892	1.121	.892
	Web calidad	.120[c]	2.368	.020	.235	.940	1.064	.937
	Web desempeño	.094[c]	1.683	.096	.169	.799	1.252	.797
3	Web tecnologia	.030[d]	.386	.701	.040	.403	2.482	.403
	Web precio del servicio	-.029[d]	-.401	.690	-.041	.462	2.162	.462
	Web imagen	-.001[d]	-.009	.993	-.001	.357	2.804	.357
	Web desempeño	.071[d]	1.277	.205	.130	.769	1.301	.769

a. Variable dependiente: Web contrataciones de clientes

b. Variables predictoras en el modelo: (Constante), Web experiencia usuario

c. Variables predictoras en el modelo: (Constante), Web experiencia usuario , Web planeación estrategica

d. Variables predictoras en el modelo: (Constante), Web experiencia usuario , Web planeación estrategica, Web calidad

Fuente: SPSS 20 IBM

Modelo 1

- Se observa que la variable $\mathbf{X_5}$ (Web experiencia del usuario) no aparece porque es la primera en usarse. Así, el que sigue para entrar con B mayor es **X** (Web planeación estratégica) es la siguiente a entrar: **0.515**

Modelo 2

- El que sigue para entrar con B mayor es $\mathbf{X_6}$ (Web calidad) es la siguiente a entrar: **0.120**

Modelo 3

- Todas las B han sido agotadas

Es importante hacer notar que los valores *t*, tolerancia, FIV aparecerán por variable; sin embargo, nos interesan los valores a nivel modelo

SPSS produce la **tabla Diagnósticos de colinealidad.Ver Figura 5.88.**

Figura 5.88. Tabla Diagósticos de colinealidad

Diagnósticos de colinealidad[a]

Modelo	Dimensión	Autovalores	Índice de condición	Proporciones de la varianza			
				(Constante)	Web experiencia usuario	Web planeación estrategica	Web calidad
1	1	1.969	1.000	.02	.02		
	2	.031	7.928	.98	.98		
2	1	2.941	1.000	.00	.01	.00	
	2	.046	8.000	.03	.85	.19	
	3	.013	14.778	.97	.14	.80	
3	1	3.882	1.000	.00	.00	.00	.00
	2	.060	8.048	.01	.02	.11	.85
	3	.045	9.246	.02	.91	.14	.04
	4	.012	17.723	.97	.07	.75	.10

a. Variable dependiente: Web contrataciones de clientes

Fuente: SPSS 20 IBM

El Diagnóstico de colinealidad no aporta información ya que lo interesante son las multicolinealidades.

Dados los 0s presentados indican que el modelo es el adecuado de la tabla Estadísticos sobre los residuos.**Ver Figura 5.89.**

Figura 5.89. Estadísticos sobre los residuos

Estadísticos sobre los residuos[a]

	Mínimo	Máximo	Media	Desviación típica	N
Valor pronosticado	23.371	60.592	46.100	7.8784	100
Residual	-12.5428	7.5725	.0000	4.3277	100
Valor pronosticado tip.	-2.885	1.839	.000	1.000	100
Residuo típ.	-2.854	1.723	.000	.985	100

a. Variable dependiente: Web contrataciones de clientes

Fuente: SPSS 20 IBM

Paso 5: Interpretación de los Resultados.

En cualquier interpretación de los resultados de un análisis de regresión, el investigador debe prestar especial atención a analizar el efecto de la multicolinealidad, esto es, la posible correlación entre las variables independientes.

Aunque este es un problema de los datos, no de la especificación del modelo, puede tener importantes consecuencias:

1. Limita el valor del coeficiente de determinación
2. Hace difícil determinar la contribución de cada variable individualmente.
3. Dado que sus efectos se enmascaran en las correlaciones de unas con otras, pudiendo ocasionar que los coeficientes de cada variable sean incorrectamente estimados y tengan signos equivocados.
4. Dos de las medidas más habituales para establecer la existencia de multicolinealidad, son los llamados **valor de tolerancia y su inversa, el factor de inflación de la varianza (FIV).**

- **El valor de tolerancia y su inversa, el factor de inflación de la varianza (FIV)** nos indican en qué medida una variable independiente está explicada por otras variables independientes.
- **FIV=0** No hay problemas; **1.5 a 2.0** puede haber problemas; >2.0 hay problemas de multicolinealidad.
- En términos más sencillos, **cada variable independiente es considerada como dependiente y regresada contra el resto de independientes.**
- La tolerancia es la cantidad de variación de la variable independiente seleccionada que no es explicada por el resto de las variables independientes.
- Por lo tanto, **valores muy pequeños de tolerancia (y por lo tanto grandes de FIV) denotan una alta colinealidad.**
- Un punto de corte bastante común es **0.10**, que corresponde a valores de **FIV superiores a 10.** Este valor se da cuando el coeficiente de determinación de la regresión señalada es de **0.95.**

En el caso en que la **multicolinealidad sea muy elevada**, se proponen normalmente las siguientes soluciones:

1. Eliminar una o más de las variables que estén altamente correlacionadas e identificar otras posibles variables independientes para ayudar en la predicción.
2. Utilizar el modelo con todas las variables sólo con fines predictivos y no intentar en ningún momento interpretar los coeficientes de regresión.
3. Utilizar los coeficientes de correlación simples entre la variable dependiente y las independientes para entender la relación entre ambas variables.}
4. Recurrir a procedimientos más sofisticados de análisis de regresión, como la bayesiana o la regresión en componentes principales que, evidentemente, se alejan del objetivo de este curso.

SPSS , genera por último la tabla **Estaísiticos descriptivos . Vera Figura 5.90**

Figura 5.90. Taba Estadísiticos descriptivos.

Regresión

[Conjunto_de_datos1] C:\Users\Juan\Desktop\proy lib

Estadísticos descriptivos

	Media	Desviación típica	N
Web contrataciones de clientes	46.100	8.9888	100
Web tecnologia	3.515	1.3207	100
Web precio del servicio	2.364	1.1957	100
Web planeación estrategica	7.894	1.3865	100
Web imagen	5.248	1.1314	100
Web experiencia usuario	2.916	.7513	100
Web calidad	2.665	.7706	100
Web desempeño	6.971	1.5852	100

Fuente: SPSS 20 IBM

Una vez estimado el modelo y llevados a cabo los diagnósticos que confirman la validez de los resultados, podemos escribir nuestra recta de regresión como sigue:
Modelo 1: $X9= 21.653+8.384X$
Modelo 2: $X9= -3.489+7.974X_5+3.336X$
Modelo 3: $X9= -6.514+7.623X_5+3.376X_3+1.4X_6$

Suponiendo usar con 2 puntos al menos, el **modelo 2:**
Modelo 2: $X_9= -3.489+7.974X_5+3.336X$

Utilizando **medias del reporte de descriptivos** (ver **Figura. 5.71**):
$X_9= -3.489 + 7.974(2.92)+ 3.336 (7.894)= 25.83\%$
Con variables X_5 y X_3 se tiene el 25.83% / 46% de X_9

Si se usa **modelo 3, con un valor fijo de 4 (arbitrario para ver qué** sucede:
$X9= -6.514+7.623(4)+3.376(4)+1.4 (4)= 43\%$

O la inclusión de la media de la 3er. variable
$X9= -6.514+7.623(2.92)+3.376(7.894)+1.4 (2.66)=46.125\%$ **o que sucede si aumenta 1%**
$X9= -6.514+7.623(3.92)+3.376(8.894)+1.4 (3.66)=58.53\%$, **incrementa 12.3%**

Paso 6: Validación de los Resultados

- Una vez estimado y analizado un modelo, el paso siguiente es **establecer su generabilidad**, esto es, que represente realmente al conjunto de la población y no sólo a la muestra que lo ha generado.
- La mejor forma de hacerlo sería ver en qué medida los resultados se comportan con modelos teóricos previos o trabajos ya validados sobre el mismo tema.
- El procedimiento más indicado para la validación empírica de los resultados de una regresión, pasa por volver a estimar el modelo en una nueva muestra extraída de la población.
- El investigador puede dividir su muestra en **dos partes: una submuestra para estimar el modelo y una submuestra de validación usada para evaluar la ecuación.**

Referencias

Belsley, D. A., Kuh E., y Welsch R. E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

BMDP Statistical Software, Inc. (1991), SOLO Power Analysis. Los Angeles: BMDP.

Box, G. E., y Cox D. R. (1964), An Analysis of Transformations. *Journal of the Royal Statistical Society* 8 26:211-43.

Cohen, J., y Cohen P. (1983), applied *Multiple Regression/Correlation Analysis for the Behavioral Sciences, 2d ed. flillsdale*, N.J.: Lawrence Erlbaum Associates.

Daniel, C., y Wood F. S. (1980), Fitting *Equations to Data*, 2d Ed. New York: Wiley-Interscience

Hair , J.F.; Anderson, R.E.; Tatham, R.L.; Black W.C. (1999).*Análisis Multivariante.*5a. Ed. España. Prentice Hall.

IBM (2011a). *IBM SPSS Statistics Base 20.* EUA. .Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Base.pdf

IBM (2011b).*Guía breve de IBM SPSS Statistics 20.* EUA.Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Brief_Guide.pdf

IBM (2011c). IBM SPSS Missing Values 20. EUA. .Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Missing_Values.pdf

Jaccard, J., Turrisi R., y Wan, C. K. (1990), Interaction *Effects in Multiple Regression*. Beverly Hills, Ca lif.: Sage Publications.

Johnson, R. A., y Wichem, D. W. (1982), Applied *Multivariate Statistical Analysis*. Upper Saddle River, N.J., Prentice Hall.

Levin, R.,I.; Rubin, D.S. (2004). Estadística para Administración y Economía (7ª. Edición). México: Prentice-Hall

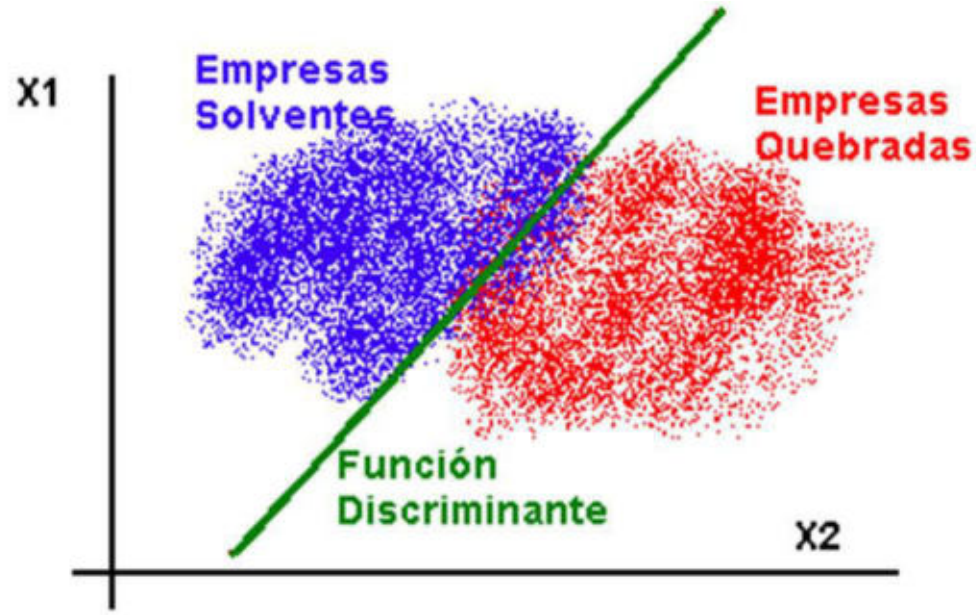
Masan, C. H., y Perreault, W. D. Jr. (1991), Colli nearity, Power, and Interpretation of Multiple Regression Analysis. *Journal of Marketing Research 28* (August): 268-80.

Mosteller, F., y Tukey L. W. (1977), *Data Analysis and Regression*. Reading, Mass.: Addison-Wesley.

Neter, J., Wasserrnann W., y Kutner, M. H. (1989), *Applied Linear Regression Models*. Homewood, Ill.: Irwin.

Seer, G. A. F. (1984), *Multivariate Observations*. New York: Wiley.

Capítulo 6. Análisis Discriminante Múltiple



6.1. Análisis Discriminante Múltiple: ¿Qué es?

La regresión múltiple es la técnica de dependencia multivariante utilizada más exten-samente. Su popularidad de la se basa en su capacidad para predecir y explicar las variables métricas. Pero, **¿qué decir sobre las variables no métricas?** La regresión multivariante no es adecuada en este contexto por lo que en este apartado, **se presentan 2 técnicas: el análisis discriminante y la regresión logística,** que tratan la situación cuando **la variable dependiente es no métrica.** En esta situación, Usted está interesado **en la predicción y explicación** de las relaciones que influyen en la categoría en que un objeto está situado. Por ejemplo, **¿por qué una persona es o no cliente?, o si un emprendimiento se logrará o no**. Para saber más, ver: IBM, 2011a; IBM 2011b; IBM 2011c.

En esta sección, se tienen **2** objetivos:

1. Presentar la naturaleza, filosofía y condiciones tanto del análisis discriminante múltiple como de la regresión logística**; y**
2. Demostrar la aplicación e interpretación de estas técnicas con un ejemplo ilustrativo.

El propósito básico del análisis discriminante (**Capítulo 2**) es estimar la relación entre una única **variable dependiente no métrica (categórica)** y un conjunto de **variables independientes métricas**, en esta forma general:

$$\boldsymbol{Y_1 = X_1 + X_2 + X_3 + \ldots X}$$

El análisis discriminante múltiple y la regresión logística cuentan con amplias aplicaciones en situaciones donde el primer objetivo es identificar el grupo al cual un objeto (por ejemplo, una persona, una empresa, o un producto) pertenece. Las aplicaciones potenciales incluyen la predicción de éxitos o fracasos de un nuevo producto, decidir si un estudiante

debe ser admitido en una universidad, clasificar a los estudiantes por sus intereses vocacionales, determinar en qué categoría de riesgo de crédito se encuentra una persona o predecir si una empresa tendrá éxito. En cada caso los objetos están incluidos en grupos y se desea que la pertenencia a cada grupo de cada objeto pueda predecirse o explicarse por un conjunto de variables independientes seleccionadas por el investigador con una de las dos técnicas presentadas en este capítulo.

En el proceso de elegir la técnica analítica apropiada, algunas veces encontraremos un problema que incluye una variable **dependiente categórica y varias variables independientes métricas**. **Por ejemplo,** si requiere **distinguir entre riesgo de crédito alto y bajo.** Si tuviéramos una **medida métrica** del riesgo de crédito, podía utilizarse la **regresión multivariante.** Pero ¿qué sucede **si requiere que conocer si alguien se encuentra en una categoría de riesgo bueno o malo? Esta no es la medida de tipo métrico** requerida para el **análisis de regresión múltiple**. **El análisis discriminante y la regresión logística** son las técnicas estadísticas apropiadas cuando la **variable dependiente es categórica (nominal o no métrica)** y **las variables independientes son métricas**, siendo algunos de los casos ejemplo a considerar:

1. La **variable dependiente consta de dos grupos o clasificaciones**, **por ejemplo**, masculino vs. femenino; alto vs. bajo.
2. Situaciones que incluyen más de **2** casos, como en una **clasificación de tres grupos que comprenda clasificaciones bajas, medias y altas**. El **análisis discriminante** tiene la capacidad de tratar tanto **2** grupos como grupos múltiples (**3** o más).
3. Cuando se incluyen **2** clasificaciones, la técnica es conocida como **análisis discriminante de dos grupos.**
4. Cuando se identifican **3** o más clasificaciones, la técnica es conocida como **análisis discriminante múltiple (MDA)**.
5. La **regresión logística**, también conocida como **análisis *logit*, está restringida en su forma básica a 2 grupos**, aunque en formulaciones alternativas puede considerar más de **2** grupos.

El análisis discriminante implica obtener un valor teórico, es decir, una combinación lineal de dos (o más) variables independientes que discrimine mejor entre los grupos definidos a priori. La discriminación se lleva a cabo estableciendo las **ponderaciones del valor teórico** para cada variable de tal forma que **maximicen la varianza entre-grupos frente a la varianza intra-grupos**. La combinación lineal para **el análisis discriminante**, también conocida como **función discriminante**, se deriva de una ecuación que adopta la siguiente forma:

$$Z_{jk}=a+W_1X_{1k}+W_2X_{2k}+W_3X_{3k}+\ldots W_nX_{nk}$$

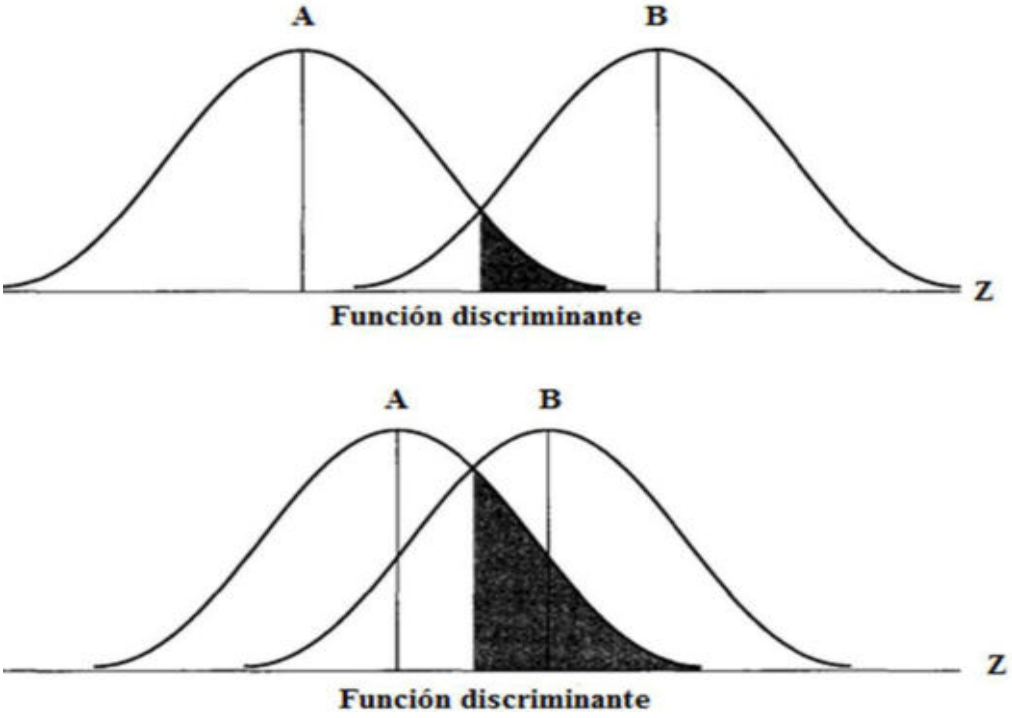
Z_{jk}**=puntuación *z*** discriminante de la función discriminante *j* para el objeto *k*
a= constante
W_i**=**ponderación discriminante para la variable independiente *i*
X_{ik}**=**variable independiente *i* para el objeto *k*

El análisis discriminante nos sirve para **contrastar la hipótesis de que las medias de los grupos de un conjunto de variables independientes para dos o más grupos son iguales.** Para logarlo, la técnica **multiplica** cada **variable independiente** por su correspondiente **ponderación** y **suma estos productos**.

El resultado es una única **puntuación *z* discriminante** compuesta para cada individuo en el análisis. **Promediando las puntuaciones discriminantes** para todos los individuos dentro de un grupo particular, obtenemos la **media del grupo (**llamada **centroide).** Cuando el análisis engloba **2 grupos**, existen **2 centroides**; con **3 grupos**, hay **3 centroides**, y así sucesivamente, logrando:

1. Los **centroides indican la situación más común de cualquier individuo de un determinado grupo,** y **una comparación de los centroides de los grupos muestra lo apartados que se encuentran los grupos** a lo largo de la dimensión que se está contrastando.
2. El **contraste** para la significación estadística de la función discriminante es una **medida generalizada de la distancia entre los centroides de los grupos.** Se **calcula comparando las distribuciones de las puntuaciones discriminantes para los grupos**.
3. Si el **solapamiento en la distribución es pequeño**, la **función discriminante separa bien los grupos.** Si el **solapamiento es grande**, la función es un **mal discriminador** entre los grupos. Ver **Figura 6.1** donde el diagrama de arriba representa la distribución de puntuaciones discriminantes para una función que **separa bien** los grupos, mientras que el diagrama de abajo muestra la distribución de puntuaciones discriminantes de una función que es un discriminador relativamente **malo** entre los grupos **A y B**.
4. Las **áreas sombreadas** son la probabilidad de clasificar erróneamente objetos del grupo **A** en el **B.**

Figura 6.1. Representación univariante de las puntuaciones Z discriminantes.



Fuente: propia

5. El **análisis discriminante múltiple** resulta único en una característica sobre la relación de dependencia: **si existen más de dos grupos en la variable dependiente, el análisis discriminante calculará más de una función discriminante.** Calculará **NG - 1** funciones, donde **NG** es el número de grupos.
6. **Cada función discriminante calculará** una puntuación **discriminante Z**. En el caso de una **variable dependiente de 3 grupos**, cada objeto tendrá una puntuación para las funciones discriminantes una y dos, permitiendo que los objetos puedan **dibujarse en dos dimensiones**, donde cada dimensión representa una función discriminante. De esta forma, el **análisis discriminante no está limitado a un sólo valor teórico**, como la **regresión múltiple**, sino que obtiene valores teóricos múltiples que representan dimensiones de discriminación entre los grupos.
7. La **regresión logística** es un tipo especial de regresión que se utiliza para predecir y explicar una **variable categórica binaria (dos grupos)** en lugar de una **medida dependiente métrica.** La forma del valor teórico de la **regresión logística** es similar a la del **valor teórico en la regresión múltiple**. El valor teórico representa una única relación multivariante con coeficientes como los de la regresión que indican la influencia relativa de la variable predictor. **Las diferencias entre la regresión logística y el análisis discriminante resultarán más claras cuando presentemos las caracterís-ticas específicas de la regresión logística.** Aun así existen también semejanzas entre ambos métodos.
8. Cuando se conocen los supuestos básicos de ambas, es-tas técnicas proporcionan resultados predictivos y clasificatorios comparables y emplean medidas de validación **similares**. Sin embargo la **regresión logística** tiene la **ventaja** de verse **menos afectada que el análisis discriminante cuando no se cumplen los supuestos básicos**, concretamente **la normalidad** de las variables.
9. Además puede **permitir** la utilización de **variables no métricas** por medio de su **codificación con variables ficticias**, tal como puede hacerse en la **regresión**. La **regresión logística** está **limitada**, sin embargo, a la predicción de tan sólo la medida dependiente de dos grupos. Por tanto, en casos donde la medida **dependiente** está formada por **2** o más grupos se adecúa mejor el **análisis discriminante.**

6.2. Análisis Discriminante Múltiple: Analogía con la regresión y MANOVA

La función discriminante es una **combinación lineal (valor teórico) de medidas métricas de dos o más variables independientes** y se usa para **describir o predecir una única variable dependiente.**

La **principal diferencia** es que el **análisis discriminante** es apropiado en trabajos de investigación en donde la **variable dependiente es categórica (nominal o no métrica),** mientras que la **regresión** se utiliza cuando la **variable dependiente es métrica.**

Como se explicó, **la regresión logística es una variante de la regresión**, contando con semejanzas excepto por el tipo de **variable dependiente**.

El **análisis discriminante** también es comparable a **"*invertir*"** el análisis multivariante de la varianza **(MANOVA)**, que presentamos en el **Capítulo 7**. En el **análisis discriminante,** la **variable dependiente única es categórica**, y las **variables independientes son métricas.**

El caso opuesto es el de **MANOVA**, donde se incluyen **variables dependientes métricas y variable(s) independiente(s) categórica(s)**

6.3. Análisis Discriminante Múltiple: Ejemplo hipotético

El **análisis discriminante** se aplica a cualquier problema de investigación que tenga por **objetivo la comprensión de la pertenencia a un grupo** como individuos (clientes vs. no clientes, etc.), empresas (rentables vs. no rentables), productos (exitosos comercialmente vs. fracasados comercialmente), o cualquier otro objeto que pue-da evaluarse sobre un conjunto de variables independientes. Para ejemplificar los supuestos básicos del **análisis discriminante**, examinamos dos ámbitos de investigación, uno que incluye **2** grupos (compradores frente vs. no compradores) y otro de **3 grupos** (niveles o comportamiento de cambio).

La **regresión logística** opera de forma **similar al análisis discriminante** de **2** grupos.

6.3.1. Análisis discriminante de dos grupos: compradores frente a no compradores

Suponga que **MKT Digital** quiere averiguar si uno de sus nuevos servicios (una nueva y mejorada plantilla web)-- tendrá éxito comercial. Para llevar a cabo la investigación, **MKT Digital** se interesa en identificar a aquellos consumidores que comprarían el nuevo servicio y los que no. En terminología estadística, **MKT Digital** debe **minimizar el número de errores** al momento de predecir qué consumidores comprarían la nueva plantilla y cuáles no. Para ayudar a identificar a los compradores potenciales, **MKT Digital** ha ideado escalas de valoración para **3 atributos (accesibilidad, desempeño y diseño)** que son utilizadas por los consumidores para evaluar la nueva plantilla. En lugar de considerar cada escala como una medida separada, **MKT Digital** espera que una combinación ponderada de las tres mejore la predicción de si es probable que un consumidor compre el nuevo producto.

El **análisis discriminante** obtiene una **combinación ponderada de las tres escalas** que se utilice para predecir la **verosimilitud** de que un consumidor compre el servicio. Además de determinar los clientes que compren el nuevo servicio, puede distinguir a los que no lo comprarán, y **MKT Digital** querría saber qué **atributos** de su **nuevo servicio** son útiles para **diferenciar** compradores de los que no lo son; es decir, **¿cuál de las tres características del nuevo servicio separa mejor a los compradores de los no compradores?** Así, si la respuesta **"*compraría*"** se asocia siempre con una valoración alta

de **accesibilidad** y la respuesta "***no compraría***" está siempre asociada con una valoración de **accesibilidad baja, MKT Digital** concluirá que el atributo de **accesibilidad diferenciaría a los compradores de los no que no.** Por el contrario, si **MKT Digital** descubre que muchas personas con una valoración alta sobre el **diseño** comprarían la plantilla tanto como aquellas que no, entonces el atributo **diseño** sería una característica que **discrimina pobremente entre compradores y no compradores.**
La **Figura 6.2** refleja las valoraciones sobre esos **3 atributos** de la nueva plantilla con un precio especificado por un grupo de **10 compradores potenciales**.

Figura 6.2 Resultados de la encuesta MKT Digital sobre la evaluación de un nuevo servicio

	Evaluación del nuevo servicio*		
Grupos basados en la intención de compra	X_1	X_2	X
Grupo 1 : Compraría			
Sujeto 1	8	9	6
Sujeto 2	6	7	5
Sujeto 3	10	6	3
Sujeto 4	9	4	4
Sujeto 5	4	8	2
Media del grupo	**7.4**	6.8	4.0
Grupo 2: No compraría			
Sujeto 6	5	4	7
Sujeto 7	3	7	2
Sujeto 8	4	5	5
Sujeto 9	2	4	3
Sujeto 10	2	2	2
Media del grupo	**3.2**	4.4	3.8
Diferencias entre las medias de los grupos	4.2	2.4	0.2

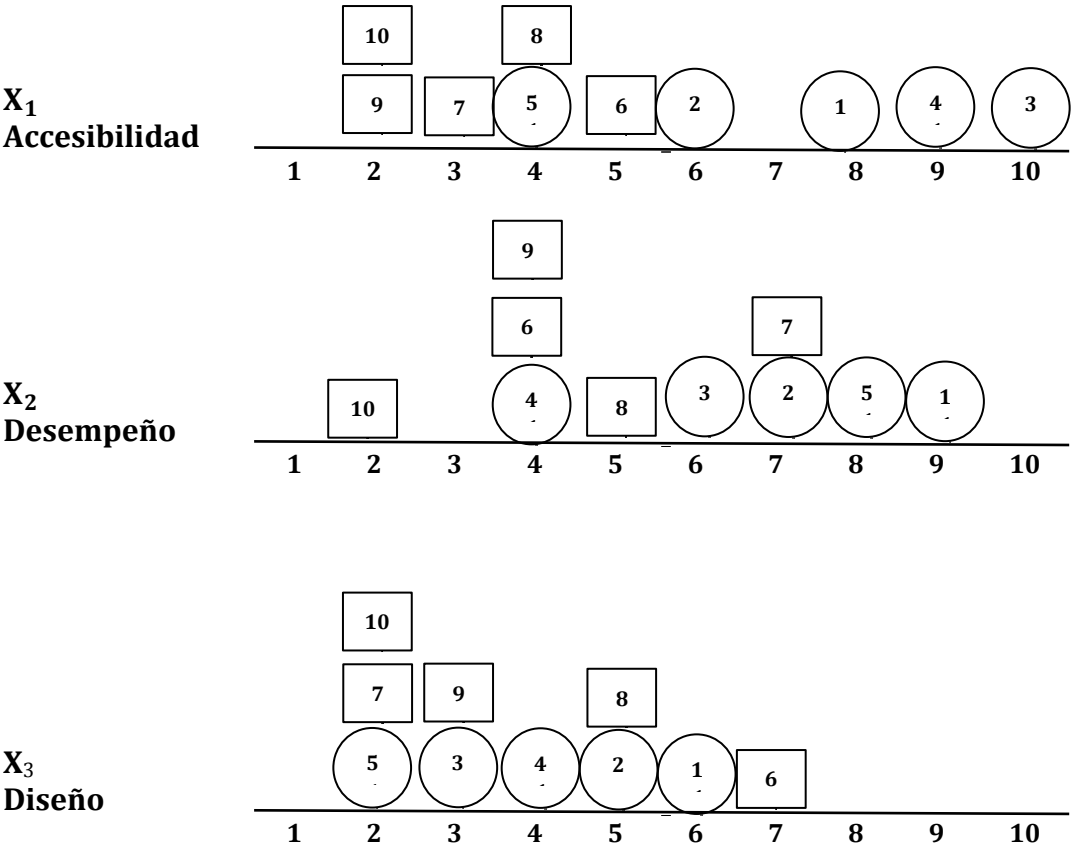
Fuente: propia
Las evaluaciones están hechas sobre una escala de 10 puntos (1 =muy baja a 10= excelente.

En la valoración de la plantilla, cada miembro de un panel podría estar comparándola implícitamente con servicios existentes en el mercado. Después de que el servicio sea evaluado, se solicitó a los evaluadores establecieran sus intenciones de compra **("*compraría*" o "*no compraría*")**. De ellos, **5** determinaron que com-prarían la nueva plantilla y **5** dijeron que no. **La Figura 6.2** identifica potencialmente **varias variables discriminantes**, de la siguiente manera:

1. Existe una diferencia sustancial entre las valoraciones **medias de los grupos "*compraría*" y "*no compraría*" sobre X_1** , el atributo **accesibilidad (7.4- 3.2 = 4.2)** al parecer discrimina bien entre los grupos **"*compraría*" y "*no compraría*"** y es probable que sea una característica importante para los compradores potenciales.

2. El atributo **diseño (X_3)** presenta una diferencia mucho más pequeña de **0.2** entre las valoraciones medias **(4.0- 3.8 = 0.2)** para los grupos "***compraría***" **y** "***no compraría***". Por tanto, se espera que sea **menos discriminante** en términos de la decisión de comprar o no.
3. Antes de que podamos llegar a conclusiones definitivas, debemos examinar **las distribuciones de las puntuaciones para cada grupo.** Grandes **desviaciones estándar** dentro de uno o de ambos grupos pueden hacer que **la diferencia entre medias sea no significativa e inoperante** para discriminar entre grupos. Ya que solamente tenemos **10 encuestados** en **dos grupos y tres variables independientes**
4. También se observan los datos gráficamente para determinar **qué análisis discriminante se intentará llevar a cabo.** La **Figura 6.3** muestra a los **10 encuestados** sobre cada una de las **3 variables.**

Figura 6.3. Representación gráfica de 10 compradores potenciales sobre 3 posibles variables discriminantes.



Fuente: propia

El grupo **"*compraría*"** está representado por **círculos** y el grupo **"*no compraría*"** por **cuadrados.** Los números de identificación de los encuestados están dentro de las figuras. Observando primero $\mathbf{X_1}$ (**la accesibilidad**) que cuenta con una diferencia sustancial en las puntuaciones medias, vemos que podríamos discriminar casi perfectamente entre los grupos usando solamente esta variable. Si estableciéramos el valor de **5.5** como punto de corte para discriminar entre los dos grupos, entonces clasificaríamos de forma incorrecta al encuestado número cinco, uno de los miembros del grupo **"*compraría*".**

Esto es indicativo de la gran diferencia en las medias de los dos grupos en lo que se refiere a duracíon y de la ausencia de solapamiento entre las distribuciones de los dos grupos (vea **Figura 6.1**). Examinando $\mathbf{X_2}$ **(el desempeño)**, vemos que hay una distinción menos clara entre los dos grupos. Sin embargo, esta variable proporciona una alta discriminación para el encuestado número cinco, que fue **mal clasificado al usar** $\mathbf{X_1}$**.** Además, los encuestados que estuviesen mal clasificados usando $\mathbf{X_2}$ están bien separados con $\mathbf{X_1}$**.** Por ello $\mathbf{X_1}$ y $\mathbf{X_2}$ podrían utilizarse de forma bastante eficiente en combinación para predecir la pertenencia a un grupo. Finalmente el diseño, $\mathbf{X_3}$**, muestra pequeñas diferencias entre los grupos.** Por ello, construyendo un valor teórico con ellos $\mathbf{X_1}$ y $\mathbf{X_2}$ y omitiendo $\mathbf{X_3}$ se podría formar una función discriminante que maximizara la separación de los grupos en base a la puntuación discriminante.

La **Figura 6.4** contiene los resultados de **3 funciones discriminantes diferentes.** La primera función discriminante contiene sólo $\boldsymbol{X_1}$**,** igualando el valor de $\boldsymbol{X_1}$**,** a la **puntuación** $\boldsymbol{z}$ discriminante. Como se explicó, el uso exclusivo de $\boldsymbol{X_1}$**,** el mejor discriminador, implicaba la **clasificación errónea** del **sujeto 5.** Como se muestra en la **matriz de clasificación** en la **Figura 6.4,** cuatro de los cinco sujetos en el **grupo 2** son clasificados **correctamente** (esto es, **caen en la diagonal de la matriz de clasificación**). El porcentaje correctamente clasificado es por tanto del **90 por ciento (9/10 sujetos).** Debido a que $\boldsymbol{X_2}$**,** proporciona discriminación para el sujeto 5, podemos formar una función discriminante combinando igualmente $\boldsymbol{X_1}$ $\boldsymbol{y}$ $\boldsymbol{X_2}$para utilizar las capacidades discriminatorias únicas de cada variable. Como vemos en la **Figura 6.4** usando **una puntuación de corte de 11** con esta nueva función discriminante alcanzamos una clasificación perfecta de los grupos (**100%** correctamente clasificado). Por tanto, la combinación de $\boldsymbol{X_1}$ $\boldsymbol{y}$ $\boldsymbol{X_2}$es capaz de hacer mejores predicciones de pertenencia al grupo que cada una de ellas separadamente.

Figura 6.4. Construcción de funciones discriminantes para predecir a los compradores y a los no compradores

	Evaluación del nuevo servicio		
Grupos basados en la intención de compra	**Función 1:** $z = \boldsymbol{X_1}$	**Función 2:** $\boldsymbol{z = X_1 + X_2}$	**Función 3:** $\boldsymbol{z = -4.53 + 0.476X_1 + 0.359X_2}$
Grupo 1 : Compraría			
Sujeto 1	8	9	2.51
Sujeto 2	6	7	0.84
Sujeto 3	10	6	2.38
Sujeto 4	9	4	1.19
Sujeto 5	4	8	0.25
Media del grupo	7.4	6.8	2.51
Grupo 2: No compraría			
Sujeto 6	5	9	-0.71

Sujeto 7	3		7		-0.59	
Sujeto 8	4		5		-0.83	
Sujeto 9	2		4		-2.14	
Sujeto 10	2		2		-2.86	
Media del grupo	3.2		4.4		0.0	
Puntuación de corte	**5.5**		**11**		**0.0**	
	Grupo predicción		**Grupo predicción**		**Grupo predicción**	
Grupo real	**1**	**2**	**1**	**2**	**1**	**2**
1: Compraría	**4**	**1**	**5**	**0**	**5**	**0**
2: No compraría	**0**	**5**	**0**	**5**	**0**	**5**

Fuente: propia

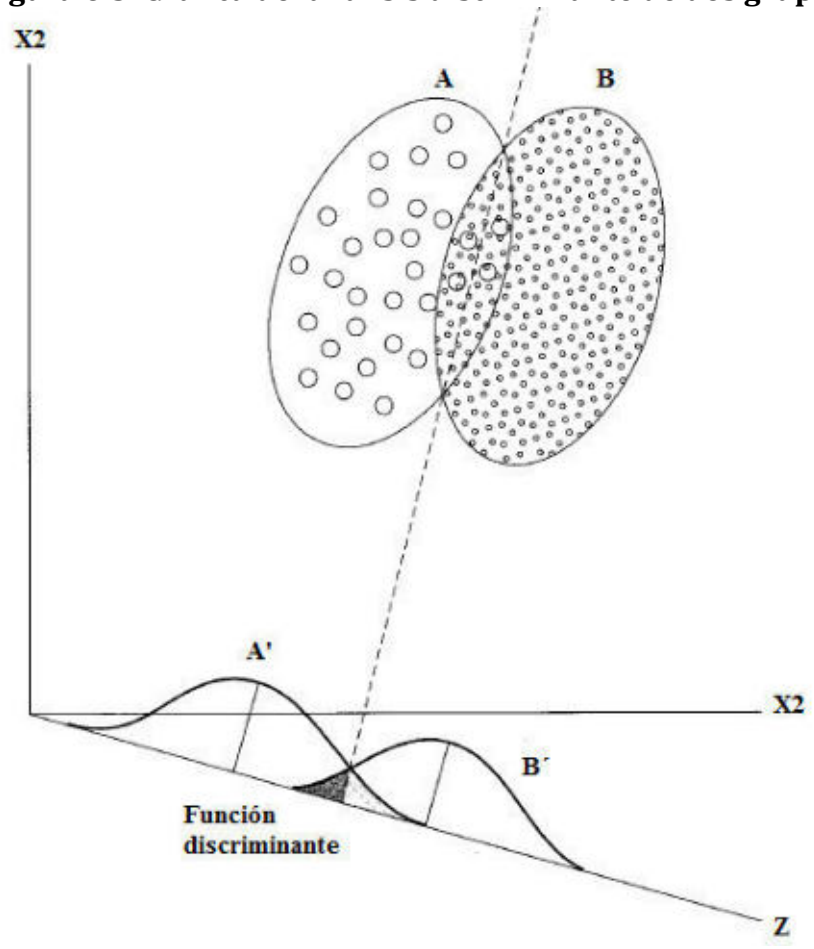
El análisis discriminante sigue un procedimiento muy similar al que está indicado en el anterior ejemplo hipotético al estimar empíricamente la función discriminante. Identifica las variables con las diferencias más grandes entre los grupos y obtiene un coeficiente de ponderación para cada variable para reflejar estas diferencias.
La tercera funcio n discriminante en la **Figura 6.4** representa la verdadera funcio n discriminante estimada (***z=-4.53+0.476*X_1*+0.359*X_2**).
Empleando una puntuación de **corte de 0.** La tercera función también alcanza un porcentaje de clasificación correcta del **100%** con la máxima separación posible entre grupos.

6.3.2 Una representación geométrica de la función discriminante de dos grupos.

Una gra fica adicional de otro caso de ana lisis de dos grupos explicara mejor la naturaleza del ana lisis discriminante **[**Green, et al. 1988**].** La **Figura 6.5** demuestra lo que ocurre cuando se calcula una funcio n discriminante de dos grupos. Suponga que tiene **2** grupos, **A y B,** y dos medidas, X_1 y X_2por cada uno de los dos grupos. Al dibujar en un diagrama de dispersio n la relacio n de las variables para cada miembro de los dos grupos. Como se observa, los **puntos pequeños** representan las medidas de las variables para los miembros del **grupo B** y los **grandes** para los del **grupo A.** Las **elipses** alrededor de los puntos grandes y pequen os encerrar an algunas proporciones pre-especificadas de los puntos, generalmente el **95%** o ma s en cada grupo. Si dibuja **una línea recta** a trave s de los **2 puntos** donde las elipses se cortan y proyectamos la l nea a un **nuevo eje *z*,** se afirma que la superposicio n entre las distribuciones univariantes **A'y B'** (la superposicio n de las a reas sombreadas) **es más pequeña** que la que se obtendr a si se dibujase cualquier otra l nea a través de las elipses formadas por los diagramas de dispersión **[**Green, et al. 1988**].** Lo que conviene destacar, es que el **eje *z*** representa los perfiles de **2 variables** de los **grupos A y B** como únicos números **(puntuaciones discriminantes).** Encontrando una combinacio n lineal de las variables originales X_1 y X_2**,** se proyectan los resultados como una funcio n **discriminante.** Así, si los puntos grandes y pequen os se proyectan sobre el nuevo eje ***z*** como puntuaciones ***z*** discriminantes, el resultado la informacio n sobre las diferencias de los grupos (mostrando en el plano X_1 y X_2) dentro de un conjunto de puntos (**puntuaciones *z***) sobre el u nico eje, mostrado por las distribuciones **A'y B'.** En **resumen,** en un determinado problema de ana lisis discriminante, se halla una combinacio n lineal de las **variables independientes,** obtenie ndose una serie de puntuaciones discriminantes para cada objeto en cada grupo. Las **puntuaciones**

discriminantes se calculan de acuerdo a la regla estadística de maximizar la varianza entre los grupos y minimizar la varianza intragrupos. Si la **varianza entre los grupos** es grande con relación a la **varianza intragrupos,** diremos que **la función discriminante separa bien los grupos. Ver Figura 6.5**

Figura 6.5. Gráfica del análisis discriminante de dos grupos.



Fuente: propia

6.3.3. Un ejemplo de análisis discriminante de tres grupos: Propósitos de cambio

El ejemplo de **2 grupos** recién examinado demuestra la razón de ser y las **ventajas de combinar variables independientes en un valor teórico para discriminar entre grupos**. Pero el **análisis discriminante** también cuenta con otros medios de discriminación:

1. **La estimación y**
2. **El uso de valores teóricos múltiples en casos donde hay tres o más grupos.**

En estos casos, estas funciones discriminantes llegan a ser **dimensiones de discriminación: cada dimensión se separa y se diferencia de las otras.** Por ello, además de **la mejora de la explicación de la pertenencia a un grupo,** estas funciones discriminantes adicionales **aportan ideas entre las varias combinaciones de variables independientes que discriminan entre los grupos**. **Por ejemplo, al** examinar la

investigación llevada a cabo por **MKT Digital** referente a **la posibilidad de que los clientes de un competidor cambien de proveedor.** En una primera encuesta a pequeña escala se entrevistó a 15 clientes de un competidor importante. En la entrevista, se les preguntó a los clientes por **la probabilidad de cambiar de proveedor en una escala de tres categorías.** Las tres respuestas posibles fueron **"*con toda seguridad cambiar*", "*indeciso*" y "*con toda seguridad no cambiar*"**. Los **clientes** fueron asignados a los **grupos 1, 2 y 3,** respectivamente, de acuerdo con su respuesta y también valoraron al competidor en las características de **competitividad en el precio y nivel de servicio**. El objetivo se centró ahora en **determinar si las valoraciones de los clientes sobre su proveedor habitual pueden predecir la probabilidad de abandonar a ese proveedor**. Dado que la **variable dependiente de la probabilidad de cambiar** se midió como una **variable categórica (no métrica)** y las valoraciones **del precio y del servicio son métricas**, el **análisis discriminante es apropiado.**

Con **3 categorías** para la **variable dependiente**, el **análisis discriminante** puede estimar **2** funciones discriminantes, cada una **representando una dimensión diferente** de discriminación. La **Figura 6.6** contiene los resultados del estudio para los **15 clientes, 5 en cada categoría de la variable dependiente.** Al igual que hicimos en el **ejemplo de 2 grupos,** podemos examinar las **puntuaciones medias** de cada grupo para ver si una de las variables discrimina bien entre todos los grupos. Para $\boldsymbol{X_1}$, la **competitividad en el precio**, vemos una diferencia en la media bastante más grande entre el **grupo 1 y los grupos 2 o 3** (**0.0 vs. 1 .6 y 2.25**).

Se observa que $\boldsymbol{X_1}$ discrimina **bien entre el grupo 1 y los grupos 2 o 3** pero probablemente **será mucho menos efectiva para discriminar entre los grupos 2 y 3**. Para $\boldsymbol{X_2}$, el nivel de servicio, vemos que la **diferencia entre los grupos 1 y 2 es muy pequeña (2.0 vs a 2.2**), mientras que existe una **gran diferencia entre el grupo 3 y los grupos y 2 (6.2 vs 2.0 y 2.2**). Así, $\boldsymbol{X_1}$ diferencia el **grupo l de los grupos 2 y 3,** mientras que $\boldsymbol{X_2}$ diferencia el **grupo 3 de los grupos 1 y 2**. Como resultado, vemos que $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$ proporcionan diferentes **"*dimensiones*"** de discriminación entre grupos.

Figura 6.6. Resultados de la encuesta de MKT Digital sobre las intenciones de cambio de compradores potenciales

	Evaluación del nuevo servicio*	
Grupos basados en la intención de cambio	$\boldsymbol{X_1}$ **Competitividad en el precio**	$\boldsymbol{X_2}$ **Nivel de servicio**
Grupo 1 : Se cambiarán con seguridad		
Sujeto 1	2	2
Sujeto 2	1	2
Sujeto 3	3	2
Sujeto 4	2	1
Sujeto 5	2	3
Media del grupo	**2.0**	**2.0**
Grupo 2: Indecisos		
Sujeto 6	4	2
Sujeto 7	4	3

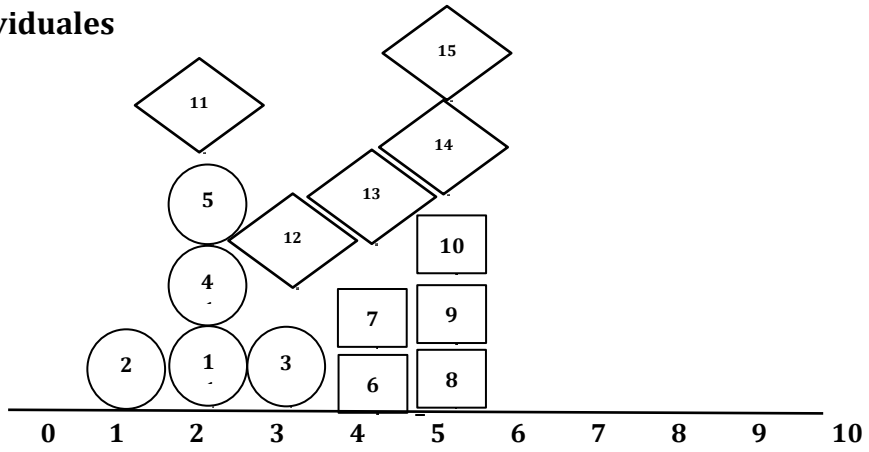
Sujeto 8	5	1
Sujeto 9	5	2
Sujeto 10	5	3
Media del grupo	**4.6**	**2.2**
Grupo 2: Indecisos		
Sujeto 11	2	6
Sujeto 12	3	6
Sujeto 13	4	6
Sujeto 14	5	6
Sujeto 15	5	7
Media del grupo	**3.8**	**6.2**

* Las evaluaciones están hechas sobre una escala de 10 puntos (l = muy baja a 10 = excelente).
Fuente: propia

Reflejando lo anterior gráficamente, la **Figura 6.7a y 6.7b** representa los tres grupos de cada una de las variables independientes de forma separada.

Figura 6.7a. Representación gráfica de variables discriminantes potenciales para un análisis discriminante de tres grupos.

A.Variables individuales



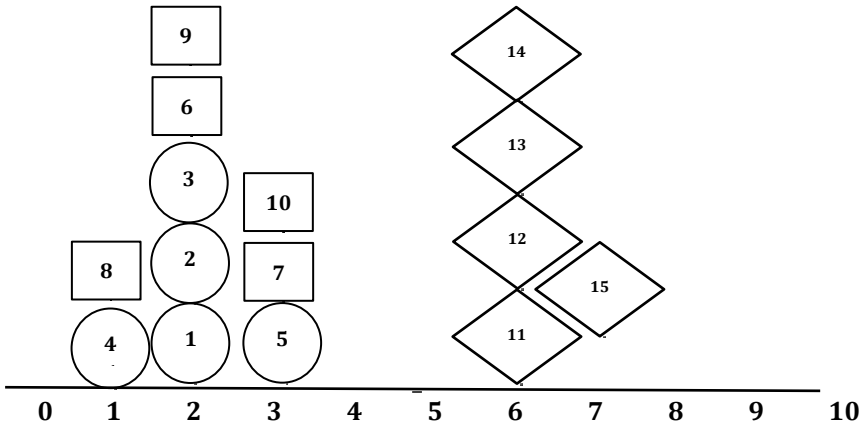
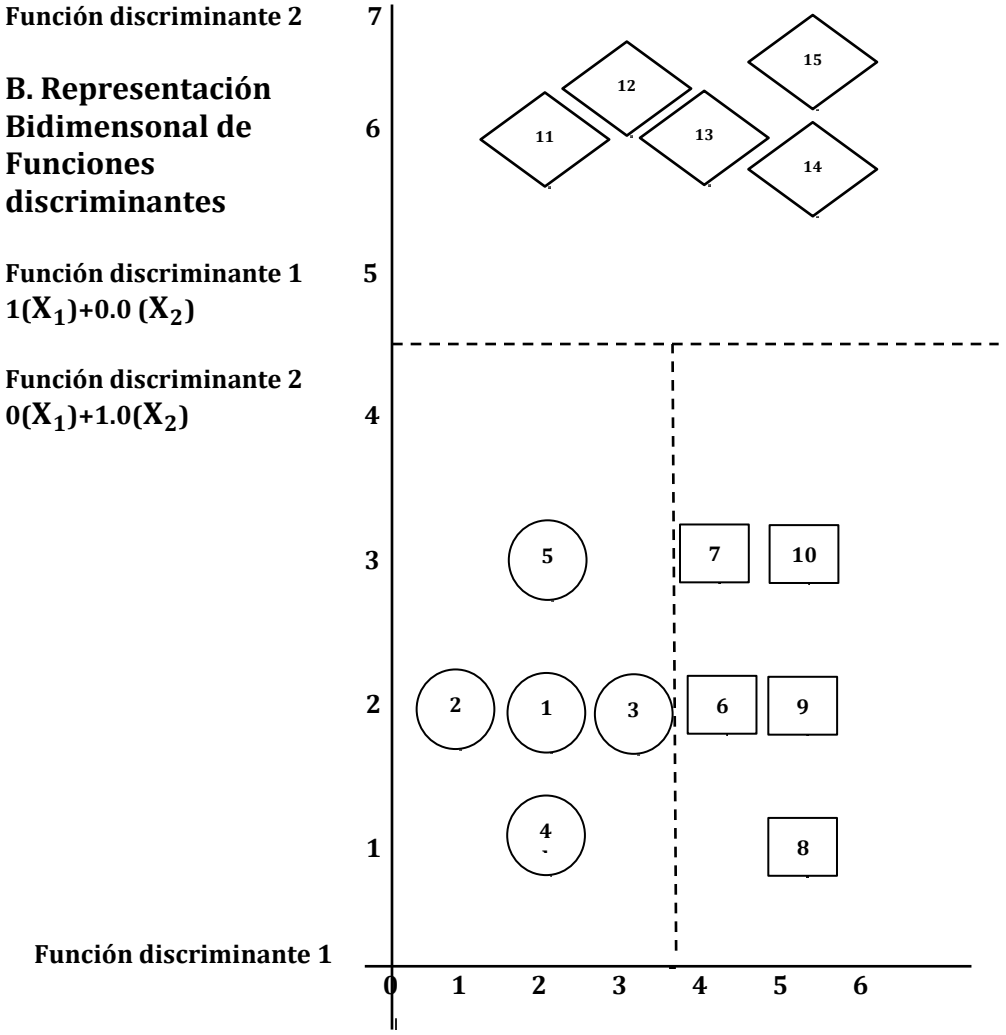


Figura 6.7b. Representación gráfica de variables discriminantes potenciales para un análisis discriminante de tres grupos.



Fuente: propia

Al ver los miembros del grupo en cada variable, se aprecia que **ninguna variable discrimina** bien entre todos los grupos. Si se construyen dos funciones discriminantes simples, los resultados llegan a ser mucho más claros. Con fines ilustrativos, calculamos dos funciones discriminantes con ponderaciones de **0.0 o 1.0** para las variables**. La función discriminante 1 da a X_1** una ponderación de **1.0**, mientras que a X_2 se le da una ponderación de **0.0**. De la misma manera, la **función discriminante 2** da a X_2 una ponderación de **1.0**, y a X_1 una ponderación de **0.0.** Las funciones pueden establecerse matemáticamente como:

Función discriminante 1: 1.0 (X_1) + 0.0 (X_2)

Función discriminante 2: 0.0 (X_1) + 1.0 (X_2)

Así, en términos sencillos se explica cómo el proceso del **análisis discriminante estima ponderaciones para maximizar la discriminación**. Con las dos funciones, **se calculan dos puntuaciones discriminantes para cada encuestado**. La **Figura 6.8** también contiene un punto para cada encuestado en una representación **bidimensional.** La

separación entre los grupos llega a ser ahora bastante evidente, y cada grupo puede ser fácilmente diferenciado.

Al establecer valores sobre cada dimensión que definan **regiones que contengan a cada grupo** (por ejemplo, todos los miembros del **grupo l** están en la **región menor que 3.5** sobre la **dimensión 1** y **< 4, .5** sobre la **dimensión 2**). Cada uno de los otros grupos puede estar definido de forma similar en términos de los tamaños de las puntuaciones de sus funciones discriminantes. Además, las dos funciones proporcionan las **dimensiones de la discriminación,** de la siguiente forma:

1. La primera función discriminante, la **competitividad en el precio**, distingue entre **clientes indecisos de los que han decidido cambiar, pero no diferencia a aquellos que han decidido no cambiar.**
2. La segunda función discriminante, la **percepción del nivel de servicio, predice si un cliente decidirá no cambiar, frente a si un cliente está indeciso o determinado a cambiar de proveedor.**
3. Usted deberá presentar la **dirección de las influencias separadas** tanto de la competitividad en el precio como del nivel de servicio al tomar esta decisión.

La estimación de más de una función discriminante, cuando sea posible, le proporciona a Usted tanto una mejora en la discriminación como perspectivas adicionales sobre las **características y combinaciones** que mejor discriminan entre los grupos.

Las siguientes secciones detallan los pasos necesarios para realizar un análisis discriminante, para valorar su nivel de ajuste predictivo y, al final, para interpretar la influencia de variables independientes al llevar a cabo esa predicción.

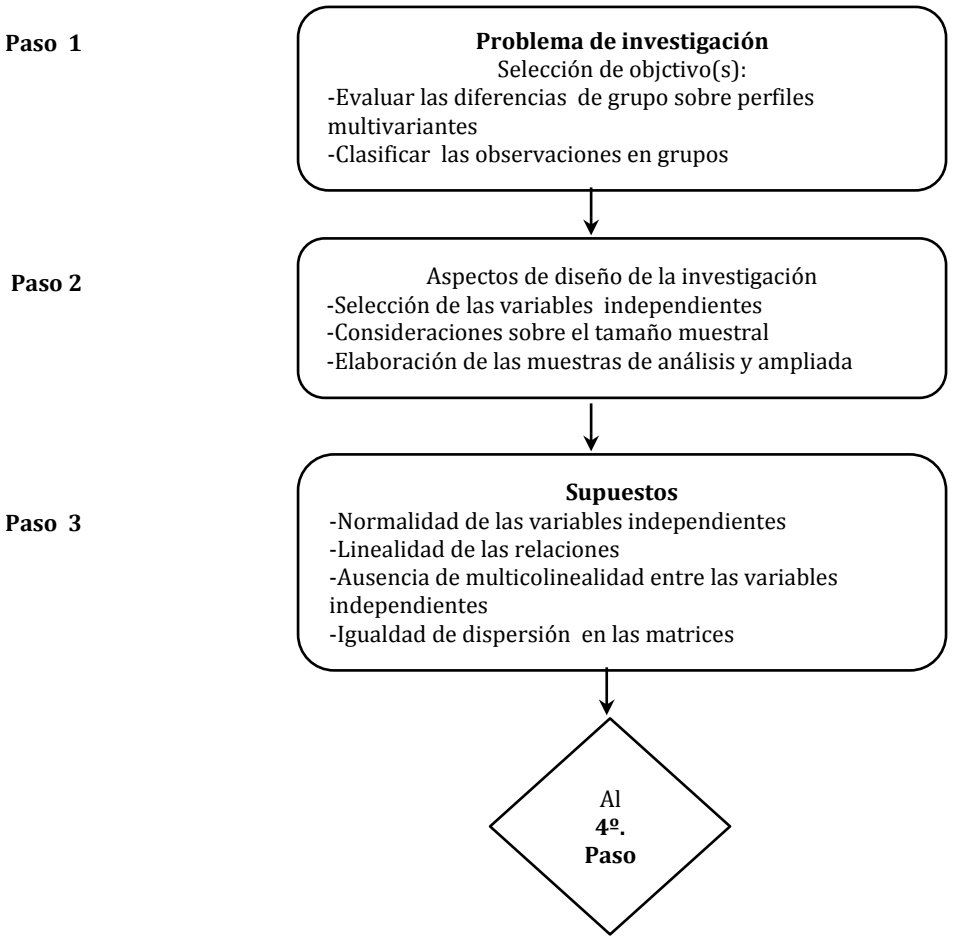
6.4. Análisis Discriminante Múltiple: Proceso

Se puede considerar la aplicación del análisis discriminante desde el punto de vista de la construcción de un modelo en seis etapas, introducido en el **Capítulo 2** y representado en la **Figura 6.8** (etapas **1 a 3**) y la **Figura 6.9** (etapas **4 a 6**).

Como en todas las aplicaciones multivariantes, el primer paso del análisis consiste en fijar los objetivos. Después, deberá centrarse en cuestiones de diseño específicas y estar seguro de que se cumplen los supuestos básicos. El análisis continúa con la derivación de la función discriminante y la determinación de si puede obtenerse o no una función estadísticamente significativa que separe los dos (o más) grupos. Los resultados discriminantes se evalúan entonces para ver la capacidad predictiva construyendo para ello una matriz de clasificación. El siguiente paso, la interpretación de la función discriminante, determina cuál de las variables independientes es la que más contribuye a discriminar entre los grupos. Finalmente, la función discriminante debería ser validada mediante una ampliación de la muestra. Cada uno de estos pasos se trata en las siguientes secciones.

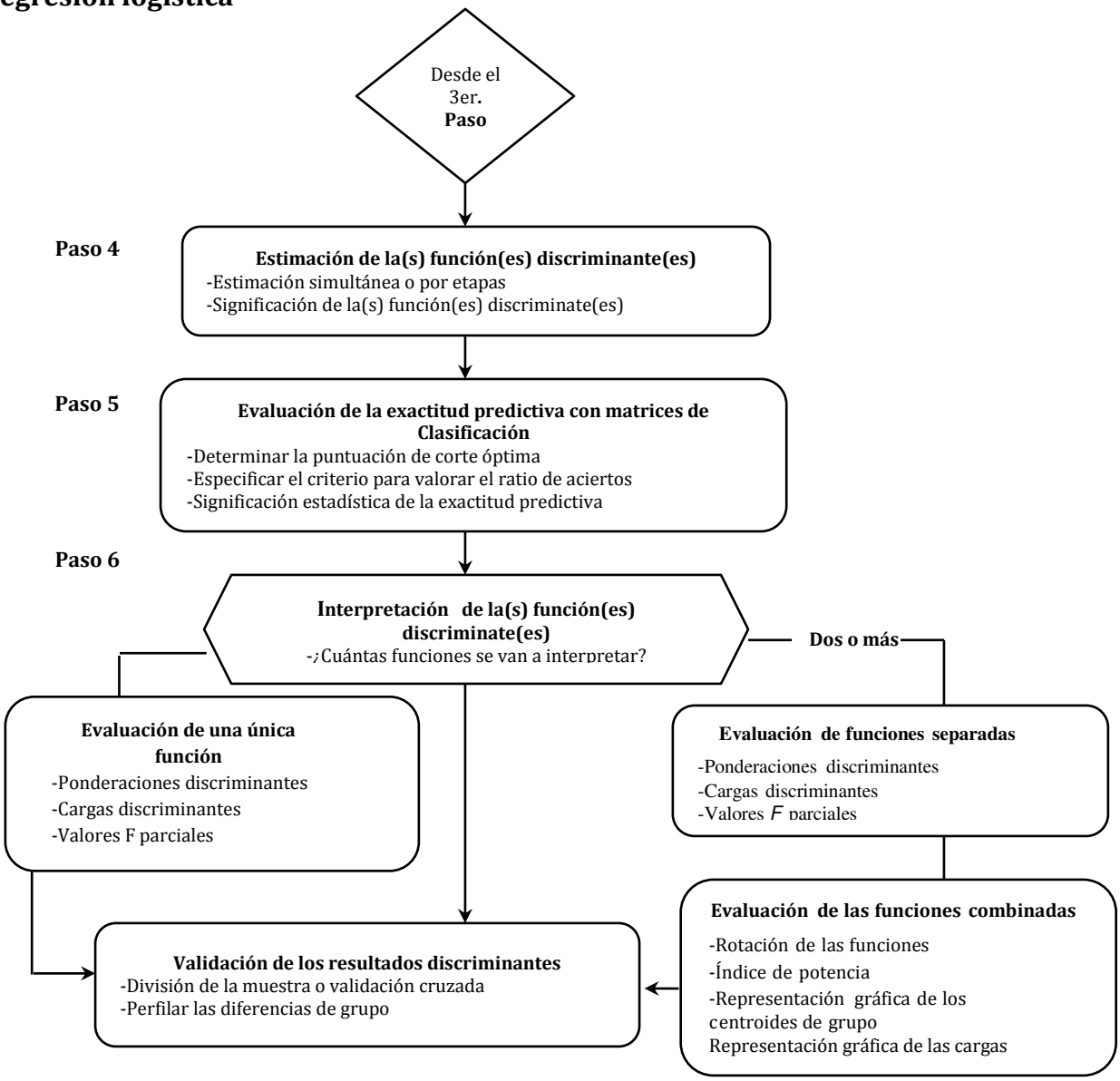
La **regresión logística** se analizará en una sección aparte tras examinar el proceso de decisión en el análisis discriminante. De esta forma, podrán aclararse las diferencias y semejanzas entre estas dos técnicas.

Figura 6.8. Diagrama de flujo pasos 1-3 del análisis discriminante múltiple y regresión logística



Fuente: Hair et al. (1999)

Figura 6.9. Diagrama de flujo pasos 4-6 del análisis discriminante múltiple y regresión logística



Fuente: Hair et al. (1999)

6.5. Análisis Discriminante Múltiple: Objetivos

Paso 1: Establecimiento de objetivos

Un repaso de los objetivos para la utilización del análisis discriminante debería clarificar bastan te su naturaleza. El análisis discriminante puede tratar cualquiera de los siguientes objetivos de investigación:

1. Determinar si existen diferencias estadísticamente significativas entre los perfiles de las puntuaciones medias sobre un conjunto de variables de dos (o más) grupos definidos a priori.
2. Determinar cuál de las variables independientes cuantifica mejor de las diferencias en los perfiles de las puntuaciones medias de dos o más grupos.
3. Establecer los procedimientos para clasificar objetos (individuos, empresas, productos, etc.), dentro de los grupos, en base a sus puntuaciones sobre un conjunto de variables independientes.
4. Establecer el número y la composición de las dimensiones de la discriminación entre los grupos formados a partir del conjunto de variables independientes.

Así, el **análisis discriminante** es útil cuando **Usted está interesado en comprender las diferencias de los grupos o en clasificar correcta mente objetos en grupos o clases**. Por tanto, se puede considerar ésta técnica tanto como **tipo de análisis de perfil** como de **técnica predictiva analítica**. En cualquier caso, la técnica es la más apropiada **cuando existe una única variable dependiente categórica y varias variables independientes escaladas métricamente.** Como en el **análisis de perfil**, el análisis discriminante proporciona una **valoración objetiva de las diferencias entre grupos** sobre un conjunto de variables independientes. En este caso, la técnica es bastante similar al análisis multivariante de la varianza (vea **Capítulo 7)** Para comprender las diferencias del grupo, ésta técnica tiene en cuenta tanto el papel de las **variables independiente**s como las combinaciones que se construyen con estas variables que representan dimensiones de discriminación entre los grupos. Estas **dimensiones** son los **efectos conjuntos** de varias variables que trabajan unidas para **diferenciar entre grupos**. El uso de los **métodos de estimación secuencial** permite también **identificar subconjuntos** de variables con la mayor capacidad discriminante.

Finalmente, para fines de **clasificación**, el **análisis discriminante** proporciona una base, no sólo para **clasificar la muestra** utilizada para estimar la función discriminante, sino también **cualesquiera otras observaciones** que puedan tener valores para todas las variables independientes. De esta forma, el análisis discriminante **puede utilizarse para clasificar otras observaciones dentro de los grupos definidos.**

6.6. Análisis Discriminante Múltiple: Diseño

Paso 2: Diseño

El éxito en la aplicación del análisis discriminante requiere tener en cuenta varias cuestiones. Éstas incluyen la selección tanto de la variable dependiente como de las independientes, el tamaño muestral necesario para la estimación de las funciones discriminantes y la división de la muestra con fines de validación

6.6.1. Selección de las variables dependientes e independientes

Para aplicar el análisis discriminante, el investigador primero debe especificar qué variables van a ser independientes y qué variable va a ser dependiente. Recuerde que la variable dependiente es categórica y las variables independientes son métricas. Usted deberá centrarse primero en la variable dependiente. El número de grupos de la **variable dependiente (categorías)** puede ser de **2** o más, pero estos grupos **deben ser mutuamente excluyentes y exhaustivos.** Esto significa que cada observación debe estar colocada dentro de un grupo solamente. En algunos casos, la **variable dependiente**:

1. Consta de dos grupos (**dicotómica,** por ejemplo. alto vs. bajo).
2. Puede incluir varios grupos (**multicotómica**, como por ejemplo ocupaciones tales como físico, abogado o profesor)

Los ejemplos anteriores de **variables categóricas** constituyen verdaderas **dicotomías** (**o multicotomías**). Sin embargo, hay algunas situaciones donde el **análisis discriminante** es apropiado incluso aunque la **variable dependiente no sea una verdadera variable categórica, por ejemplo** cuando ésta es **ordinal o medida a intervalos** que queremos utilizar como **variable dependiente categórica.** En tales casos, **tendremos que crear una variable categórica. Por ejemplo,** si tuviéramos una variable que midiera el **número medio de bebidas de cola consumidas por día,** y la respuesta de los individuos se basara en una **escala de cero a ocho o más por día,** podríamos crear una tricotomía artificial **(tres grupos)** simplemente designando a aquellos individuos que:

1. Han **consumido ninguna,**
2. Una o dos bebidas al día como pequeños consumidores;
3. Aquellos que han consumido tres, cuatro o cinco al día como consumidores medios; y
4. A quienes consumieron seis, siete, ocho o más como grandes consumidores.

Este procedimiento mencionado, **crearía una variable categórica de tres grupos donde el objetivo sería discriminar entre pequeños, medios y grandes consumidores de cola.**

Se puede construir **cualquier número de grupos categóricos artificiales.** Lo más frecuente es **crear dos, tres o cuatro categorías.** Pero se podría establecer un número más grande de categorías si fuese necesario.

Cuando se crean tres o más categorías, se presenta la posibilidad de examinar **solamente los grupos extremos en un análisis discriminante de dos grupos**. A este proceso se le **denomina enfoque de los extremos polares.** Este enfoque **compara solamente los 2 grupos extremos excluyendo los grupos medios del análisis discriminante.** Por ejemplo, Usted podría examinar los pequeños y grandes consumidores de cola, y excluir a los consumidores medios. Este enfoque puede ser utilizado en cualquier momento en que desee examinar solamente los grupos extremos. **Usted puede también intentar esta aproximación cuando los resultados del análisis de la regresión no sean tan buenos como esperaba.**

Este accionar puede ser de ayuda porque es posible que las diferencias del grupo puedan aparecer incluso aunque los resultados de la regresión sean pobres; es decir, **el enfoque de los extremos polares con análisis discriminante puede revelar diferencias que no resultan claras en un análisis de regresión de un conjunto completo de datos .** Tal manipulación de los datos naturalmente **exige precaución** al interpretar los resultados.

Después de tomarse una decisión sobre la variable dependiente, Usted debe **decidir qué variables independientes incluye en el análisis.** Éstas variables generalmente se seleccionan de dos formas:

1. **Identificar las variables** tanto en la investigación previa como desde el modelo teórico que sirve de fundamento a la pregunta de la investigación.
2. **Por intuición,** utilizando el conocimiento del investigador y seleccionando intuitivamente las variables para las cuales no existe investigación previa o teoría, pero que lógicamente podrían relacionarse **para predecir** los grupos de la variable dependiente.

6.6.2. Tamaño muestral

El análisis discriminante es bastante **sensible al ratio entre el tamaño muestral y el número de variables predictoras**. Muchos estudios sugieren un **ratio de 20 observaciones por cada variable predictora.** Aunque e**ste ratio puede ser difícil de conseguir en la práctica**, Usted debe tener en cuenta que **los resultados podrían llegar a ser inestables** a medida que el **tamaño muestral disminuye** en relación con el **número de variables independientes**. El tamaño mínimo recomendado es de cinco (**5**) observaciones por variable independiente. Nótese que este ratio se aplica a todas las **variables consideradas en el análisis**, incluso si todas las variables consideradas **no entran en la función discriminante** (como en la estimación por etapas). Además del tamaño muestral total, Usted debe también considerar el tamaño muestral de cada grupo. **Como mínimo, el tamaño del grupo más pequeño debe ser mayor que el número de variables independientes.** Como una regla práctica, **cada grupo debe tener al menos 20 observaciones.** Pero incluso aunque todos los grupos excedan las **20** observaciones. Usted debe también considerar los **tamaños relativos de los grupos.** Si los grupos **varían ampliamente** en tamaño, esto puede afectar a la estimación de la función discriminante y a la clasificación de las observaciones. En la **etapa de clasificación,** los grupos más grandes tienen una posibilidad des proporcionadamente más grande de clasificación. Si los tamaños de los grupos varían de forma importante, puede que el investigador quiera muestrear aleatoriamente desde el grupo más grande, y con ello reducir su tamaño a un nivel comparable con el grupo más pequeño.

6.6.3. División de la muestra

Una observación final sobre la influencia del tamaño muestral en el análisis discriminante. Como se verá más adelante, en muchas ocasiones **la muestra se divide en dos submuestras;** una, utilizada para la **estimación de la función discriminante**, y otra con fines de **validación**. Es **esencial que cada submuestra tenga un tamaño adecuado** para apoyar las conclusiones de los resultados. Se han sugerido un conjunto de procedimientos para dividir la muestra, pero el más utilizado implica **desarrollar la función discriminante con un grupo y luego probarla con un segundo grupo.** El procedimiento habitual consiste en **dividir aleatoriamente la muestra total de encuestados en dos grupos:**

1. El primer grupo, de la **muestra de análisis**, se usa para construir la función discriminante.

2. El segundo grupo, la **ampliación de la muestra**, se usa para validar la función discriminante. Este método de validación de la función se denomina **división de la muestra o enfoque de validación cruzada [**Frank, R. et al. 1965, Green, P. E., y Carroll, J. D. 1978**,** Perreault, et al 1979**].**

No se ha establecido una manera definitiva para dividir la muestra en los grupos de análisis y ampliación (o validación). El procedimiento más común es **dividir el total del grupo, de tal forma que la mitad de los encuestados pertenezca a la muestra de análisis y la otra mitad a la ampliación de la muestra**. No obstante, **no se ha establecido ninguna regla fiable,** y algunos investigadores prefieren **una división 60-40 o 75-25** entre los **grupos de análisis y ampliación.**

Cuando se seleccionan los individuos para los **grupos de análisis y validación**, generalmente se sigue un proceso de **muestreo estratificado proporcional**. Si los grupos categóricos del análisis discriminante **están igualmente representados** en el total de la muestra, se selecciona un **número igual de individuos**. Si los grupos categóricos **son desiguales,** los tamaños de los grupos seleccionados para la ampliación de la muestra deben ser **proporcionales a la distribución total de la muestra. Por ejemplo,** si una muestra consiste en **50 hombres y 50 mujeres**, la ampliación de la muestra tendría **25 hombres y 25 mujeres.** Si la muestra contiene **70 mujeres y 30 hombres,** entonces la ampliación de la muestra consistiría en **35 mujeres y 15 hombres**.

Es necesario realizar algunos comentarios adicionales en lo referente a dividir la muestra total en los grupos de análisis y ampliación. Si Usted va a **dividir la muestra en los grupos de análisis y ampliación, la muestra debe ser suficientemente grande para realizarlo.** Una vez más**, no se ha establecido una regla fiable,** pero parece lógico que Usted querría **al menos 100 observaciones en el total de la muestra para justificar el dividirla en dos grupos.** Una solución de compromiso es que, si el tamaño muestral **es demasiado pequeño** como para justificar la división en los grupos de **análisis y ampliación**, construya la función con la **muestra entera y después utilice esta función para clasificar** el mismo grupo que sirvió para construirla. Este procedimiento **sesga al alza la capacidad predictiva de la función**, pero es ciertamente mejor que no validar la función en absoluto.

6.7. Análisis Discriminante Múltiple: Supuestos

Paso 3: Supuestos de aplicabilidad

Es **deseable encontrar** las condiciones para la correcta aplicación del **análisis discriminante.** Los supuestos clave para obtener la función discriminante son el de **normalidad multivariante de las variables independientes y el de estructuras (matrices) de covarianza y dispersión desconocida (pero iguales) para los grupos,** como se definió para la variable dependiente **[**Green, P. E. 1978**].** Aunque existe una evidencia contradictoria sobre **la sensibilidad del análisis discriminante a incumplimientos de estos supuestos**, Usted deberá examinar los datos y, si los supuestos no se cumplen, deberá **identificar los métodos alternativos disponibles y la influencia que cabría esperar sobre los resultados**. **Los datos que no cumplan** el supuesto de **normalidad multivariant**e pueden causar problemas en la estimación de la función discriminante. Por ello, se sugiere que se use la **regresión logística** como una **técnica alternativa**, si es posible. **Las matrices de covarianzas** distintas pueden afectar

desfavorablemente al proceso de clasificación. Si los **tamaños muestrales son pequeños y las matrices de covarianzas son distintas**, la **significación** estadística del proceso de estimación se ve **afectada desfavorablemente**. El caso más· probable es el de **covarianzas distintas** entre grupos de tamaño muestral adecuado, en donde las observaciones son "***sobre clasificadas***" dentro de los grupos con matrices de covarianzas más grandes.

Este efecto **puede minimizarse incrementando el tamaño muestral y también usando las matrices de covarianzas específicas de cada grupo con fines clasificatorios**, pero esta aproximación obliga a la **validación cruzada** de los resultados discriminantes. Finalmente, en muchos de los programas estadísticos están disponibles **técnicas de clasificación cuadráticas,** si existen grandes diferencias entre las matrices de covarianzas de los grupos y otras soluciones no minimizan el efecto **[**Gessner, et al**,** 1988 Huberty, 1984**,** Johnson, N., y Wichem D. 1982**]**.Otra característica que afecta los resultados es la **multicolinealidad** entre las variables **independientes** la cual consiste en que **2 o más variables independientes están altamente correlacionadas,** por lo que una variable puede venir **muy bien explicada o predicha por otras variables** y, por ello, **añadir poca capacidad explicativa al conjunto completo.** Esto es especialmente crítico cuando se emplean los procesos **por etapas.** Usted deberá interpretar la función discriminante, conociendo el **nivel de multicolinealidad** y su influencia al determinar que variables entran en la solución por etapas. (Véase **Capítulo 5**).Al igual que con alguna de las técnicas multivariantes que emplean un valor teórico, **un supuesto implícito es que todas las relaciones son lineales**. **Las relaciones no lineales no están reflejadas en la función discriminante** a menos que se realicen **transformaciones específicas** de la variable para representar los efectos no lineales. Finalmente, **los casos atípicos** pueden tener una influencia sustancial en la precisión clasificatoria de cualquier resultado del análisis discriminante. **Examine todos los resultados por la presencia de casos atípicos y elimine si fuera necesario.** (Vea **Capítulo 3.**)

6.8. Análisis Discriminante Múltiple: Estimación

Paso 4: Estimación y ajuste

Para obtener la función discriminante, Usted deberá **decidir el método de estimación y determinar después el número de observaciones que se van a mantener** (véase **Figura 6.9**). Una vez que se han estimado las funciones, puede valorarse el ajuste global del modelo de varias formas. Puede iniciar al calcular las **puntuaciones discriminantes *z***, también conocidas como **puntuaciones *z***. La **comparación de las medias de los grupos sobre las puntuaciones *z*** ofrece una medida de la discriminación entre grupos.

La capacidad en la **predicción** se valora por el **número de observaciones clasificadas** dentro de los grupos adecuados. Se dispone de varios criterios para valorar si el proceso de **clasificación alcanza significación estadística y/o práctica. Finalmente, la validación por casos** puede identificar la precisión en la clasificación de cada caso y su influencia relativa sobre la estimación global del modelo.

6.8.1. Método de cálculo

Se tienen los métodos de cálculo para derivar una función discriminante:

1. **El método simultáneo (directo).** Implica el **cálculo de la función discriminante donde todas las variables independientes son consideradas simultáneamente**. Por ello, la(s) función(es) discriminante(s) se calculan basándose en el **conjunto completo de variables independientes**, sin considerar la capacidad discriminante de cada variable independiente. Este método es apropiado cuando, por **razones teóricas**, Usted quiere introducir todas las variables independientes en el análisis y **no está interesado en observar resultados intermedios** basados solamente en las variables que discriminan mejor
2. **El método por etapas.** Es una alternativa al enfoque simultáneo. Incluye las **variables independientes** dentro de la función discriminante **de una en una**, según su capacidad discriminatoria. **Comienza eligiendo la variable que mejor discrimina**. La variable inicial se empareja entonces con cada una de las variables independientes (de una en una), y se elige la variable que más consigue incrementar la capacidad discriminante de la función en combinación con la primera variable.
3. **La tercera y posteriores variables se seleccionan de una manera similar.** Mientras se incluyen variables adicionales, algunas variables seleccionadas previamente pueden ser eliminadas si la información que contienen sobre las diferencias del grupo está contenida en alguna combinación de otras variables incluidas en posteriores etapas. Al final, o bien todas las variables habrán sido incluidas en la función, o se habrá considerado que las variables excluidas no contribuyen significativamente a una mejor discriminación. Este método es útil cuando Usted **quiere considerar un número relativamente grande de variables independientes para incluir en la función.** Seleccionando **secuencialmente** la siguiente variable que mejor discrimina en cada paso, **se eliminan las variables que no son útiles para discriminar** entre los grupos y se identifica un conjunto reducido de variables. El conjunto reducido es generalmente tan bueno como, y algunas veces mejor que, el conjunto completo de variables. Pero Usted deberá darse cuenta de que ésta técnica **puede llegar a ser menos estable y generalizable en tanto que el ratio del tamaño muestral respecto a las variables independientes se reduce por debajo de las 20 observaciones por variable independiente.** Valide los resultados de tantas formas como sea posible.

6.8.2. Significación estadística.

Después de calcular la función discriminante, usted debe **valorar el nivel de significación.** Se dispone de varios criterios estadísticos:

1. Las medidas del ***lambda de Wilks,*** *la* ***traza de Hotelling*** *y el* ***criterio de Pillai*** evalúan la **significación estadística de la capacidad discriminatoria de la función(es) discriminante(s).**
2. **La mayor raíz característica de Roy evalúa solamente la primera función discriminante.** Para un tratamiento más detallado de las ventajas y desventajas de cada criterio, se remite al lector a la discusión del contraste de significación en el **análisis multivariante de la varianza** del **Capítulo 7.**
3. Si se utiliza un **método por etapas** para estimar la función discriminante, son más apropiadas las medidas **de *Mahalanobis* y V de *Rao*.** Ambas son medidas de

distancia generalizada. El procedimiento de la D^2 **de** ***Mahalanobis*** se basa en la **distancia euclídea al cuadrado generalizada que se adecúa a varianzas desiguales.** La **principal ventaja** de este procedimiento es que se calcula en el **espacio original de las variables predictoras,** en lugar de alguna otra versión obsoleta utilizada en otras medidas.

El procedimiento D^2 **de** ***Mahalanobis*** llega a ser particularmente **critico** a medida que el número de variables predictoras se incrementa dado que no se da ninguna reducción en la dimensionalidad. La **pérdida de dimensionalidad causa una pérdida de información,** porque hace disminuir la varianza de las variables independientes. En general, el procedimiento de la D^2 **de** ***Mahalanobis*** **es el indicado** cuando se está interesado en **aprovechar al máximo la información disponible.** El procedimiento D^2 **de** ***Mahalanobis*** lleva a cabo un análisis discriminante por etapas similar al análisis de regresión por etapas. Este **procedimiento por etapas** está diseñado para obtener **el mejor modelo de una variable, seguido por el mejor modelo de dos variables,** y así sucesivamente hasta que ninguna otra variable cumpla la regla de selección deseada. La regla de selección en este procedimiento **es maximizar la distancia** D^2 **de** ***Mahalanobis*** entre los grupos. Tanto el **método por etapas como el simultáneo** están disponibles en los principales programas estadísticos. El criterio convencional de **0.05** o superior se utiliza a menudo. Muchos investigadores creen que si la función no es significativa a ese nivel o más, existe poca justificación para seguir adelante.Algunos investigadores, sin embargo, no están de acuerdo. Su regla de decisión para continuar hasta un nivel de significación mayor (**0.10 o más**) resulta de **comparar el coste de la in formación frente a su valor**. Si son aceptables mayores **niveles de riesgo** por **incluir resultados no significativos** (por ejemplo, **niveles de significación mayores a 0.5**), se pueden mantener funciones discriminantes que son significativas al nivel del **0.2 o incluso al 0.3**. Si el número de **grupos es de tres o más**, entonces deberá decidir no solamente si la discriminación entre el **total de los grupos** es estadísticamente significativa, sino también **si cada una de las funciones discriminantes estimadas es estadísticamente significativa**. **El análisis discriminante estima una función discriminante menos que grupos existentes. Si se analizan tres grupos, se estimarán dos funciones discriminantes** y así sucesivamente. Todo el software estadístico proporciona al investigador la información necesaria para averiguar el **número de funciones necesarias para obtener significación estadística**, **sin incluir funciones discriminantes que no incrementen la capacidad discriminatoria significativamente.** Con una o más funciones que no son estadísticamente significativas, el modelo deberá reestimarse con el número de funciones que se hayan obtenido, limitado por el número de funciones significativas. Así, la valoración de la precisión en la predicción e interpretación de las funciones discriminantes **estarán basadas solamente en funciones significativas**

6.8.3. Valoración del ajuste global.

Una vez **identificadas las funciones discriminantes significativas,** deberá **averiguar el ajuste global** de la(s) función(es) discrimínate(s) considerada(s). Esta valoración conlleva tres tareas:

1. **Calcular la puntuación *z* discriminante** para cada observación

2. **Evaluar diferencias de grupos sobre las puntuaciones *z* discriminantes,** y
3. **Valorar la precisión en la predicción** de pertenencia al grupo.

6.8.4. Cálculo de las puntuaciones *Z* discriminantes

Definidas las funciones discriminantes, **se han establecido las bases para el cálculo de las puntuaciones *z* discriminantes.** La **puntuación *z*** discriminante de cualquier función discriminante puede calcularse para cada observación mediante la siguiente fórmula: $$Z_{jk} = a + W_1 X_{1k} + W_2 X_{2k} + W_3 X_{3k} + \ldots W_n X_{nk}$$

Donde:

Z_{jk}**= puntuación *z*** discriminante de la función discriminante j para el objeto k

a= constante

W_i**=** ponderación discriminante para la variable independiente i

X_{ik}**=** variable independiente i para el objeto k

Esta puntuación, una **medida métrica**, ofrece unas **medias directas para comparar observaciones para cada función.** Las observaciones con **puntuaciones *z* similares** se suponen más parecidas sobre las variables que constituyen esta función que aquellas con puntuaciones dispares. Estas son versiones de la **función discriminante** que emplean valores y **ponderaciones estandarizadas o no estandarizadas**. La **versión estandarizada** es más útil en la interpretación, pero la **versión no estandarizada** es más fácil de utilizar en el cálculo de la **puntuación *z* discriminante**. Debemos damos cuenta de que **la función discriminantes difiere de la función de clasificación, también conocida como la función discriminante lineal de Fisher.** Las **funciones de clasificación**, una para cada grupo, pueden utilizarse al **clasificar observaciones**. En este método, unos valores de la observación para las variables independientes se incluyen en las **funciones de clasificación** y se calcula una **puntuación de clasificación para cada grupo para esa observación.** La observación se clasifica entonces en el grupo con la mayor puntuación de clasificación. Se utiliza la **función discriminante** como el **medio de clasificar** porque ofrece una **representación resumida y simple** de cada función discriminante, simplificando la interpretación y la valoración de las variables independientes**.**

6.8.5. Valorando la exactitud en la predicción de pertenencia al grupo

Dado que la variable dependiente no es métrica, no es posible utilizar una medida como el R^2, como se hace la regresión múltiple, para valorar la **exactitud predictiva.** En su lugar, **cada observación debe valorarse como si fuera correctamente clasificada.** Al hacer esto, deben realizarse una serie de consideraciones: **la razón de ser práctica y estadística** para elaborar **matrices de clasificación,** la determinación de la **puntuación de corte,** la construcción de **matrices de clasificación** y los **estándares para valorar la exactitud clasificatoria,** que implica cuestionarse:

¿Por qué se elaboran matrices de clasificación? Los contrastes estadísticos para valorar la significación de las funciones discriminantes **no informan sobre lo que correctamente que predice la fundición. Por ejemplo,** suponga dos grupos que son **significativamente diferentes** por encima del nivel del **0.01.** Con tamaños de la muestra suficientemente grandes, las **medidas de los grupos (centroides)** podrían ser **virtualmente idénticas** y todavía tendríamos **significación estadística.**

En resumen, estos **contrastes adolecen de los mismos inconvenientes que los contrastes clásicos de hipótesis.** Por esto, **el nivel significación de estos estadísticos es una indicación muy pobre de la capacidad de la función para discriminar entre los dos grupos.** Para determinar **la capacidad predictiva** de una funcio n discriminante, **deberá construir matrices de clasificación.** Aclarando mejor la utilidad del procedimiento de las matrices de clasificacio n, lo relacionaremos con el concepto de $\boldsymbol{R^2}$ en el ana lisis de regresio n. La mayor a de las lecturas de art culos acade micos el/los autor(es) ha encontrado relaciones estad sticamente significativas, **sin embargo ha explicado solamente el 10% (o menos) de la varianza ($\boldsymbol{R^2}$ = 0.10).** Generalmente este $\boldsymbol{R^2}$ es significativamente **distinto de cero** simplemente **por el tamaño muestral es grande.** En el ana lisis discriminante mu ltiple, el **ratio de aciertos** (porcentaje correctamente clasificado) es ana logo al $\boldsymbol{R^2}$ de la regresio n. El **ratio de aciertos revela lo correctamente** que la funcio n discriminante clasifico los objetos; el R^2 indica cua nta varianza explico la ecuacio n de regresio n. **El contraste *F*** para la significación estadística del R^2 es por tanto análogo al contraste ***Chi-cuadrado*** (o D^2) de significación en el análisis discriminante. Claramente, **con un tamaño muestral suficientemente grande** el análisis discriminante, podr amos tener una diferencia estad sticamente significativa entre los dos (o ma s) grupos sin embargo **clasificar correctamente solamente el 53% (cuando la probabilidad es el 50 % con tamaños de grupo iguales) [**Morrison, 1967**]**

Determinación de la puntuación de corte. Si los contrastes estadísticos indican que la función discrimina significativamente, es usual elaborar matrices de clasificación para proporcionar una valoración más precisa de la calidad discriminatoria de la función. Sin embargo, antes de que pueda ser construida una matriz de clasificación el investigador debe determinar la puntuación de corte. La puntuación de corte es el criterio (puntuación) frente al cual cada puntuación discriminante individual es comparada para determinar dentro de qué grupo debe ser clasificado cada objeto.
Al construir las matrices de clasificación, el investigador querrá determinar la puntuación de corte óptima (también llamada **valor *z* crítico**). La puntuación de corte óptima diferirá dependiendo de si los tamaños de los grupos son iguales o distintos. Si los grupos son de igual tamaño, la puntuación de corte óptima estará a mitad de camino entre los centroides de los dos grupos. El punto de corte para dos grupos de igual tamaño

se define tanto como: $$\boldsymbol{Z_{CE} = (Z_A + Z_B)}$$

Donde:

$\boldsymbol{Z_{CE}}$= valor de la puntuación de corte crítica para los grupos de igual tamaño

$\boldsymbol{Z_A}$= centroide del grupo A

$\boldsymbol{Z_B}$= centroide del grupo B

Especificación de las probabilidades de clasificación para tamaños de grupo distintos. Para calcularlo correctamente, cuando los tamaños de grupo son distintos, deberá también determinar si los tamaños de los grupos poblacionales se deben **considerar iguales. El supuesto por defecto es que las probabilidades sean iguales:** en otras palabras, se supone que cada grupo tiene una misma probabilidad de ocurrir incluso aunque los tamaños de los grupos en la muestra sean distintos. Si el investigador no está seguro de si las proporciones observadas en la muestra son representativas de las proporciones de la población, la **solución conservadora es la igualdad en las probabilidades.**

Sin embargo, **si la muestra está tomada aleatoriamente de la población por lo que los grupos sí representan las proporciones poblacionales en cada grupo,** entonces la mejor estimación de las probabilidades anteriores no es la de igualdad, sino la de las proporciones muestrales. La influencia de especificar las anteriores probabilidades como iguales a las proporciones muestrales varía según la diferencia que exista entre las proporciones muestrales y las proporciones poblacionales. Pero Usted deberá determinar las probabilidades en todos los análisis (bien como iguales o bien basadas en los tamaños muestrales) para asegurar que los supuestos adecuados están presentes en los procesos de clasificación.

Determinación de la puntuación de corte para grupos de tamaño desigual. Si los **grupos no son de igual tamaño** y se supone que son representativos de las proporciones de la población, **una media ponderada de los centroides de los grupos proporcionará una puntuación de corte óptima para una función discriminante.** Se calcula como sigue:

$$\boldsymbol{Z_{CU} = (N_A Z_B + N_B Z_A)/(N_A + N_B)}$$

Donde:

$\boldsymbol{Z_{CU}}$**=** valor de la puntuación de corte crítica para tamaños de grupo distintos

$\boldsymbol{N_A}$**=** número del grupo A

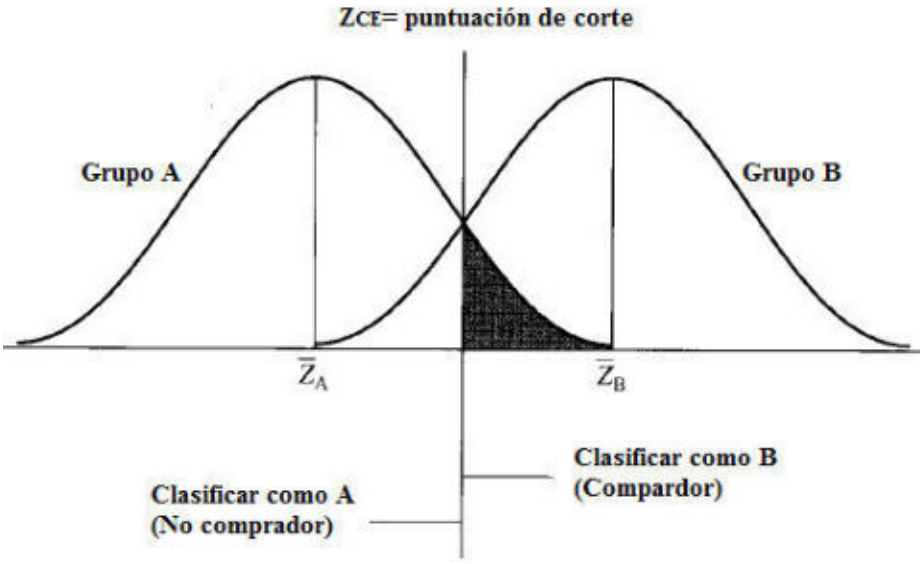
$\boldsymbol{N_B}$= número del grupo B

$\boldsymbol{Z_A}$= centroide del grupo A

$\boldsymbol{Z_B}$= centroide del grupo B

Las dos formas con las que se calcula la **puntuación de corte óptima** suponen que **las distribuciones están distribuidas normalmente** y se conocen las estructuras de dispersión de los grupos. Las **Figuras 6.10 y 6.11** ilustran el concepto de una **puntuación de corte óptima** para grupos iguales y distintos.

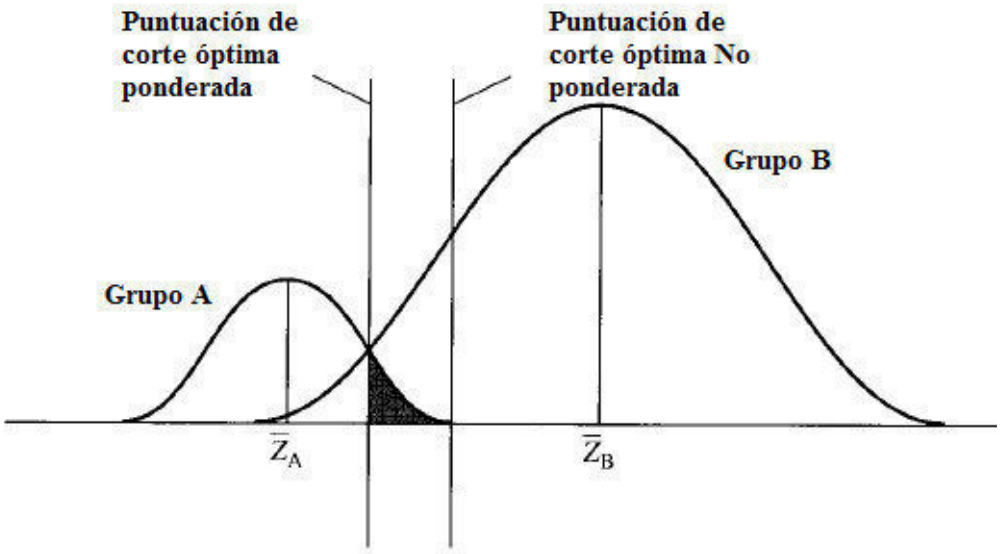
Figura 6.10. Puntuación de corte óptima con tamaños muestrales iguales



Fuente: propia

Se muestran tanto las **puntuaciones corte ponderadas como no ponderadas.** Es evidente que si el **grupo A** es mucho más pequeño que el **grupo B,** el **punto de corte óptimo estará más cercano al centroide del grupo A que al centroide del grupo B.** Además, si se utilizase **una puntuación de corte no ponderada,** ninguno de los objetos del **grupo A** estaría mal clasificado, pero una parte importante de los que están en el **grupo B** sí lo estaría.

Figura 6.11. Puntuación de corte óptima con tamaños muestrales distintos.



Fuente: propia

Costes de la clasificación errónea. La puntuación de **corte óptima** también debe tener en **cuenta el coste de clasificar de forma incorrecta** un objeto dentro de un grupo erróneo. Si los costes de clasificar incorrectamente a un individuo son aproximadamente iguales para todos los grupos, la puntuación de corte óptima será aquella que clasifique mal el menor número de objetos entre todos los grupos. Si los costes de clasificación errónea son distintos, la puntuación de **corte óptima** será aquella que minimice dichos costes. Se consideran enfoques ma s sofisticados para determinar las puntuaciones de corte (Dillon y Golstein 1984 ; Huberty 1987). Estos enfoques esta n **basados en un modelo estadístico bayesiano** y son apropiados cuando los costes de clasificar mal en ciertos grupos son muy altos, cuando los grupos de tamaños enormemente diferentes o cuando se quiere aprovechar un conocimiento a priori de las probabilidades desde la pertenencia a un grupo. En la práctica, al calcular la puntuación de corte generalmente no es necesario incluir las medidas de la variable primaria para cada individuo dentro de la función discriminante, y obtener la situación discriminante para cada persona para utilizarlo en el cálculo de $\boldsymbol{Z_A}$ y $\boldsymbol{Z_B}$ **(centroides de los grupos A y B).** En muchos casos el programa de computador proporciona las puntuaciones discriminantes, al igual que $\boldsymbol{Z_A}$ y $\boldsymbol{Z_B}$, como output habitual. Cuando el investigador tiene los centroides de los grupos y los tamaños muestrales, debe

únicamente sustituir los valores en la fórmula adecuada para obtener la puntuación de corte óptima.

Construcción de las matrices de clasificación. Para **validar** la función discriminante por medio de matrices de clasificación, **la muestra debe dividirse aleatoriamente en dos grupos.** Uno de los grupos (**la muestra de análisis**) se utiliza para calcular la función discriminante. El otro grupo (**la muestra de validación** o **ampliación de la muestra**) se usa para la elaboración de la matriz de clasificación. El proceso consiste en **multiplicar las ponderaciones generadas por la muestra de análisis por las medidas de la variable primaria de la ampliación de la muestra.** Después, las puntuaciones discriminantes individuales para la ampliación de la muestra se comparan con el valor de la puntuación de corte crítica y se clasifican de la siguiente forma:

Clasificar a un individuo dentro del **grupo A si** $Z_n < Z_{ct}$

Clasificar a un individuo dentro del **grupo B si** $Z_n > Z_{ct}$

Donde:

Z_n= **puntuación** ***z*** discriminante para el individuo n-ésimo

Z_{ct}= valor de la puntuación de corte crítica

Los resultados del proceso de clasificación se presentan de **forma matricial,** como se muestra en la **Figura 6.12** los elementos de la diagonal de la matriz representan el número de individuos correctamente clasificados.

Figura 6.12. Matriz de clasificación para el análisis discriminante de dos grupos

	Grupo predicho			
Grupo real	1	2	Tamaño del grupo real	Porcentaje correctamente clasificado
1	22	3	25	88
2	5	20	25	80
Tamaño del grupo predicho	27	23	50	**84***

*Porcentaje correctamente clasificado= (Número correctamente clasificado/Número total de observaciones) X 100= [(22 + 20)/50] X 100= **84%**

Fuente: propia

Los números fuera de la diagonal representan las clasificaciones incorrectas. Los números de la columna denominada "***tamaño del grupo real***" representan el número de individuos que realmente hay en uno de los dos grupos. Los números que están al final de las columnas representan el número de individuos asignados a los grupos por la función discriminante. El porcentaje correctamente clasificado en cada grupo aparece en el lado derecho de la matriz y el porcentaje total correctamente clasificado, también conocido como **ratio de aciertos**, aparece al final. En el ejemplo, el número de individuos correctamente asignado en el grupo 1 es **de 22,** mientras que **3 miembros del grupo 1** fueron asignados incorrectamente al grupo 2. De modo similar, el número de clasificaciones correctas en el **grupo 2 es de 20 y el número de asignaciones incorrectas en el grupo 1 es de 5**. Con ello el porcentaje de precisión clasificatoria de la función discriminante para los **verdaderos grupos 1 y 2 sería del 88 y del 80 %,** respectivamente. La precisión total de la clasificación (**ratio de aciertos**) sería **del 84 %**

Un último tema referido al procedimiento de clasificación es el contraste t. Se dispone de un **contraste *t*** para determinar el nivel de significación para la precisión clasificatoria. La expresión para un análisis de dos grupos (con tamaño muestral igual) es:

$$t= (p^*-0.5)/\sqrt{0.5(1-0.5)/N}$$

Donde:
p= proporción clasificada correctamente
N= tamaño muestra

Esta expresión puede adaptarse para su uso cuando existen más grupos y para tamaños muestrales distintos.

6.8.6. Medición de la capacidad predictiva mediante la aleatoriedad.

Como se mencionó anteriormente, la capacidad predictiva de la función discriminante se mide con el **ratio de aciertos**, el cual se obtiene de la **matriz de clasificación**. El investigador podría preguntarse sobre qué se considera un nivel aceptable de capacidad predictiva para una función discriminante. **Por ejemplo**, **¿es el 60%un nivel aceptable o debería esperarse un 80 o un 90 % de capacidad predictiva?** Para responder a esta pregunta, deberá determinar primero el porcentaje que podría ser clasificado correctamente de forma aleatoria (sin la ayuda de la función discriminante).

Determinación del criterio basado en la aleatoriedad Cuando los tamaños muestrales son iguales, la determinación de la clasificación aleatoria es bastante simple; se obtiene dividiendo **1** por el número de grupos. La fórmula es **C = 1/(número de grupos).** Por ejemplo, en una función de dos grupos la probabilidad sería de **0.50**; para una función de **tres grupos** la probabilidad sería de **0.33** y así sucesivamente.

El establecimiento de la clasificación aleatoria en situaciones donde los **tamaños de los grupos son distintos es algo más complicado.** Supongamos que tenemos una muestra en la que **75 sujetos pertenecen a un grupo y 25 al otro**. Podríamos asignar arbitrariamente a todos los sujetos al grupo más grande y conseguir un **75% de capacidad predictiva** sin la ayuda de la función discriminante. Se podría concluir que a menos que la función discriminante lograse una capacidad predictiva de más del **75 %**, no se tendría en consideración ya que no nos ayudaría a mejorar nuestra capacidad predictiva. Determinar la clasificación aleatoria basándose en el tamaño muestral del grupo más grande se conoce como criterio de máxima aleatoriedad. Se determina calculando el porcentaje de la muestra completa representado por el más grande de los dos (o más) grupos. Por ejemplo, si los tamaños de los grupos son **65 y 35**, el criterio de máxima aleatoriedad es **65% de clasificaciones correctas**. Por tanto, si el ratio de aciertos para la función discriminante no excedió el **65%**, entonces no nos ayudaría a predecir basamos en este criterio.

El **criterio de máxima aleatoria** debería utilizarse cuando el único objetivo del análisis discriminante es **maximizar el porcentaje clasificado correctamente [**Morrison, 1967**].** Sin embargo, son escasas las situaciones en que sólo nos importa la maximización del porcentaje correctamente a los miembros de los grupos. En casos donde los tamaños muestrales son distintos y el investigador quiere clasificar a los miembros de los grupos, la función discriminante desafía lo extraño clasificando a un sujeto en un grupo(s) más pequeño(s). Pero el criterio de la aleatoriedad no tiene este hecho en cuenta **[**Morrison, 1967**].** Por lo que otro criterio de aleatoriedad **–el criterio de aleatoriedad proporcional-**

debe emplearse en tales situaciones y debe emplearse cuando los tamaños de los grupos son distintos y el investigador desea identificar adecuadamente a los miembros de los dos (o más) grupos. La fórmula para este criterio es:

$$C_{PRO} = p \ \ (1 - p)$$

Donde:
p= proporción de individuos del grupo 1
1-p= proporción de individuos del grupo 1

Empleando los tamaños de los grupos de nuestro ejemplo anterior (**75 y 25**), se tiene que el **criterio de aleatoriedad proporcional** sería del **62.5 %** frente al **75%**. Por lo tanto, en este ejemplo, una **capacidad predictiva del 75 %** sería aceptable porque **es > 62,5 %** del **criterio de aleatoriedad proporcional.** Estos criterios de aleatoriedad son útiles sólo cuando se calculan con **ampliación** de la muestra (**enfoque de división de la muestra**). Si los individuos utilizados para calcular la función discriminante son los que están siendo clasificados, el resultado estará sesgado al alza en su capacidad predictiva. En tales casos, estos dos criterios tendrían que ser ajustados a su vez al alza para tener en cuenta este sesgo.

Comparación del ratio de aciertos con el criterio de aleatoriedad. La cuestión de la precisión en la clasificación es crucial. Si el porcentaje de clasificaciones correctas es significativamente más grande que el que cabría esperar de forma aleatoria, se puede llevar a cabo un ejercicio de interpretación de las funciones discriminantes con la finalidad de elaborar perfiles de grupo. Sin embargo, si la precisión clasificatoria no es más grande que lo que se podría esperar aleatoriamente, sean cuales sean las diferencias estructurales que parezcan existir, apenas se aportaría nada a la interpretación; es decir, las diferencias en los perfiles de las puntuaciones no proporcionan una información significativa para identificar la pertenencia a un grupo. Entonces, la cuestión es, ¿cómo ha de ser la precisión en la clasificación en relación al **criterio de aleatoriedad**? Por ejemplo, si la probabilidad es del **50 %** (dos grupos, mismo tamaño muestral) ¿una precisión clasificatoria (**predictiva**) del **60%** justifica el pasar a la etapa de interpretación? No existen directrices generales para responder a esta pregunta. En última instancia, la decisión depende de los costes relacionados con el valor de la información. Si los costes asociados con un **60 %** de capacidad predictiva (en relación con un **50 %** de forma aleatoria) son más grandes que el valor que se deriva de los resultados, no hay justificación para pasar a la interpretación. Si el valor es alto en relación a los costes, el **60 %** de precisión justificaría pasar a la etapa de interpretación. Si el valor es alto en relación a los costes, el **60%** de precisión justificaría pasar a la etapa de interpretación. El argumento de costes versus valor ofrece escasa ayuda al investigador de datos poco experimentado; se sugiere el siguiente criterio: **la precisión clasificatoria debería ser por lo me nos un cuarto mayor** que aquella obtenida por aleatoriedad. Por ejemplo, si la precisión aleatoria es del **50%** la precisión clasificatoria debe ser del **62.5 % (62.5% = 1.25 * 50%).** Si la precisión aleatoria es del **30%,** la precisión clasificatoria debe ser del **37.5%** Este criterio proporciona solamente una estimación burda del nivel aceptable de capacidad predictiva. El criterio es fácil de aplicar con grupos de igual tamaño. Con grupos de tamaño diferente, se alcanza una cota superior cuando se utiliza el modelo de aleatoriedad máxima para determinar la precisión aleatoria. Esto no representa un gran

problema sin embargo, ya que en la mayo ría de las circunstancias el modelo de aleatoriedad máxima no se utilizaría con grupos de distinto tamaño.

Medidas de precisión clasificatoria fundamentadas estadísticamente relacionadas con la aleatoriedad. Un contraste estadístico para contrastarla capacidad discriminatoria de una matriz de clasificación cuando se compara con un modelo de aleatoriedad es el estadístico Q de Press. Esta medida sencilla compara el número de clasificaciones correctas con el tamaño muestral total y el número de grupos. Se compara el valor crítico, la matriz de clasificación puede considerarse estadísticamente mejor que la aleatoriedad. El estadístico Q se calcula mediante la siguiente fórmula:

Q de press= $(N - nK)^2/N\ (K-1)$

Donde:

N= tamaño muestra total

n= número de observaciones correctamente clasificadas

K= número de grupos

Por ejemplo, en la **Figura 6.11,** el **estadístico Q** se calcularía con una **muestra total** de **N= 50, n= 42 observaciones clasificadas correctamente y K= 2 grupos.**

El estadístico calculado sería

Q de press= $(50 - 40 * 2)^2/\ (50(2-1))$

Q de Press= 23.12

El valor crítico a un nivel de significación de 0.01 es de 6.63. Por ello concluiríamos que en el ejemplo, **las predicciones fueron significativamente mejores que las obtenidas aleatoriamente, lo cual darían una tasa correcta de clasificación del 50%.**

El contraste es sensible al tamaño muestral, de tal forma que con muestras grandes es más probable que resulte significativo que con tamaño muestral se incrementa hasta **100** en el ejemplo y la **tasa de clasificación permanece en el 84%, el estadístico Q se incrementa a 46.24.** Sin embargo, se debe tener cuidado en derivar conclusiones basadas solamente en este estadístico ya que según el tamaño muestra crece, una tasa de clasificación más baja seguiría considerándose significativa.

6.8.7. Diagnosis mediante casos

A fin de evaluar el ajuste del modelo se deberán examinar los resultados predictivos basándonos en un **análisis caso por caso.** Similar al análisis de los residuos en la regresión múltiple, se trata de entender qué observaciones:

1. Han sido mal clasificadas, y
2. No son representativas del resto de los miembros del grupo.

Aunque **la matriz de clasificación** ofrece la exactitud de la clasificación global, **no detalla los resultados para casos individuales**. Además, incluso si pudiéramos saber qué casos están correctamente clasificados y cuáles no, seguiríamos necesitando una medida de similitud de la observación con respecto al resto del grupo.

Clasificación errónea de casos individuales.-Al analizar los residuos en el modelo de regresión múltiple, una decisión importante es la relativa al establecimiento del nivel de los residuos considerado sustantivo y digno de atención. En el análisis discriminante este asunto es algo más sencillo, pues una observación está bien clasificada o mal clasificada. La finalidad de identificar y analizar las observaciones mal clasificadas es identificar

cualesquiera características de estas observaciones que podrían incorporarse en el análisis discriminante para mejorar su capacidad predictiva. Este análisis puede adoptar la forma de perfilar los casos mal clasificados, bien sobre las variables independientes o sobre otras variables no incluidas en el modelo. Examinando estos casos sobre las variables independientes se pueden identificar tendencias no lineales u otras relaciones o atributos que conducen a la clasificación errónea. Inspeccionar otras variables por sus diferencias entre estos casos sería el primer paso para su inclusión posible en el análisis discriminante. Aunque no existe un análisis preestablecido, como el encontrado en la regresión múltiple, se recomienda al investigador que evalúe estos casos mal clasificados desde di versas perspectivas en un intento por descubrir los rasgos únicos con que cuentan en comparación con otros miembros del grupo.
Usted puede realizar alguna valoracio n sobre la similitud de una observacio n con los otros miembros del grupo elevando la distancia $\boldsymbol{D^2}$ **de** ***Mahalanobis*** de la observacio n a la **centroide** del grupo. Las observaciones cercanas al **centroide** se suponen representativas del grupo en una mayor medida. En el ana lisis gra fico de las observaciones, usted puede identificar **observaciones atípicas** y realizar alguna valoracio n sobre su influencia en los resultados. **Por ejemplo,** en una situacio n con dos grupos, un miembro del **grupo A** puede tener una gran distancia $\boldsymbol{D^2}$ **de** ***Mahalanobis*****,** indicando as que es **poco representativo del grupo.** Sin embargo si esa distancia esta lejos de la **centroide** del **grupo B,** incrementar a realmente la aleatoriedad de una clasificacio n correcta, incluso siendo la **menor representativa del grupo.** Una menor distancia que situ e una observacio n entre **dos centroides** probablemente contar a con una **menor probabilidad de clasificación correcta,** incluso entando ma s cerca de si **centroide** de grupo que en la situacio n anterior. Una **representación gráfica** de las observaciones es otra forma de examinar las caracter sticas de las observaciones, concretamente de **las mal clasificadas.** Una forma habitual de proceder es representar la observacio n basada en sus **puntuaciones *z* discriminantes** y reflejar el **solapamiento entre grupos** y los casos mal clasificados. Si se mantienen dos o ma s funciones, los **puntos de corte óptimos** pueden tambie n representarse dando lugar a lo que se conoce como **mapa territorial** que delimita las regiones correspondientes a cada grupo. Representar los casos individuales junto con los **centroides de grupo**, como se sen alo ma s arriba, muestra no so lo las caracter sticas genera les del grupo representadas por los **centroides**, sino tambie n la variacio n entre los miembros del grupo. Esto es ana logo a las a reas definidas en el ejemplo con tres grupos del comienzo de este cap tulo, en el que las puntuaciones de corte sobre ambas funciones defin an a reas correspondientes a predicciones de clasificacio n para cada grupo.

6.9. Análisis Discriminante Múltiple: Interpretación

Paso 5: interpretación

Si la función discriminante es **estadísticamente significativa y la precisión en la clasificación es aceptable**, deberá realizar adecuadas interpretaciones de los resultados. Dentro de este proceso se examinan las funciones discriminantes para determinar la importancia relativa de cada variable independiente en la discriminación de los grupos. Se han propuesto 3 métodos para determinar la importancia relativa:

1. Las ponderaciones discriminantes estandarizadas
2. Las cargas discriminantes (correlaciones de estructura), y

3. Los %.valores parciales de la ***F*.**

6.9.1. Ponderaciones discriminantes

El enfoque tradicional para interpretar las funciones discriminantes estudia el **signo y la magnitud de la ponderación discriminante estandarizada** (denominado algunas veces **coeficiente discriminante**) asignada a cada variable para calcular las funciones discriminantes. Si se **ignora el signo**, cada ponderación representa la contribución relativa de su variable asociada a esa función. Las **variables independientes** con ponderaciones relativamente grandes contribuyen más a la capacidad discriminante de la función que las variables con ponderaciones más pequeñas. **El signo solamente denota que la variable ofrece una contribución positiva o negativa [**Dillon y Goldstein, 1984**]**. La interpretación de las ponderaciones discriminantes es **análoga a la interpretación de las ponderaciones beta en el análisis de regresión y por ello está sujeta a las mismas críticas. Por ejemplo**, una ponderación pequeña puede indicar:

1. Bien que su correspondiente variable es **irrelevante** para determinar una relación, o
2. Que ha sido apartada de la relación debido a un alto grado de **multicolinealidad**.

Otro problema en la utilización de las ponderaciones discriminantes es que están sujetas a una **considerable inestabilidad.** Estos problemas implican que se tenga **precaución en el uso** de las ponderaciones para interpretar los resultados del análisis discriminante.

6.9.2. Cargas discriminantes

En los últimos años se ha incrementado la utilización de las cargas como fundamento de la interpretación debido a las deficiencias al utilizar ponderaciones. Las **cargas discriminantes,** denominadas también **correlaciones de estructura,** miden la correlación lineal simple entre cada variable independiente y la función discriminante. Las cargas discriminantes reflejan la varianza que las variables independientes comparten con la función discriminante, y pueden ser interpreta das como cargas de los factores para valorar la contribución relativa de cada variable independiente a la función discriminante. (En el **Capítulo 12** se trata más profundamente la interpretación de la **carga del factor**.)Las cargas discriminantes (**al igual que las ponderaciones**) **están sujetas a inestabilidad**. Se considera que **las cargas son relativamente más válidas que las ponderaciones** como medio de interpretación de la capacidad discriminante de las variables independientes debido a su naturaleza de correlación. Usted deberá ser cauto cuando utilice las cargas para interpretar las funciones discriminantes.

6.9.3. Valores parciales de la *F*

Como se mencionó, se pueden aplicar dos enfoques de cálculo (**simultáneo y por etapas**) para derivar las funciones discriminantes. Cuando se selecciona el **método por etapas**, se cuenta con un medio adicional de interpretar la capacidad discriminatoria de las **variables independientes** por medio del uso de los **valores parciales de la *F*.** Este se realiza examinando los **tamaños absolutos** de los valores significativos de la ***F*** y clasificándolos. **Valores de la *F* grandes indican una capacidad discriminante mayor**. En la práctica, las clasificaciones que emplean el enfoque de los valores de la ***F*** son las mismas que la clasificación derivada al utilizar las ponderaciones, pero los valores de la **F** indican además los niveles de significación asociados a cada variable.

6.9.4. Interpretación de dos o más funciones

Cuando existen dos o más funciones discriminantes significativas, los problemas de interpretación, son:

1. ¿Podemos simplificar las ponderaciones o las cargas discriminantes para facilitar la definición de cada función?
2. ¿Cómo representamos la influencia de cada variable entre las funciones? Estos problemas se dan tanto en la medida de los efectos discriminantes totales entre funciones como al valorar el papel de cada variable en perfilar cada función de forma separada. Tratamos estas dos cuestiones en los siguientes apartados introduciendo los conceptos de rotación de las funciones, índice de potencia y vectores de atributos extendidos en las representaciones gráficas.

-Rotación de las funciones discriminantes. Después de haber construido las funciones discriminantes, éstas pueden ser **"*rotadas*"** para **redistribuir la varianza.** (Ver **Capítulo 12**.) Básicamente, la **rotación** mantiene la estructura original y la fiabilidad de la solución discriminante mientras que al mismo tiempo hace que las funciones sean más fáciles de interpretar de forma sustancial. En la mayoría de los casos, se hace uso de la rotación **VARIMAX** como fundamento de la rotación.

Índice de potencia. En las secciones anteriores, se expuso el uso de las **ponderaciones estandarizadas y de las cargas discriminantes** como medidas de la contribución de la variable a la función discriminante. Sin embargo, cuando se obtienen dos o más funciones, una **medida resumen o compuesta** es útil para describir las contribuciones de la variable entre todas las funciones significativas. El **índice de potencia** es una medida relativa entre todas las variables que señala la capacidad discriminante de cada variable **[**Perreault, et al 1979**].** Incluye tanto la contribución de la variable a la función discriminante (su carga discriminante) como la contribución relativa de la función a la solución global (una medida relativa entre los autovalores de las funciones). La composición es simplemente la suma de los índices de potencia individuales entre todas las funciones discriminantes significativas. Sin embargo, la interpretación de la **medida compuesta** está limitada por el hecho de que es útil solamente para describir la posición relativa (como el orden en la clasificación) de cada variable, y el valor absoluto no tiene un verdadero significado. El **índice de potencia se calcula a partir del siguiente proceso en dos etapas:**

Etapa 1: Calcular un valor de potencia para cada función significativa. En el primer paso, la capacidad discriminante de la variable, representada por el valor al cuadrado de la carga discriminante, es ***"ponderada"*** por la contribución relativa de la función discriminante en la solución global. Primero, la medida del autovalor relativa a cada función significativa se calcula simplemente como:

$$\text{Autovalor relativo a la función discriminante i} = \frac{\text{autovalor de la función discriminante i}}{\text{suma de autovalores entre todas las funciones significativas}}$$

El valor de potencia de cada variable en una función discriminante es:

$$\text{Valor de potencia de la variable i} = (\text{carga discriminante ij}^2) * \text{autovalor de la función}$$

para cada función j

Etapa 2: Calcular un **índice de potencia** compuesto entre todas las funciones significativas. Calculados el **valor de potencia para cada función,** el índice de potencia compuesto se calcula como la suma de los valores de potencia de cada función discriminante significativa. El índice de potencia representa ahora el efecto discriminante total de la variable entre todas las funciones discriminantes significativas. Recuérdese que es solamente una medida relativa, y su valor absoluto no tiene un significado sustantivo.

-**Representación gráfica de las cargas discriminantes.** Para reflejar las diferencias de los grupos en las **variables predictoras,** Usted puede representar gráficamente las cargas discriminantes. El enfoque más sencillo es dibujar las verdaderas cargas rotadas y no rotadas en un gráfico. Lo ideal sería dibujar las cargas rotadas. Sin embargo, un enfoque incluso más preciso incluye lo que se denomina extender los vectores. Antes de explicar el proceso de extensión, primero debemos definir un vector en este contexto. Un vector es meramente una línea recta dibujada desde el origen (centro) de un gráfico hasta las coordenadas de una determinada carga de una variable. La longitud de cada vector indica la importancia relativa de cada variable para discriminar entre los grupos. Para extender un vector, el investigador multiplica la carga discriminante (preferiblemente después de la rotación) por su respectivo **valor univariante de la F**. El proceso de representación engloba a todas las variables incluidas en el modelo como significativas. Pero el investigador también puede representar las otras **variables con ratios univariantes de la F** significativos que no lo fueron en la función discriminante. Este proceso muestra la importancia de variables colineales que no están incluidas, al igual que en la solución por etapas.

Empleando este procedimiento, notamos que los vectores señalan a los grupos que tienen las medias más altas sobre el respectivo predictor, y están lejos de los grupos que tienen las puntuaciones medias más bajas. Los **centroides** de grupo son también extendidos en este procedimiento multiplicándolas por el **valor aproximado de la F** asociado a cada función discriminante. Si lascar gas son extendidas, los **centroides** deben ser también extendidas para representarlas exactamente en el mismo gráfico. Los **valores de la F** apropiados para cada función discriminante se obtienen a partir de la siguiente expresión:

Valor de la función i=Autovalor función i (Tamaño muestral utilizado en la estimación – No de grupos)/(No. de grupos-1)

Como ejemplo, suponga que la muestra de **50** observaciones se dividió en tres grupos. El múltiplo de cada autovalor sería **(50- 3)/(3- 1) = 23,5**. Para más detalles sobre este procedimiento, vea **[**Dillon, y Goldstein, 1984**].** Para quienes no desean extender los **centroides** y los vectores de atributos, se cuenta con los "***mapas territoriales***" que ofrecen la mayoría de los programas. No incluye los vectores, pero sí representa los **centroides** y las cotas para cada grupo.

6.9.5. Qué método usar

Se consideran varios métodos para interpretar la naturaleza de las funciones discriminantes, tanto para soluciones únicas como múltiples. ¿Qué métodos deberán emplearse? El **enfoque de las cargas** es algo más válido que el uso de las ponderaciones y debería ser utilizado siempre que fuera posible. El uso **de valores de la F parciales y**

univariantes permite al investigador la utilización de varias medidas y la búsqueda de algo de consistencia en las evaluaciones de las variables. Si se estiman dos o más funciones, el investigador puede emplear varias **herramientas gráficas y el índice de potencia**, el cual ayuda a interpretar la solución multidimensional. **Lo más importante es que Usted emplee todos los métodos disponibles para llegar a la interpretación más precisa.**

6.10. Análisis Discriminante Múltiple: Validación de resultados

Paso 6: Validación

El último paso del análisis discriminante comprende la validación de los resultados discriminantes para asegurar que los resultados tienen validez tanto externa como interna. Dada la propensión del análisis discriminante a aumentar el ratio de aciertos si se evalúa solamente utilizando la muestra de análisis, la validación cruzada es una etapa fundamental. A menudo la validación cruzada se realiza con la muestra original, pero es posible emplear una muestra adicional como la ampliación de la muestra. Además de la validación cruzada, el investigador debe llevar a cabo el diseño de grupos que aseguren que las medias de estos grupos son indicadores válidos del modelo conceptual empleado para seleccionar las variables independientes. Estos dos enfoques se consideran a continuación.

6.10.1. División de la muestra o procedimientos de validación cruzada

Recuérdese que el procedimiento más frecuentemente utilizado para validar la función discriminante es dividir los grupos aleatoriamente en la muestra de análisis y en una ampliación de la muestra. Esto implica tener que construir una función discriminante con la muestra de análisis y después validarla con la ampliación de la muestra. La justificación para dividir la muestra total en dos grupos es que aparecerá un sesgo al alza en la capacidad predictiva de la función discriminante si los individuos incluidos en la construcción de la matriz de clasificación son los mismos que aquellos incluidos para calcular la función; es decir, la precisión clasificatoria será más alta que lo que es válido para la función discriminante, si fuese utilizada para clasificar una muestra separada. Las implicaciones de este sesgo al alza son particularmente importantes cuando el investigador está interesado en la validez externa de los resultados.

Sin embargo, otros investigadores han sugerido que se podría tener más confianza en la validez de la función siguiendo este proceso varias veces **[**Morrison,1967**].** En lugar de dividir el total de la muestra aleatoriamente en muestra de análisis y ampliación de la muestra una sola vez, el investigador dividiría la muestra completa aleatoriamente en muestra de análisis y ampliación de la muestra varias veces, cada vez comprobando la validez de la función a través de la construcción de matrices de clasificación y de un ratio de aciertos. Entonces los diferentes ratios de aciertos serían promediados para obtener una medida única. Se han sugerido métodos más sofisticados basados en la estimación con subconjuntos múltiples de la muestra para validar las funciones discriminantes **[**Crask, y Perreault. 1977**,** Dillon y Goldstein, 1984**]**. Los dos enfoques más ampliamente empleados son el ***método-U*** **y el método** ***jackknife.***

Ambos métodos están basados en el principio "***dejar-uno-fuera***", donde la función discriminante es ajustada con muestras tomadas repetidamente de la muestra original. El uso más extendido de este método ha sido estimar **k- 1** muestras, eliminando una observación cada vez de una muestra de **k** casos. La diferencia principal entre los dos

métodos es que el ***método-U*** se centra en la precisión clasificadora, mientras que el enfoque ***jackknife*** centra su atención en la estabilidad de los coeficientes discriminantes. Ambos métodos son bastante sensibles a tamaños muestrales pequeños. Se suele sugerir el uso de cualquiera de estos dos métodos solamente cuando el tamaño del grupo más pequeño sea al menos tres veces tan grande como el número de variables predictoras, y la mayoría de los investigadores proponen un ratio de cinco a uno **[**Huberty, 1984**].** A pesar de estas limitaciones, ambos métodos proporcionan la estimación más válida y consistente de la tasa de precisión clasificatoria. El uso de los ***métodos-U* y *jackknife*** ha sido limitado porque sólo uno de los paquetes informáticos principales los incluye.

6.10.2. Perfilar las diferencias entre los grupos

Otra técnica de validación consiste en perfilar los grupos de las variables independientes para asegurar su correspondencia con los fundamentos conceptuales empleados en la formulación del modelo original. Cuando el investigador ha identificado las variables independientes que contribuyen de forma más importante a la discriminación entre los grupos, el próximo paso es perfilar las características de los grupos atendiendo a las medias de los grupos. Este perfil le permite a Usted comprender el carácter de cada grupo de acuerdo con las variables predictoras. Por ejemplo, refiriéndonos al conjunto de datos de **MKT Digital** presentado en la **Figura 6.2**, vemos que la tasa media de **"*duración*"** para el grupo **"*compraría*"** es de **7.4**, mientras que la tasa media comparable de "duración" para el grupo **"*no compraría*"** es de **3.2**. Por ello un perfil de estos dos grupos muestra que el grupo "compraría" clasifica la duración percibida del nuevo producto de forma sustancialmente más alta que el grupo **"*no compraría*"**. Otro enfoque es perfilar los grupos sobre un conjunto separado de variables que deben reflejar las diferencias de los grupos observados. Este perfil separador proporciona una valoración de validez externa en donde los grupos varían tanto en variable(s) independiente(s) como en el con junto de variables asociadas. Esta es una característica parecida a la **validación de los *clúster*** estimados descrita en el **Capítulo 9**.

6.10.3. Análisis Discriminante Múltiple: Resumen para aplicar

- El análisis discriminante es una técnica multivariable **de dependencia**, que permite encontrar **funciones capaces de separar dos o más grupos de individuos**, tomando como base un conjunto de medidas sobre los mismos representadas por una serie de variables.
- Dichas funciones (**combinaciones lineales de variables independientes**), discriminan o identifican los grupos, definidos por una variable dependiente.
- Por lo tanto, el análisis discriminante puede **ser considerada como una técnica de reducción de datos,** ya que ofrece, al desarrollar un pequeño número de funciones discriminantes (**nuevos ejes**), **una nueva visión de los factores que contribuyen a las diferencias entre los grupos.**
- En el análisis discriminante **cada individuo puede pertenecer a un solo grupo,** ya que la pertenencia a uno u otro grupo se introduce en el análisis mediante **una variable categórica** que toma tantos valores como grupos existentes.

En el análisis discriminante juega el papel de **variable dependiente** las

variables que se utilizan para realizar la clasificación de los individuos, las cuales también se denominan variables **clasificadoras.**

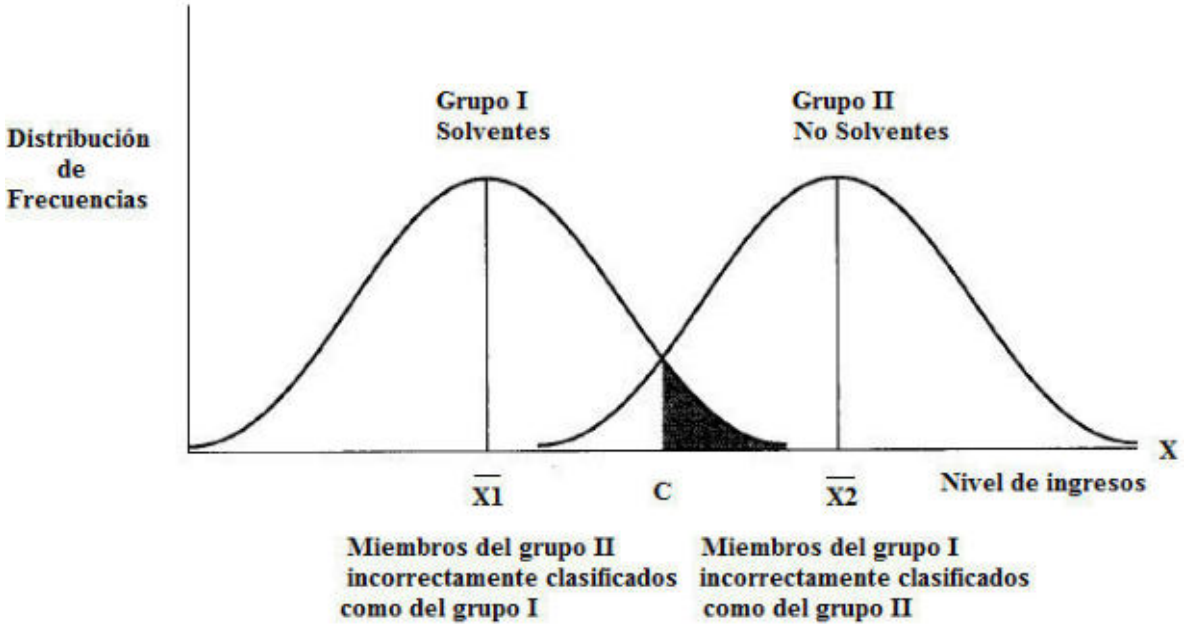
- También se emplean las denominaciones de **variables criterio o variables predictoras,** o la denominación genérica de variables **explicativas.**
- En el análisis discriminante, la información de las **variables clasificadoras** se sintetiza en unas funciones, denominadas **funciones discriminantes**, que son las que finalmente se utilizan en el proceso de clasificación. Asimismo, el análisis discriminante se aplica para fines **explicativos y predictivos.**
- En la utilización **explicativa** se trata de **determinar la contribución de cada variable clasificadora a la clasificación correcta** de cada uno de los individuos. En una **aplicación predictiva**, se trata de determinar el **grupo al que pertenece un individuo** para el que se conocen los valores que toman las **variables clasificadoras**.
- Es una **técnica estadística similar a la regresión lineal**, y es la más apropiada cuando la **variable dependiente es categórica (nominal) y las variables independientes son métricas**. El análisis discriminante permite **determinar cuáles son las variables,** de entre la serie de variables seleccionadas previamente por el investigador, **que mejor explican la pertenencia de un individuo a un grupo determinado.**
- **Fines explicativos:**

  Determinar **porqué una población está de hecho clasificada en diversos grupos** (clientes que solicitaron un préstamo en solventes e insolventes), atendiendo a un conjunto de variables explicativas.
- **Fines predictivos:**

  Estimar la probabilidad con que un individuo de esa población que no esté clasificado, pertenecerá a algunos de los grupos en que se divide la misma (un cliente que solicita un nuevo préstamo, con qué probabilidad se le considerará solvente o insolvente).
- **La muestra total deberá tener una relación óptima de 20 a 1, y como mínima una relación de 5 a 1.**
- El análisis discriminante se utiliza para **clasificar a distintos individuos en grupos o poblaciones alternativos** a partir de los valores de un conjunto de variables sobre los individuos a los que se pretende clasificar.
- Por ejemplo, el director de una sucursal bancaria **necesita establecer algún criterio para conceder o no los préstamos que le son solicitados**. Su misión es detectar si el solicitante pertenecerá en el futuro al grupo de los que **devuelven los préstamo**s o si, por el contrario, será de aquellos **que no lo hacen.**
- Ese director tiene el historial de todos aquellos individuos que, en el pasado, solicitaron préstamos. En ese historial figura**, si finalmente el préstamo fue devuelto o no**, es decir, **el director tiene clasificados a los individuos en solventes e insolventes.**
- Lo que se plantea ahora es si se puede obtener algún tipo de función que le permita, ante una nueva solicitud**, predecir a cuál de los dos grupos va a pertenecer el solicitante.**
- El aplicar intuición geométrica del análisis discriminante, nos servirá, además, para introducir algunos conceptos necesarios.
- Suponga que tiene una población que puede dividirse en dos grupos.

- En el caso del director de banco: clientes solventes e insolventes. Suponga, también, que requiere ser capaz de **explicar esa clasificación** atendiendo a una única variable, por ejemplo, **el nivel de ingresos del cliente.**
- Como el director del banco tiene el historial de los créditos pasados que concedió, sabe qué nivel de ingresos tenían los solventes y los insolventes. De esta información podría obtenerse fácilmente la **Figura 6.13.**

Figura 6.13. Visión geométrica del análisis discriminante



Fuente: propia

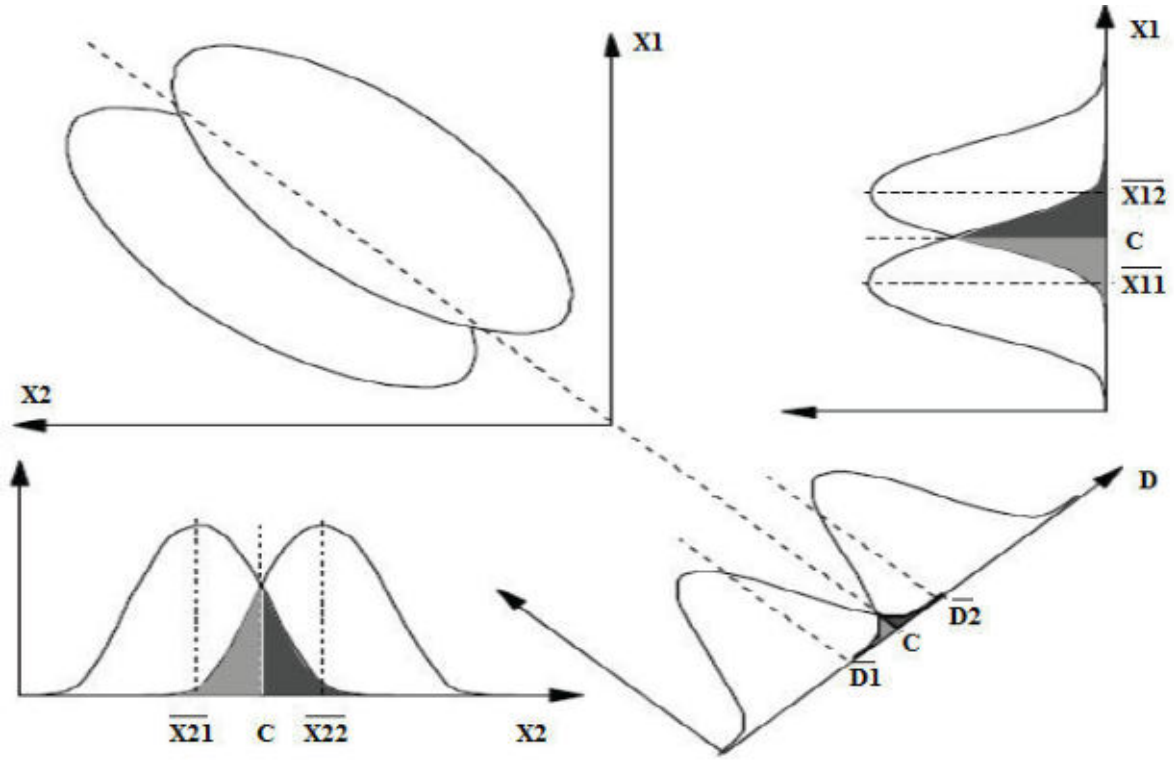
- Un criterio que podría adoptar el director de banco para **conceder o no un préstamo**, podría ser calcular la **media de ingresos de los dos grupos**. La media de ambas medias **(C)** sería un buen punto de corte como se ilustra en la anterior **Figura 6.12**. Si el nuevo solicitante tiene unos ingresos **(X)** superiores a **C**, se le concede el préstamo y si los tiene inferiores no se le concede:

$$C = (X1+X2)/2$$

- Es decir, si **X > C** al individuo se le clasifica en el grupo de los solventes y si **X< C** en el de los probables insolventes. Este criterio, como también se observa en la anterior figura, no es infalible, dado que en la base de datos del director del banco hay clientes con unos ingresos inferiores a **C** que sí devolvieron sus créditos y, por el contrario, hay clientes que tenían ingresos superiores a esa cantidad y que acabaron siendo insolventes.
- La misión del análisis discriminante es **obtener un criterio de clasificación que reduzca ese error**. Es decir, encontrar **una función discriminante** que **separe lo mejor posible las dos poblaciones**. La **Figura 6.13**, ilustra el caso anterior cuando utilizamos **no una variable explicativa (los ingresos), sino dos, por ejemplo, los ingresos y la edad del solicitante.**

- Un criterio que podría adoptar el director de banco para **conceder o no un préstamo**, podría ser calcular la **media de ingresos de los dos grupos**. La media de ambas medias **(C)** sería un buen punto de corte como se ilustra en la anterior **Figura 6.12**. Si el nuevo solicitante tiene unos ingresos **(X)** superiores a **C**, se le concede el préstamo y si los tiene inferiores no se le concede:
- Es decir, si **X > C** al individuo se le clasifica en el grupo de los solventes y si **X< C** en el de los probables insolventes. Este criterio, como también se observa en la anterior figura, no es infalible, dado que en la base de datos del director del banco hay clientes con unos ingresos inferiores a **C** que sí devolvieron sus créditos y, por el contrario, hay clientes que tenían ingresos superiores a esa cantidad y que acabaron siendo insolventes.
- La misión del análisis discriminante es **obtener un criterio de clasificación que reduzca ese error**. Es decir, encontrar **una función discriminante** que **separe lo mejor posible las dos poblaciones**. La **Figura 6.14**, ilustra el caso anterior cuando utilizamos **no una variable explicativa (los ingresos), sino dos, por ejemplo, los ingresos y la edad del solicitante.**

Figura 6.14. Visión geométrica del análisis discriminante



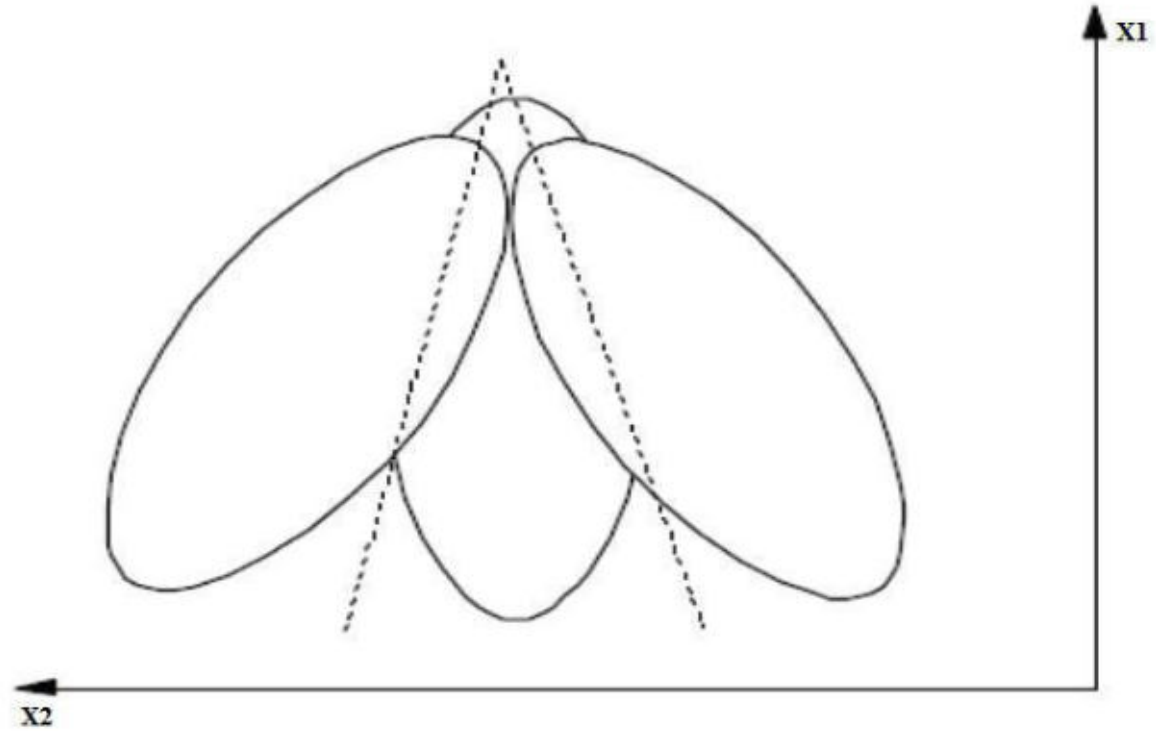
Fuente: propia

- En la figura 6.13 se intenta ilustrar cómo, si en lugar de utilizar para clasificar una de las dos **variables X_1 y X_2** por separado, se utiliza una **combinación de ambas (D), el área que recoge el error es mucho menor.** En síntesis, el análisis discriminante pretende encontrar aquella función discriminante:

$$D = u_1 X_1 + u_2 X_2 + \ldots u_k X_k$$

- Que menor error de clasificación produzca, donde $X_1 \ldots X_k$ son las **k variables explicativas** y $u_1 \ldots u_k$ son **coeficientes de ponderación.**
- Cuando a los individuos se les clasifica en dos grupos, bastará con una **función discriminante (D),** pero si se les quiere clasificar en **tres grupos, harán falta dos funciones discriminantes. Ver Figura 6.145.**
- En general serán necesarias **G-1** funciones discriminantes donde **G** es el número de grupos en que se divide la población.
- **Para conocer de los pasos vea las Figuras 6.8 y 6.9.**

Figura 6.15. Visión geométrica del análisis discriminante



Fuente: propia

6.11. Análisis Discriminante Múltiple: Ejemplos

Paso 1: Objetivos

-Problema 1: la empresa **MKT Digital** con su base de datos **BM_MKT_Digital.sav**, requiere saber la percepción de sus clientes de diseño en software 8empresarial y de juegos) de acuerdo a la variable **X_6 Calidad de servicio y X_4 país** cómo se comportan.

Paso 2: Diseño

Análisis previo de datos. Anteriormente ya se ha indicado que es necesario dar una serie de pasos previos antes de aplicar una técnica multivariable determinada. Algunos de ellos tienen que ver con la propia técnica y la comprobación del cumplimiento de sus hipótesis subyacentes: **normalidad, homoscedasticidad y linealidad;** otras comprobaciones son, incluso, previas al uso de la técnica y tienen que ver con la fiabilidad de los datos de partida: existencia de valores perdidos y de observaciones anómalas. Asimismo, debe señalarse que

algunas de las técnicas de análisis que se expondrán en apartados posteriores, tienen sus propios procedimientos para la comprobación del cumplimiento de sus hipótesis o, por ejemplo, la detección de las observaciones anómalas, y así serán presentadas en su momento (por ejemplo, **en la regresión lineal múltiple**).

Base de Datos

- Para explicar de manera más sencilla y clara las diversas técnicas multivariables, se utilizará la base de datos de **BM_MKT_Digital.sav,** la cual es una empresa manufacturera de diversos productos de papel. Para la obtención de la información se aplicó una encuesta a los Gerentes de Compras de las empresas clientes de **BM_MKT_Digital.sav**
- La base de datos contiene **200 observaciones** con un total de **23 variables**, de las cuales las primeras **18 variables** se refieren a la segmentación del mercado. **BM_MKT_Digital.sav** vende productos de papel a dos segmentos básicos: **la industria del sw empresarial y a la industria del sw de juegos**
- Asimismo, se recolectó información de tres tipos, uno acerca de la percepción del desempeño de **BM_MKT_Digital.sav** medido mediante **13 atributos**, desarrollados a través de focus groups y entrevistas a profundidad aplicados a los principales clientes de la industria software.
- Los Gerentes de Compras de las empresas clientes registrados en **BM_MKT_Digital.sav** contestaron los**13 atributos** utilizando una escala de 0 – 10, siendo **10 "*Excelente*" y 0 "*Pésimo*" con decimales**.
- El segundo tipo de información está relacionado con los resultados de las compras y la relación entre los negocios (satisfacción y posibilidad de recomendación de **BM_MKT_Digital.sav**).
- El tercer tipo de información es la información disponible del almacén de **BM_MKT_Digital.sav** el cual incluye información como tamaño de los clientes y la cantidad de veces que compran sus clientes.
- Con estos datos, **BM_MKT_Digital.sav** puede desarrollar un mejor entendimiento tanto de las características como de las percepciones que tienen sus clientes, y generar acciones que mejoren la empresa. **Ver Figura 6.16**

Ver Figura 6.16. Estructura de la base de datos BM_MKT_Digital.sav

	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	id	Numérico	3	0	ID	Ninguna	Ninguna	4	Centrado	Nominal	Entrada
2	X1	Numérico	2	0	X1 - Antiguedad del cons...	{1, < a 1 año}...	Ninguna	11	Centrado	Nominal	Entrada
3	X2	Numérico	2	0	X2 - Tipo de industria	{0, Software empresarial}...	Ninguna	15	Centrado	Nominal	Entrada
4	X3	Numérico	2	0	X3 - Tamaño de la empre...	{0, PyME (0 to 499)}...	Ninguna	17	Centrado	Nominal	Entrada
5	X4	Numérico	2	0	X4 - País	{0, MEX/Norteamérica}...	Ninguna	18	Centrado	Nominal	Entrada
6	X5	Numérico	2	0	X5 - Sistema de distribuc...	{0, Indirecto a través de tercer...	Ninguna	4	Centrado	Nominal	Entrada
7	X6	Numérico	5	1	X6 - Calidad del servicio	{.0, Mala}...	Ninguna	4	Centrado	Escala	Entrada
8	X7	Numérico	5	1	X7 - Comercio electrónic...	{.0, Mala}...	Ninguna	4	Centrado	Escala	Entrada
9	X8	Numérico	5	1	X8 - Soporte técnico	{.0, Mala}...	Ninguna	4	Centrado	Escala	Entrada
10	X9	Numérico	5	1	X9 - Respueta a quejas	{.0, Mala}...	Ninguna	4	Centrado	Escala	Entrada
11	X10	Numérico	5	1	X10 - Publicidad	{.0, Mala}...	Ninguna	4	Centrado	Escala	Entrada
12	X11	Numérico	5	1	X11 - Línea de servicios	{.0, Mala}...	Ninguna	4	Centrado	Escala	Entrada
13	X12	Numérico	5	1	X12 - Imagen de la fuerz...	{.0, Mala}...	Ninguna	4	Centrado	Escala	Entrada
14	X13	Numérico	5	1	X13 - Precio competitivo	{.0, Mala}...	Ninguna	4	Centrado	Escala	Entrada
15	X14	Numérico	5	1	X14 - Garantías	{.0, Mala}...	Ninguna	4	Centrado	Escala	Entrada
16	X15	Numérico	5	1	X15 - Nuevos productos ...	{.0, Mala}...	Ninguna	4	Centrado	Escala	Entrada
17	X16	Numérico	5	1	X16 - Ordenes y facturac...	{.0, Mala}...	Ninguna	4	Centrado	Escala	Entrada
18	X17	Numérico	5	1	X17 - Flexibilidad de prec...	{.0, Mala}...	Ninguna	4	Centrado	Escala	Entrada
19	X18	Numérico	5	1	X18 - Velocidad de entrega	{.0, Mala}...	Ninguna	4	Centrado	Escala	Entrada
20	X19	Numérico	5	1	X19 - Satisfacción	{.0, Not At All Satisfied}...	Ninguna	4	Centrado	Escala	Entrada
21	X20	Numérico	5	1	X20 - Prbabilidad de reco...	{.0, Definitivamente NO}...	Ninguna	4	Centrado	Escala	Entrada
22	X21	Numérico	5	1	X21 - Probabilidad de co...	{.0, Deinitivamente NO}...	Ninguna	4	Centrado	Escala	Entrada
23	X22	Numérico	5	1	X23 - Nievl de compra	{.0, Cero por ciento}...	Ninguna	4	Centrado	Escala	Entrada
24	X23	Numérico	5	0	X24 - Consideració de ali...	{0, NO esta considerado}...	Ninguna	4	Centrado	Nominal	Entrada

Fuente: SPSS 20 IBM

Paso 3: Condiciones de aplicabilidad (supuestos)

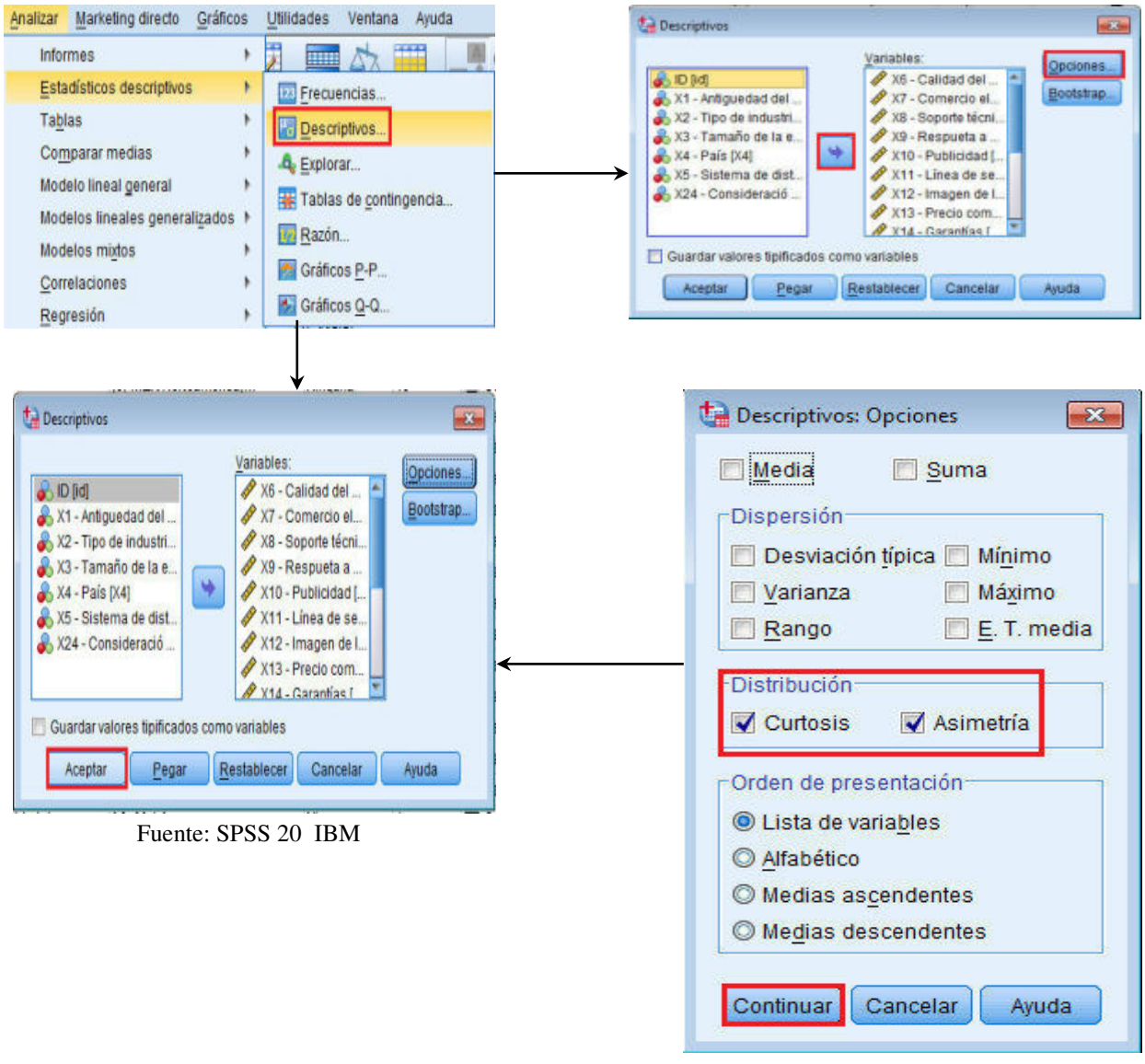
- **Valores Perdidos**: La existencia de valores perdidos en un estudio de las ciencias de la administración es algo prácticamente inevitable. Las consecuencias para la investigación dependerá del patrón que sigan estos datos ausentes, cuántos son y por qué están perdidos.
- **El patrón de los valores perdidos** es más importante que su cuantía; pues si su distribución **es aleatoria en la matriz de datos no pueden causar mucho daño al análisis**; sin embargo, si responden a un patrón determinado sí.
- **Valores perdidos menores o iguales al 10%** del total de los casos por lo general son ignorados y sustituidos por la media o la moda, según sea el caso, excepto cuando los valores perdidos se concentran en una pregunta determinada o un grupo de preguntas del cuestionario.
- **Valores Atípicos u Outliers**: Son aquellos casos para los que una, dos o múltiples variables de una investigación determinada toman valores extremos que los hace diferir del comportamiento del resto de la muestra, y permiten al investigador sospechan que han sido alterados o generados por mecanismos distintos al resto de los datos.
- ¿Por qué es importante detectar los valores atípicos? Por las consecuencias que generan:
- **1.** Distorsionan los resultados al oscurecer el patrón de comportamiento del resto de casos y obtenerse conclusiones que, sin ellos, serían completamente distintas y,
- **2.** Pueden afectar gravemente a una de las condiciones de aplicabilidad más habituales de la mayor parte de las técnicas multivariables: **la normalidad.**
- **Normalidad:** La condición básica que debe asumirse en el análisis multivariable es la normalidad, y se refiere a que todos los datos de las variables métricas deben de seguir una distribución normal. Si la variación de la distribución normal es demasiado amplia,

- todos los resultados del análisis multivariable serán inválidos, porque la normalidad es un requisito esencial para los estadísticos ***F* y *t*.**
- La prueba de normalidad para una sola variable es fácil de realizar y existen diversos métodos para estimarse, pero la normalidad multivariable es más complicada de realizar y existen relativamente pocos métodos para estimarla.
- **Métodos:** Análisis de la Asimetría y Curtosis, Gráficos q-q de residuos y Estadísticos de prueba de ***Kolmogorov-Smirnov-Lillieforce*** (**KSL**).
- **Solución:** Transformación potencial y logarítmica

Paso 4: Estimación y ajuste

Teclear: Analizar->Descriptivos->Seleccionar variables métricas-> Opciones->Curtosis; Asimetría->Continuar ->Aceptar. Figura 6.17.

Figura 6.17. Curtosis y asimetría variable X_6



Fuente: SPSS 20 IBM

Paso 5: Interpretación

Para comprobar si, por ejemplo, la variable **X_6. Calidad de servicio** es o no simétrica y mesocúrtica (normal), hay que comparar los valores de **ZA** y **ZC**

ZA *= (-0.287 – 0)/0.172=* ***-1.668***

ZC *= (-1.076 – 0)/0.342 =* ***-3.146***

Como criterio general para considerar que la distribución responde a una normal, los indicadores calculados deberán caer en el intervalo **-1.96 a 1.96** para un nivel de significatividad del **5%**, o en el intervalo **-2.58 a 2.58** para el **1%.**

Ver Figura 6.18.

Figura 6.18. Resultados: Curtosis y asimetría variable 6

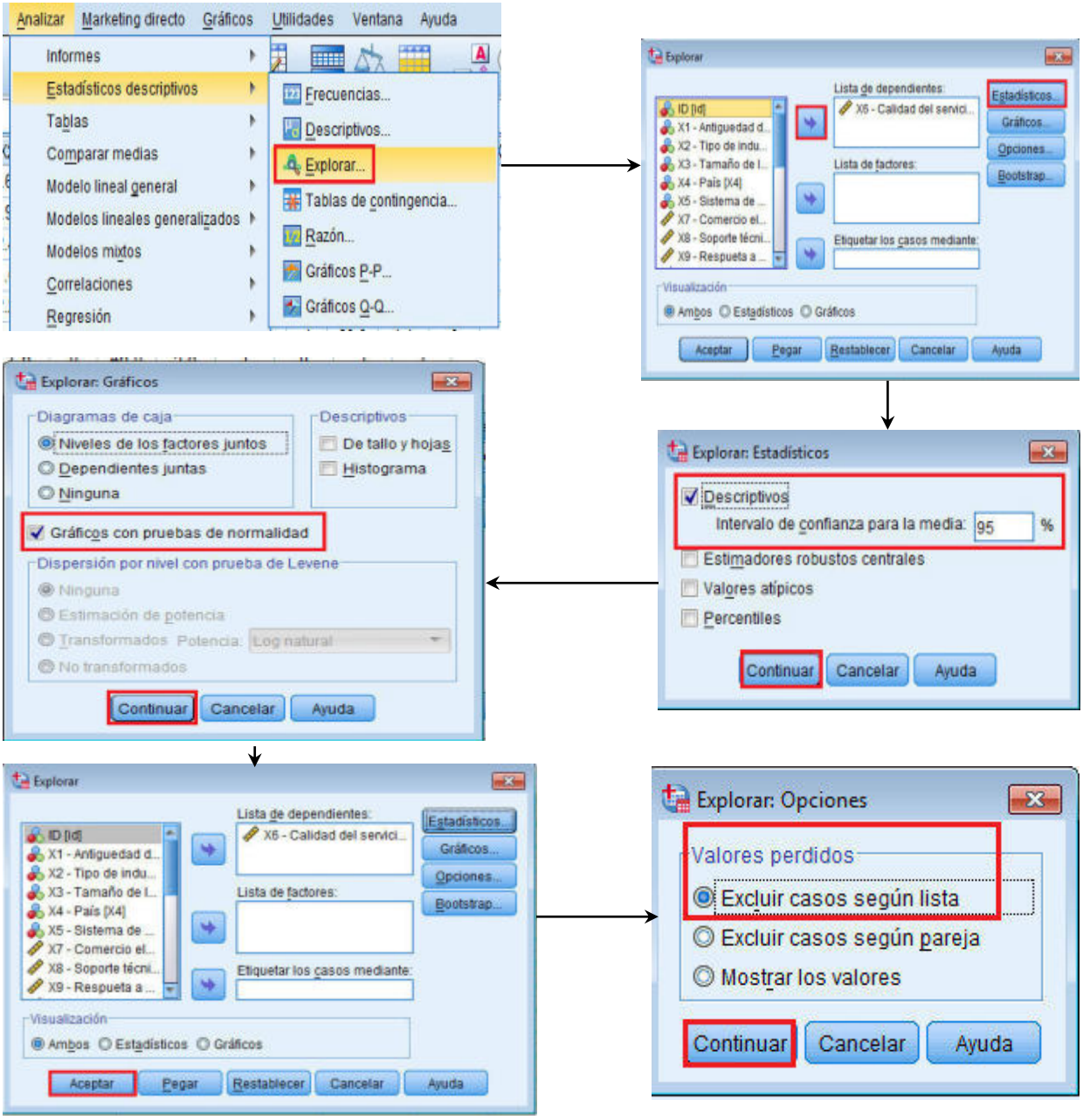
Estadisticos descriptivos

	N	Asimetría		Curtosis	
	Estadístico	Estadístico	Error típico	Estadístico	Error típico
X6 - Calidad del servicio	200	-.287	.172	-1.076	.342
X7 - Comercio electrónico (e-Commerce)	200	.490	.172	.074	.342
X8 - Soporte técnico	200	-.343	.172	.134	.342
X9 - Respueta a quejas	200	-.117	.172	-.064	.342
X10 - Publicidad	200	.068	.172	-.794	.342
X11 - Línea de servicios	200	-.085	.172	-.529	.342
X12 - Imagen de la fuerza de ventas	200	.216	.172	.052	.342
X13 - Precio competitivo	200	-.227	.172	-.859	.342
X14 - Garantías	200	-.119	.172	-.069	.342
X15 - Nuevos productos y servicios	200	.061	.172	-.040	.342
X16 - Ordenes y facturación	200	-.307	.172	.576	.342
X17 - Flexibilidad de precios	200	.466	.172	-.527	.342
X18 - Velocidad de entrega	200	-.370	.172	.107	.342
X19 - Satisfacción	200	.090	.172	-.769	.342
X20 - Prbabilidad de recomendación	200	.070	.172	-.226	.342
X21 - Probabilidad de compra	200	-.206	.172	.584	.342
X23 - Nievl de compra	200	-.062	.172	-.737	.342
N válido (según lista)	200				

Fuente: SPSS 20 IBM

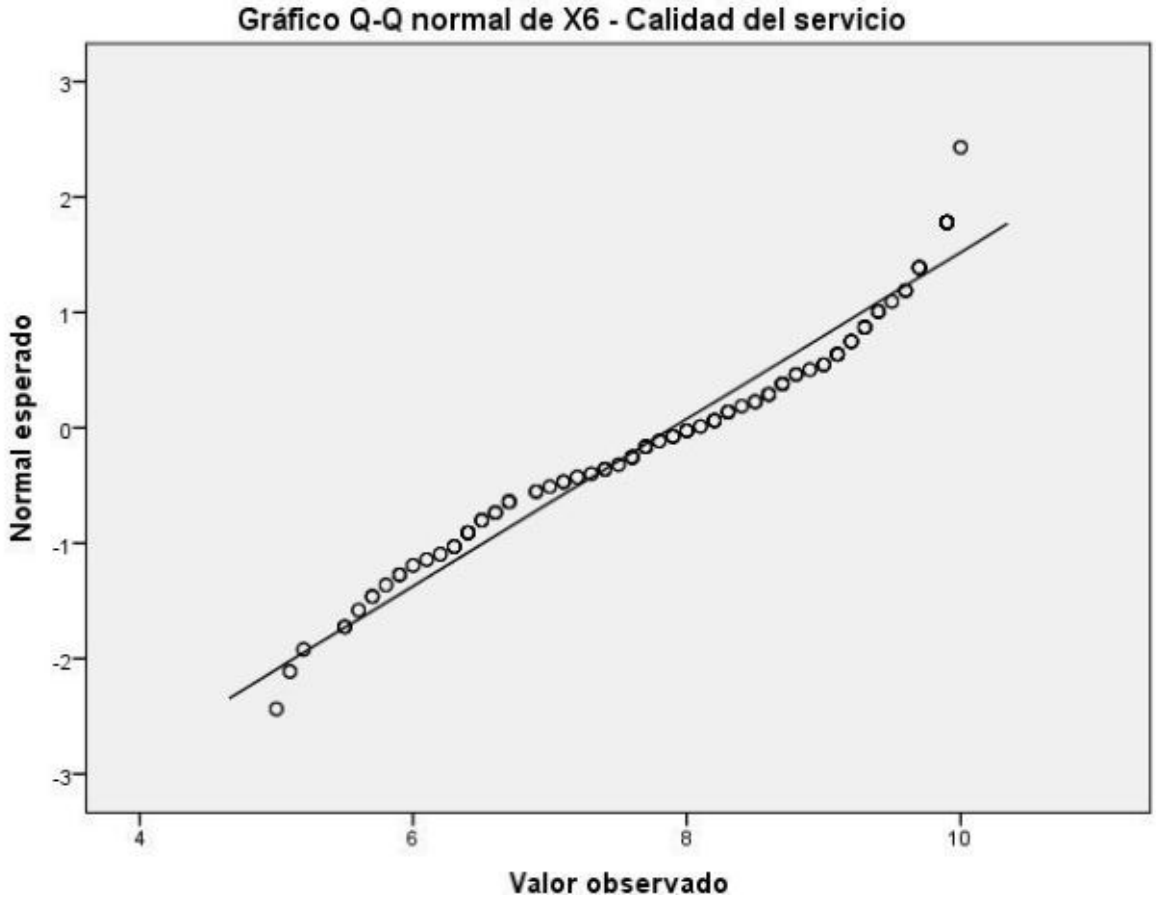
Paso 1: Objetivos; Paso 2: Diseño: Paso 3: Condiciones e aplicabilidad: caso de normalidad,
Paso 4: Estimación y ajuste
-Teclear: Analizar->Estadísticos descriptivos-Explorar->Seleccionar lista de variables dependientes: X_6 Calidad de servicio->Estadísticos->Seleccionar: Descriptivos a intervalo de confianza para la media: 95% ->Continuar->Seleccionar: Gráficos con prueba de normalidad->Continuar->Opciones->Valores perdidos: Excluir casos según lista->Continuar ->Aceptar. Ver **Figuras 6.19 y 6.20.**

Figura 6.19. Proceso para verificar distribución normal de la variable $_6$



Fuente: SPSS 20 IBM

Figura 6.20. Resultados para verificar distribución normal de la variable



Fuente: SPSS 20 IBM

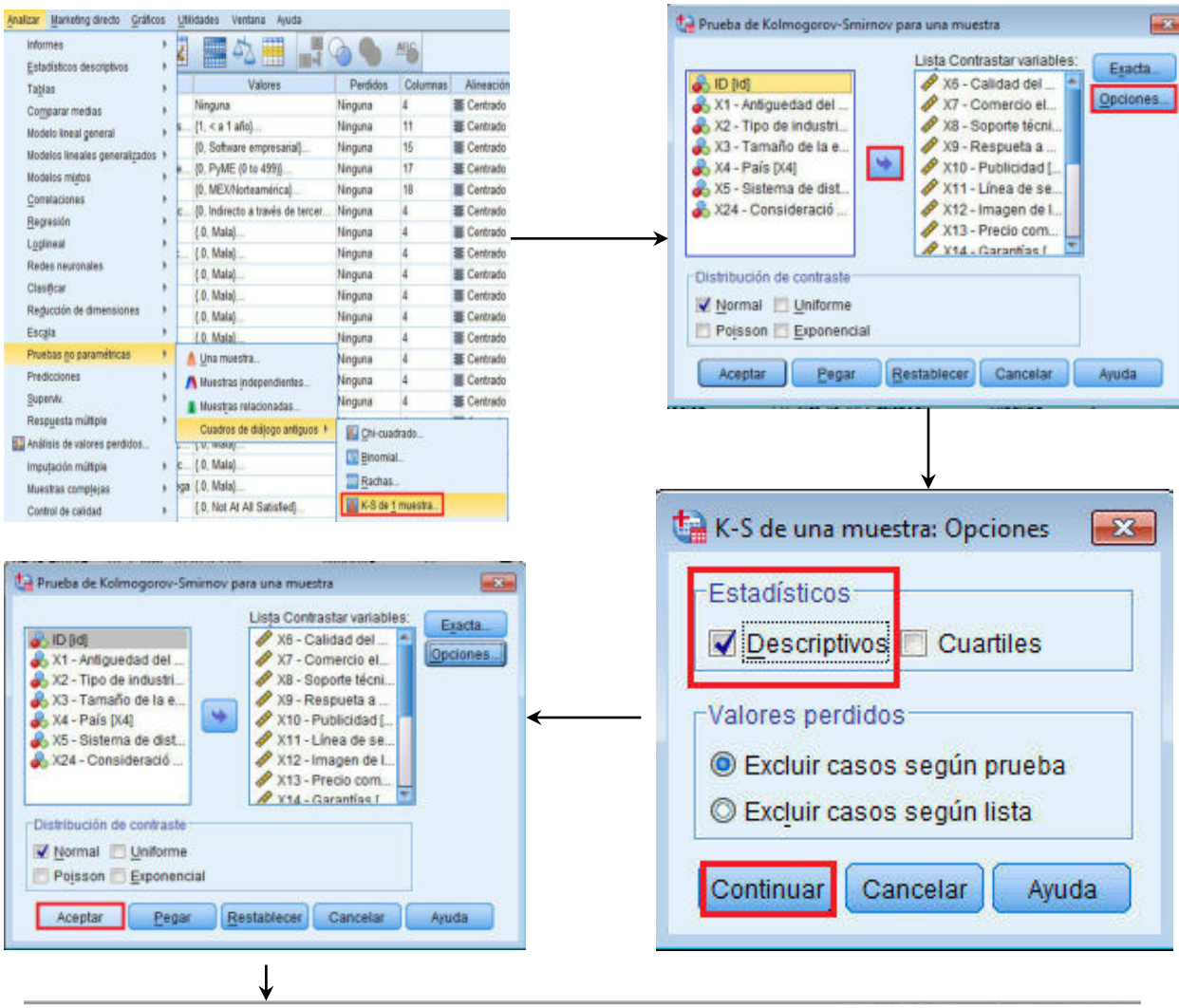
Paso 1: Objetivos; Paso 2: Diseño y Paso 3: Condiciones de aplicabilidad: Normalidad
-Problema 2: De la base de datos **BM_MKT_Digital.sav** donde **N>50 muestras**, pruebe qué Hipótesis es aprobada:

- **Ho.-**Las variables **a $_{23}$** tienen una población con distribución normal
- **H1.-**La variable **a $_{23}$ O** tienen una población con distribución normal

Dado que **N>50** muestras, la normalidad se analizará de acuerdo a ***Kolmogorov-Smirnov.***

Paso 4: Estimación y ajuste
-Teclear: Analizar->Pruebas no paramétricas->Cuadro de diálogo antiguos->K-S de 1 muestra; Seleccionar lista de variables de prueba (X_6 a X_{23}); Seleccionar Distribución de contraste (Normal) ->Opciones->Estadísticos->Descriptivos->Continuar->Aceptar.
Ver **Figura 6.21.**

Figura 6.21. Proceso para verificar distribución normal de la variable [6]



Prueba de Kolmogorov-Smirnov para una muestra

		X6 - Calidad del servicio	X7 - Comercio electrónico (e-Commerce)	X8 - Soporte técnico	X9 - Respueta a quejas	X10 - Publicidad	X11 - Línea de servicios	X12 - Imagen de la fuerza de ventas	X13 - Precio competitivo	X14 - Garant
N		200	200	200	200	200	200	200	200	
Parámetros normales[a,b]	Media	7.894	3.765	5.243	5.368	4.061	5.815	5.248	6.971	
	Desviación típica	1.3830	.7689	1.6552	1.2100	1.1471	1.3174	1.1286	1.5813	
Diferencias más extremas	Absoluta	.095	.122	.046	.045	.078	.063	.107	.091	
	Positiva	.086	.122	.021	.043	.078	.055	.107	.075	
	Negativa	-.095	-.061	-.046	-.045	-.052	-.063	-.104	-.091	
Z de Kolmogorov-Smirnov		1.346	1.723	.657	.631	1.100	.897	1.513	1.294	
Sig. asintót. (bilateral)		.054	.005	.782	.821	.178	.397	.021	.070	

a. La distribución de contraste es la Normal.

b. Se han calculado a partir de los datos.

Fuente: SPSS 20 IBM

- **Paso 1: Objetivos; Paso 2: Diseño y Paso 3: Homoscedasticidad**
- Debe definirse de manera distinta según se estén analizando datos no agrupados (caso de una regresión lineal múltiple), o datos agrupados (caso de un análisis de la varianza de un factor).
- En el primer caso la hipótesis de homoscedasticidad puede definirse como la asunción de que cada uno de los valores que puede tomar la distribución se mantiene constante para todos los valores de la otra variable continua.
- En el caso de datos agrupados la homoscedasticidad implica que la varianza de la variable continua es más o menos la misma en todos los grupos que conforman la variable no métrica que es la que determina los grupos.
- En resumen, se puede decir que la homoscedasticidad es la igualdad de varianza entre las variables independientes.
- *Métodos:* **Test de *Levene***
- *Solución:* **Transformación logarítmica y potencial**

Paso 3: Condiciones de Aplicabilidad

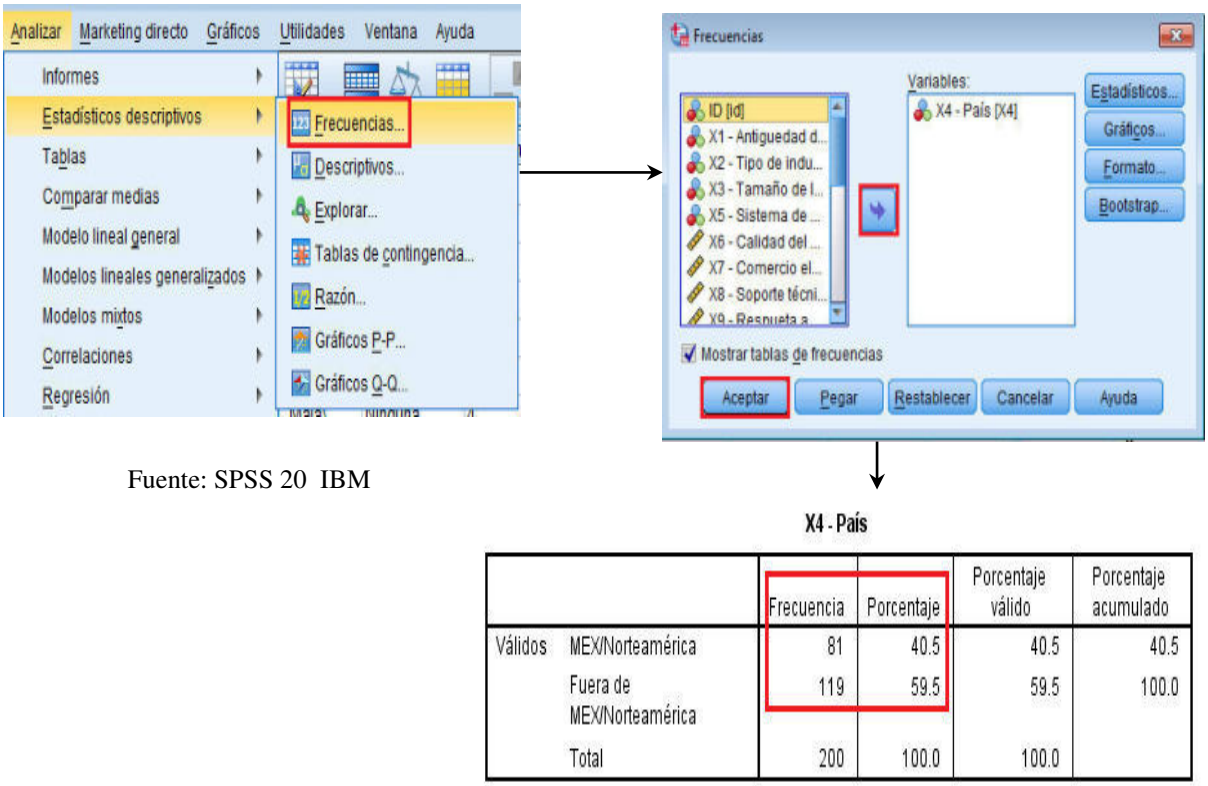
- **Linealidad:** La asunción de linealidad es fundamental para todas aquellas técnicas que se centren en el análisis de las matrices de correlación o de varianzas – covarianzas, como el análisis factorial, regresión lineal o los modelos de ecuaciones estructurales. La razón es sencilla: el **coeficiente de correlación de *Pearson*** sólo podrá captar una relación si ésta es lineal.
- Si la relación existe y es intensa pero, por ejemplo, es curvilínea, el **coeficiente de correlación de *Pearson*** tomará un valor relativamente bajo y el investigador puede interpretarlo como ausencia de relación cuando, de hecho, ésta existe sólo que no es lineal.
- Cuando la técnica empleada tiene una variable dependiente, como ocurre en el caso de la regresión lineal múltiple, existen diversos procedimientos para contrastar la linealidad de las relaciones basadas en el análisis de los residuos o residuales.
- *Métodos:* Gráficos: de dispersión entre variables y Estadísticos: coeficientes de correlación bivariados.

Paso 1: Objetivos: Paso 2. Diseño; Paso 3: Condiciones de aplicabilidad
-Problema 3: Para el caso de la variable X4, tenemos:

- En nuestra base de datos, disponemos de **200 observaciones** y de **13 variables independientes,** lo que da un ratio **de 15 a 1,** no muy alejado de la cifra óptima.
- Además, el tamaño de los **dos grupos 81 y 119** excede por mucho el tamaño mínimo de **20 observaciones por grupo**.
-

Paso 4: Ejecución y ajuste
-Teclear: Analizar->Estadísticos descriptivos->Frecuencias->Selección variable categórica (X4) ->Aceptar. Ver Figura 6.22.

Figura 6.22. Estadística descriptiva caso variable X4.



Fuente: SPSS 20 IBM

X4 - País

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	MEX/Norteamérica	81	40.5	40.5	40.5
	Fuera de MEX/Norteamérica	119	59.5	59.5	100.0
	Total	200	100.0	100.0	

Paso 1. Objetivos; Paso 2: Diseño; Paso 3; Condiciones de aplicabilidad: Normalidad
Paso 4: Ejecución y ajuste
-Teclear: Analizar->Estadísticos descriptivos->Explorar->Selección variable categórica (X4) ->Estadísticos->Descriptivos 95%->Continuar->Gráficos->Gráficos con pruebas de normalidad->Continuar->Opciones->Valores perdidos: Excluir casos según lista->Continuar->Aceptar.
Ver Figura 6.23.

Figura 6.23. Prueba de normalidad de la variable X4.

Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
X4 - País	.390	200	.000	.623	200	.000

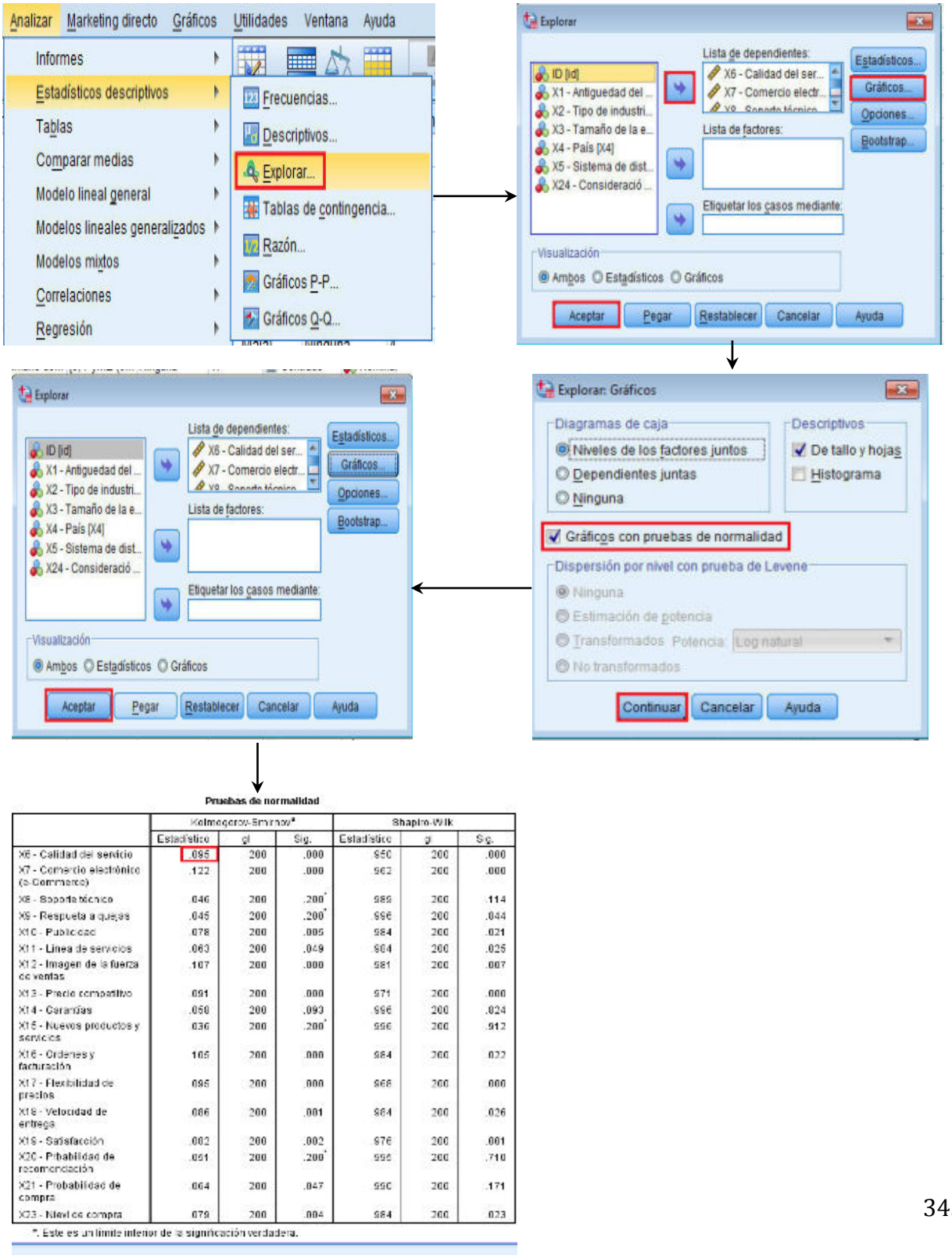
a. Corrección de la significación de Lilliefors

Fuente: SPSS 20 IBM

Paso 4: Ejecución y ajuste

Teclear: Analizar->Estadísticos descriptivos->Explorar->Selección de variables métricas->Gráficos->Seleccionar: Gráficos con pruebas de normalidad->Continuar->Aceptar. Ver Gráfico 6.24.

Figura 6.24. Proceso prueba de normalidad variables métricas.



Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
X6 - Calidad del servicio	.095	200	.000	.950	200	.000
X7 - Comercio electrónico (e-Commerce)	.122	200	.000	.962	200	.000
X8 - Soporte técnico	.046	200	.200*	.989	200	.114
X9 - Respuesta a quejas	.045	200	.200*	.996	200	.044
X10 - Publicidad	.078	200	.005	.984	200	.021
X11 - Línea de servicios	.063	200	.049	.984	200	.025
X12 - Imagen de la fuerza de ventas	.107	200	.000	.981	200	.007
X13 - Precio competitivo	.091	200	.000	.971	200	.000
X14 - Garantías	.059	200	.093	.996	200	.024
X15 - Nuevos productos y servicios	.036	200	.200*	.996	200	.912
X16 - Ordenes y facturación	.105	200	.000	.984	200	.022
X17 - Flexibilidad de precios	.095	200	.000	.968	200	.000
X18 - Velocidad de entrega	.086	200	.001	.984	200	.026
X19 - Satisfacción	.092	200	.002	.976	200	.001
X20 - Probabilidad de recomendación	.051	200	.200*	.995	200	.710
X21 - Probabilidad de compra	.064	200	.047	.990	200	.171
X23 - Nivel de compra	.079	200	.004	.984	200	.023

*. Este es un límite inferior de la significación verdadera.

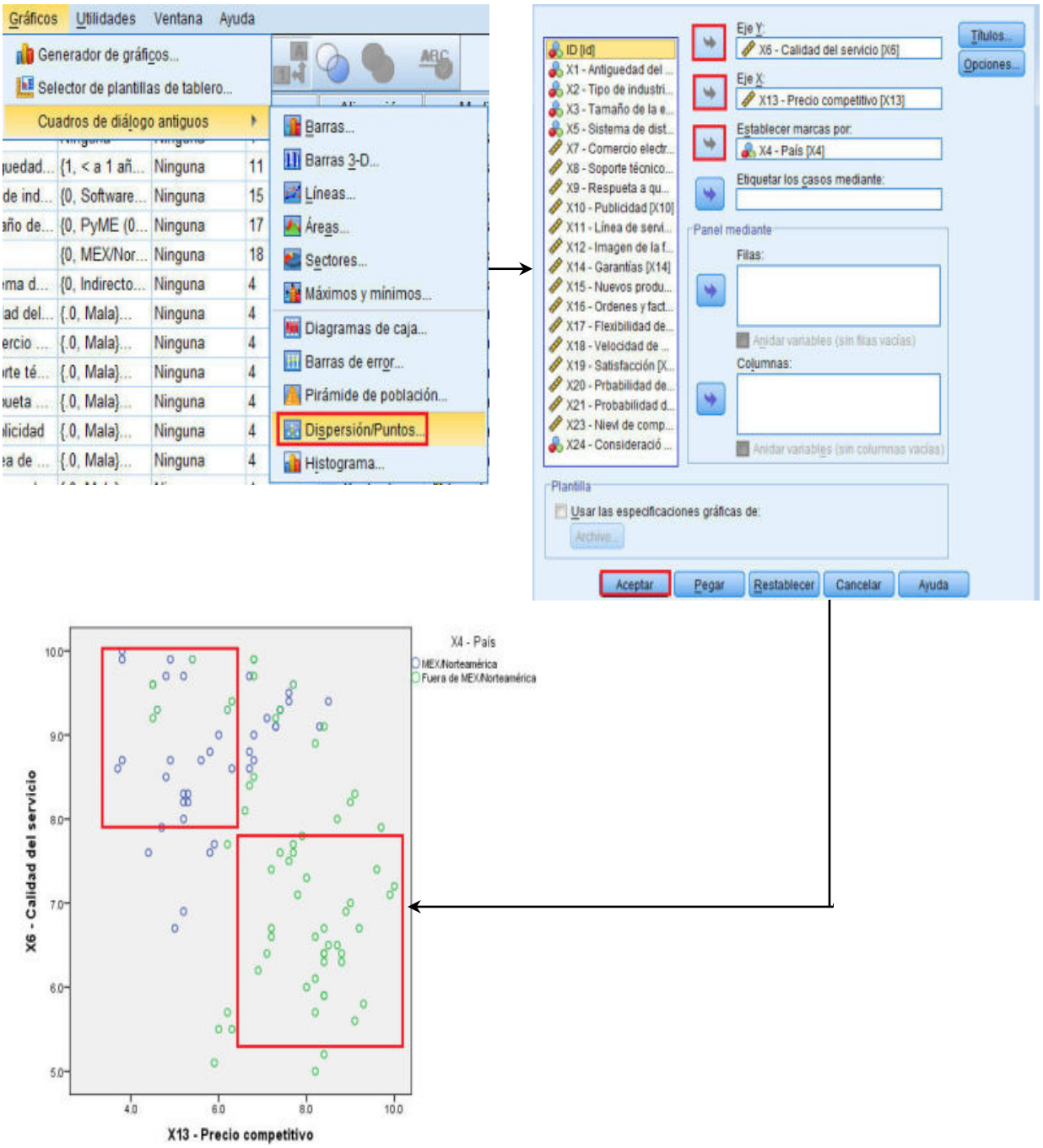
Fuente: SPSS 20 IBM

Paso 1: Objetivos

-Problema 4: gráficamente determine si existen grupos diferenciados. Para resolver, se sugiere hacerlo a través de un gráfico de dispersión de 3 ejes: 2 variables métricas (**X_6.-Calidad/X_6.-Precio**) **vs** 1 nominal (**X_4.-País**).

Teclear: Gráficos->Cuadro de diálogos antiguos->Dispersión/Puntos->Dispersión simple->Definir->Eje X: X_6->Eje: X_{13}->Establecer marcas por: X_4->Aceptar. Ver Figura 6.25.

Figura 6.25. Proceso prueba de normalidad variables métricas.



Fuente: SPSS 20 IBM

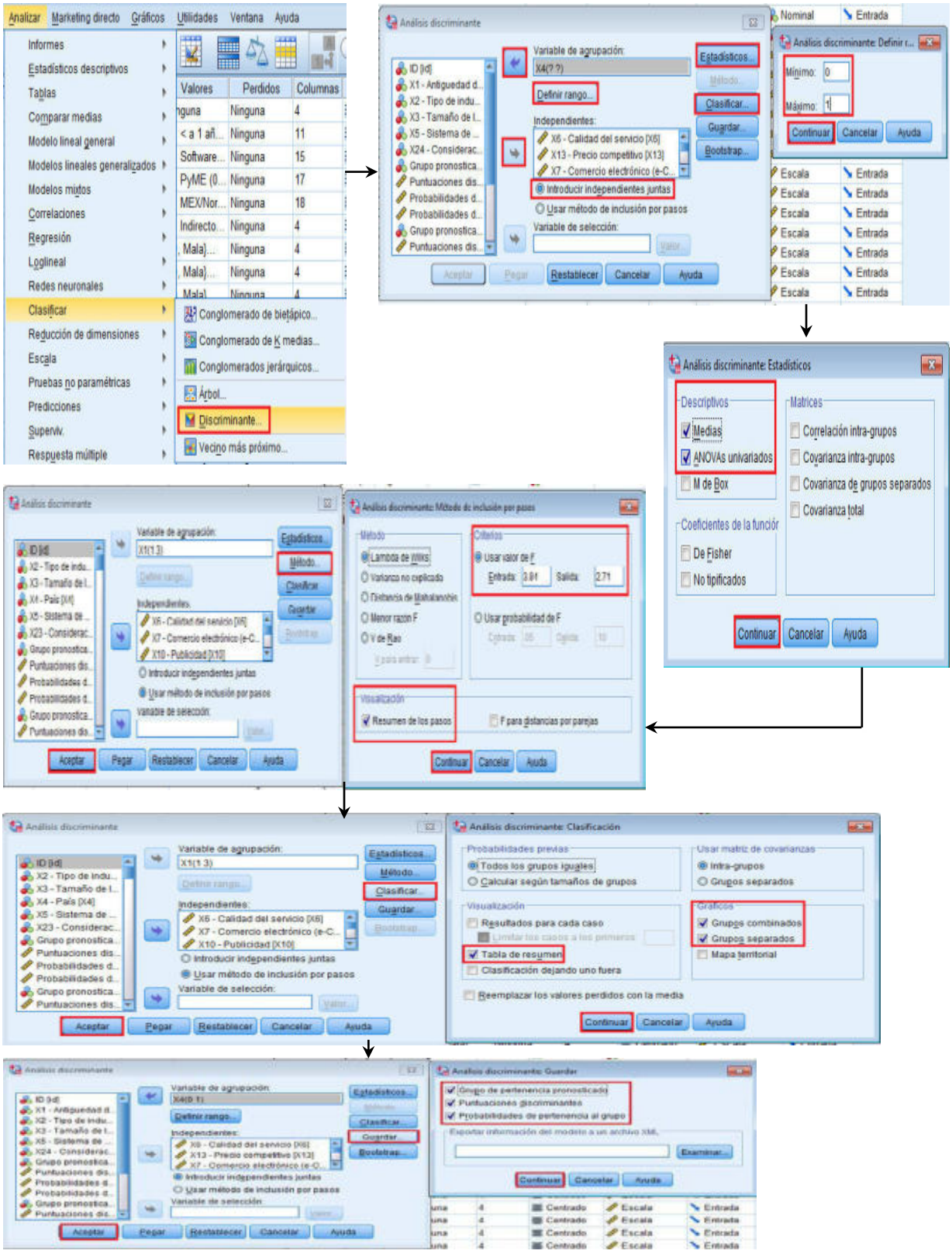
Paso 2; Diseño; Paso 3: Condiciones de aplicabilidad.
Se observan 2 grupos diferenciados, donde los clientes fuera de México/Norteamérica perciben los servicios más por precio competitivo que calidad. Si las variables están muy mezcladas, éstas no discriminan

Paso 4: Estimación y ajuste
Dado que el objetivo esencial del análisis discriminante es identificar el
grupo de variables independientes (percepciones de los clientes de **MKT Digital**) que maximizan las diferencias entre los dos grupos de clientes.

- Asimismo, al igual que ocurría con la regresión lineal, se puede optar por dos procedimientos de estimación y ajuste:

Estimación Simultánea: Introducir todas las variables explicativas o independientes, por lo cual primará la precisión en la clasificación, esto es, no nos importa tanto explicar por qué la función discriminante clasifica como lo hace, sino que clasifique bien.
Estimación Paso a Paso: Sólo entrarán aquellas variables independientes que superen ciertos niveles mínimos de poder explicativo, por lo cual primará la necesidad de explicar la clasificación.
-Teclear: Analizar->Clasificar->Discriminante->Variable de agrupación X_4. nominal->Definir rango->Mínimo: 0; Máximo: 1->Continuar->Independientes: X_6 a X_{23} métricas->Introducir independientes juntas->Estadísticos->Descriptivos: Medias; ANOVAS->Continuar->Método: Lambda de Wilks->Criterios: usar valor de F entrada: 3.84; salida: 2.71-> Visualización: Resumen de pasos->Continuar->Clasificar->Probabilidades previas: Todos los grupos iguales->Gráficos: Grupos combinados; Grupos separados->Visualización: Tabla de resumen-> Continuar->Guardar->Grupo de pertenencia pronosticado; Puntuaciones discriminantes; Probabilidades de pertenencia al grupo ->Continuar-Aceptar. Ver Figura 6.26.

Figura 6.26.- Proceso análisis discriminante variables: X_4 vs. X_6-X_{23}. Estimación simultánea



Fuente: SPSS 20 IBM

Paso 5: Interpretación
SPSS genera diversas tablas para análisis. **Ver Figura 6.27 , 6.28 y 6.29.**

Figura 6.27.- Tabla Estadísticos de grupo. Estimación simultánea

Estadísticos de grupo

X4 - País		Media	Desv. típ.	N válido (según lista)	
				No ponderados	Ponderados
MEX/Norteamérica	X6 - Calidad del servicio	8.804	.7987	81	81.000
	X13 - Precio competitivo	5.844	1.2739	81	81.000
	X7 - Comercio electrónico (e-Commerce)	3.577	.7433	81	81.000
	X8 - Soporte técnico	5.236	1.6070	81	81.000
	X9 - Respueta a quejas	5.373	1.0829	81	81.000
	X10 - Publicidad	3.752	1.0821	81	81.000
	X11 - Línea de servicios	6.633	.9313	81	81.000
	X12 - Imagen de la fuerza de ventas	4.731	1.0187	81	81.000
	X14 - Garantías	6.067	.8624	81	81.000
	X15 - Nuevos productos y servicios	5.305	1.4837	81	81.000
	X16 - Ordenes y facturación	4.237	.7916	81	81.000
	X17 - Flexibilidad de precios	3.635	.6491	81	81.000
	X18 - Velocidad de entrega	3.804	.6254	81	81.000
	X19 - Satisfacción	7.211	.9901	81	81.000
	X20 - Probabilidad de recomendación	7.040	.8729	81	81.000
	X21 - Probabilidad de compra	7.798	.6458	81	81.000
	X23 - Nivel de compra	60.137	8.8536	81	81.000
Fuera de	X6 - Calidad del servicio	7.275	1.3559	119	119.000

Autovalores

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	1.743[a]	100.0	100.0	.797

a. Se han empleado las 1 primeras funciones discriminantes canónicas en el análisis.

Lambda de Wilks

Contraste de las funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	.365	191.222	17	.000

Fuente: SPSS 20 IBM

De la tabla **Estadísticos de grupo,** es posible establecer discusión de cómo las medias se relacionan las variables de cada grupo como análisis descriptivo.
De las **tablas Autovalores y *Lambda de Wilks*,** se afirma que el autovalor es de difícil interpretación por sí sólo, por eso, se prefiere analizar ***Lambda de Wilks***. La correlación canónica es la correlación entre la función discriminante (mientras más alto, mejor) y la variable dependiente (0,1), esto es: **La función discriminante calculada que tan correlacionada está con las categorías de la variable dependiente (X_4). *Lambda de Wilks* mientras más pequeña mejor ya que justifica diferencia entre grupos (discrimina más); la Sig *p*<0.05, por lo que si hay diferencia**

SPSS también genera la tabla **Pruebas de igualdad de las medias delos grupos. Ver Figura 6.28.**

Figura 6.28. Tabla Pruebas de igualdas de las medias de los grupos

Pruebas de igualdad de las medias de los grupos

	Lambda de Wilks	F	gl1	gl2	Sig.
X6 - Calidad del servicio	.704	83.242	1	198	.000
X7 - Comercio electrónico (e-Commerce)	.959	8.485	1	198	.004
X11 - Línea de servicios	.736	71.007	1	198	.000
X12 - Imagen de la fuerza de ventas	.856	33.211	1	198	.000
X13 - Precio competitivo	.653	105.323	1	198	.000
X17 - Flexibilidad de precios	.669	97.904	1	198	.000
X8 - Soporte técnico	1.000	.003	1	198	.960
X9 - Respueta a quejas	1.000	.003	1	198	.959
X10 - Publicidad	.950	10.389	1	198	.001
X14 - Garantías	1.000	.058	1	198	.809
X15 - Nuevos productos y servicios	.997	.541	1	198	.463
X16 - Ordenes y facturación	1.000	.004	1	198	.950
X18 - Velocidad de entrega	1.000	.036	1	198	.849
X19 - Satisfacción	.970	6.085	1	198	.014
X20 - Probabilidad de recomendación	.996	.878	1	198	.350
X21 - Probabilidad de compra	.985	3.027	1	198	.083
X22 - Nivel de compra	.968	6.530	1	198	.011

Fuente: SPSS 20 IBM

En este caso, las variables con **Sig. p>0.05** indican que **NO** intervienen en la diferencia de grupo X_6- $_{23}$ para la variable X_4. Se requiere etirar para realizar sólo el estudio con: $_6$-X_7-X_{10}-X_{11}-X_{12}-X_{13}-X_{17}-X_{19}, por lo que habrá que recargar para evaluar. **Ver Figura 6.29.**

Figura 6.29. Tablas Autovalores y *Lambda de Wilks*. Estimación simultánea

Autovalores

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	1.520[a]	100.0	100.0	.777

a. Se han empleado las 1 primeras funciones discriminantes canónicas en el análisis.

Lambda de Wilks

Contraste de las funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	.397	179.271	8	.000

Fuente: SPSS 20 IBM

Se sigue conservando la congruencia

Paso 5: Interpretación

Coeficientes Estandarizados:

- Contribución relativa de la variable asociada a la función discriminante.
  -A mayores coeficientes, mayor contribución
  -Problemas con la multicolinealidad: Un coeficiente bajo implica o que su contribución no es relevante o que ha sido eliminado por su elevada correlación con las demás variables.
- **Matriz de Estructura:**
  -Las puntuaciones discriminantes miden la correlación simple entre cada variable independiente y la función discriminante.
  -Reflejan la varianza que la variable independiente comparte con la función discriminante.
- ***F* Univariante:**
  -Poder discriminante relativo de cada variable independiente
  -Tiene asociados niveles de significación

SPSS genera diversas tablas para análisis. **Ver Figura 6.30 .**

Figura 6.30.- Tablas de Coeficientes estandarizados de las funciones discriminantes canónicas, Matriz de estructura y Resultados de clasificación . Estimación simultánea.

Coeficientes estandarizados de las funciones discriminantes canónicas

	Función
	1
X6 - Calidad del servicio	-.267
X7 - Comercio electrónico (e-Commerce)	-.409
X10 - Publicidad	-.006
X11 - Línea de servicios	-.411
X12 - Imagen de la fuerza de ventas	.816
X13 - Precio competitivo	.411
X17 - Flexibilidad de precios	.372
X19 - Satisfacción	-.025

Matriz de estructura

	Función
	1
X13 - Precio competitivo	.592
X17 - Flexibilidad de precios	.570
X6 - Calidad del servicio	-.526
X11 - Línea de servicios	-.486
X12 - Imagen de la fuerza de ventas	.332
X10 - Publicidad	.186
X7 - Comercio electrónico (e-Commerce)	.168
X19 - Satisfacción	-.142

Correlaciones intra-grupo combinadas entre las variables discriminantes y las funciones discriminantes canónicas tipificadas
Variables ordenadas por el tamaño de la correlación con la función.

Resultados de la clasificación[a]

			Grupo de pertenencia pronosticado		
		X4 - País	MEX/Norteamérica	Fuera de MEX/Norteamérica	Total
Original	Recuento	MEX/Norteamérica	80	1	81
		Fuera de MEX/Norteamérica	20	99	119
	%	MEX/Norteamérica	98.8	1.2	100.0
		Fuera de MEX/Norteamérica	16.8	83.2	100.0

a. Clasificados correctamente el 89.5% de los casos agrupados originales.

Fuente: SPSS 20 IBM

- Para la **tabla Coeficientes estandarizados de las funciones discriminantes canónicas**, responde al caso a **qué variable es más discriminante o diferenciador de los 2 grupos de X_4.** Se afirma que la variable de mayor importancia, por valor absoluto **es X_{12}**.-Imagen de la fuerza de ventas al momento de definir los grupos, respecto a los valores de **X_4.-**País, aunque el comercio electrónico sea inversamente proporcional **X_7.** Es posible explicar la diferencia de los grupos tomando en cuenta los valores absolutos.

- Para la **Matriz de estructura,** responde al caso a **qué variable independiente (X_6-$_{23}$) tiene mayor correlación con la dependiente X_4.** Son correlaciones entre las variables y la función discriminante estandarizada. En este caso es la variable **X_{13}** de precio competitivo.

- Para la **tabla Resultados de la clasificación** Explica el porcentaje de aciertos en los grupos, el cual es altamente adecuado (**89.5%**) a pesar de las variables excluidas

Paso 6: Validación

- El último paso del análisis discriminante pasa por validar los resultados.
- La mejor forma de hacerlo consiste en reservar parte de la muestra cuando se estima la función discriminante.
- Una vez obtenida esta, se clasifica mediante el procedimiento que acaba de describirse a los individuos que no se utilizaron para estimarla. Si el porcentaje de acierto es similar al de la muestra de estimación, el análisis sería válido.
- Otra forma de validar el modelo del Análisis Discriminante es **dividir la muestra en dos submuestras, y se realiza el mismo procedimiento para ambas submuestras.** Si los resultados don similares en las dos submuestras entonces se valida el modelo.
- **Problema 5:** ¿se puede validar aplicando otro método de ingreso de datos? ... Se puede realizar también con el metodo de inclusión por pasos.

Fuente: SPSS 20 IBM

Paso 5: interpretación

SPSS genera diversas tablas para análisis. **Ver Figura 6.31, Figura 6.32, Figura 6.33 y Figura 6.34 .**

Figura 6.31. Tablas : Estadísiticos de grupo , Autovalores y *Lambda de Wilks.* Estimación simultánea vs. Inclusión por pasos.

Estimación simultánea

Estadísticos de grupo

X4 - País		Media	Desv. típ.	N válido (según lista) No ponderados	N válido (según lista) Ponderados
MEX/Norteamérica	X6 - Calidad del servicio	8.804	.7987	81	81.000
	X13 - Precio competitivo	5.844	1.2739	81	81.000
	X7 - Comercio electrónico (e-Commerce)	3.577	.7433	81	81.000
	X8 - Soporte técnico	5.236	1.6070	81	81.000
	X9 - Respuesta a quejas	5.373	1.0829	81	81.000
	X10 - Publicidad	3.752	1.0821	81	81.000
	X11 - Línea de servicios	6.633	.9313	81	81.000
	X12 - Imagen de la fuerza de ventas	4.731	1.0187	81	81.000
	X14 - Garantías	6.067	.8624	81	81.000
	X15 - Nuevos productos y servicios	5.305	1.4837	81	81.000
	X16 - Ordenes y facturación	4.237	.7916	81	81.000
	X17 - Flexibilidad de precios	3.635	.6491	81	81.000
	X18 - Velocidad de entrega	3.804	.6254	81	81.000
	X19 - Satisfacción	7.211	.9901	81	81.000
	X20 - Probabilidad de recomendación	7.040	.8729	81	81.000
	X21 - Probabilidad de compra	7.798	.6458	81	81.000
	X23 - Nivel de compra	60.137	8.8536	81	81.000
Fuera de	X6 - Calidad del servicio	7.275	1.3559	119	119.000

Inclusión por pasos

Estadísticos de grupo

X4 - País		Media	Desv. típ.	N válido (según lista) No ponderados	N válido (según lista) Ponderados
MEX/Norteamérica	X6 - Calidad del servicio	8.804	.7987	81	81.000
	X7 - Comercio electrónico (e-Commerce)	3.577	.7433	81	81.000
	X10 - Publicidad	3.752	1.0821	81	81.000
	X11 - Línea de servicios	6.633	.9313	81	81.000
	X12 - Imagen de la fuerza de ventas	4.731	1.0187	81	81.000
	X13 - Precio competitivo	5.844	1.2739	81	81.000
	X17 - Flexibilidad de precios	3.635	.6491	81	81.000
	X19 - Satisfacción	7.211	.9901	81	81.000
Fuera de MEX/Norteamérica	X6 - Calidad del servicio	7.275	1.3559	119	119.000
	X7 - Comercio electrónico (e-Commerce)	3.893	.7625	119	119.000
	X10 - Publicidad	4.272	1.1465	119	119.000
	X11 - Línea de servicios	5.258	1.2515	119	119.000
	X12 - Imagen de la fuerza de ventas	5.600	1.0658	119	119.000
	X13 - Precio competitivo	7.738	1.2854	119	119.000
	X17 - Flexibilidad de precios	5.029	1.1487	119	119.000
	X19 - Satisfacción	6.776	1.3623	119	119.000

Autovalores

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	1.743[a]	100.0	100.0	.797

a. Se han empleado las 1 primeras funciones discriminantes canónicas en el análisis.

Autovalores

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	1.519[a]	100.0	100.0	.777

a. Se han empleado las 1 primeras funciones discriminantes canónicas en el análisis.

Lambda de Wilks

Contraste de las funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	.365	191.222	17	.000

Lambda de Wilks

Contraste de las funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	.397	180.178	6	.000

Fuente: SPSS 20 IBM

SPSS produce diversas tablas, las cuales se pueden comparar en los métodos: **inclusión por pasos y estimación simultánea. Ver Figura 6.32.**

Figura 6.32.- Tablas Pruebas de igualdad de las medias de los grupos, Autovalores y *Lambda de Wilks*. Inclusión por pasos vs. Estimación simultánea

Inclusión por pasos

Pruebas de igualdad de las medias de los grupos

	Lambda de Wilks	F	gl1	gl2	Sig.
X6 - Calidad del servicio	.704	83.242	1	198	.000
X7 - Comercio electrónico (e-Commerce)	.959	8.485	1	198	.004
X10 - Publicidad	.950	10.389	1	198	.001
X11 - Línea de servicios	.736	71.007	1	198	.000
X12 - Imagen de la fuerza de ventas	.856	33.211	1	198	.000
X13 - Precio competitivo	.653	105.323	1	198	.000
X17 - Flexibilidad de precios	.669	97.904	1	198	.000
X19 - Satisfacción	.970	6.085	1	198	.014

Estimación simultánea

Pruebas de igualdad de las medias de los grupos

	Lambda de Wilks	F	gl1	gl2	Sig.
X6 - Calidad del servicio	.704	83.242	1	198	.000
X7 - Comercio electrónico (e-Commerce)	.959	8.485	1	198	.004
X11 - Línea de servicios	.736	71.007	1	198	.000
X12 - Imagen de la fuerza de ventas	.856	33.211	1	198	.000
X13 - Precio competitivo	.653	105.323	1	198	.000
X17 - Flexibilidad de precios	.669	97.904	1	198	.000
X8 - Soporte técnico	1.000	.003	1	198	.960
X9 - Respueta a quejas	1.000	.003	1	198	.959
X10 - Publicidad	.950	10.389	1	198	.001
X14 - Garantías	1.000	.058	1	198	.809
X15 - Nuevos productos y servicios	.997	.541	1	198	.463
X16 - Ordenes y facturación	1.000	.004	1	198	.950
X18 - Velocidad de entrega	1.000	.036	1	198	.849
X19 - Satisfacción	.970	6.085	1	198	.014
X20 - Probabilidad de recomendación	.996	.878	1	198	.350
X21 - Probabilidad de compra	.985	3.027	1	198	.083
X22 - Nivel de compra	.968	6.530	1	198	.011

Nota: Se conserva congruencia

Autovalores

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	1.519[a]	100.0	100.0	.777

a. Se han empleado las 1 primeras funciones discriminantes canónicas en el análisis.

Autovalores

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	1.520[a]	100.0	100.0	.777

a. Se han empleado las 1 primeras funciones discriminantes canónicas en el análisis.

Lambda de Wilks

Contraste de las funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	.397	180.178	6	.000

Lambda de Wilks

Contraste de las funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	.397	179.271	8	.000

Fuente: SPSS 20 IBM

Figura 6.33.- Tablas Coeficientes estandarizados de las funciones discriminantes canónicas y Matriz de estructura. Inclusión por pasos vs. Estimación simultánea.

Inclusión por pasos.

Coeficientes estandarizados de las funciones discriminantes canónicas

	Función 1
X6 - Calidad del servicio	-.267
X7 - Comercio electrónico (e-Commerce)	-.409
X10 - Publicidad	-.006
X11 - Línea de servicios	-.411
X12 - Imagen de la fuerza de ventas	.816
X13 - Precio competitivo	.411
X17 - Flexibilidad de precios	.372
X19 - Satisfacción	-.025

Matriz de estructura

	Función 1
X13 - Precio competitivo	.592
X17 - Flexibilidad de precios	.570
X6 - Calidad del servicio	-.526
X11 - Línea de servicios	-.486
X12 - Imagen de la fuerza de ventas	.332
X10 - Publicidad[a]	.189
X7 - Comercio electrónico (e-Commerce)	.168
X19 - Satisfacción[a]	-.137

Correlaciones intra-grupo combinadas entre las variables discriminantes y las funciones discriminantes canónicas tipificadas
Variables ordenadas por el tamaño de la correlación con la función.

a. Esta variable no se emplea en el análisis.

Estimación simultanea

Coeficientes estandarizados de las funciones discriminantes canónicas

	Función 1
X6 - Calidad del servicio	-.276
X7 - Comercio electrónico (e-Commerce)	-.405
X11 - Línea de servicios	-.422
X12 - Imagen de la fuerza de ventas	.799
X13 - Precio competitivo	.413
X17 - Flexibilidad de precios	.364

Matriz de estructura

	Función 1
X13 - Precio competitivo	.592
X17 - Flexibilidad de precios	.570
X6 - Calidad del servicio	-.526
X11 - Línea de servicios	-.486
X12 - Imagen de la fuerza de ventas	.332
X10 - Publicidad	.186
X7 - Comercio electrónico (e-Commerce)	.168
X19 - Satisfacción	-.142

Correlaciones intra-grupo combinadas entre las variables discriminantes y las funciones discriminantes canónicas tipificadas
Variables ordenadas por el tamaño de la correlación con la función.

Fuente: SPSS 20 IBM

Figura 6.34. Tablas Resultados de clasificación. Inclusión por pasos y Estimación simultánea

Inclusión por pasos

Resultados de la clasificación[a]

		X4 - País	Grupo de pertenencia pronosticado: MEX/Norteamérica	Grupo de pertenencia pronosticado: Fuera de MEX/Norteamérica	Total
Original	Recuento	MEX/Norteamérica	80	1	81
		Fuera de MEX/Norteamérica	20	99	119
	%	MEX/Norteamérica	98.8	1.2	100.0
		Fuera de MEX/Norteamérica	16.8	83.2	100.0

a. Clasificados correctamente el 89.5% de los casos agrupados originales.

Fuente: SPSS 20 IBM

Estimación simultánea

Resultados de la clasificación[a]

		X4 - País	Grupo de pertenencia pronosticado: MEX/Norteamérica	Grupo de pertenencia pronosticado: Fuera de MEX/Norteamérica	Total
Original	Recuento	MEX/Norteamérica	80	1	81
		Fuera de MEX/Norteamérica	20	99	119
	%	MEX/Norteamérica	98.8	1.2	100.0
		Fuera de MEX/Norteamérica	16.8	83.2	100.0

a. Clasificados correctamente el 89.5% de los casos agrupados originales.

Fuente: SPSS 20 IBM

En éstas tablas Explica el porcentaje de aciertos en los grupos, el cual es altamente adecuado (**89.5%**) a pesar de las variables excluidas. Ver **Figura 6.35.**

Figura 6.35.- Tabla de variables introducidas/ excluídas. Inclusión por pasos.

Variables introducidas/excluidas[a,b,c,d]

Paso	Introducidas	Lambda de Wilks							
						F exacta			
		Estadístico	gl1	gl2	gl3	Estadístico	gl1	gl2	Sig.
1	X13 - Precio competitivo	.653	1	1	198.000	105.323	1	198.000	.000
2	X17 - Flexibilidad de precios	.539	2	1	198.000	84.411	2	197.000	.000
3	X11 - Línea de servicios	.494	3	1	198.000	66.923	3	196.000	.000
4	X12 - Imagen de la fuerza de ventas	.431	4	1	198.000	64.412	4	195.000	.000
5	X6 - Calidad del servicio	.412	5	1	198.000	55.347	5	194.000	.000
6	X7 - Comercio electrónico (e-Commerce)	.397	6	1	198.000	48.872	6	193.000	.000

En cada paso se introduce la variable que minimiza la lambda de Wilks global.

a. El número máximo de pasos es 16.

b. La F parcial mínima para entrar es 3.84.

c. La F parcial máxima para salir es 2.71

d. El nivel de F, la tolerancia o el VIN son insuficientes para continuar los cálculos.

Fuente: SPSS 20 IBM

¿Qué sucede con el análisis cuando hay 3 grupos, Suponga X_1.-Antigüedad del consumidor?

- Dado que la mayoría de los pasos anteriores son idénticos, para el caso de los tres grupos nos centraremos, sobre todo, en la interpretación de las funciones discriminantes, que es el elemento novedoso, al haber más de una función.
- El problema que analizamos ahora es similar que en el caso anterior, con la diferencia de que la variable dependiente que se utilizará es $\mathbf{X_1}$ (antigüedad del consumidor) y las variables independientes serán las mismas $\mathbf{X_6}$ a $\mathbf{X_{23}}$ (percepción de los clientes de **MKT Digital**).
- Nótese que la variable $\mathbf{X_1}$ es distinta de la variable $\mathbf{X_4}$, ya que aquella tiene **3 categorías** en los cuales se clasifican los clientes de **MKT Digital** (**1 = menos de un año; 2 = 1 a 5 años; 3 = Más de 5 años**).
- **El objetivo es el mismo: establecer los determinantes de este uso y predecir a qué grupo pertenecerán los nuevos clientes o empresas.** El proceso es el mismo.
- Obviamos el detalle del proceso paso a paso por ser análogo al anterior.

Paso 1: Objetivos

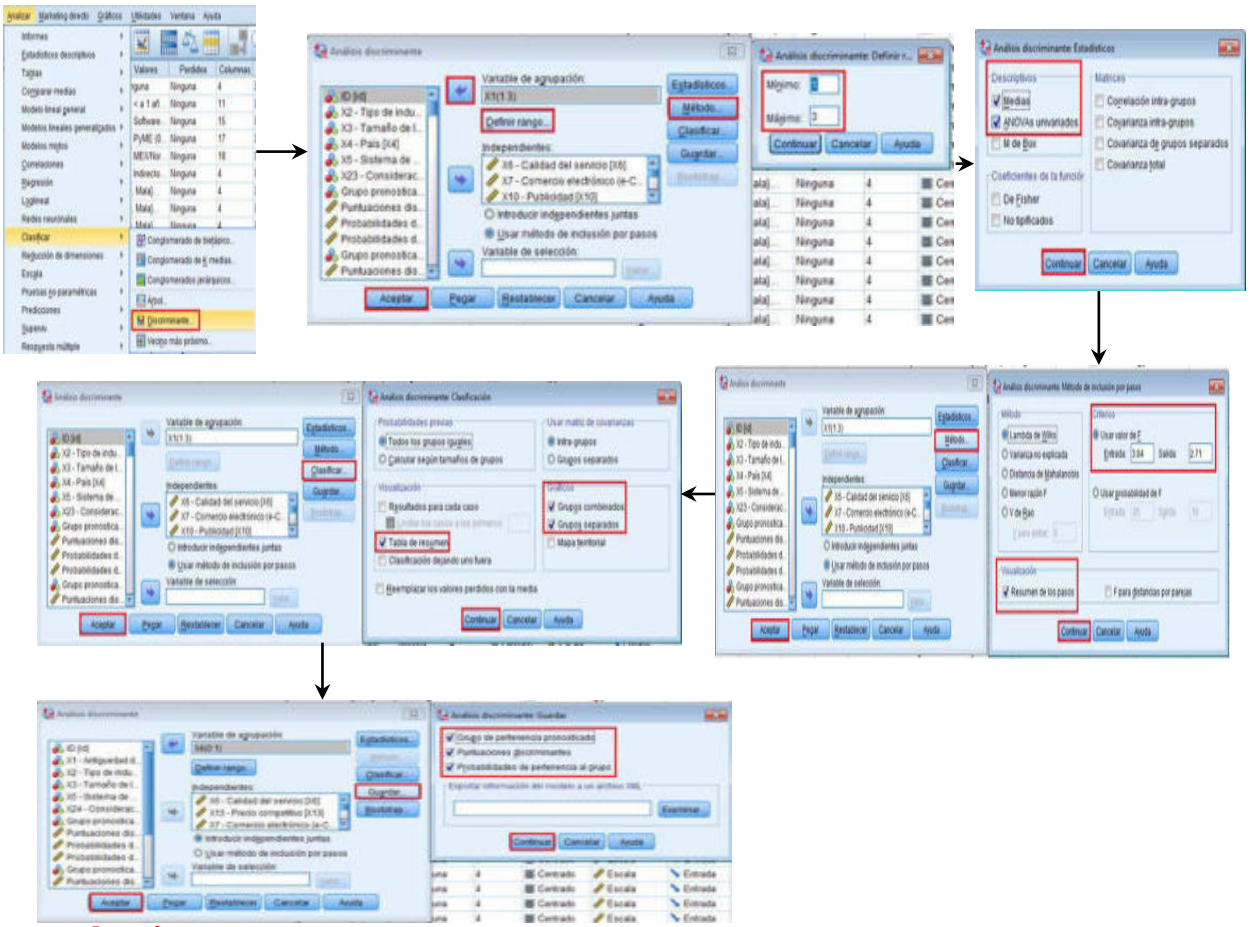
-Problema 6: ¿qué sucede con el análisis cuando hay 3 grupos, Suponga X_1.- Antigüedad del consumidor?.

Paso 2: Diseño y Paso 3: Condiciones de aplicabilidad, se asumen sin cambios.

Paso 4: Ejecución y ajuste

Teclear: Analizar->Clasificar->Discriminante->Variable de agrupación X_1. nominal->Definir rango->Mínimo: 1; Máximo: 3->Continuar->Independientes: X_6 a X_{23} métricas->Introducir independientes juntas->Estadísticos->Descriptivos: Medias; ANOVAS->Continuar->Método: *Lambda de Wilks*->Criterios: usar valor de F entrada: 3.84; salida: 2.71-> Visualización: Resumen de pasos->Continuar->Clasificar->Probabilidades previas: Todos los grupos iguales->Gráficos: Grupos combinados; Grupos separados->Visualización: Tabla de resumen-> Continuar->Guardar->Grupo de pertenencia pronosticado; Puntuaciones discriminantes; Probabilidades de pertenencia al grupo ->Continuar-Aceptar. Ver Figura 6.36 ,6.37, 6.38, 6.39, .6.40.

Figura 6.36.- Proceso análisis discriminante variables: X_1 vs. X_6-X_{23}



. Inclusión por pasos.

3 grupos de X2. Antigüedad del consumidor. Se generan 3 grupos-1= 2 funciones discriminantes

Fuente: SPSS 20 IBM

Paso 5: interpretación

SPSS genera varias tablas para su análisis. Ver **Figura 6.37.**

Figura 6.37.- Estadísticos de grupo, Pruebas de igualdad de las medias de los grupos, Autovalores y *Lambda de Wilks*.

Estadísticos de grupo

X1 - Antiguedad del consumidor		Media	Desv. típ.	N válido (según lista) No ponderados	N válido (según lista) Ponderados
< a 1 año	X6 - Calidad del servicio	7.135	1.0157	68	68.000
	X7 - Comercio electrónico (e-Commerce)	3.715	.7057	68	68.000
	X8 - Soporte técnico	4.816	1.7668	68	68.000
	X9 - Respueta a quejas	4.340	.9498	68	68.000
	X10 - Publicidad	3.703	1.0237	68	68.000
	X11 - Línea de servicios	4.782	.9854	68	68.000
	X12 - Imagen de la fuerza de ventas	4.959	1.0160	68	68.000
	X13 - Precio competitivo	7.615	1.2989	68	68.000
	X14 - Garantías	5.803	.9087	68	68.000
	X15 - Nuevos productos y servicios	4.810	1.5033	68	68.000
	X16 - Ordenes y facturación	3.541	.8646	68	68.000
	X17 - Flexibilidad de precios	4.194	.9441	68	68.000
	X18 - Velocidad de entrega	3.129	.6203	68	68.000
	X19 - Satisfacción	5.729	.7643	68	68.000
	X20 - Probabilidad de recomendación	6.141	.9949	68	68.000
	X21 - Probabilidad de compra	6.962	.7598	68	68.000
	X22 - Nivel de compra	49.012	5.0216	68	68.000
1 a 5 años	X6 - Calidad del servicio	7.297	1.2950	64	64.000

Pruebas de igualdad de las medias de los grupos

	Lambda de Wilks	F	gl1	gl2	Sig.
X6 - Calidad del servicio	.526	88.905	2	197	.000
X7 - Comercio electrónico (e-Commerce)	.997	.267	2	197	.766
X8 - Soporte técnico	.965	3.594	2	197	.029
X9 - Respueta a quejas	.626	58.793	2	197	.000
X10 - Publicidad	.939	6.439	2	197	.002
X11 - Línea de servicios	.541	83.458	2	197	.000
X12 - Imagen de la fuerza de ventas	.952	4.959	2	197	.008
X13 - Precio competitivo	.800	24.621	2	197	.000
X14 - Garantías	.953	4.905	2	197	.008
X15 - Nuevos productos y servicios	.950	5.180	2	197	.006
X16 - Ordenes y facturación	.689	44.532	2	197	.000
X17 - Flexibilidad de precios	.772	29.156	2	197	.000
X18 - Velocidad de entrega	.565	75.795	2	197	.000
X19 - Satisfacción	.464	113.794	2	197	.000
X20 - Probabilidad de recomendación	.696	43.112	2	197	.000
X21 - Probabilidad de compra	.663	50.121	2	197	.000
X22 - Nivel de compra	.313	216.631	2	197	.000

En esta tabla se aprecian los valores por cada categoría

En este caso, las variables con **Sig p>0.05** indican que no intervienen en la diferencia de grupo X_6-X_{23} para la variable X_1. En este caso se sugieren se requiere **retirar** X_7 por lo que habrá que recargar para evaluar

Autovalores

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	3.593[a]	87.9	87.9	.884
2	.496[a]	12.1	100.0	.576

a. Se han empleado las 2 primeras funciones discriminantes canónicas en el análisis.

Observar las correlaciones, ***Lambda de Wilks*** y Sig. De las 2 funciones discriminantes

Lambda de Wilks

Contraste de las funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1 a la 2	.146	374.874	12	.000
2	.668	78.368	5	.000

Fuente: SPSS 20 IBM

Coeficientes Estandarizados:

- Contribución relativa de la variable asociada a la función discriminante.

-A mayores coeficientes, mayor contribución

-Problemas con la multicolinealidad: Un coeficiente bajo implica o que su contribución no es relevante o que ha sido eliminado por su elevada correlación con las demás variables.

- **Matriz de Estructura:**

-Las puntuaciones discriminantes miden la correlación simple entre cada variable independiente y la función discriminante.

-Reflejan la varianza que la variable independiente comparte con la función discriminante.

- ***F* Univariante:**

-Poder discriminante relativo de cada variable independiente

-Tiene asociados niveles de significación

SPSS genera la tabla Coeficientes estandarizados se las funciones discriminantes canónicas y Matriz de estructura. Ver **Figura 6.38.**

Figura 6.38.- Coeficientes estandarizados de las funciones discriminantes canponicas y Matriz de estructura

Coeficientes estandarizados de las funciones discriminantes canónicas

	Función	
	1	2
X6 - Calidad del servicio	.114	-.547
X12 - Imagen de la fuerza de ventas	-.247	-.341
X13 - Precio competitivo	-.421	-.022
X17 - Flexibilidad de precios	.037	.642
X19 - Satisfacción	.392	.784
X22 - Nivel de compra	.824	-.180

Responde al caso a **qué variable independiente tiene mayor correlación con la dependiente X_1.** Son correlaciones entre las variables y la función discriminante estandarizada a partir de En este caso es la variable X_{22} nivel de compra y X_{17} flexibilidad de precios. Hacer **parsimonia** del modelo discriminante

Matriz de estructura

	Función	
	1	2
X22 - Nivel de compra	.782*	.072
X19 - Satisfacción	.549*	.386
X11 - Línea de servicios[b]	.502*	-.110
X20 - Probabilidad de recomendación[b]	.366*	.195
X21 - Probabilidad de compra[b]	.252*	.139
X13 - Precio competitivo	-.248*	.242
X17 - Flexibilidad de precios	-.046	.762*
X18 - Velocidad de entrega[b]	.400	.647*
X6 - Calidad del servicio	.451	-.588*
X9 - Respueta a quejas[b]	.335	.546*
X16 - Ordenes y facturación[b]	.283	.436*
X12 - Imagen de la fuerza de ventas	.056	.281*
X10 - Publicidad[b]	.064	.208*
X15 - Nuevos productos y servicios[b]	-.018	.112*
X14 - Garantías[b]	-.015	.077*
X8 - Soporte técnico[b]	-.032	.040*

Correlaciones intra-grupo combinadas entre las

Responde al caso a qué variable es más discriminante o diferenciador de **los 3 grupos de X_1.** Se afirma que la variable de mayor importancia, por valor absoluto es X_{22} **y** X_{13}, respecto a los valores de X_1.-Antigüedad del consumidor, así como una segunda función dependiente de X_{19} **y** X_{17} para X_1.Hacer **parsimonia** del modelo discriminante.

Fuente: SPSS 20 IBM

SPSS genera la tabla Resultados de la clasificación y Variables introducidas/extraídas.Ver **Figura 6.39.**

Figura 6.39.- Resultados de la clasificación

Resultados de la clasificación[a]

		X1 - Antiguedad del consumidor	Grupo de pertenencia pronosticado			Total
			< a 1 año	1 a 5 años	Más de 5 años	
Original	Recuento	< a 1 año	66	2	0	68
		1 a 5 años	8	50	6	64
		Más de 5 años	0	6	62	68
	%	< a 1 año	97.1	2.9	.0	100.0
		1 a 5 años	12.5	78.1	9.4	100.0
		Más de 5 años	.0	8.8	91.2	100.0

a. Clasificados correctamente el 89.0% de los casos agrupados originales.

Explica el porcentaje de aciertos en los grupos, el cual es altamente adecuado **(89.0%)** a pesar de las variables excluidas, como se observa en la **Figura 6.39**

Figura 6.39. Variables introducidas / excluídas.

Variables introducidas/excluidas[a,b,c,d]

Paso	Introducidas	Lambda de Wilks							
						F exacta			
		Estadístico	gl1	gl2	gl3	Estadístico	gl1	gl2	Sig.
1	X22 - Nivel de compra	.313	1	2	197.000	216.631	2	197.000	.000
2	X17 - Flexibilidad de precios	.225	2	2	197.000	108.504	4	392.000	.000
3	X13 - Precio competitivo	.186	3	2	197.000	85.693	6	390.000	.000
4	X19 - Satisfacción	.163	4	2	197.000	71.741	8	388.000	.000
5	X6 - Calidad del servicio	.153	5	2	197.000	60.193	10	386.000	.000
6	X12 - Imagen de la fuerza de ventas	.146	6	2	197.000	51.883	12	384.000	.000

En cada paso se introduce la variable que minimiza la lambda de Wilks global.

a. El número máximo de pasos es 32.

b. La F parcial mínima para entrar es 3.84.

c. La F parcial máxima para salir es 2.71

d. El nivel de F, la tolerancia o el VIN son insuficientes para continuar los cálculos.

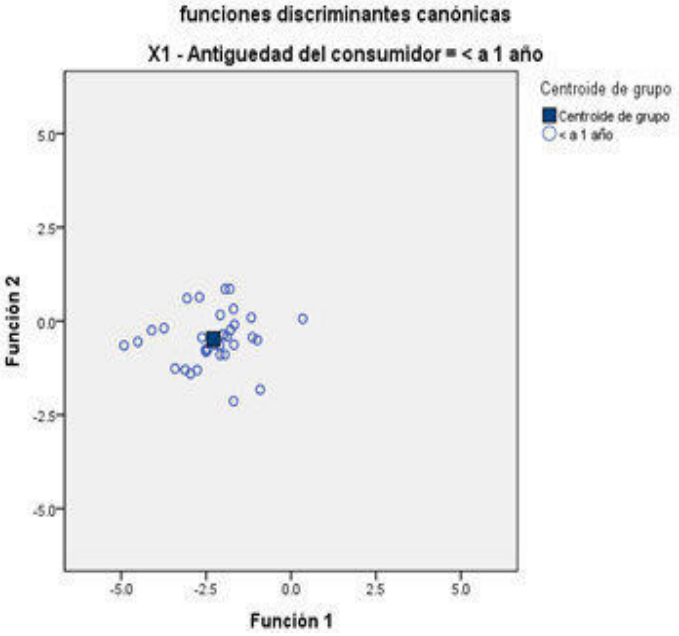
Fuente: SPSS 20 IBM

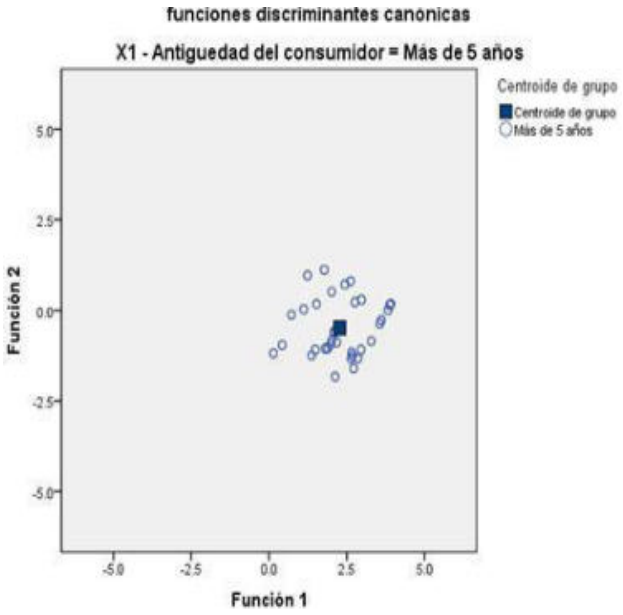
Figura 6.40.- Centroides de grupos

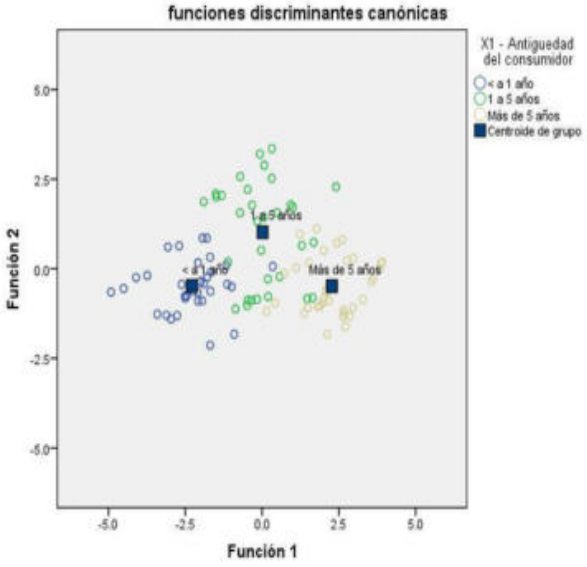
Funciones en los centroides de los grupos

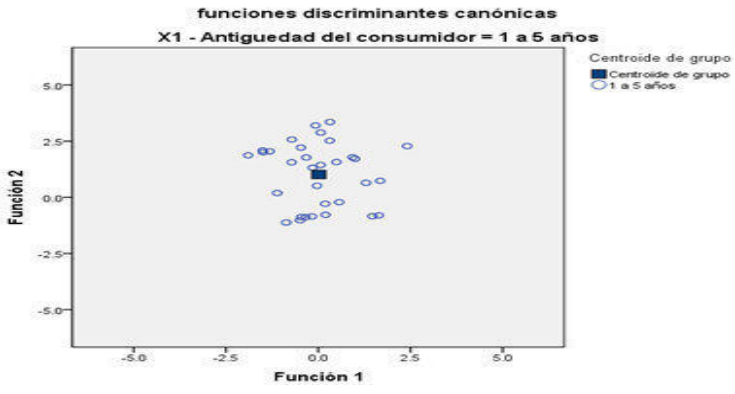
X1 - Antiguedad del consumidor	Función	
	1	2
< a 1 año	-2.289	-.475
1 a 5 años	.016	1.019
Más de 5 años	2.274	-.484

Funciones discriminantes canónicas no tipificadas evaluadas en las medias de los grupos









Fuente: SPSS 20 IBM

6.11.1. Visión gerencial final

El **análisis discriminante**, orientado a la comprensión de las diferencias perceptivas de los clientes basadas en su tipo habitual de situación de compra con **MKT Digital y su base de datos BM_MKT_Digital.sav**, es el origen de varios resultados:

1. Hay dos dimensiones de discriminación entre los tipos de situación de compra. La primera dimensión está tipificada por $\mathbf{X_{22}}$**. nivel de compra y** $\mathbf{X_{13}}$ **precio competitivo**, que puede ser indicativa de **cualidades relacionales**.
2. La segunda dimensión está mejor caracterizada por altas percepciones sobre $_{19}$ **Satisfacción y** $_{17}$ **flexibilidad de precios**, características más **objetivas de la transacción**.

Los 3 grupos de $\mathbf{X_1}$ pueden también perfilarse sobre estas dos dimensiones y las variables asociadas con cada dimensión para comprender las diferencias perceptivas entre ellas. El grupo de **cualidades relacionales**, tiene unas percepciones bastante más elevadas de **MKT Digital** que los otros dos grupos sobre la primera dimensión, quizá indicativo de una relación desde hace mucho tiempo. Estos patrones generales pueden extenderse al perfilar los grupos sobre variables independientes separadas y centrarse sobre variables diferenciadoras clave, $\mathbf{X_{22}}$**,** $\mathbf{X_{13}}$**,** $\mathbf{X_{19}}$ **y** $\mathbf{X_{17}}$**.** El análisis discriminante **identifica las variables con más influencia** para que la dirección de la empresa pueda desarrollar programas más concisos incorporando un pequeño con junto de variables.

Los resultados también identifican un conjunto de observaciones de los grupos de nueva tarea y recompra modificada, que potencialmente representa bien un cuarto tipo de situación de compra o están caracterizadas por una variable todavía no incluida en el modelo. Tienen percepciones bastante precisas, pero no son habituales de los otros clientes en esos tipos de situaciones de compra. En muchos casos, son más parecidos al grupo de recompra directa en lo relativo a sus percepciones. Por tanto, a la dirección se le presenta un input para la gestión para la planificación estratégica y táctica a partir, no sólo de los resultados directos del análisis discriminante, también a partir de los errores en la clasificación.

6.12. Análisis Regresión logística: ¿qué es?

Se explica como:

6.12.1. Regresión con una variable dependiente binaria

Como se ha expuesto, **el análisis discriminante es apropiado cuando la variable dependiente es no métrica.** Sin embargo, **cuando la variable dependiente tiene sólo dos grupos, puede preferirse la regresión logística por varios motivos:**

1. El **análisis discriminante** descansa sobre un cumplimiento estricto de los supuestos de **normalidad multivariante** y la **igualdad de matrices de varianzas covarianzas** entre los grupos, supuestos que no siempre se verifican.
2. La **regresión logística no se enfrenta a estos supuestos tan estrictos**, y es mucho más **robusta** cuando estos supuestos **no se cumplen**, haciendo muy apropiada su aplicación en muchas situaciones.
3. Incluso si se cumplieran los supuestos, muchos investigadores prefieren la **regresión logística por que es similar a la regresión.** Ambas cuentan con **contrastes estadísticos directos,** capacidad para incorporar **efectos no lineales** y permitir una **amplia variedad de diagnósticos.** Por estas razones y otras más técnicas, **la regresión logística es equivalente al análisis discriminante de dos grupos** y puede considerarse más apropiada en muchas situaciones.

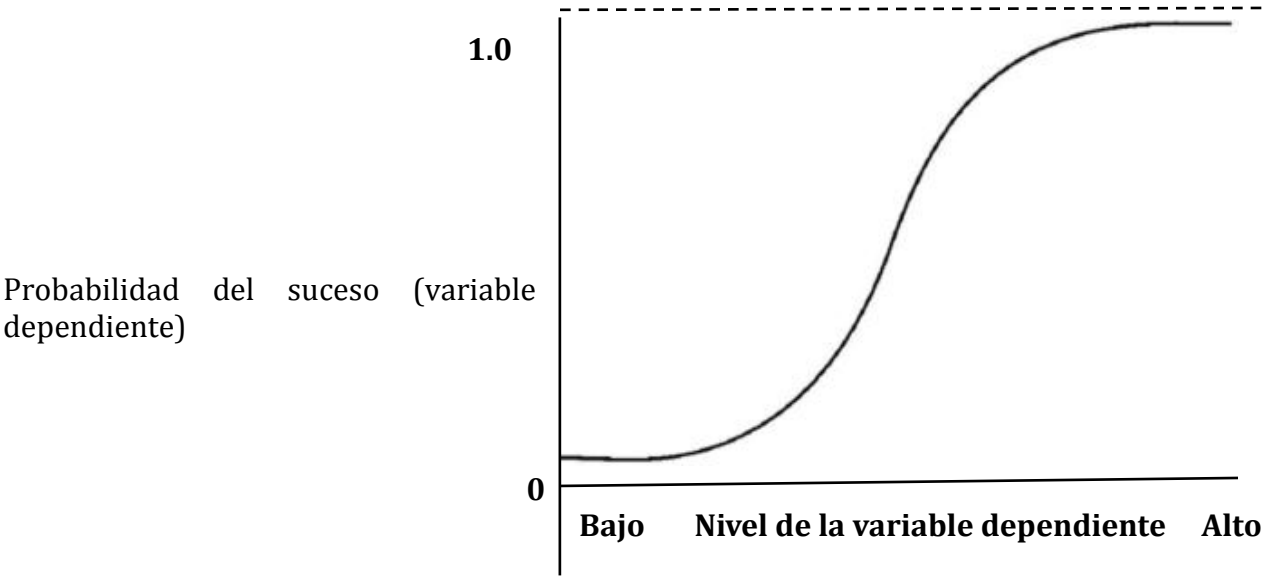
Nuestra presentación de la regresión logística no cubre cada una de las seis etapas del proceso de decisión; en su lugar se resaltarán **las diferencias y parecidos** entre la **regresión logística y el análisis discriminante o la regresión múltiple.**

6.12.2. Representación de la variable dependiente binaria

Para lograrlo, considere:

1. En el **análisis discriminante**, el carácter **no métrico** de una **variable dependiente dicotómica se adecúa** haciendo **predicciones de pertenencia al grupo** basadas en sus **puntuaciones *z* discriminantes.** Esto requiere el **cálculo de puntuaciones de corte** y la **asignación de observaciones a grupos.**
2. La **regresión logística** afronta esta tarea de forma algo parecida a la **regresión múltiple**, aunque se diferencia de ésta, en que **predice** directamente la **probabilidad de ocurrencia de un suceso.**
3. Aunque el valor de la **probabilidad sea una medida métrica,** existen **diferencias** fundamentales entre la **regresión múltiple** y la **logística.**
4. Los **valores de la probabilidad** pueden ser cualesquiera entre **0 y 1**, pero el **valor predicho** debe estar acotado para que caiga en el **rango de 0 y 1.**
5. Para definir una relación acotada por **0 y 1**, la regresión logística utiliza una relación supuesta entre las **variables dependientes e independientes** que recuerda a una **curva en forma de S (véase Figura 6.41).**

Figura 6.41. Forma de la relación logística entre las variables independiente y dependiente



Para **niveles muy bajos de la variable independiente**, la **probabilidad** se aproxima a **cero.** Según **crece la variable independiente,** la **probabilidad crece** a lo largo de la curva, pero como la **pendiente empieza a decrecer para cierto nivel de la variable in dependiente**, la **probabilidad se acercará a 1** sin llegar a excederlo. En la discusión de la **regresión**, los modelos de **regresión lineal no permitían captar tal relación, al ser inherentemente no lineal**. Además, tales situaciones no pueden estudiarse mediante la regresión ordinaria, porque al hacerlo se incumplen varios supuestos:

1. El **término de error** de una **variable discreta** sigue una **distribución binomial** en lugar de la **distribución normal**, **invalidando todos los contrastes estadísticos** basados en el supuesto de **normalidad**.
2. La **varianza** de una **variable dicotómica no es constante**, creando en consecuencia situaciones de **heterocedasticidad.**

La regresión logística se desarrolló para tratar precisamente estas situaciones. La única relación entre variables dependientes e independientes requiere de una aproximación algo diferente en la estimación, la evaluación de la bondad de ajuste y la interpretación de los coeficientes.

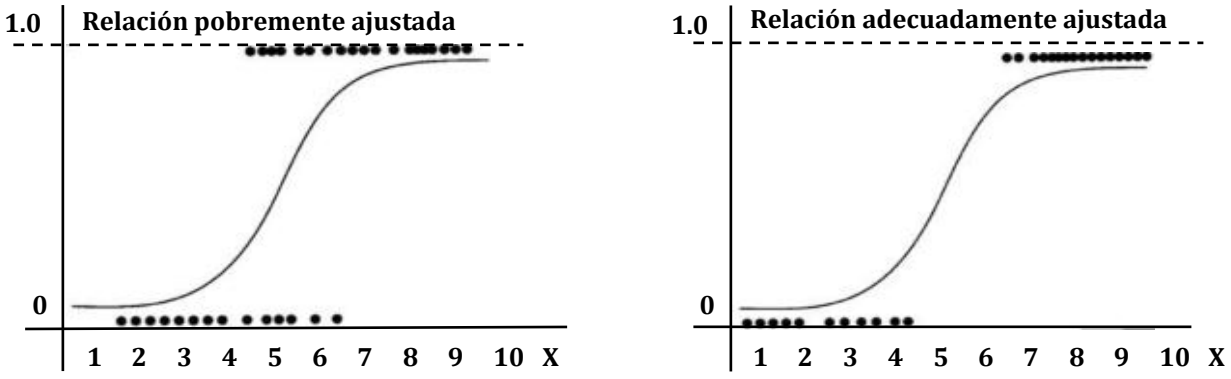
6.12.3. Estimación del modelo de regresión logística

La **regresión logística**, aunque incluya un único valor teórico resumen de los coeficientes estimados para cada variable independiente (como se encontró en la regresión múltiple) se estima de forma totalmente diferente:

1. La **regresión múltiple** emplea el **método de los mínimos cuadrados**, que **minimiza las sumas de las diferencias al cuadrado entre los valores reales y predichos de la variable dependiente.**
2. Debido a la naturaleza no lineal de la transformación logística requiere que otro procedimiento, el de máxima verosimilitud, se utilice de forma **iterativa** para encontrar la **estimación "*más probable*" de los coeficientes**.
3. Por lo anterior, se usa el **valor de la verosimilitud en lugar de la suma de los cuadrados al calcular la medida de ajuste global del modelo.**
4. El proceso de **estimación de los coeficientes** es, aun así, **bastante parecido en muchos aspectos al de regresión lineal**.
5. Como se vio antes, el **modelo logístico** tiene la forma concreta de una **curva logística.** Para estimar un modelo de regresión logística, se ajusta esta curva a los datos reales.

La **Figura 6.42** representa **dos ejemplos hipotéticos** de ajuste de una relación logística a **datos muestrales**. Los **datos reales**, que un suceso tenga o no lugar (**1 o 0**), se representan como observaciones en lo alto o en lo bajo del gráfico. Estos son los sucesos que ocurren para cada valor de la variable independiente (**eje X**). En la **parte A**, la **curva logística** no puede ajustar los datos bien porque hay varios valores de la **variable independiente** que cuentan tanto con **sucesos** como con **no sucesos** (esto es, un importante **solapamiento de las distribuciones**). Sin embargo, en la **parte B,** existe una relación mucho más definida, y **la curva logística** se ajusta a los datos bastante bien. Este sencillo ejemplo, **similar a una nube de puntos** entre las **variables dependiente e independiente** de la regresión con una línea que representa el **"*mejor ajuste*"** de la correlación, puede extenderse para incluir múltiples variables independientes como en la regresión.

Figura 6.42. Forma de la relación logística entre las variables independiente y dependiente.



6.12.4. Interpretación de los coeficientes

Una de las **ventajas** de la **regresión logística** es que sólo necesitamos saber si un suceso ocurrió (**comprar o no, riesgo de crédito o no, quiebra de la empresa o éxito**) para entonces utilizar un **valor dicotómico** como nuestra **variable dependiente.** A partir de este **valor dicotómico**, el procedimiento **predice su estimación de la probabilidad** de que el suceso tenga o no lugar. Si la predicción de la probabilidad es mayor que **0.50,** entonces la **predicción es sí, y no en otro caso**.

La **regresión logística** deriva su nombre de la **transformación logística utilizada con la variable dependiente**, la que sin embargo, al emplearse **sus coeficientes tienen un sentido diferente del que encontramos en la regresión con una variable de pendiente métrica.**

El procedimiento que calcula **el coeficiente logístico compara la probabilidad de la ocurrencia de un suceso con la probabilidad de que no ocurra** y que se expresa como:

$$(Prob_{evento}/Prob_{no\ evento}) = e^{Bo+B1X1+B2X2+\cdots BnXn}$$

Los coeficientes estimados (B_0 , B_1, B_2,..., B_n) son en realidad medidas de los cambios en el **ratio de probabilidades** (**odds ratio)**, expresados en **logaritmos**, por lo que necesitaríamos **re transformarlos (tomando los valores del antilogaritmo)** de tal forma que **se evalúe más fácilmente su efecto sobre la probabilidad**. **SPSS** lo hace automáticamente calculando tanto el **coeficiente real como el transformado**. Utilizar este procedimiento no cambia en modo alguno la forma de interpretar el signo del coeficiente. Un **coeficiente positivo aumenta la probabilidad,** mientras que u**n valor negativo disminuye la probabilidad predicha.** Por ejemplo:

1. Si ***B;*** es **positivo**, su transformación (**antilog**) será **>1**, y el ***odds ratio*** **aumentará.** Este aumento se produce cuando la **probabilidad de ocurrencia** de un suceso **aumenta** y la probabilidad prevista de su **no ocurrencia disminuye. Por tanto, el modelo tiene una elevada probabilidad de ocurrencia.**
2. De la misma forma, si ***B;*** **es negativo**, el **antilogaritmo** es **<1** **y el *odds* ratio disminuye.**
3. Un **coeficiente cero equivale a un valor de 1.0** lo que no produce cambios en el **odds.** En varios textos **[**Hosmer, D. W., y Lemcshow, S. 1989**]** puede encontrarse una exposición más detallada de la interpretación de los coeficientes, la transformación logística y los procedimientos de estimación.

Como se expuso, la **distribución supuesta** de las posibles **variables dependientes**, describimos una **curva con forma de S o logística**. Para representar esa relación entre las variables dependiente e independiente, los coeficientes deben representar efectivamente **relaciones no lineales entre las variables dependientes e independientes**. Aunque el proceso de transformación de tomar **logaritmos proporciona una linealización** de la relación, el investigador debe recordar que **los coeficientes** representan en realidad **diferentes pendientes** en la relación entre los valores de la variable independiente. De esta forma, puede estimarse la relación en **forma de S**. Si Usted está interesado en la

pendiente de la relación para varios niveles de la variable independiente, se pueden calcular los coeficientes y evaluar la relación **[**Gessner, et al.1988**].**

6.12.5. Valoración de la bondad del ajuste del modelo estimado

La **regresión logística** es **similar** a la **regresión múltipl**e en muchos otros resultados, pero es **diferente** en el **método de estimación de los coeficientes**. En lugar de **minimizar** la **desviación** de los **cuadrados (mínimos cuadrados)**, la regresión logística maximiza la **"*verosimilitud*"** de que un suceso tenga lugar. La utilización de esta técnica de estimación alternativa requiere también **que evaluemos el ajuste del modelo de varias formas**.

La medida global de cómo se **ajusta el modelo**, **similar** al valor de la **suma de errores o residuos al cuadrado en la regresión múltiple**, viene dada por el **valor de la verosimilitud.** (Que es **-2 veces el logaritmo del valor de verosimilitud** y se representa por **-2LL o -2 veces el logaritmo de la verosimilitud**), con las siguientes características:

1. Un modelo con un **buen ajuste** tendrá un **valor pequeño para -2LL**.
2. El **valor mínimo para -2LL es cero**.
3. Un **ajuste perfecto** tiene una **verosimilitud de 1 y -2LL es cero.**
4. El **valor de la verosimilitud** puede compararse asimismo entre ecuaciones, donde la diferencia representa **el cambio en el ajuste predictivo desde una ecuación a otra**.
5. Los programas como **SPSS** cuentan con **contrastes automáticos** para evaluar la **significación** de estas diferencias.
6. El **contraste *Chi-cuadrado*** para la reducción en el logaritmo del valor de verosimilitud proporciona una medida de mejora debida a la introducción de **variable(s) independiente(s).**
7. Un **modelo nulo**, que es similar a **calcular el total de la suma de los cuadrados utilizando sólo la media**, proporciona el punto de partida para la comparación.
8. Adema s de las contrastaciones estad sticas de los **test *Chi-cuadrado***, se han construido varias medidas diferentes tipo R^2 para representar el **ajuste global** del modelo, como lo hace el coeficiente de determinación de la **regresión múltiple**. Usted puede construir un valor "***pseudo*** R^2" para la regresión logística similar al valor R^2 del **análisis de regresión [**Gessner, et al.1988**]**.
9. El R de un **modelo *logit*** (R^2 ***logit***) se calcula como:

*(**R logit**)= (-2**LL**$_{nulo}$-(-2**LL**$_{modeo}$))/-2**LL**$_{nulo}$*

10. Podemos evaluar **el ajuste global** de forma similar a la **regresión múltiple**, y podemos hacer uso de varios métodos **que utilizan la característica no métrica** de la variable dependiente:

    -Utilice el método de las **matrices de clasificación del análisis discriminante** para evaluar la **exactitud predictiva** en términos de pertenencia al grupo. Todas las medidas relacionadas con la aleatoriedad utilizadas previamente son de aplicación también aquí.

    -Hosmer y Lebeshow **[**1989**]** han desarrollado otros contrastes de clasificación. **Los casos se dividen primero en 10 clases aproximadamente iguales**. Luego, el **número de sucesos reales y predichos se compara en cada clase con el estadístico *chi-cuadrado*.** Este contraste proporciona una medida global de exactitud **predicitiva** que **no se basa en el valor de verosimilitud**, sino en la **predicción** real de la **variable dependiente**. **El uso correcto de este contraste** requiere un tamaño de muestra adecuado para asegurar que cada grupo cuenta al menos con **5 observaciones** y **nunca**

es <1. Además, el estadístico ***Chi-cuadrado*** **es sensible al tamaño muestral**, permitiendo, por tanto, que esta medida **encuentre diferencias estadísticamente muy pequeñas cuando el tamaño muestral crece**. Usted deberá hacer uso de todas estas medidas de ajuste para valorar esta técnica, que cuenta con aspectos tanto de la **regresión múltiple como del análisis discriminante.**

6.12.6. Contrastación de la significación de los coeficientes

La **regresión logística** puede contrastar también **la hipótesis de que un coeficiente sea <>0** (el 0 significa que el **odds ratio no cambia y que la probabilidad no se ve afectada**), como se hizo en la regresión múltiple. En ésta, el **valor del t se utiliza para valorar la significatividad de cada coeficiente.** La **regresión logística** utiliza un estadístico diferente, **el estadístico de Wald**. Este proporciona la **significación estadística para cada coeficiente** estimado de tal forma que se pueden contrastar hipótesis igual que en la regresión múltiple.

Otras semejanzas con la regresión múltiple A pesar del hecho de utilizar una medida de **pendiente binaria** y de que el resultado sea la predicción de pertenencia al grupo, el formato de la regresión logística **es bastante parecido al de la regresión múltiple.** Al igual que en la regresión, los **datos categóricos y nominales** pueden incluirse como **variables independientes** por medio de su codificación como **variables ficticias**. También encontramos los procedimientos de selección de modelo al igual que en la regresión múltiple (por etapas hacia delante y hacia atrás). Finalmente, para examinar con mayor claridad los resultados, también se cuenta con muchas **medidas de diagnóstico, como los residuos, los gráficos de los residuos y medidas de influencia.**

Usted que se enfrentará con una **variable dicotómica y no necesita recurrir a métodos diseñados que dan cabida a las limitaciones de la regresión múltiple ni se ve obligado a emplear el análisis discriminante, especialmente si no se verifican sus supuestos estadísticos.** La **regresión logística** salva estos problemas y proporciona un método para tratar directamente con esta situación de la forma más eficiente posible.

6.13. Análisis Regresión logística: Ejemplos

Regresión con una variable dependiente binaria

Hay varios motivos por los que la **regresión logística** es una alternativa atractiva al análisis **discriminante** en tanto que la **variable dependiente cuente sólo con dos categorías:**

1. La **regresión logística** está **menos influida** que el **análisis discriminante** por las diferencias de **varianzas-covarianzas** entre los grupos, un supuesto básico del análisis discriminante.
2. La **regresión logística** puede tratar con **variables independientes categóricas** fácilmente, mientras que en el **análisis discriminante el uso de variables ficticias** creaba **problemas** con las **igualdades de varianzas-covarianzas.**
3. Los resultados de la **regresión logística** son **paralelos** a los de la **regresión múltiple** en lo relativo a su **interpretación** y las **medidas de diagnóstico caso a caso disponibles para el examen de los residuos**.

El siguiente ejemplo es idéntico al del **análisis discriminante de dos grupos** presentado previamente, utilizando esta vez la **regresión logística** para estimar el modelo. Como

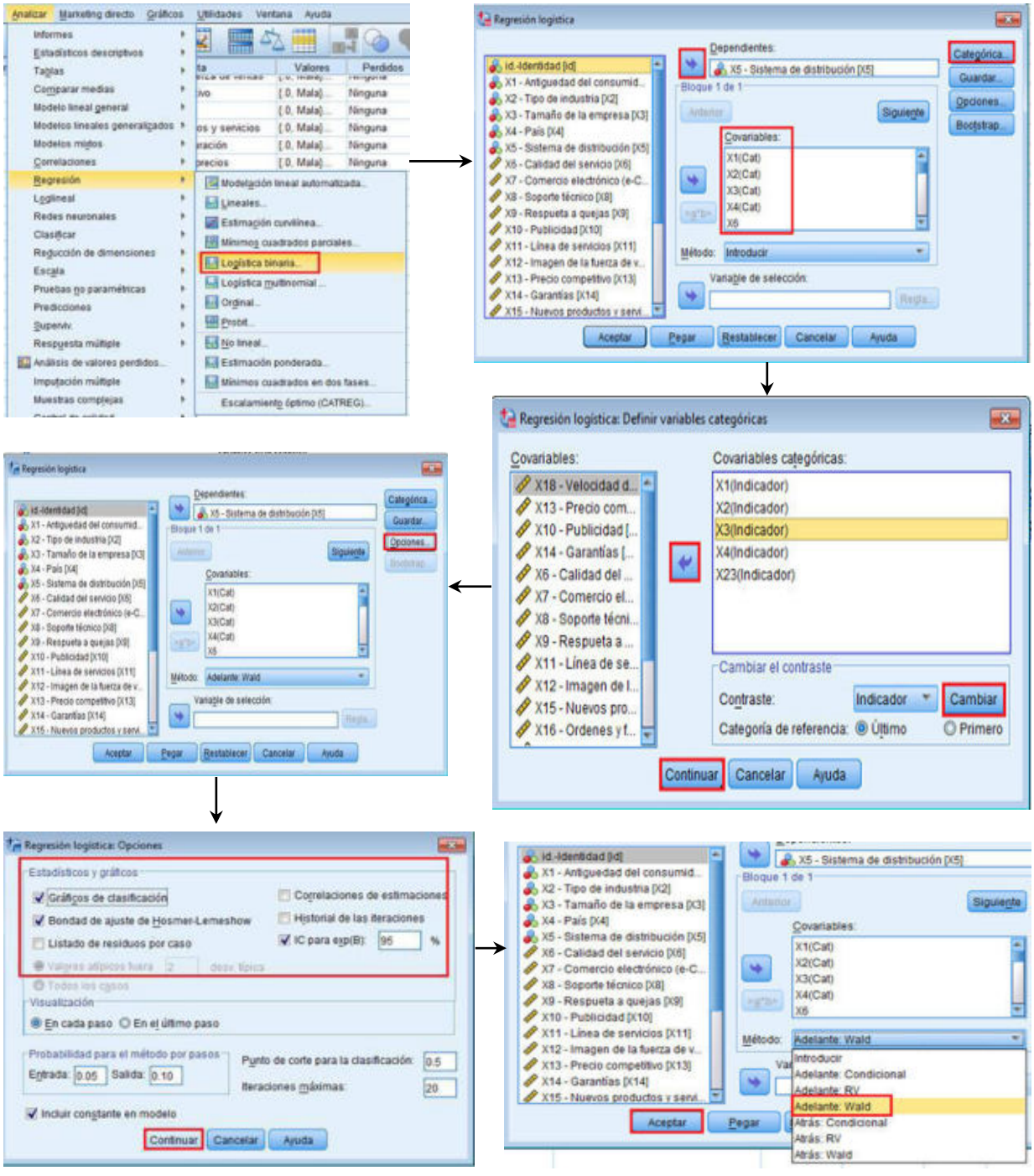
veremos, la **regresión logística** como **alternativa al análisis discriminante** cuenta con muchas **ventajas**, pero el investigador debe también examinar los resultados cuidadosamente para **evitar el sobreajuste** de los datos u otros problemas de estimación del modelo como los encontrados en este ejemplo.

Paso 1: Objetivo;
-Problema 7: la empresa **MKT Digital** con su base de datos **BM_MKT_Digital.sav**, requiere saber las posibilidades de que $\mathbf{X_5}$. **Sistema de distribución (con 2 valores),** sea explicada y predicha con base al resto de las variables de la base de datos.
$\mathbf{H_1}$: $\mathbf{X_5}$. es capaz de predecirse con las variables del modelo de negocios de MKT_Digital con un 80% de éxito

Paso 2: Diseño; Paso 3: Condiciones de aplicabilidad
Se asumirán correctos dados los análisis previos de la base de datos

Paso 4: Estimación y ajuste de la función logística
-Teclear: Regresión->Logística binaria->Selección de variable categórica ($\mathbf{X_5}$)->Selección de covariables ($\mathbf{X_1}$-$\mathbf{X_{23}}$) restantes->Categórica->Selección de variables categóricas en Covariables categóricas ($\mathbf{X_1, X_2, X_3, X_4, X_{23}}$)->Categoría de referencia: Último->Cambiar->Continuar->Opciones->Estadísticos y gráficos, Seleccionar: Gráficos de clasificación, Bondad de ajuste de Hosmer-Lemeshow->Intervalo de confianza (IC para exp(B)): 95%->Continuar->Guardar->Valores pronosticados, seleccionar: Probabilidades, Grupo de pertenencia->Continuar->Acepto. Ver Figura 6.43.

Figura 6.43. Proceso de regresión logística de X_5. vs. Resto de la base de datos BM_MKT_Digital



Fuente: SPSS 20 IBM

Paso 5: Interpretación
SPSS genera la **tabla de Clasificación.Ver Figura 6.44.**

Figura 6.44. Tabla de clasificación

Bloque 0: Bloque inicial

Tabla de clasificación[a,b]

Observado			Pronosticado		
			X5 - Sistema de distribución		
			Indirecto a través de terceros	Directo al consumidor	Porcentaje correcto
Paso 0	X5 - Sistema de distribución	Indirecto a través de terceros	108	0	100.0
		Directo al consumidor	92	0	.0
	Porcentaje global				54.0

a. En el modelo se incluye una constante.

b. El valor de corte es .500

Fuente: SPSS 20 IBM

En **Bloque 0**, identifica el modelo base o de comparación de predicción más sencillo para afirmar que el valor de la variable categórica con mayor frecuencia (**X_5**). con valor= indirecto a través de terceros) tiene un **54%** de acierto de ser cierta. Lo que se espera es que con el agregado de mayor cantidad de variables aumente su % de éxito. Así se asume que los clientes de **MKT_Digital realizan la mayoría de ellos la adquisición de sus servicios a través de terceros**. Este se considera el modelo de base para comparar contra las demás variables independientes; es decir, sólo se tiene la presencia de la variable dependiente. Con esto, lo que se espera es que el nivel de acierto de **54% inicial, aumente conforme se agreguen las variables independientes a fin de mejorar la predicción. En Bloque 0, nos reporta que con la única variable X_5, la afirmación planteada anteriormente tiene una significancia baja aunque probable. Se requiere indagar sobre lo que sucedería si se integran las variables independientes. En Bloque 0, se reportan las variables que tienen oportunidad de generar mejoras de predicción al modelo dada su significancia (*p*<=0.05). Por lo que se observa, existe una gran oportunidad d mejorar el modelo predictivo de insertar el resto de las variables independientes. Ver Figura 6.45 y Figura 6.46.**

Figura 6.45. Variables en la ecuación

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 0	Constante	-.160	.142	1.277	1	.258	.852

Fuente: SPSS 20 IBM

Figura 6.46. Variables que no están en la ecuación

Variables que no están en la ecuación

			Puntuación	gl	Sig.
Paso 0	Variables	X1	21.508	2	.000
		X1(1)	20.943	1	.000
		X1(2)	8.454	1	.004
		X2(1)	3.945	1	.047
		X3(1)	5.052	1	.025
		X4(1)	13.558	1	.000
		X6	29.209	1	.000
		X7	18.132	1	.000
		X8	5.281	1	.022
		X9	7.886	1	.005
		X10	5.575	1	.018
		X11	41.711	1	.000
		X12	16.864	1	.000
		X13	21.989	1	.000
		X14	11.355	1	.001
		X15	.170	1	.680
		X16	13.821	1	.000
		X17	9.385	1	.002
		X18	12.907	1	.000
		X19	60.221	1	.000
		X20	43.422	1	.000
		X21	31.990	1	.000
		X22	15.080	1	.000
		X23(1)	24.979	1	.000
	Estadísticos globales		124.402	23	.000

Fuente: SPSS 20 IBM

SPSS genera en **Bloque 1**, la tabla **Pruebas ómnibus sobre coeficientes del modelo** donde la prueba de ***Chi-cuadrada*** es similar a la prueba de **ANOVA en regresión lineal**, aquí se verifica la **bondad de ajuste del modelo**. Se comprueba que las variables de introducción al modelo, mejoran la predicción del mismo en la ocurrencia de la variable categórica independiente. Se observa que cada paso contiene **3 etapas** que significa cómo el modelo está mejorándose al integrar las variables independientes en cada modalidad. Se recomienda ver la etapa Modelo. **La prueba también se le llama prueba de eficiencia estadística de ROA.** Así, la puntuación de eficiencia estadística de **ROA** indica que hay una mejora significativa en la predicción de la probabilidad de ocurrencia de la categoría de la

variable dependiente categórica $_5$**(Chi-cuadrado: 145.435; gl: 6; *p*<0.001). Ver Figura 6.47.**

Figura 6.47. Pruebas ómnibus sobre coeficientes del modelo

Pruebas omnibus sobre los coeficientes del modelo

		Chi cuadrado	gl	Sig.
Paso 1	Paso	68.690	1	.000
	Bloque	68.690	1	.000
	Modelo	68.690	1	.000
Paso 2	Paso	22.110	1	.000
	Bloque	90.800	2	.000
	Modelo	90.800	2	.000
Paso 3	Paso	28.287	2	.000
	Bloque	119.087	4	.000
	Modelo	119.087	4	.000
Paso 4	Paso	17.896	1	.000
	Bloque	136.983	5	.000
	Modelo	136.983	5	.000
Paso 5	Paso	8.452	1	.004
	Bloque	145.435	6	.000
	Modelo	145.435	6	.000

Fuente: SPSS 20 IBM

SPSS produce en **Bloque 1**, la tabla **Resumen del modelo,** que serpia equivalente al reporte de R^2de los modelos de regresión lineal simple/mútiple, aunque con sus debidas restricciones ya que aquí la variable es categórica. Las R^2 rportadas, aún así, tienen la misma interpretación, como **% de varianza que explica a la variable dependiente.**
¿Cuál de los 2 escoger (R^2 de Cox y Snell o R^2 de Negelkerke?. Se sugiere aplicar criterio de **utilidad**, mediante el que reporte mayor puntaje. **Así, se afirma: que el modelo propuesto explica el 69% de la varianza de la variable dependiente X_5. Ver Figura 6.46**

Figura 6.46. Resumen del modelo

Resumen del modelo

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	207.287[a]	.291	.388
2	185.178[a]	.365	.488
3	156.891[b]	.449	.600
4	138.995[b]	.496	.663
5	130.543[c]	.517	.690

a. La estimación ha finalizado en el número de iteración 5 porque las estimaciones de los parámetros han cambiado en menos de .001.

b. La estimación ha finalizado en el número de iteración 6 porque las estimaciones de los parámetros han cambiado en menos de .001.

c. La estimación ha finalizado en el número de iteración 7 porque las estimaciones de los parámetros han cambiado en menos de .001.

Fuente: SPSS 20 IBM

SPSS genera la Tabla **Prueba de *Hosmer y Lemeshow* y la Tabla de Contingencia para la prueba de *Hosmer y Lemeshow*. Ver Figura 6.47.**

Figura 6.47. Tabla Prueba de *Hosmer y Lemeshow* y Tabla de contingencias para la prueba de *Hosmer y Lemeshow*

En **Bloque 1**, en la t**abla Prueba de *Hosmer y Lemeshow*** se prueba que el último paso tiene una ***Chi-cuadrada*= 25.079; *p*<0.01**, por lo que la inserción paulatina de variables independientes a nivel de varianza explicada por el modelo, es significativa para definir $_5$

Prueba de Hosmer y Lemeshow

Paso	Chi cuadrado	gl	Sig.
1	22.701	8	.004
2	8.861	8	.354
3	16.063	8	.041
4	14.244	8	.076
5	25.079	8	.002

En **Bloque 1**, en la tabla de contingencias para la prueba de ***Hosmer y Lemeshow*** da cuenta de los valores observados y predichos así como su diferencia

Tabla de contingencias para la prueba de Hosmer y Lemeshow

		X5 - Sistema de distribución = Indirecto a través de terceros		X5 - Sistema de distribución = Directo al consumidor		Total
		Observado	Esperado	Observado	Esperado	
Paso 1	1	18	16.756	0	1.244	18
	2	17	15.907	1	2.093	18
	3	15	17.798	7	4.202	22
	4	13	17.229	11	6.771	24
	5	13	10.662	5	7.338	18
	6	6	9.661	14	10.339	20
	7	14	9.514	10	14.486	24
	8	11	6.390	13	17.610	24
	9	1	3.054	19	16.946	20
	10	0	1.029	12	10.971	12
Paso 2	1	21	20.411	0	.589	21
	2	22	20.288	0	1.712	22
	3	18	16.925	2	3.075	20
	4	10	13.381	9	5.619	19
	5	9	9.983	8	7.017	17
	6	10	9.085	8	8.915	18
	7	6	8.127	17	14.873	23
	8	7	5.446	16	17.554	23
	9	3	3.381	17	16.619	20
	10	2	.972	15	16.028	17
Paso 3	1	21	20.763	0	.237	21
	2	18	18.189	1	.811	19

Fuente: SPSS 20 IBM

SPSS genera la **tabla Clasificación. Ver Figura 6.48**

Figura 6.48. Tabla Clasificación

Tabla de clasificación[a]

			Pronosticado		
			X5 - Sistema de distribución		
	Observado		Indirecto a través de terceros	Directo al consumidor	Porcentaje correcto
Paso 1	X5 - Sistema de distribución	Indirecto a través de terceros	79	29	73.1
		Directo al consumidor	31	61	66.3
	Porcentaje global				70.0
Paso 2	X5 - Sistema de distribución	Indirecto a través de terceros	82	26	75.9
		Directo al consumidor	25	67	72.8
	Porcentaje global				74.5
Paso 3	X5 - Sistema de distribución	Indirecto a través de terceros	85	23	78.7
		Directo al consumidor	15	77	83.7
	Porcentaje global				81.0
Paso 4	X5 - Sistema de distribución	Indirecto a través de terceros	96	12	88.9
		Directo al consumidor	16	76	82.6
	Porcentaje global				86.0
Paso 5	X5 - Sistema de distribución	Indirecto a través de terceros	95	13	88.0
		Directo al consumidor	9	83	90.2
	Porcentaje global				89.0

a. El valor de corte es .500

Fuente: SPSS 20 IBM

En **Bloque 1,** la **tabla de Clasificación**, nos reporta un **89.0%** de probabilidad de acierto en el resultado de la variable dependiente $\mathbf{X_5}$ a nivel de predicción cuando intervienen en ella el resto de las variables de **BM_MKT_Digital,** incrementando del **54%** (**del Bloque 0, sólo $\mathbf{X_5}$**) al **89%**(**resto de las variables de la base de datos**).

A este momento se tiene en resumen:

- Con la sola variable dependiente $\mathbf{X_5}$, hay de la **tabla de clasificación: 54%** de acierto de afirmar que la entrega delos servicios por terceros , con alta significancia
- Que hay una explicación del **69%** de la varianza
- Que cuando se insertan el resto de las variables de se tiene de la tabla de clasificación: **89%** de acierto la predicción de afirmar que la entrega delos servicios por terceros , con alta significancia.Así que:
- $\mathbf{H_1}$**:** $_5$ es capaz de predecirse con las variables del modelo de negocios de MKT_Digital con un **80% de éxito**, **es afirmativa**.

SPSS produce la **tabla Variables en la ecuación. Ver Figura 6.49**

Figura 6.49. Tabla Variables en la ecuación.

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95% para EXP(B)	
								Inferior	Superior
Paso 1[a]	X19	1.180	.175	45.606	1	.000	3.253	2.310	4.582
	Constante	-8.426	1.243	45.929	1	.000	.000		
Paso 2[b]	X3(1)	1.808	.421	18.411	1	.000	6.100	2.671	13.932
	X19	1.483	.212	48.727	1	.000	4.406	2.905	6.682
	Constante	-11.530	1.647	48.985	1	.000	.000		
Paso 3[c]	X1			20.984	2	.000			
	X1(1)	3.637	.878	17.138	1	.000	37.972	6.787	212.446
	X1(2)	2.372	.566	17.575	1	.000	10.723	3.537	32.508
	X3(1)	2.842	.549	26.772	1	.000	17.151	5.845	50.333
	X19	2.663	.399	44.458	1	.000	14.343	6.556	31.380
	Constante	-22.248	3.349	44.124	1	.000	.000		
Paso 4[d]	X1			24.783	2	.000			
	X1(1)	4.124	.976	17.863	1	.000	61.777	9.127	418.125
	X1(2)	4.024	.846	22.617	1	.000	55.926	10.651	293.660
	X3(1)	2.483	.595	17.389	1	.000	11.975	3.728	38.465
	X17	-.918	.240	14.646	1	.000	.399	.249	.639
	X19	2.866	.452	40.212	1	.000	17.573	7.246	42.620
	Constante	-20.000	3.643	30.131	1	.000	.000		
Paso 5[e]	X1			25.758	2	.000			
	X1(1)	4.647	1.056	19.368	1	.000	104.303	13.166	826.322
	X1(2)	4.215	.874	23.239	1	.000	67.710	12.200	375.799
	X3(1)	2.940	.643	20.881	1	.000	18.919	5.361	66.770
	X17	-.969	.252	14.799	1	.000	.379	.232	.622
	X19	2.525	.468	29.088	1	.000	12.492	4.990	31.274

Fuente: SPSS 20 IBM

En **Bloque 1,** sobre la tabla de variables de la ecuación, informa cuales son los valores que se debe tener en cuenta para la ec. de la regresión a fin de calcular la predicción de $\mathbf{X_5}$ en funcio n al resto de las variables de la base de datos **BM_MKT_Digital.** Cada una de las variables muestran constantes ($\mathbf{B_0}$ en regresio n lineal). Se debera tomar en cuenta **EXP (B)** es el resultado de la ec. de tener ciertos valores de las variables independientes con base a las variables independientes. Si la probabilidad de ocurrencia de que se de uno de los valores de $\mathbf{X_5}$, con **EXP (B):**

- **<1** significa que si aumenta el valor de las **variables X independiente**s, disminuye la **dependiente $\mathbf{X_5}$ (1/0. Inversa)**
- **>1** significa que si **aumenta** el valor de las variables **X independientes**, **también aumenta la dependiente $\mathbf{X_5}$ (1/0. Directa)**
- a **puntuación de Wald** para el modelo probado indica que las variables independientes ($\mathbf{X_s}$) aportan significativamente a la predicción de la variable dependiente ($\mathbf{X_5}$), los resultados obteniéndos se pueden generalizar a la población, a partir del **paso 5**.

SPSS produce tablas **Codificación de variable dependiente, Codificación de la variable categórica, Pruebas ómnibus sobre los coeficientes del modelo y Resumen del modelo.Ver Figura 6.50.**

Figura 6.50. Tablas Codificación de variable dependiente, Codificación de la variable categórica, Pruebas ómnibus sobre los coeficientes del modelo y Resumen del modelo.

Codificación de la variable dependiente

Valor original	Valor interno
Indirecto a través de terceros	0
Directo al consumidor	1

Codificaciones de variables categóricas

		Frecuencia	Codificación de parámetros (1)	(2)
X1 - Antiguedad del consumidor	< a 1 año	68	1.000	.000
	1 a 5 años	64	.000	1.000
	Más de 5 años	68	.000	.000
X2 - Tipo de industria	Software empresarial	100	1.000	
	Software para juegos	100	.000	
X23 - Consideración de alianza estratégica	NO esta considerado	114	1.000	
	SI está considerado	86	.000	
X4 - Pais	MEX/Norteamérica	81	1.000	
	Fuera de MEX/Norteamérica	119	.000	
X3 - Tamaño de la empresa	PyME (0 to 499)	98	1.000	
	Grande (500+)	102	.000	

Sólo variables categóricas declaradas

Pruebas omnibus sobre los coeficientes del modelo

		Chi cuadrado	gl	Sig.
Paso 1	Paso	68.690	1	.000
	Bloque	68.690	1	.000
	Modelo	68.690	1	.000
Paso 2	Paso	22.110	1	.000
	Bloque	90.800	2	.000
	Modelo	90.800	2	.000
Paso 3	Paso	28.287	2	.000
	Bloque	119.087	4	.000
	Modelo	119.087	4	.000
Paso 4	Paso	17.896	1	.000
	Bloque	136.983	5	.000
	Modelo	136.983	5	.000
Paso 5	Paso	8.452	1	.004
	Bloque	145.435	6	.000
	Modelo	145.435	6	.000

Los valores indican alta significatividad

Resumen del modelo

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	207.287[a]	.291	.388
2	185.178[a]	.365	.488
3	156.891[b]	.449	.600
4	138.995[b]	.496	.663
5	130.543[c]	.517	.690

a. La estimación ha finalizado en el número de iteración 5 porque las estimaciones de los parámetros han cambiado en menos de .001.

b. La estimación ha finalizado en el número de iteración 6 porque las estimaciones de los parámetros han cambiado en menos de .001.

c. La estimación ha finalizado en el número de iteración 7 porque las estimaciones de los parámetros han cambiado en menos de .001.

De acuerdo al criterio de utilidad, se determina mejor ***R*** **cuadrado de Negelkerke** que en este caso el **69%**

SPSS genera tablas Pruebas de ***Hosmer y Lemeshow***, tabla de Clasificación y tabla Variables en la ecuación. Ver **Figura 6.51.**

Fuente: SPSS 20 IBM

Figura 6.51. Pruebas de *Hosmer y Lemeshow*, tabla de Calsificación y tabla Variables en la ecuación.

Prueba de Hosmer y Lemeshow

Paso	Chi cuadrado	gl	Sig.
1	22.701	8	.004
2	8.861	8	.354
3	16.063	8	.041
4	14.244	8	.076
5	25.079	8	.002

En la prueba, hubieron pasos con baja significancia que al ajustarse al paso final, el modelo en general **sí cumple.**

Tabla de clasificación[a]

			Pronosticado		
			X5 - Sistema de distribución		
	Observado		Indirecto a través de terceros	Directo al consumidor	Porcentaje correcto
Paso 1	X5 - Sistema de distribución	Indirecto a través de terceros	79	29	73.1
		Directo al consumidor	31	61	66.3
	Porcentaje global				70.0
Paso 2	X5 - Sistema de distribución	Indirecto a través de terceros	82	26	75.9
		Directo al consumidor	25	67	72.8
	Porcentaje global				74.5
Paso 3	X5 - Sistema de distribución	Indirecto a través de terceros	85	23	78.7
		Directo al consumidor	15	77	83.7
	Porcentaje global				81.0
Paso 4	X5 - Sistema de distribución	Indirecto a través de terceros	96	12	88.9
		Directo al consumidor	16	76	82.6
	Porcentaje global				86.0
Paso 5	X5 - Sistema de distribución	Indirecto a través de terceros	95	13	88.0
		Directo al consumidor	9	83	90.2
	Porcentaje global				89.0

a. El valor de corte es .500

En **Bloque 1,** la **tabla de Clasificación**, nos reporta un **89.0%** de probabilidad de acierto en el resultado de la variable dependiente $\mathbf{X_5}$a nivel de predicción cuando intervienen en ella el resto de las variables de **BM_MKT_Digital,** incrementando del **54%** (**del Bloque 0, sólo $\mathbf{X_5}$**) al **89%**(**resto de las variables de la base de datos**). Valores predichos, muy cercanos a los verdaderos.Ver **Figura 6.48.**

Variables en la ecuación

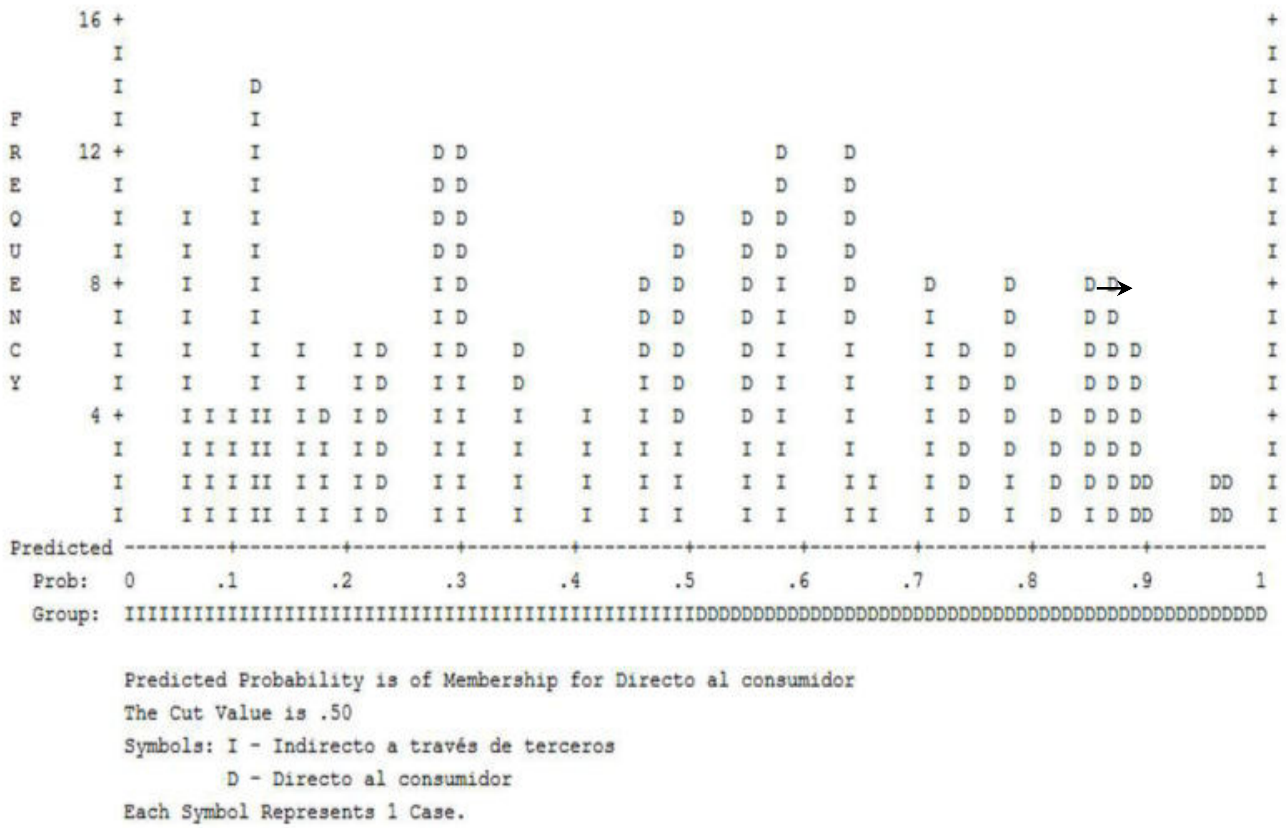
		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95% para EXP(B) Inferior	Superior
Paso 1[a]	X19	1.160	.175	45.606	1	.000	3.253	2.310	4.582
	Constante	-8.426	1.243	45.929	1	.000	.000		
Paso 2[b]	X3(1)	1.808	.421	18.411	1	.000	6.100	2.671	13.932
	X19	1.483	.212	48.727	1	.000	4.406	2.905	6.682
	Constante	-11.530	1.647	48.985	1	.000	.000		
Paso 3[c]	X1			20.984	2	.000			
	X1(1)	3.637	.878	17.138	1	.000	37.972	6.787	212.446
	X1(2)	2.372	.566	17.575	1	.000	10.723	3.537	32.508
	X3(1)	2.842	.549	26.772	1	.000	17.151	5.845	50.333
	X19	2.663	.399	44.458	1	.000	14.343	6.556	31.380
	Constante	-22.248	3.349	44.124	1	.000	.000		
Paso 4[d]	X1			24.783	2	.000			
	X1(1)	4.124	.976	17.863	1	.000	61.777	9.127	418.125
	X1(2)	4.024	.846	22.617	1	.000	55.926	10.651	293.660
	X3(1)	2.483	.595	17.389	1	.000	11.975	3.728	38.465
	X17	-.918	.240	14.646	1	.000	.399	.249	.639
	X19	2.866	.452	40.212	1	.000	17.573	7.246	42.620
	Constante	-20.000	3.643	30.131	1	.000	.000		
Paso 5[e]	X1			25.758	2	.000			
	X1(1)	4.647	1.056	19.368	1	.000	104.303	13.166	826.322
	X1(2)	4.215	.874	23.239	1	.000	67.710	12.200	375.799
	X3(1)	2.940	.643	20.881	1	.000	18.919	5.361	66.770
	X17	-.969	.252	14.799	1	.000	.379	.232	.622

Valores predichos, muy cercanos a los verdaderos. Modelo altamente significativo.

Fuente: SPSS 20 IBM

SPSS genera Gráfico punto de corte. **Ver Figura 6.52.**

Figura 6.52. Gráfico punto de corte



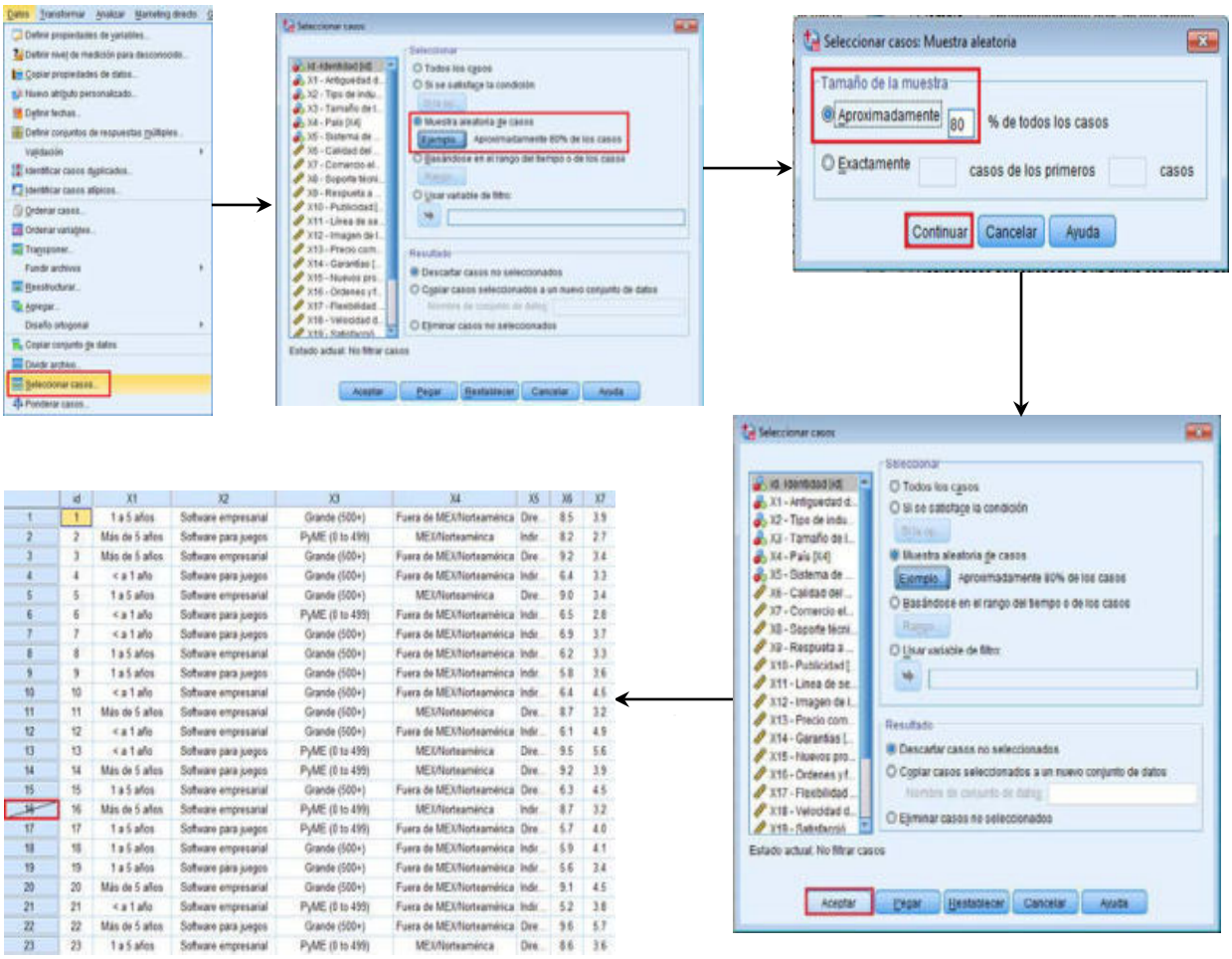
Fuente: SPSS 20 IBM

Paso 1: Objetivos

Problema 8: Valide el poder de predicción de su modelo partiendo del 80% de su muestra

Teclear: Datos->Seleccionar casos->Muestra aleatoria de casos->Ejemplo: Seleccionar tamaño de la muestra: Aproximadamente: 80%->Continuar->Aceptar. Ver Figura 6.53.

Figura 6.53. Proceso de selección del 80% de casos de la base de datos



Fuente: SPSS 20 IBM

Teclear de nueva cuenta, el proceso descrito de la Figura 6.41 y analizar los resultados. Ver Figura 6.54.

Figura 6.54.Tablas Resumen del modelo, Tabla Prueba de *Hosmer y Lemeshow* y Tabla de porcentaje pronosticado vs. calculado
Nota: Resultados de selección del 80% de casos de la base de datos.

Resumen del modelo

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	167.980[a]	.326	.435
2	148.499[b]	.399	.533
3	126.692[b]	.471	.629
4	108.648[c]	.525	.700
5	101.678[c]	.544	.726
6	97.014[c]	.556	.742

Prueba de Hosmer y Lemeshow

Paso	Chi cuadrado	gl	Sig.
1	19.672	8	.012
2	13.624	8	.092
3	6.634	8	.577
4	7.700	8	.463
5	33.950	8	.000
6	59.886	8	.000

Resultados muy similares al 100% de datos de la Figura 6.48.

			Pronosticado		
			X5 - Sistema de distribución		
	Observado		Indirecto a través de terceros	Directo al consumidor	Porcentaje correcto
Paso 1	X5 - Sistema de distribución	Indirecto a través de terceros	66	24	73.3
		Directo al consumidor	21	59	73.8
	Porcentaje global				73.5
Paso 2	X5 - Sistema de distribución	Indirecto a través de terceros	70	20	77.8
		Directo al consumidor	23	57	71.3
	Porcentaje global				74.7
Paso 3	X5 - Sistema de distribución	Indirecto a través de terceros	73	17	81.1
		Directo al consumidor	15	65	81.3
	Porcentaje global				81.2
Paso 4	X5 - Sistema de distribución	Indirecto a través de terceros	81	9	90.0
		Directo al consumidor	13	67	83.8
	Porcentaje global				87.1
Paso 5	X5 - Sistema de distribución	Indirecto a través de terceros	82	8	91.1
		Directo al consumidor	7	73	91.3
	Porcentaje global				91.2
Paso 6	X5 - Sistema de distribución	Indirecto a través de terceros	79	11	87.8
		Directo al consumidor	7	73	91.3
	Porcentaje global				89.4

Fuente: SPSS 20 IBM

Referencias

Crask, M., y Perreault W. (1977), Validation of Discriminant Analysis in Marketing Research. *Journal of Marketing Research 14* (February): 60-68.

Dillon, W. R., y Goldstein M. (1984), Multivariate *Analysis: Methods and Applications*. New York: Wiley.

Gessner, G., Maholtra, N. K., Kamakura, W. A., y Zmijewski M. E. (1988), Estimating Models with Binary Dependent Variables: Sorne Theoretical and Empirical Observations. *Journal of Business Research 16*(1):49-65.

Green, P. E., Tull, D., y Albaum G. (1988), *Research for Marketing Decisions*. Upper Saddle River, N.J.: Prentice Hall.

Green, P. E. (1978*), Analyzing Multivariate Data*. Hinsdale, Ill.: Holt, Rinehart, and Winston.

Green, P. E., y Carroll, J. D. (1978), *Mathematical Tools for Applied Multivariate Analysis*. New York: Academic Press.

Hair , J.F.; Anderson, R.E.; Tatham, R.L.; Black W.C. (1999).*Análisis Multivariante.*5a. Ed. España. Prentice Hall.

IBM (2011a). *IBM SPSS Statistics Base 20.* EUA. .Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Base.pdf

IBM (2011b).*Guía breve de IBM SPSS Statistics 20.* EUA.Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Brief_Guide.pdf

IBM (2011c). IBM SPSS Missing Values 20. EUA. .Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Missing_Values.pdf

Hosmer, D. W., y Lemcshow, S. (1989), *Applied Logistic Regression*. New York: Wiley.

Huberty, C. J. (1984), lssues in the Use and Interpretation of Discriminant Analysis. *Psychological Bulletin* 95: 156-71.

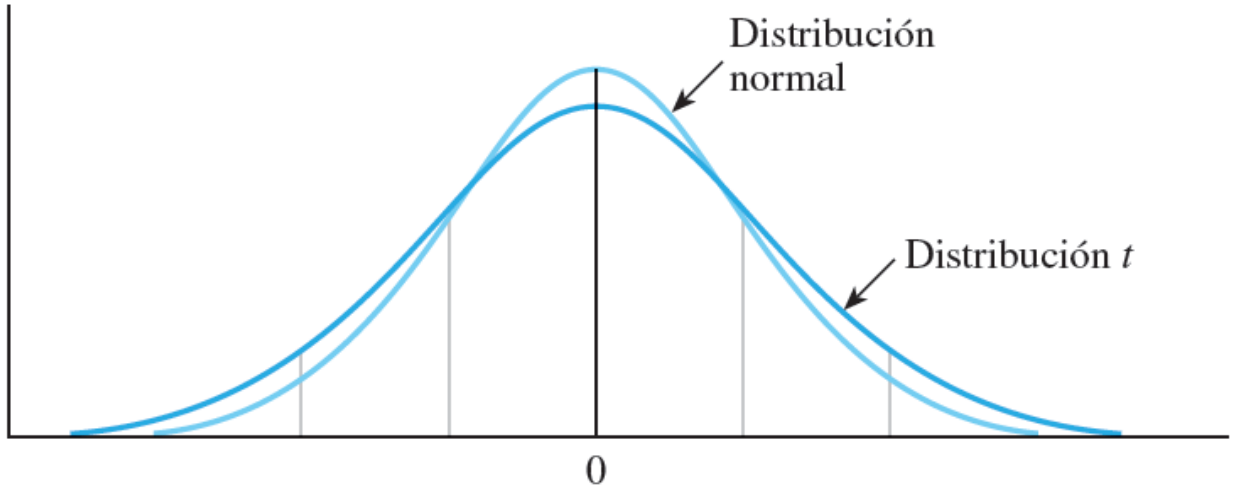
Huberty, C. J., Wisenbaker, J. W. y Smith, J. C. (1987), Assessing Predictive Accuracy in Discriminant Analysis. *Multivariate Behavioral Research* 22 (July): 307-29.

Johnson, N., y Wichem D. (1982), *Applied Multivariate Statistical Analysis.* Upper Saddle River, N.J.: Prentice Hall.

Morrison, D. G. (1967), Ün the interpretation of Discriminant Analysis. *Journal of Marketing Research* 6(2): 156-63.

Perreault, W. D., Behrman, D. N., y Armstrong, G. M. (1979), Aitemative Approaches for Interpretation of Multiple Discriminant *Analysis in Marketing Research. Jornal*

Capítulo 7. Pruebas No Paramétricas de Dos Muestras



7.1. Prueba estadística *t*: ¿Qué es?

Las **prueba *t*** son una de las pruebas más populares para **comparar dos muestras**. Hay muchas situaciones en las que queremos ver si **una manipulación experimental tiene un efecto o no:** ¿un nuevo proceso de gestión del conocimiento mejora el rendimiento de la empresa?, ¿una nueva forma de organización mejora los ingresos? En estos casos estamos comparando un **grupo de control**, que realizan la tarea de la forma habitual, con **un grupo experimental**, realizan la tarea en las mismas condiciones que el **grupo de control** con una excepción, nuestra **manipulación experimenta**l (es decir, el nuevo esquema de enseñanza o la nueva práctica de trabajo). Si el **grupo 2** funciona de manera diferente, entonces podemos atribuir la diferencia al efecto de nuestra manipulación.

Comparamos a menudo dos grupos: ¿varían los hombres y las mujeres en una tarea particular? ¿La tarea se realiza con más precisión con música que sin música? Aquí estamos de nuevo **comparando dos muestras** para ver si una diferencia surge de nuestra manipulación experimental: ¿El género de los gerentes (hombres vs. mujeres) **afecta el desempeño en la tarea**? ¿El tiempo de día (**mañana versus tarde**) afecta la precisión del desempeño en la tarea? En todos estos casos **la prueba *t*** nos permite ver si hay una **diferencia entre los resultados** de los dos grupos. (Hinton et al. 2004; Levin y Rubin, (2004)). Para saber más, vea: IBM 2011a; IBM, 2011b; IBM, 2011c.

Necesitamos tener cuidado de que hemos realizado el estudio apropiadamente **porque la prueba *t* es simplemente una técnica estadística y no nos dirá cuándo hemos cometido un error**. Por ejemplo, **la prueba *t*** nos indicará si hay una diferencia en las los dos grupos, **pero no lo que causó la diferencia.** Si configura el estudio para que sólo la diferencia entre los grupos **sea su manipulación experimental**, entonces Usted puede estar seguro de que esto causó la diferencia en el rendimiento. Sin embargo, si lo diseñó mal y los grupos difieren en un número de maneras que Usted no sabría cuál de estas diferencias es responsable de cualquier diferencia en el desempeño. Necesitamos hacer que los grupos sean seleccionados de modo que la única diferencia entre ellos sea que están interesados en (la manipulación experimental) y no hay otra confusión variable. Además, hay una serie de supuestos que subyacen a la **prueba *t*** y necesitamos asegurarnos de que

nuestros datos cumplen con estos supuestos, de lo **contrario el resultado puede no ser significativo**.

La **prueba *t*** se basa en una serie de supuestos**, ya que es una prueba NO paramétrica:**

• Las **muestras son seleccionadas aleatoriamente e independientemente de sus poblaciones**. Esto es importante, ya que la **prueba *t*** utiliza las muestras para estimar detalles de las poblaciones que (como la media de la población y la varianza, que se denominan **parámetros**) y si las muestras se seleccionan de manera sesgada afectará la precisión de las estimaciones.

Los datos recolectados deben ser de **intervalos o de razón**, de distribuciones continuas y poblaciones normalmente distribuidas. Existe cierto debate sobre la importancia de estos supuestos pero subyacen en la lógica de la prueba. Siempre y cuando las distribuciones sean **aproximadamente normales**, los resultados pueden ser significativos, sobre todo para muestras grandes (**más de 30**), así que Usted puede seguir adelante con la prueba. Sin embargo, violaciones flagrantes de este supuesto, pueden conducir a resultados sin sentido.

• Los datos de las dos muestras provienen de poblaciones con igualdad de varianzas (**homocedasticidad**). La **prueba *t*** utiliza este supuesto en su cálculo al agrupar las varianzas de las dos muestras para estimar la varianza homogénea de la población. **Si el supuesto no es válido, el cálculo no es significativo**. Pero, como la **prueba *t*** "***robusta***", todavía es probable que tenga sentido incluso cuando las variancias tiendan a diferir un "***poco***" (**Por ejemplo, incluso en muestras tan pequeñas como 10, una muestra de varianza de hasta tres veces la otra, la prueba *t* todavía puede interpretarse correctamente**). Por lo tanto, usualmente, se reporta el resultado de la **prueba *t*** para varianzas supuestas como iguales, sin embargo, grandes diferencias en las varianzas de las muestras no deben ser ignoradas. En muchos estudios, se asume que nuestra manipulación tiene un efecto constante, por lo que las varianzas de la muestra deben ser las mismas. Debe advertirse que, cuando encuentre **varianzas desiguales**, si puede entender el por qué las varianzas difieren y tiene sentido, continúe con la **prueba *t*** y posteriormente elija el resultado de la prueba **donde No se asumen varianzas iguales**. (Hinton et al. (2004); Levin y Rubin, (2004))

Esencialmente, la **prueba *t*** compara dos resultados:

1. El de la diferencia entre la media de las dos muestras.
2. La estimación de lo que cabría esperar de la diferencia en medias para cuando la **hipótesis nula es verdadera**, con dos resultados posibles:

-Si la diferencia en las medias es No más grande que la diferencia esperada entonces no podemos rechazar la hipótesis nula, es decir, No se ha encontrado evidencia de que nuestra manipulación experimental esté teniendo un efecto.

-Si la diferencia en las medias es mayor que la diferencia esperada, entonces podemos ver si es suficientemente grande como para rechazar la hipótesis nula y afirmar que nuestra manipulación experimental Está teniendo un efecto estadísticamente significativo. Se realiza cuando **las muestras no están relacionadas**, con diferentes participantes en cada muestra, como las tareas diurnas y nocturnas se discute más adelante. Esta prueba **también se denomina prueba *t* no relacionada** o **prueba *t*** de mediciones independientes (nota: con **muestras relacionadas** tales como el mismo grupo de participantes probado en **2 momentos del día**, debe utilizarse la **prueba *t* de muestras pareadas**.). (Hinton et al. (2004); Levin y Rubin, (2004))

Paso 1: Objetivos

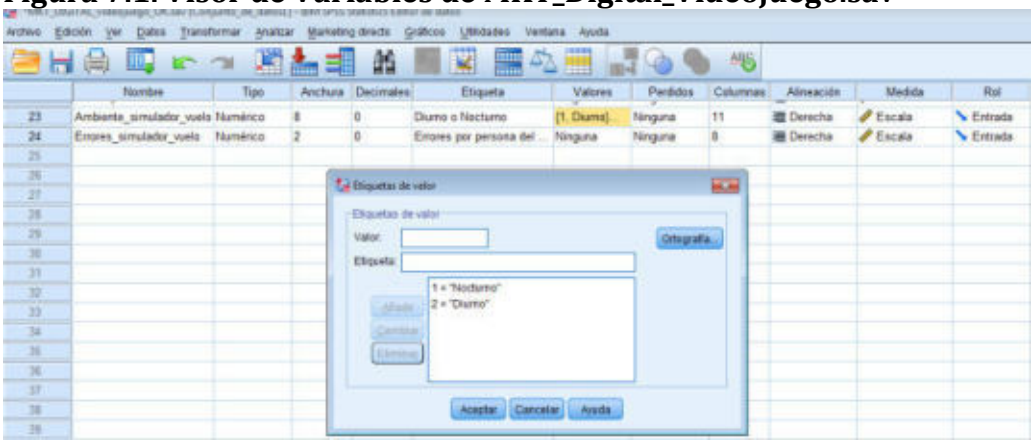
Problema 1: La empresa **MKT Digital** tiene un software de simulación de vuelo y desea investigar el número de errores que produce en los usuarios tanto en los **ambientes diurno y nocturno** para la mejora de producto, utilizando **14 personas divididas en las 2 ambientaciones**. Se predijo que se cometerían más errores en la prueba nocturna. El número de errores cometidos se calculó para cada persona.

H_1= El número de errores que se producen en el simulador de vuelo son iguales en la ambientación diurna y nocturna

H_2= El número de errores que se producen en el simulador de vuelo No son iguales en la ambientación diurna y nocturna y se producen más en la diurna.

H_3= El número de errores que se producen en el simulador de vuelo No son iguales en la ambientación diurna y nocturna y se producen más en la nocturna. **Ver Figura 7.1 y 7.2**

Figura 7.1. Visor de Variables de MKT_Digital_Videojuego.sav



Fuente: SPSS 20 IBM

Figura 7.2 Visor de Datos de MKT_Digital_Videojuego.sav

MKT_DIGITAL_videojuego_OK.sav [Conjunto_de_datos1] - IBM SPSS Statistics Editor de datos

Archivo Edición Ver Datos Transformar Analizar Marketing directo Gráficos Utilidades Ventana Ayuda

14 : Errores_simulador... 6

	Puntaje_Sucursal_Sur	Estimulo_económico1	Estimulo_económico2	Desempeño	Ambiente_simulador_vuelo	Errores_simulador_vuelo
1	47	.	.	.	Nocturno	9
2	49	.	.	.	Nocturno	10
3	46	.	.	.	Nocturno	8
4	48	.	.	.	Nocturno	9
5	51	.	.	.	Nocturno	9
6	48	.	.	.	Nocturno	7
7	50	.	.	.	Diurno	8
8	45	.	.	.	Diurno	7
9	.	.	.	.	Diurno	5
10	.	.	.	.	Diurno	6
11	.	.	.	.	Diurno	5
12	.	.	.	.	Diurno	4
13	.	.	.	.	Diurno	5
14	.	.	.	.	Diurno	6

Fuente: SPSS 20 IBM

Paso 2: Diseño

- En adelante para efectos de aprendizaje, sólo se tomarán de forma directa los datos con el fin de aprender a utilizar la técnica.
- Para muestras independientes tenemos que introducir una variable de agrupación para indicar la muestra y su puntuación. En este caso hemos incluido la variable **Ambiente_simulador_vuelo**, y han dado a los jugadores nocturnos la etiqueta de valor de '1' y a los jugadores diurnos la etiqueta de valor de '2'. El número de errores en el simulador se identifica con la variable **Errores_simulador_vuelo**. Así que la **fila 1** muestra un **jugador nocturno** que hizo **9 errores** y la **fila 7** muestra un **jugador diurno** que hizo **8 errores.**
- **La asignación aleatoria de grupos** debe ser abordada al calcular una **prueba *t*** de muestras independientes Para asegurar esto, la variable independiente debe permitir la **asignación aleatoria**. En el ejemplo anterior, los participantes son aleatoriamente asignados a una condición de ambientación: diurna o nocturna.
- El objetivo es llevar a cabo una investigación de **prueba *t*** de muestras. Todas la estadísticas inferenciales se encuentran en la sección **Analizar** como comando.
- La razón de ser de la **prueba *t*** es probar las diferencias significativas en las medias de dos muestras, por lo tanto, elija **Comparar Medias**.
- Nuestro estudio es de diseño de medición de medidas independiente (comparando del simulador el juego diurno vs juego nocturno), por lo tanto seleccione: **Prueba T para muestras independientes.**
- La variable dependiente es **Errores_simulador_vuelo** y esto debe ser enviado al cuadro **Variable para contrastar.**
- La variable independiente es **Ambiente_simulador_vuelo** y esto debe ser enviado al cuadro **Variable de agrupación.**
- Una vez que realizado, se le pide que defina **Grupos**, que se realiza basado en los valores dados a **Nocturno (1) y Diurno (2).**
- Para completar la prueba, haga ***click*** en **Continuar** y entonces **Aceptar**.

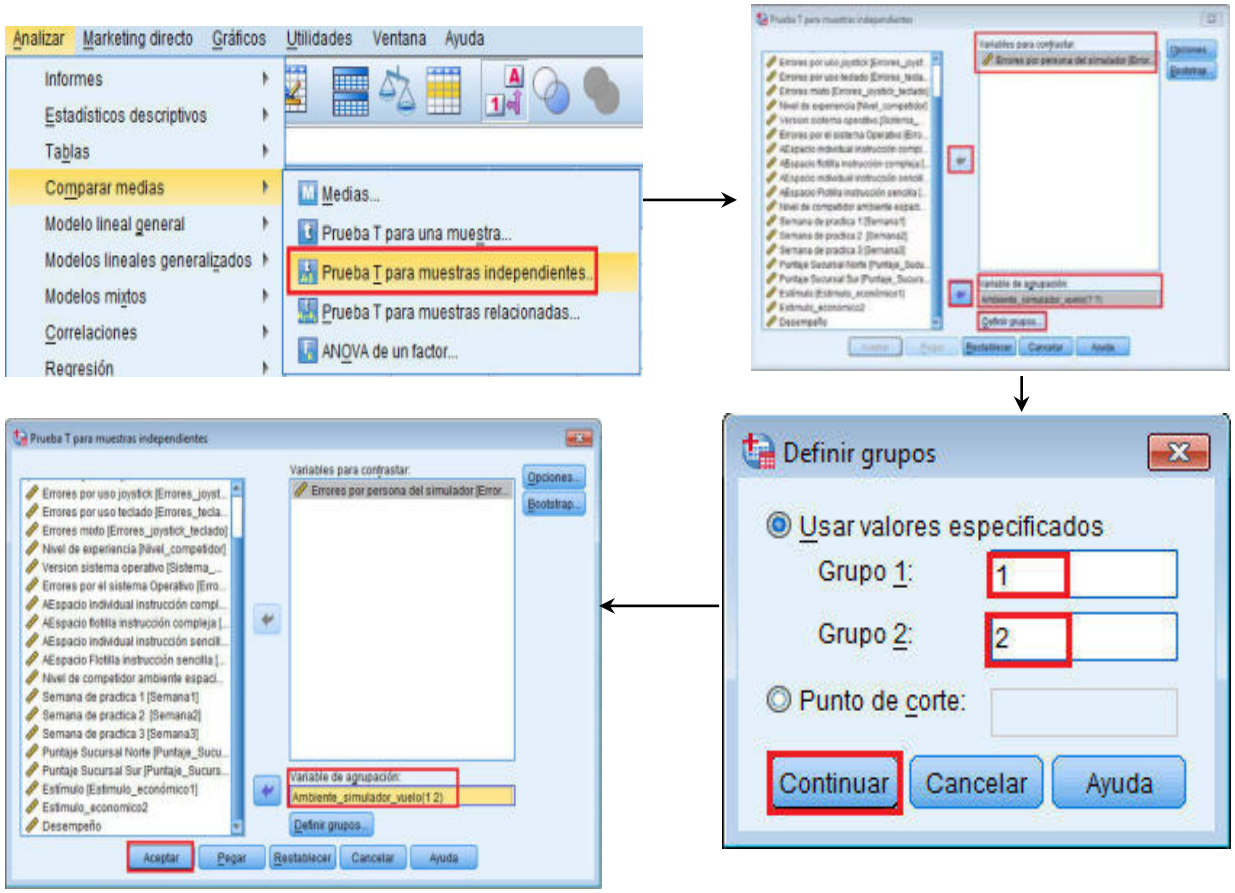
Paso 3: Condiciones de Aplicabilidad

- No olvidar realizar los análisis previos de inspección visual de la base de datos para detectar ausentes y/o atípicos (corregir en su defecto), confiabilidad, normalidad, homocedasticidad y linealidad

Paso 4: Estimación y ajuste

-Teclear: Analizar->Comparar medias->Prueba T para muestras independientes->Variables para contrastar: Errores por persona del simulador->Variable de agrupación: Ambiente del simulador de vuelo-> Definir grupos-> (Diurno: 1, Nocturno: 2) ->Continuar->Aceptar. Ver figura 7.3

Figura 7.3. Proceso para Prueba *t* de muestras independientes



Fuente: SPSS 20 IBM

La primera tabla que genera el **SPSS** es la de **Estadísticos descriptivos**, Ver **Figura 7.4**

Figura 7.4. Tabla de Estadísticos descriptivos

Estadísticos de grupo

	Diurno o Nocturno	N	Media	Desviación típ.	Error típ. de la media
Errores por persona del simulador	Nocturno	6	8.67	1.033	.422
	Diurno	8	5.75	1.282	.453

Variable dependiente **Variable independiente**

Fuente: SPSS 20 IBM

- El número de participantes (**N**) se incluye en los resultados estadísticos descriptivos.
- Observando las medias se puede ver que el juego nocturno produjo más errores en el simulador de vuelos, pero esta diferencia puede no ser significativa. **Para determinar si el resultado es significativo o debido al azar**, la **tabla de Pruebas de muestras independientes** debe ser examinado.

- La desviación estándar muestra que los participantes del **simulador diurno tienen una amplia dispersión que el grupo de operación nocturna.**
- El error estándar de la media es una estimación de la desviación estándar de la distribución muestral de la media basada en la muestra que probamos. **Es decir, es la distancia estándar, o error, que una muestra media es de la media de la población.**
- El **Error típ.de la media** es una cifra útil al utilizarse en el cálculo de los **intervalos de confianza** y las pruebas de significatividad, **como la prueba *t*.**
- En nuestro ejemplo el **Error típ. de la media** muestra que si hubiéramos obtenido las medias de cada muestra de **6 conductores nocturnos** y los analizamos, estimar que la desviación estándar de esos medias sería **0.422. Similarmente**, si tomamos todas las muestras de jugadores de ocho días que estimamos la desviación estándar de los medias, sería de **0.453.**

SPSS genera la tabla de **Prueba de muestras independientes** despliega los resultados de la estadística inferencial. Ver **Figura 7.5.**

Figura 7.5. Pruebas de muestras independientes

Prueba de muestras independientes

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
									95% Intervalo de confianza para la diferencia	
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	Inferior	Superior
Errores por persona del simulador	Se han asumido varianzas iguales	.390	.544	4.560	12	.001	2.917	.640	1.523	4.310
	No se han asumido varianzas iguales			4.712	11.990	.001	2.917	.619	1.567	4.267

Prueba estadística

p valor

Fuente: SPSS 20 IBM

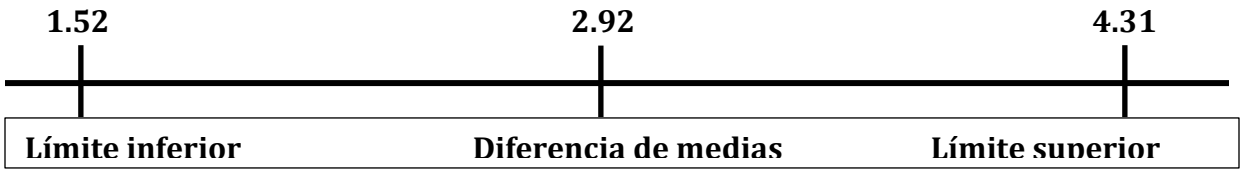
- El ***p* valor** en una tabla de **Prueba de muestras independientes** es para prueba de hipótesis de dos colas, mostrada en la columna '**Sig. (bilateral)**'. Si su prueba de hipótesis es de 1 cola, divida entre 2.
- Se puede ver en la tabla anterior que dos líneas de valores son generadas por **SPSS**:
- La **Prueba de *Levene*** para la igualdad de varianzas que indica cuál debe ser usado. Uno de los criterios para el uso de una **prueba *t* paramétrica** es la suposición de que ambas poblaciones, tienen **homocedsticidad.** Si la **prueba estadística F** tiene significatividad, la **prueba de *Levene*** habrá encontrado que las dos varianzas difieren significativamente, en cuyo caso deberemos **utilizar los valores parte inferior**. En el ejemplo, se puede ver que ***F* = 0.390, *p*=0.544> 0.05** por lo tanto, como las varianzas **NO** son significativamente diferentes, podemos aceptar el supuesto de las varianzas iguales y utilizar los valores de la **parte superior.**
- Si la **prueba estadística de *Levene* es significativa**, es cuestión de juicio académico ya sea aceptar los valores de la parte inferior, o si tomar esta violación de los supuestos de la prueba paramétrica, como una justificación para realizar en su lugar, la prueba no-paramétrica de ***Mann-Whitney U***. Si Usted encuentra una violación inesperada del supuesto de la **homocedasticidad** (por ejemplo, otros estudios de

- este tipo en su campo No lo han encontrado), entonces tenga cuidado al interpretar los resultados de su estudio, y si Usted no reporta un **valor de *t***, asegúrese que sea la opción de "***Variaciones iguales no asumido***" en la tabla **SPSS**.
- Para nuestra **hipótesis de dos colas**, la forma convencional de reportar los hallazgos, es establecer la **prueba estadística (*t*) y su probabilidad en el nivel de significatividad (*p*) elegido**. Por ejemplo, si: ***t* = 4.560** con **12 grados de libertad** y una probabilidad de **0.001**, entonces se reportaría esto como ***t* (12) = 4.560, *p* <0.001.**
- En nuestra prueba, **se pronosticó que se producirían más errores en el simulador de vuelo por prueba nocturna, el cual es una hipótesis de una cola.** Para una **hipótesis de dos colas**, Usted establecerá que habrá diferencia entre los medias pero **NO predecirá la dirección de la diferencia.**
- Si hace una predicción de una **cola**, deberá dividir la **'Sig. (bilateral)** '(***p* valor**) a la mitad. Si existe una diferencia significativa entre las medias necesitará asegurarse de si las medias están mostrando la diferencia en la dirección que usted predijo. **En nuestro ejemplo, debe asegurarse de que la puntuación de las medias de error para los participantes de las pruebas nocturnas del simulador sea de hecho mayor que para los diurnos**
- Nuestra hipótesis es de una cola, por lo que el cálculo de ***p* valor es= (0.001/2)=0.0005.** Como nuestra probabilidad de **0.0005** es menor que nuestro nivel de significatividad de **0.001**, usamos el **signo (<)** para indicar que nuestro resultado es significativo en **0.001**.
- **No se preocupe si tiene un valor t negativo para una hipótesis de dos colas.** Si el valor es positivo o negativo depende de las puntuaciones del grupo en el que fueron ingresadas al principio en la ecuación de la **prueba *t***. En nuestro caso, al ingresar primero los errores de los jugadores del simulador nocturnos, que eran los valores más grandes, **nuestro valor de t era positivo.**
- La **Diferencia de las medias** es la diferencia entre los medias de nuestros dos grupos.
- Suponga que **la hipótesis nula es verdadera**, entonces la diferencia real entre las medias de las poblaciones es **cero.** Si seleccionamos todas las **muestras del tamaño 6 y el tamaño 8** y calculamos la diferencia de sus medias podríamos averiguar cuáles serían las diferencias en las medias originadas por **la casualidad.**
- **El Error típico de la diferencia** estima la desviación estándar de todas las diferencias en las medias de la muestras cuando la **hipótesis nula es verdadera**. Esto es, indica la diferencia de las medias que esperamos se **originen por casualidad,** si la **hipótesis nula es verdadera**. En nuestro caso el **Error típico de la diferencia** se estima en **0.64**. La **prueba *t*** compara la **diferencia en nuestros medias vs diferencia de error estándar**:

  ***t*= (Diferencia de medias/ Error típico de la diferencia)=4.557= 2.917/0.64 aprox. 4.560**
- Como nuestra **diferencia en las medias** de **2.917,** al calcular observamos que es **4.560 veces mayor** que el **Error típico de la diferencia**, entonces nuestra diferencia de medias es lo suficientemente grande como para ser ***p* <0.001.**

- **El 95% del intervalo de confianza para la diferencia,** nos indica que estamos el **95 %** confiados en que la verdadera diferencia de la media de la población estará entre los límites superiores e inferiores.
- **La prueba *t* nos indica si nuestra diferencia es significativa o no**. Sin embargo, **el intervalo de confianza** nos proporciona más información sobre **el tamaño de la diferencia.**
- **El intervalo de confianza** nos proporciona una estimación de la diferencia real en la población. Observando nuestro reporte podemos ver que el límite inferior es **1.52** y el **l**ímite superior es **4.31**, lo que indica que podemos estar seguros de que la verdadera diferencia media de la población se encuentra entre estos dos valores. Por lo tanto, en el peor de los casos, los jugadores del simulador nocturno siguen haciendo **1.5 errores** más que los jugadores del simulador diurno, que sigue siendo un diferencia importante. Ver **Figura 7.6**

Figura 7.6. Valores producto del intervalo de confianza



- **El intervalo de confianza** es usado a menudo como indicador estadístico de la significatividad suplementaria o alternativa, por ejemplo:

  Diferencia de las medias=2.92 (95% CI; 1.52 a 4.31)
- **El intervalo de confianza** predeterminado dado por SPSS es **95%.** Esto puede ser manipulados entrando en el comando **Opciones** en el **procedimiento t de prueba** y seleccionando el nivel requerido. (Hinton et al. (2004); Levin y Rubin, (2004)).

Conclusión: H_2**=** El número de errores que se producen en el simulador de vuelo **No son iguales** en la ambientación diurna y nocturna y se producen más en la diurna.

7.3. Prueba *t* de muestras pareadas. Ejemplo

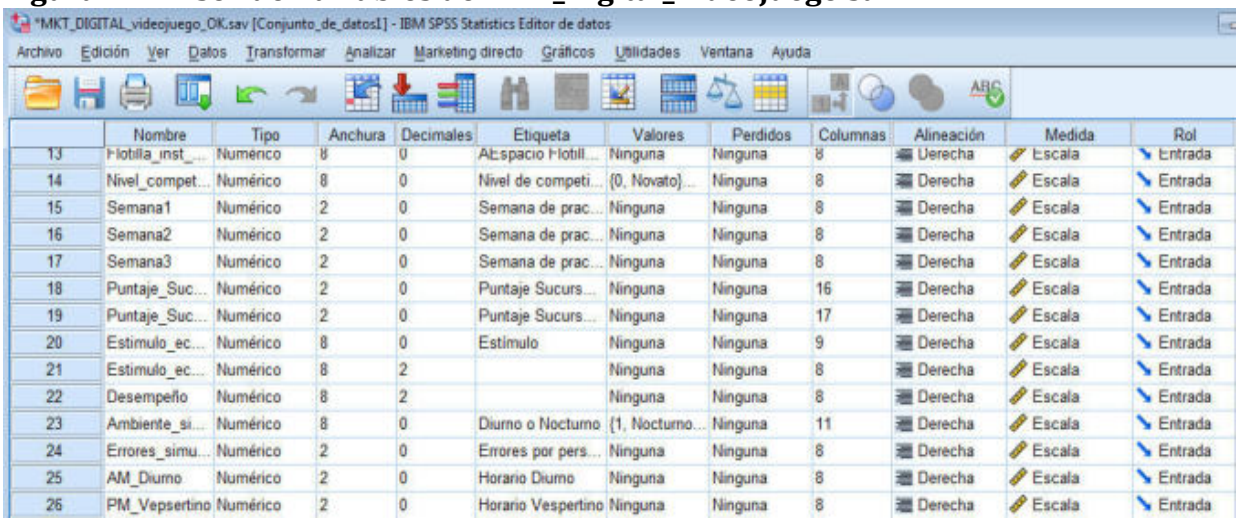
Paso 1: Objetivos

-Problema 1. La empresa **MKT Digital** del mismo software de simulación de vuelo desea investigar el horario en el que los participantes del simulador de vuelo tienen mayores calificaciones clasificando los horarios en: AM (diurno) o PM (nocturno) para la mejora de producto. Utilizando 8 personas, mismas que participaron en cada uno de los horarios. Se predijo que se cometerían más errores en el horario PM (nocturno). La calificación se hace en escala de 0-10 por cada participante.

H_1**=** Con los datos anteriores, **SÍ** estamos en posibilidad de afirmar que hay diferencia de desempeño por los **Horarios Diurno vs Horario Vespertino**

H_2**=** Con los datos anteriores, **NO** estamos en posibilidad de afirmar que hay diferencia de desempeño por los **Horarios Diurno vs Horario Vespertino** Ver **Figuras 7.7 y 7.8**

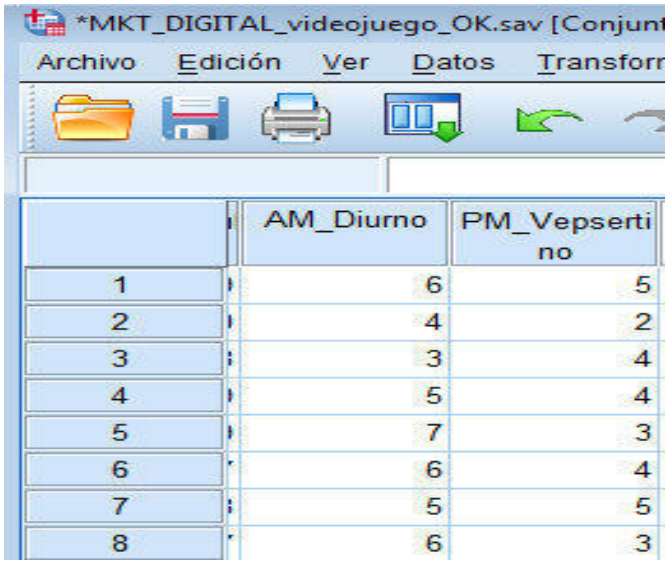
Figura 7.7. Visor de Variables de MKT_Digital_Videojuego.sav



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
13	Motilla_inst_...	Numérico	8	0	Espacio Motill...	Ninguna	Ninguna	8	Derecha	Escala	Entrada
14	Nivel_compet...	Numérico	8	0	Nivel de competi...	{0, Novato}...	Ninguna	8	Derecha	Escala	Entrada
15	Semana1	Numérico	2	0	Semana de prac...	Ninguna	Ninguna	8	Derecha	Escala	Entrada
16	Semana2	Numérico	2	0	Semana de prac...	Ninguna	Ninguna	8	Derecha	Escala	Entrada
17	Semana3	Numérico	2	0	Semana de prac...	Ninguna	Ninguna	8	Derecha	Escala	Entrada
18	Puntaje_Suc...	Numérico	2	0	Puntaje Sucurs...	Ninguna	Ninguna	16	Derecha	Escala	Entrada
19	Puntaje_Suc...	Numérico	2	0	Puntaje Sucurs...	Ninguna	Ninguna	17	Derecha	Escala	Entrada
20	Estimulo_ec...	Numérico	8	0	Estímulo	Ninguna	Ninguna	9	Derecha	Escala	Entrada
21	Estimulo_ec...	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala	Entrada
22	Desempeño	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala	Entrada
23	Ambiente_si...	Numérico	8	0	Diurno o Nocturno	{1, Nocturno...	Ninguna	11	Derecha	Escala	Entrada
24	Errores_simu...	Numérico	2	0	Errores por pers...	Ninguna	Ninguna	8	Derecha	Escala	Entrada
25	AM_Diurno	Numérico	2	0	Horario Diurno	Ninguna	Ninguna	8	Derecha	Escala	Entrada
26	PM_Vepsertino	Numérico	2	0	Horario Vespertino	Ninguna	Ninguna	8	Derecha	Escala	Entrada

Fuente: SPSS 20 IBM

Figura 7.8 Visor de Datos de MKT_Digital_Videojuego.sav



	AM_Diurno	PM_Vepserti no
1	6	5
2	4	2
3	3	4
4	5	4
5	7	3
6	6	4
7	5	5
8	6	3

Fuente: SPSS 20 IBM

- Un **contrapeso** debe aplicarse a las pruebas de muestras pareadas. Para el ejemplo las muestras se balancearon en las dos versiones de la prueba, con la mitad de los jugadores del simulador que se someten a prueba en la mañana y después en la tarde como control de efectos. Aunque esto no se muestra en el conjunto de datos en SPSS, se supone que las muestras son imparciales.

Paso 2: Diseño

- En adelante para efectos de aprendizaje, sólo se tomarán de forma directa los datos con el fin de aprender a utilizar la técnica.

- Ingrese los datos, donde se muestran los contenidos de las muestras como **AM_Diurno y PM_Vespertino y etiquételos como Horario Diurno y Horario Vespertino.**
- El objetivo es llevar a cabo la **prueba *t* de muestras pareadas**, mediante el comando **Analizar**, la justificación es la de observar la significatividad de las diferencias entre las medias de las dos muestras, por lo tanto, seleccione **Comparar Medias.**
- El diseño del estudio es por mediciones repetidas **es decir,** por comparación de los puntajes de un grupo de participantes bajo dos condiciones, por lo tanto seleccione **prueba T para muestras relacionadas.**

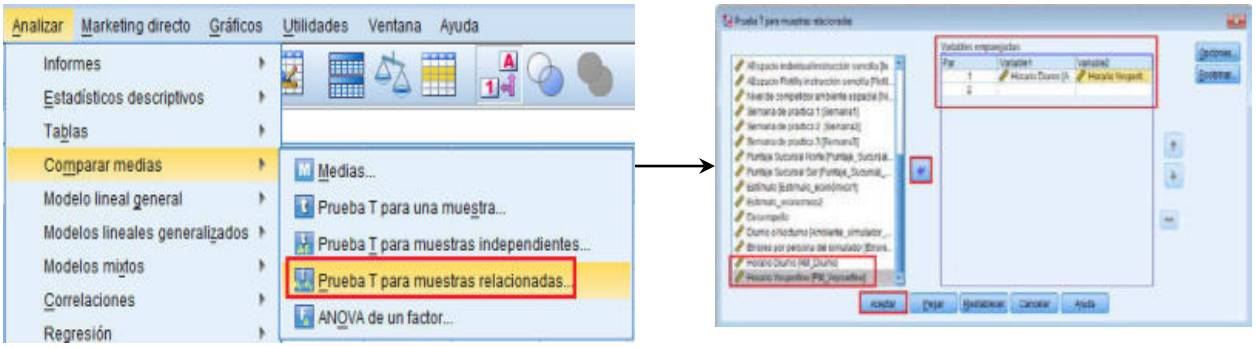
Paso 3: Condiciones de Aplicabilidad

- No olvidar realizar los análisis previos de inspección visual de la base de datos para detectar ausentes y/o atípicos (corregir en su defecto), confiabilidad, normalidad, homocedasticidad y linealidad
- Si **SPSS** no le permitiera seleccionar ambas variables al mismo tiempo, haga ***click*** en una variable primero, suelte el botón del ratón y luego haga ***click*** en la segunda variable. Ambas variables deberían estar ahora visibles en el cuadro de diálogo y SPSS ahora permitirá que se envíen al **cuadro de Variables emparejadas**. La razón por la que **SPSS** sólo permitiera que las variables sean enviadas a través de pares es porque **puede haber más de dos variables para elegir y ya que le permite ejecutar múltiples pares para pruebas de muestras**. Sólo asegúrese de seleccionar en **SPSS** las que Usted requiere ya que **si desea realizar más de una prueba *t* a la vez**, existen varias otras combinaciones de los pares se pueden enviar sobre este punto. (Hinton et al. (2004); Levin y Rubin, (2004)).
- A menudo puede ser difícil seleccionar dos variables que no están próximas entre sí. **Haga *click* en una variable, mantenga presionada la tecla Ctrl y haga *click* en la otra variable**

Paso 4: Interpretación y ajuste

-Teclear: Analizar->Comparar medias->Prueba T para muestras relacionadas->Variables emparejadas: Horario Diurno (AM) ->Horario Vespertino (PM) ->Aceptar. Ver figura 7.9

Figura 7.9. Proceso para Prueba *t* de muestras pareadas



Fuente: SPSS 20 IBM

- La primera tabla producida por **SPSS** es la de **Estadísticos de muestras relacionadas** que detalla las estadísticas descriptivas. Ver **Figura 7.10.**

Figura 7.10. Tabla de Estadísticos de muestras relacionadas

Estadísticos de muestras relacionadas

		Media	N	Desviación típ.	Error típ. de la media
Par 1	Horario Diurno	5.25	8	1.282	.453
	Horario Vespertino	3.75	8	1.035	.366

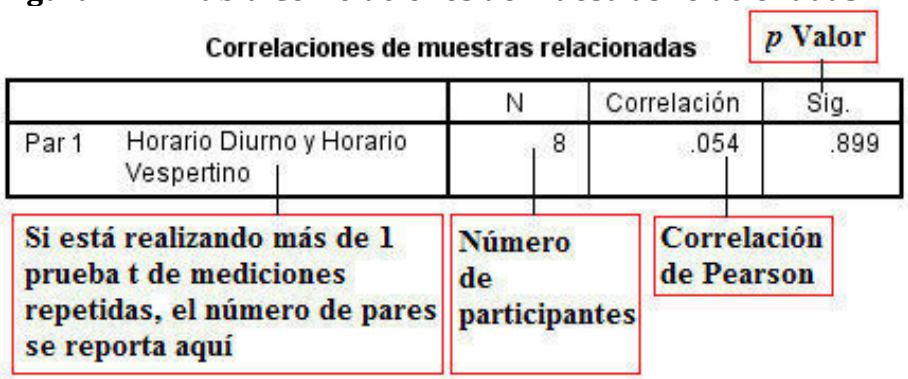
Variable independiente **Puntaje de la variable dependiente**

Fuente: SPSS 20 IBM

- El número de participantes (**N**) se incluye en los resultados de la estadística descriptiva.
- Al observar las **Medias** se puede apreciar que cuando los jugadores del simulador lo hicieron en un **Horario Diurno (AM)** mañana tuvieron mejores calificaciones (**5.25**) que cuando una prueba similar fue realizada en la tarde (**3.75**). **Estas diferencias parecen apoyar nuestra hipótesis**, pero para determinar si este resultado es significativo o debido a la **casualidad o el azar** de las muestras pareadas, se deben realizar más análisis en los resultados.
- La **Desviación típ**. (deviación estándar), muestra que la dispersión de las puntuaciones en el **Horario Diurno (1.282)** es ligeramente mayor el **Horario Vespertino (1.035)**
- El **Error típ. de la media** es una estimación de la **Desviación típ**. (desviación estándar) de la distribución muestral de la media. **Un valor pequeño nos dice que esperaríamos una media similar si hiciéramos la prueba de nuevo y valor grande nos indica una gran variabilidad predicha en la media**. Así, si fuéramos capaces de probar todas las muestras de tamaño **N = 8 de prueba en el Horario Diurno** y graficar sus **medias** estimaríamos que su distribución tendría una desviación estándar de **0.453**, y para muestras de tamaño **N = 8 de prueba en el Horario Vespertino** se estimaría que la distribución de sus **medias** tendría una desviación estándar de **0.366.**

- El **Error típ. de la media** (error estándar de la media) es una cifra útil al utilizarse en el cálculo de las **pruebas de significatividad que comparan las medias**, tales como la **prueba *t*,** y en el cálculo de los **intervalos de confianza.**
- La siguiente tabla que reporta el **SPSS** es la de **Correlaciones de muestras relacionadas**. Ver **Figura 7.11**
-

Figura 7.11. Tabla Correlaciones de muestras relacionadas.



Fuente: SPSS 20 IBM

- La **tabla Correlaciones de muestras relacionadas** muestra el **coeficiente de correlación de *Pearson*** y su valor de significatividad. Esta prueba se realiza para demostrar si los **resultados encontrados son consistentes**. Estamos prediciendo que un cambio en el horario, tendría el mismo efecto en todos los participantes, es decir, **todos serán peores en Horario Vespertino por la misma cantidad**. Habría un efecto coherente sobre ellos: esperaríamos que un participante quien se desempeñó mejor que el promedio en la prueba del **Horario Diurno** todavía sería mejor que el promedio en el **Horario Vespertino** y alguien cerca del fondo del grupo en el **Horario Diurno** todavía estaría cerca del fondo del grupo por el **Horario Vespertino.**
- De nuestro ejemplo ***r* = 0.054, p = 0.899**, el cual **NO** es significativo ya que ***p*> 0.05**. Por lo tanto, nuestros participantes **NO** se comportan de forma consistente, ya que las puntuaciones de su prueba en **Horario Diurno** no se correlacionan significativamente con sus puntuaciones en la prueba de **Horario Vespertino**. Aunque **nuestros participantes se comportan de manera inconsistente**, vale la pena considerar la diferencia en las medias. Esta diferencia nos indicará si hay una caída general en las puntuaciones entre el **Horario Diurno** y el **Horario Vespertino**, sin embargo, no será tan fácil de interpretar con participaciones no consistentes.
- Un **diagrama de dispersión** mostraría los datos en su forma gráfica. También podría utilizarse para comprobar datos atípicos, ya que estos son los pueden hacer que los datos inconsistentes aparezcan más consistentes.
- La siguiente tabla que el **SPSS** reporta es la de **Prueba de muestras relacionadas,** la cual nos informa si existe o no, una diferencia significativa entre nuestras medias. Ver **Figura 7.12.**

Figura 7.12. Tabla Prueba de muestras relacionadas.

Prueba de muestras relacionadas — **Prueba estadística** — ***p* Valor**

		Diferencias relacionadas							
					95% Intervalo de confianza para la diferencia				
		Media	Desviación típ.	Error típ. de la media	Inferior	Superior	t	gl	Sig. (bilateral)
Par 1	Horario Diurno - Horario Vespertino	1.500	1.604	.567	.159	2.841	2.646	7	.033

Fuente: SPSS 20 IBM

- El ***p* valor** en una tabla de **Prueba de muestras relacionadas** es para una **hipótesis de dos colas** mostrado en la columna "***Sig. (bilateral)***". Si su hipótesis es de una cola, realice una división de esta cifra entre 2.
- La forma convencional de informar los hallazgos es declarar la prueba estadística (***t***), los grados de libertad (**gl**) y valor de probabilidad (***p***), de la siguiente manera:

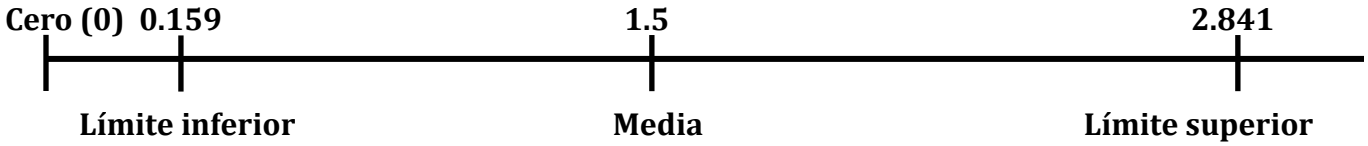
t (7)=2.646; p<0.05

- Recuerde revisar las puntuaciones medias de la **tabla de Estadística descriptiva** para asegurar que la diferencia significativa vaya en la dirección prevista en la hipótesis.
- Las **Diferencias relacionadas** entre parejas muestran las diferencias entre las puntuaciones de nuestras dos muestras.
- La **Media** muestra la **diferencia entre las medias de las dos muestras** (**5.25 -3.75 = 1.50**).
- Como nuestras muestras están pareadas, podemos encontrar una diferencia de puntuación para cada participante, restando sus puntuaciones de la primera muestra de sus puntuaciones de la segunda muestra.
- La **Desviación típ.** indica la desviación estándar de las diferencias de las puntuaciones.
- El **Error típ. de la media** estima la desviación estándar de todas las diferencias entre muestras la media de las muestras para muestras de tamaño **N = 8 cuando la hipótesis nula es verdadera**. Esto indica la **diferencia en las medias que esperaríamos por el azar si la hipótesis nula es verdadera**. Nuestra diferencia de medias es de **1.50**, que es mucho mayor que el **Error tip. de la media** de **0.567**, lo que sugiere que los datos **NO** apoyan la hipótesis nula. Nuestro valor t calculado es la relación de estos dos valores:

t= (1.50/0.567)=2.6

- **El 95% del intervalo de confianza para la diferencia,** nos indica que estamos el **95 %** confiados en que la verdadera diferencia de la media de la población estará entre los límites superiores e inferiores. La diferencia de media muestral se sitúa entre estos dos valores.
- **El intervalo de confianza no incluye cero** y, por lo tanto, incluso en el peor caso (para nuestra predicción), en el extremo inferior del intervalo de confianza, todavía hay una expectativa en la diferencia, aunque pequeña. Ver **Figura 7.13.**

Figura 7.13. Valores producto del intervalo de confianza.



- Esta medida se utiliza a menudo como un indicador **complementario o alternativo** de significatividad. Una forma sugerida de reportar estos hallazgos es la siguiente:

Diferencia en las medias *= 1.50 (95% IC: 0.16 a 2.84)*

- El intervalo de confianza predeterminado dado por **SPSS es 95 %**. Esto puede ser Cambiado al entrar en el **comando Opciones** en el procedimiento de **prueba *t*** y seleccionar el nivel requerido.

- El resultado de nuestras estadísticas podría seguir una de las 4 rutas:

1. Ambas, la correlación y la prueba estadística t son significativas, lo cual indica una consistente y significativa diferencia. Esto es fácil de explicar, ya que es el efecto que estamos prediciendo.

2. La correlación no es significativa, pero la prueba estadística t es significativa, lo cual indica que si bien se ha encontrado una diferencia significativa entre las puntuaciones de las medias, la diferencia en el puntaje de cada participante NO son consistentes. Esto es más difícil de explicar. Predijimos que habría diferencias entre las medias (las cuales encontramos), pero esto no puede atribuirse a un efecto consistente de la variable independiente (Horario Diurno/ Horario Vespertino) sobre la participantes, por lo que otros factores pueden estar influyendo en nuestro resultado.

3. La correlación es significativa pero la prueba estadística t no es significativa. Esto es fácil de explicar. Los participantes se comportan consistentemente y registran puntajes muy similares en las dos muestras, así que **NO** hemos encontrado el efecto que estamos prediciendo.

4. Tanto la correlación como la prueba estadística t no son significativas. Esto puede no ser fácil de explicar, pero con los participantes comportándose de manera inconsistente a través de las dos muestras y no haber diferencia significativa en sus puntuaciones de las medias, resulta que no vale la pena preocuparse por el estudio.

Conclusión: ya que en nuestro caso *r* = 0.054, *p* = 0.899, el cual **NO** es significativo ya que ***p> 0.05 y t (7)=2.646; p<0.05*****,** caemos en el **CASO 2,** por lo tanto, nuestros participantes **NO** se comportan de forma consistente, ya que las puntuaciones de su prueba en **Horario Diurno** no se correlacionan significativamente con sus puntuaciones en la prueba de **Horario Vespertino**. Aunque **nuestros participantes se comportan de manera inconsistente**, vale la pena considerar la diferencia en las medias. Esta diferencia nos indicará si hay una caída general en las puntuaciones entre el **Horario Diurno** y el **Horario Vespertino**, sin embargo, no será tan fácil de interpretar con participaciones no consistentes.

***H_2*=** Con los datos anteriores, **NO** estamos en posibilidad de afirmar que hay diferencia de desempeño por los **Horarios Diurno vs Horario Vespertino**

7.4. Pruebas No Paramétricas de Dos Muestras: ¿Qué es?

Cuando deseamos comparar dos muestras, se suele escoger la ***prueba estadística t***. Sin embargo, **ésta es una prueba paramétrica** que hace ciertos supuestos sobre nuestros datos que cuando NO se cumplen, la **prueba *t*** puede no ser apropiada para realizar **pruebas no paramétricas** ya que no requieren estas suposiciones. Las situaciones en las que podríamos usar una **prueba no paramétrica son cuando nuestros datos son más ordinales que de intervalo**. Si medimos situaciones tales como el **tiempo, la velocidad o la precisión**, podemos estar seguros que los datos **son intervalos**. Esto significa que estamos midiendo nuestros resultados en una **escala con iguales intervalos**. Por ejemplo, la diferencia entre 10 y 15 segundos es de 5 segundos y La diferencia entre 20 y 25 segundos es también 5 segundos. Un segundo es siempre se mide igual, dondequiera que se aplique la escala. Sin embargo, podemos generar datos que no sean intervalos. Por ejemplo, cuando solicitamos que un cliente juzgue la cortesía de los empleados en una oficina de gobierno o un gerente califique el potencial de liderazgo del personal. El cliente o el gerente no son como relojes o velocímetros, en la que basen sus percepciones para realizar las mediciones. En estas calificaciones no posible confiar más que en el orden de las calificaciones y el por qué son utilizadas las **pruebas no paramétricas** ya que no se analizan las puntuaciones directas, sino el rango de los datos y su clasificación.

7.5. Prueba *Mann–Whitney* para muestras independientes: ¿Qué es?

La **prueba de *Mann-Whitney*** es un equivalente no paramétrico de la **prueba *t*** de muestras independientes. Utilizamos una prueba no paramétrica cuando no se cumplen los supuestos de la ***prueba t***. La ***prueba t*** requiere datos de intervalo y que las muestras provengan de poblaciones normalmente distribuidas, con variaciones iguales en ambos grupos. **El uso más común de la prueba de *Mann-Whitney* es cuando nuestros datos son ordinales**. Una escala de calificación se suele tratar como ordinal, sobre todo si le preocupa que los participantes no utilicen toda la gama de la escala. Si Usted está interesado en ver si existe una diferencia entre hombres y mujeres en qué medida son felices en su ciudad en la que viven, podríamos darles una escala de 0 a 10 para determinarlo. Decidimos que la escala no es un intervalo ya que nuestros participantes (de forma cortés) no calificaran a la ciudad muy baja, incluso de vivir de forma infeliz en ella. Como no estamos suponiendo que las calificaciones vienen de una escala de intervalo no vamos a calcular medias y desviaciones estándar de los datos brutos (porque no creemos que esto nos dará valores significativos). Así que lo primero que hace la prueba de ***Mann-Whitney*** es clasificar al conjunto completo de puntajes de menor a mayor. Si todas las mujeres de la ciudad dan un puntaje alto, y los hombres no, se tendría ya una expectativa de calificación alta de las mujeres, baja por los hombres. Si los hombres prefirieran a la ciudad más que las mujeres entonces, se tendría una expectativa contraria. Si no hubiera diferencia entre hombres y mujeres, se esperaría que los resultados estuvieran dispersos en el rango de clasificación.

El problema surge cuando hay una cierta separación de las muestras (por ejemplo, las mujeres suelen dar calificaciones más altas), pero también hay una mezcla de las muestras entre las calificaciones (por ejemplo, algunos hombres también valoran altamente la ciudad). La prueba ***Mann-Whitney*** nos proporciona una estadística que nos permite decidir cuándo podemos reclamar una diferencia entre las muestras (en un nivel de

significatividad elegida). **Calcula dos valores *U***: una para los hombres y otra para las mujeres, produciendo:

1. **Si ambos valores *U* son casi iguales**, significa que las muestras están muy mezcladas entre las calificaciones y no tenemos ninguna diferencia entre ellos.
2. **Si un valor *U* es grande y el otro pequeño** entonces esto indica una separación de los grupos entre las calificaciones. De hecho, si el menor de los dos valores de **U es cero**, indica que toda una muestra está en las calificaciones superiores y toda la otra muestra en las calificaciones inferiores **Sin superposición en absoluto**. Para probar la significatividad de nuestra diferencia tomamos el más pequeño de los dos **valores de U** y al examinar la probabilidad de obtener este valor cuando no hay diferencia entre los grupos. **Si esta probabilidad es menor que nuestro nivel de significatividad (*p* <0.05 usualmente) podemos rechazar la hipótesis nula y reclamar una diferencia significativa entre nuestras muestras.**

Con muestras grandes la **distribución de *U*** se aproxima a la **distribución normal** así que **SPSS** también reporta **una puntuación *z***, también como el menor de los **dos valores *U***. Tenemos que tener cuidado si obtenemos varias calificaciones ligadas, particularmente con pequeños tamaños de muestra y que estos hacen que el resultado de la prueba de ***Mann-Whitney* sea menos válido**.

7.6. Prueba *Mann–Whitney* para muestras independientes: Ejemplo

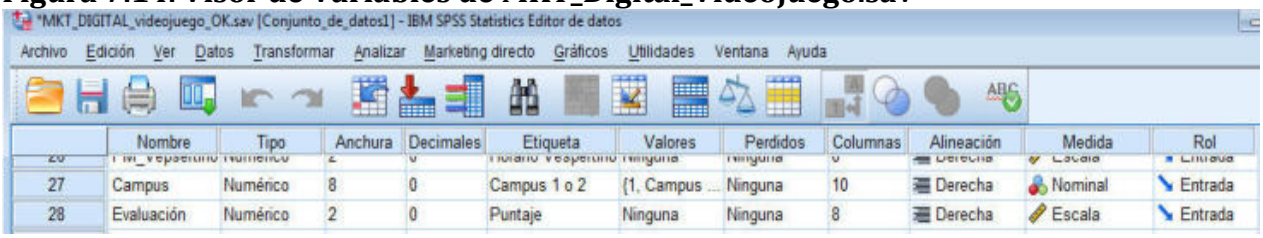
Paso 1: Objetivos

-Problema 1: La empresa **MKT Digital** cuenta con dos sedes en las cuales, ha dispuesto de mejoras ambientales así como de infraestructura para que sus programadores trabajen mejor y esto incremente la innovación y productividad, por lo que decide realizar una investigación para averiguar cuánto disfrutaban los empleados de dichas instalaciones, pidiendo se calificara su disfrute, en una escala de 0 a 100. Los miembros del campus 2 quieren verse como personas muy innovadoras y productivas por lo que se espera califiquen lo más alto su campus. **Ver figura 7.14** y **7,15**

H_1**=** El campus 1 disfruta más sus instalaciones por lo que es el más innovador y productivo

H_2**=** El campus 2 disfruta más sus instalaciones por lo que es el más innovador y productivo

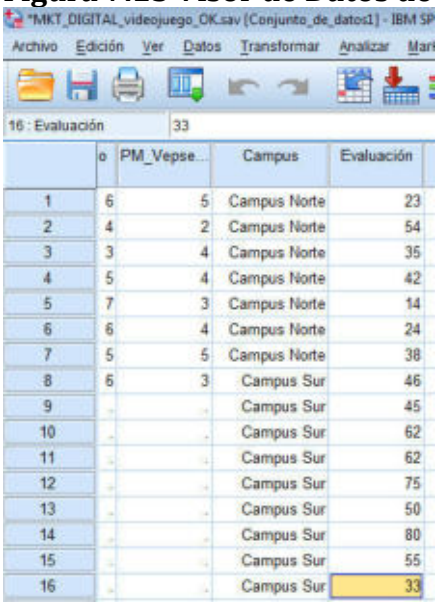
Figura 7.14. Visor de Variables de MKT_Digital_Videojuego.sav



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
27	Campus	Numérico	8	0	Campus 1 o 2	{1, Campus ...	Ninguna	10	Derecha	Nominal	Entrada
28	Evaluación	Numérico	2	0	Puntaje	Ninguna	Ninguna	8	Derecha	Escala	Entrada

Fuente: SPSS 20 IBM

Figura 7.15 Visor de Datos de MKT_Digital_Videojuego.sav



	o	PM_Vepse...	Campus	Evaluación
1	6	5	Campus Norte	23
2	4	2	Campus Norte	54
3	3	4	Campus Norte	35
4	5	4	Campus Norte	42
5	7	3	Campus Norte	14
6	6	4	Campus Norte	24
7	5	5	Campus Norte	38
8	6	3	Campus Sur	46
9	.	.	Campus Sur	45
10	.	.	Campus Sur	62
11	.	.	Campus Sur	62
12	.	.	Campus Sur	75
13	.	.	Campus Sur	50
14	.	.	Campus Sur	80
15	.	.	Campus Sur	55
16	.	.	Campus Sur	33

Fuente: SPSS 20 IBM

Paso 2: Diseño

- Al introducir los datos, debido al diseño la independencia de grupos, una columna será la variable de agrupación y la otra columna, la calificación de cada persona.
- Seleccione Analizar y Pruebas no paramétricas del menú emergente, cuadro de diálogo antiguos
- **SPSS** refiere la prueba de **Mann–Whitney** como **2 muestras independientes**.
- A través del botón Opciones, Usted puede seleccionar el cálculo de las medias y las desviaciones estándar de las variables. Si bien estas estadísticas descriptivas se pueden realizar en conjuntos de datos ordinales, recomendamos sea precavido al examinarlos e interpretarlos.

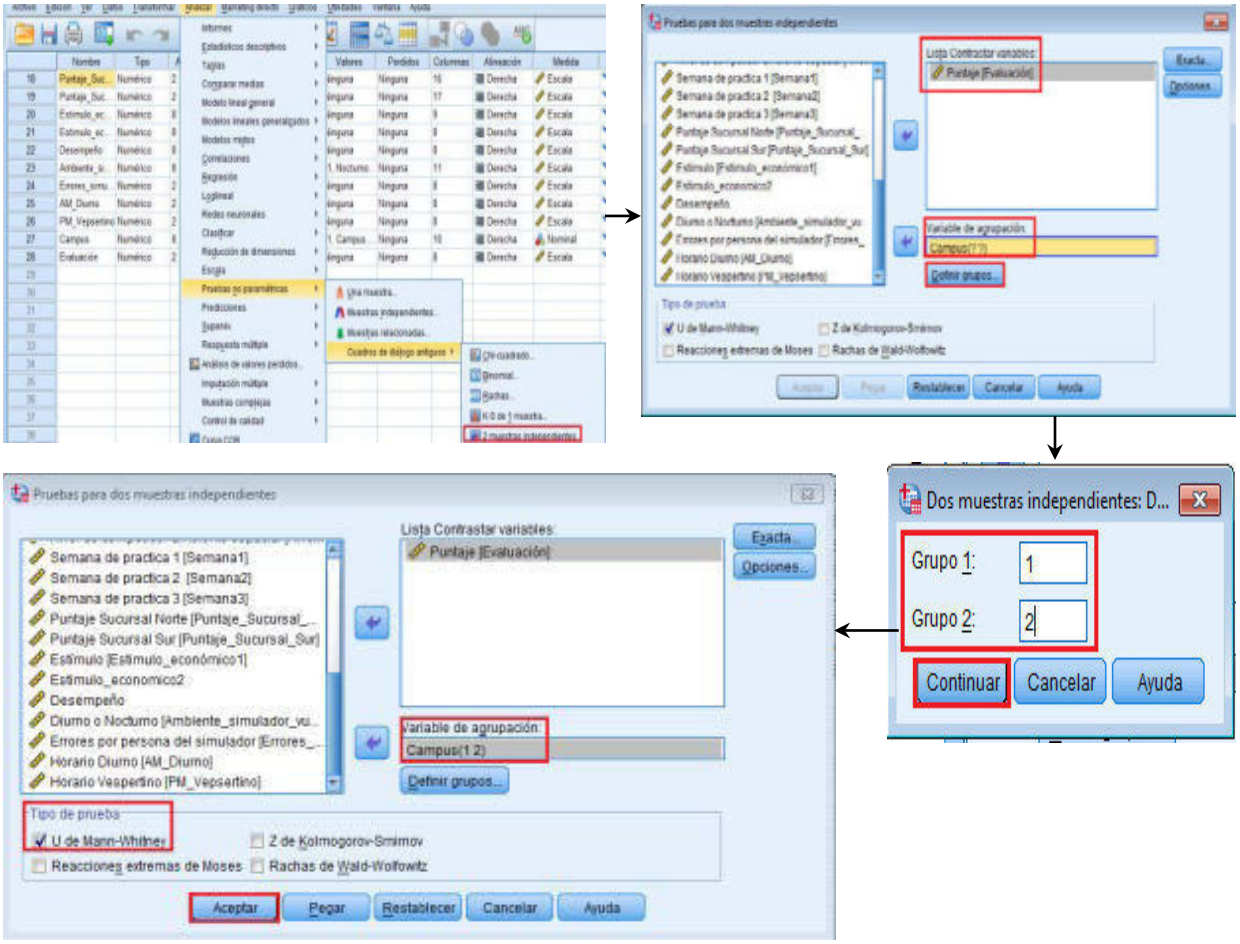
Paso 3: Condiciones de Aplicabilidad

- No olvidar realizar los análisis previos de inspección visual de la base de datos para detectar ausentes y/o atípicos (corregir en su defecto), confiabilidad, normalidad, homocedasticidad y linealidad.

Paso 4: Interpretación y Ajuste

-Teclear: Analizar->Pruebas no paramétricas->cuadro de diálogo antiguos->2 muestras independientes->Lista contrastar variables (intervalo): Puntaje (evaluación) ->Variable de agrupación (ordinal): Campus->Definir grupos: (1,2) ->Continuar->Tipo de prueba: U de Mann-Whitney->Aceptar. Ver Figura 7.16.

Figura 7.16. Proceso para Prueba *Mann-Whitney* para muestras independientes



Fuente: SPSS 20 IBM

- Es poco probable que requiera el uso de otros tipos de prueba que la ***U Mann-Whitney***. Sin embargo, puede utilizar relacionado a esta prueba:

1. La **prueba de *Kolmogorov-Smirnov*,** que se utiliza para determinar si los dos conjuntos de puntuaciones provienen de una misma distribución (normal).
2. La **prueba de corridas de *Wald-Wolfowitz*** que se usa para mostrar cuántas ***"corridas"*** tenemos en el ordenamiento de calificaciones los datos (**una corrida es una serie de calificaciones del mismo grupo**). A medida que avanzamos por las calificaciones, podemos ver si las calificaciones consecutivas provienen del mismo grupo. Si un grupo tiene todas las calificaciones inferiores y el otro grupo tiene todas las calificaciones superiores, sólo tenemos **dos corridas**. En nuestro caso de ejemplo, el número de corridas es de 6 indicando alguna mezcla de los grupos. **Con 16 participantes, la peor mezcla nos daría 16 corridas**
3. La prueba de reacciones extremas de Moisés, que toma uno de los grupos como **grupo de contro**l y a un segundo grupo como **grupo experimental**, y verifica si el grupo experimental tiene **valores más extremos** que el grupo control.

- La primera tabla **Rangos** generada por **SPSS** que reporta el rango de las medias así como las sumas de las mismas, por cada grupo. Ver Figura **7.17.**

Figura 7.17. Tabla Rangos

Prueba de Mann-Whitney

Rangos

	Campus 1 o 2	N	Rango promedio	Suma de rangos
Puntaje	Campus Norte	7	5.00	35.00
	Campus Sur	9	11.22	101.00
	Total	16		

Fuente: SPSS 20 IBM

- **N** indica el número de participantes en cada grupo, así como el total del mismo.
- El **Rango promedio** indica el rango promedio de puntajes dentro de cada grupo.
- La **Suma de rangos** indica la suma total de todos los rangos dentro de cada grupo.
- Si no hubieran diferencias entre las calificaciones de los grupos, por ejemplo, **si la hipótesis nula fuera cierta**, cabría esperar que el **Rango medio y la Suma de rangos sean aproximadamente iguales en Los dos grupos.**
- Podemos ver en nuestro ejemplo, que los dos grupos no parecen ser iguales en sus puntajes.
- Con el fin de determinar si la diferencia en estas calificaciones es significativa, la prueba de ***Mann-Whithey*** se debe analizar la tabla de Estadísticos de contraste Ver Figura **7.18.**

Figura 7.18. Tabla Estadísticos de contraste

Estadísticos de contraste[a]

	Puntaje
U de Mann-Whitney	7.000
W de Wilcoxon	35.000
Z	-2.595
Sig. asintót. (bilateral)	.009
Sig. exacta [2*(Sig. unilateral)]	.008[b]

a. Variable de agrupación: Campus 1 o 2

b. No corregidos para los empates.

Fuente: SPSS 20 IBM

- Como se observa, la prueba estadística que se reporta generalmente es la ***Mann–Whitney U***, con **7.000** en nuestro caso
- El valor de probabilidad se asegura analizando la **Sig. exacta [2*(Sig. unilateral)],** donde una cifra de menos de **0.05** es considerado a ser indicativo de diferencias significativas. Así, se tiene ***U* = 7.00; *p* = 0.008**
- **Conclusión: existe una diferencia significativa entre los puntajes de Campus Norte y Campus Sur**
- Debido a que nuestra hipótesis experimental fue de una cola, el ***p* valor** debería partirse a la mitad, para checar que la diferencia está en la dirección correcta, así que:

  ***U* = 7.00; *p* = 0.0045 < 0.001**
- Podemos ver, examinando los rangos medios que **Campus Sur** reporta una clasificación media más alta que **Campus Norte**, apoyando así nuestra hipótesis.
- **El estadístico de prueba *Mann-Whitney U*** debe ser reportado para muestras de **N <20 en cada grupo**.
- **La puntuación *z* debe ser reportada cuando N> 20 para ambos grupos,** ya que la distribución de U se aproxima a la **distribución normal**, particularmente para **tamaños de muestra de 20 y mayores**
- **SPSS** también genera la **prueba estadística de *Wilcoxon W***, que puede usarse si las **poblaciones a ser comparadas NO son normales**. En lugar de clasificar los dos grupos por separado, la prueba combina los dos grupos en uno para fines de clasificación, luego compara el total de la clasificación de cada grupo para determinar si son significativamente diferentes de cada otro. Esta prueba es la **prueba de la suma-calificación de *Wilcoxon* para dos muestras independientes** y es diferente de la **prueba de *Wilcoxon* para muestras relacionadas** discutidos más adelante en este capítulo.

- El nivel de **Sig. asintót. (bilateral)** es una aproximación que es útil cuando el conjunto de datos es grande, y se utiliza cuando **SPSS** no puede dar una cifra exacta o toma demasiado tiempo para trabajar el valor significativo.
- La **Sig. exacta [2*(sig. unilateral)]** se basa en la distribución exacta de la prueba estadística (en este caso U). Esto se reportado cuando el conjunto de datos es pequeño, mal distribuido o contiene muchos datos muy acotados. Informar este nivel de significatividad refleja un juicio más preciso de significatividad cuando se trabaja con conjuntos de datos de esta naturaleza.
- Recuerde que para realizar un **análisis no paramétrico** estamos trabajando con calificaciones (**rangos**) y **no los datos en bruto**. **Por lo tanto, es poco probable que confiemos en las medias y las desviaciones estándar como una representación adecuada de nuestros hallazgos.**

7.7. Prueba de *Wilcoxon* para muestras relacionadas: ¿Qué es?

- **La prueba de *Wilcoxon* para muestras relacionadas es el equivalente no paramétrico de la prueba *t* relacionada** y es utilizado cuando no creemos que se cumplen los supuestos de la **prueba *t***. Como las muestras son relacionadas, se tienen que parear las puntuaciones en las dos muestras. Por ejemplo, un grupo de personas tasa su estado de vigilia en una escala de diez puntos por la mañana y de también, por la tarde. Nuestras dos muestras son las calificaciones en la mañana y las calificaciones en la tarde, con cada participante proporcionando una puntuación en cada muestra. Como las muestras son pareadas, la prueba de ***Wilcoxon*** produce una "***puntuación de diferencia***" para cada participante, con la puntuación de cada persona en una muestra tomado de la puntuación de la misma persona en la segunda muestra. (Hinton et al. (2004); Levin y Rubin, (2004)).
- Si todas estas diferencias van en la misma dirección (o son todas positivas o todas negativas) y si existe una gran cantidad de muestras con tamaño suficientemente grande, esto es una prueba convincente de que hay una diferencia entre los grupos. Sin embargo, **si algunas diferencias son positivas y algunas son negativas, entonces es más difícil juzgar si existe una diferencia consistente entre los grupos**. Para resolver esto, la prueba de ***Wilcoxon*** clasifica el tamaño de las diferencias (**ignorando la signo de la diferencia**) de menor a mayor. Entonces las calificaciones de las diferencias positivas se suman y se suman las calificaciones s de las diferencias negativas. **El más pequeño de estos dos totales se toma como el valor calculado de la estadística de *Wilcoxon T*.**
- Si la mayoría de las diferencias van en una dirección con sólo unas pocas diferencias en la dirección contraria, con las diferencias discrepantes siendo pequeñas, esto dará como resultado una ***T* pequeña**. Cuando calculamos ***T*** a mano podemos comparar el valor calculado con los valores críticos de T, en un nivel de significatividad apropiada, en una **tabla de estadística**. Sin embargo, **SPSS calcula una puntuación *z***, ya que existe una **relación entre *T* y *z*** (con la **distribución de *T*** que se aproxima a la **distribución normal**, **particularmente para tamaños de muestra de 25 y superiores).** Si bien no requerimos los supuestos de una **prueba paramétrica**, necesitamos ser conscientes de dos cosas antes de decidir si la prueba es válida:

1. Cuando la diferencia entre las puntuaciones **es cero**, esto indica que una persona ha dado la misma puntuación en ambas muestras. Como esto No nos proporciona información sobre qué muestra tiene las puntuaciones más grandes, **tenemos que rechazar las puntuaciones individuales** de los datos y **reducir el tamaño de la muestra en uno**.
2. Si nuestro muestreo es demasiado pequeño no será capaz de reclamar una diferencia entre ellos.

- También, las **pruebas no paramétricas funcionan con mayor precisión sin rangos ligados,** por lo que si hay un gran número de rangos atados y si el tamaño de la muestra es pequeño debemos tener cuidado de que el análisis sea apropiado.

7.8. Prueba de *Wilcoxon* para muestras relacionadas: Ejemplo

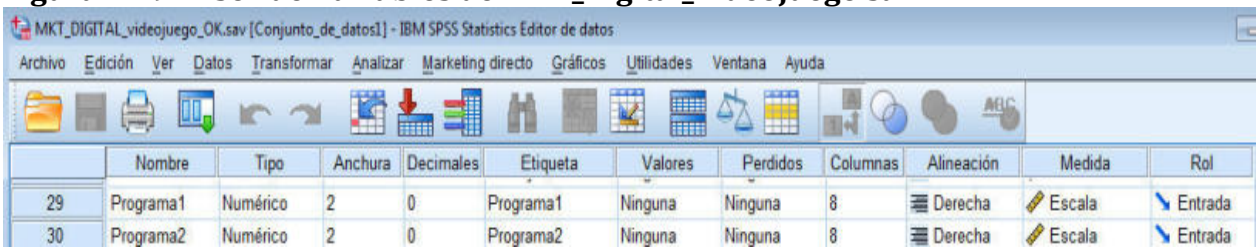
Paso 1: Objetivos

Problema 2. La empresa **MKT Digital** solicitó a 10 jugadores, dieran su opinión sobre dos videojuegos de reciente lanzamiento al mercado y de misma temática, en escala de 1 a 20 en términos de satisfacción para elegirlo como la mejor selección del verano. ¿Está no de los videojuegos evaluado significativamente más alto que el otro por los jugadores?

H_1= Ninguno de los videojuegos está evaluado significativamente más al que el otro por los jugadores

H_2=Existe uno de los videojuegos evaluado significativamente más al que el otro por los jugadores Ver Figura 7.19 y 7.20.

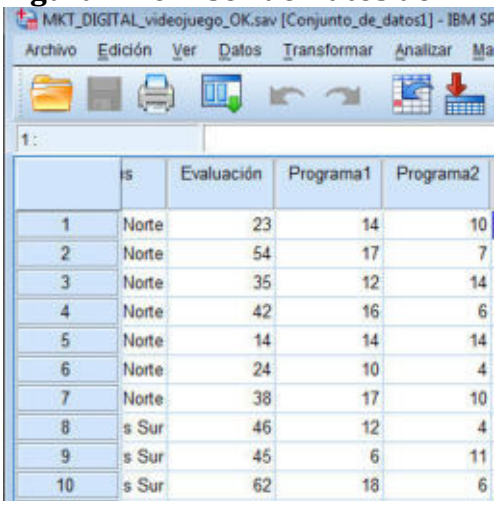
Figura 7.19. Visor de Variables de MKT_Digital_Videojuego.sav



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
29	Programa1	Numérico	2	0	Programa1	Ninguna	Ninguna	8	Derecha	Escala	Entrada
30	Programa2	Numérico	2	0	Programa2	Ninguna	Ninguna	8	Derecha	Escala	Entrada

Fuente: SPSS 20 IBM

Figura 7.20 Visor de Datos de MKT_Digital_Videojuego.sav



	s	Evaluación	Programa1	Programa2
1	Norte	23	14	10
2	Norte	54	17	7
3	Norte	35	12	14
4	Norte	42	16	6
5	Norte	14	14	14
6	Norte	24	10	4
7	Norte	38	17	10
8	s Sur	46	12	4
9	s Sur	45	6	11
10	s Sur	62	18	6

Fuente: SPSS 20 IBM

Paso 2: Diseño

- **En adelante para efectos de aprendizaje, sólo se tomarán de forma directa los datos con el fin de aprender a utilizar la técnica.**
- **Debido al diseño de medidas repetidas podemos ver que cada jugador emite dos clasificaciones.**
- **Vaya al comando Analizar , seleccione Pruebas no paramétricas del menú emergente**
- **SPSS refiere la Prueba de *Wilcoxon* como 2 muestras relacionadas.**

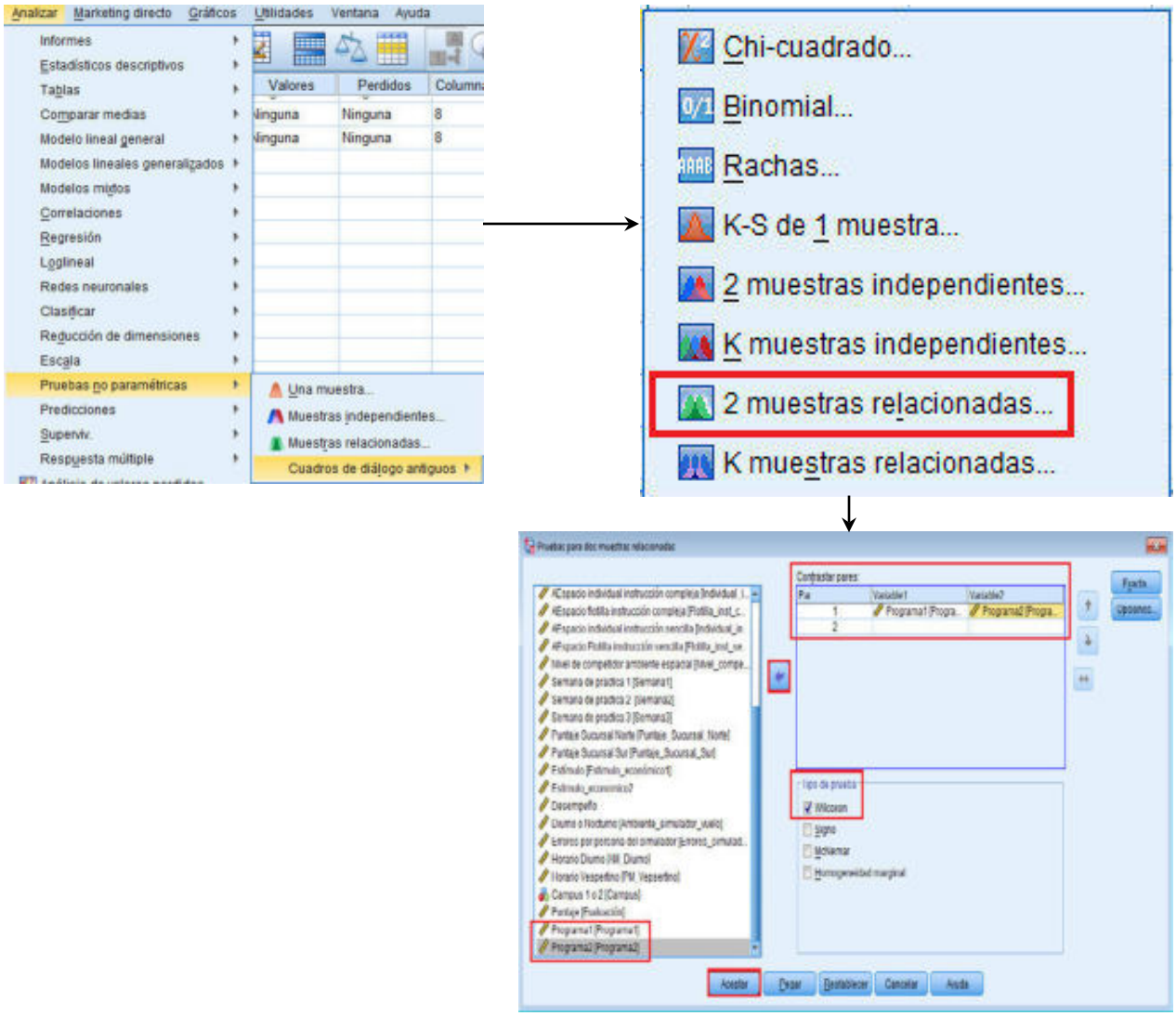
Paso 3: Condiciones de Aplicabilidad

- No olvidar realizar los análisis previos de inspección visual de la base de datos para detectar ausentes y/o atípicos (corregir en su defecto), confiabilidad, normalidad, homocedasticidad y linealidad
- Si el **SPSS** no le permite seleccionar ambas variables al mismo tiempo, haga ***click*** en una variable primero, suelte el botón del ratón y luego haga ***click*** en la segunda variable. Ambas variables deberían estar ahora visibles en el cuadro de selecciones actuales y **SPSS** permitirá que se envíen a la lista de pares de prueba.
- Si desea **realizar más de una prueba de *Wilcoxon* a la vez,** varias combinaciones de pares, se pueden proponer en este punto en este punto.
- A menudo, puede ser difícil seleccionar dos variables que no están próximas entre sí. Haga ***click*** en una variable, mantenga presionada la **tecla Ctrl** y haga ***click*** en la otra variable.
- Mediante el **botón Opciones** puede seleccionar las **medias** y las **desviaciones estándar** para las variables. Si bien estas estadísticas descriptivas pueden realizarse **en conjuntos de datos ordinales,** recomendamos precaución al examinar e interpretar.

Paso 4: Estimación y ajuste

- **Teclear: Analizar->Pruebas no paramétricas->Cuadro de diálogo antiguos->2 muestras relacionadas->Comparar medias->Contrastar pares->Variable1:Programa1->Variable2: Programa2->Tipo de prueba: Wilcoxon->Aceptar->Continuar. Ver Figura 7.21.**

Figura 7.21. Proceso para Prueba de *Wilcoxon* para muestras relacionadas



Fuente: SPSS 20 IBM

- Es poco probable que utilice otros **tipos de prueba aparte** además del de ***Wilcoxon***, pero a continuación se describen brevemente algunos de los usos para las otras pruebas estadísticas
- **La prueba de Signo** examina si la diferencia mediana entre los pares es cero. Esta prueba es menos sensible que la de ***Wilcoxon***, por lo que ésta es usualmente preferida.
- **La prueba de *McNemar*** evalúa el significado de la diferencia entre las dos muestras cuando los datos consisten en dos categorías (por ejemplo, **0 y 1**).
- **La prueba de *Homogeneidad Marginal*** es una extensión de la prueba de McNemar donde más de dos valores de categoría en los datos (por ejemplo**, 0, 1 y 2**).

Paso 5: Interpretación y ajuste

La primera tabla producida por **SPSS** es la tabla Rangos, que reporta en resumen, el número de calificaciones negativas, positivas y vinculadas junto con el **Rango promedio** y la **Suma De rangos. Ver Figura 7.22.**

Figura 7.22. Tabla Rangos

Prueba de los rangos con signo de Wilcoxon

Variable independiente

Rangos

		N	Rango promedio	Suma de rangos
Programa2 - Programa1	Rangos negativos	7[a]	5.86	41.00
	Rangos positivos	2[b]	2.00	4.00
	Empates	1[c]		
	Total	10		

a. Programa2 < Programa1

b. Programa2 > Programa1

c. Programa2 = Programa1

Muestra cupantas calificaciones de Programa1 fueon más grandes que, más pequeñas que o iguales que el Programa2

Fuente: SPSS 20 IBM

- El **Programa2** se ha introducido en la ecuación primero, y por lo tanto el cálculo de la calificación se basa en las calificaciones del **Programa2**- las calificaciones del **Programa1.**
- Los **rangos negativos** indican por lo tanto cuántas calificaciones del **Programa1** son **más grandes** que el **Programa2**.
- Los **rangos positivos** indican el número de calificaciones del **Porgrama1** que fueron **más pequeñas** que el **Programa2.**
- Finalmente, los **rangos empatados** indican cuántos de las calificaciones del **Programa1** y **Programa2** son **iguales.**
- El **Total** es el número total de calificaciones, que será **igual al número total de Jugadores.**
- Otra información que se puede obtener de esta tabla incluye el **Rango promedio** y la **Suma de rangos** tanto para las calificaciones positivas como negativas

La tabla que muestra nuestra estadística inferencial es la **tabla de Estadísticas de contraste (ver Figura 7.23)**. Lo primero que se puede ver es que **SPSS genera una puntuación *z*** más que a prueba ***Wilcoxon T*** que se generará cuando se elaboren los cálculos a mano. **Ver Figura 7.23**

Figura 7.23. Tabla Estadísticos de contraste

Estadísticos de contraste[a]

	Programa2 - Programa1
Z	-2.194[b]
Sig. asintót. (bilateral)	.028

Prueba estadística

p Valor

a. Prueba de los rangos con signo de Wilcoxon

b. Basado en los rangos positivos.

Fuente: SPSS 20 IBM

- De la tabla Estadísticos de contraste se puede ver que ***z* = -2.194**. Un análisis de dos colas es llevado a cabo por defecto, lo que ha dado ***p* = 0.028, que es significativo en *p* <0.05.**
- **De esto podemos concluir que los jugadores calificaron a los dos programas significativamente diferentes, con el Programa1 siendo significativamente favorecido por los entrevistadores, como se indica por las calificaciones positivas y negativas.**
- Los hallazgos de la **prueba de *Wilcoxon*** deben ser reportados como sigue:

z = -2,194, N = 9, p y lt; 0.05

(Observe **que N se indica como el número total de participantes menos los rangos vinculados).**

- Cuando un ***Wilcoxon*** se elabora a mano, el cálculo final suele ser un puntaje ***T***. Sin embargo, el investigador tiene la opción de convertir esto a **una puntuación *z*** (a medida que la **distribución *T*** se aproxima a la **distribución normal** con grandes tamaños de muestra), dando así un ***p* Valor,** que se basa en la **distribución normal**. Esto es particularmente común cuando se trabaja con un gran conjunto de datos (**cuando el número de participantes supera los 25**).
- **SPSS** realiza automáticamente este cálculo independientemente del tamaño del conjunto de datos.
- Recuerde que para analizar datos no paramétricos estamos trabajando con calificaciones y no con los datos en bruto. Como los datos son no paramétricos, es poco probable que confiemos en las medias y desviaciones estándar como una representación adecuada de nuestros hallazgos.

Referencias

Hinton, P.R.; Brownlow, Ch.; McMurray,I y Cozens, B. (2004). *SPSS Explained*. USA: Routledge,.Taylor y Francis Group

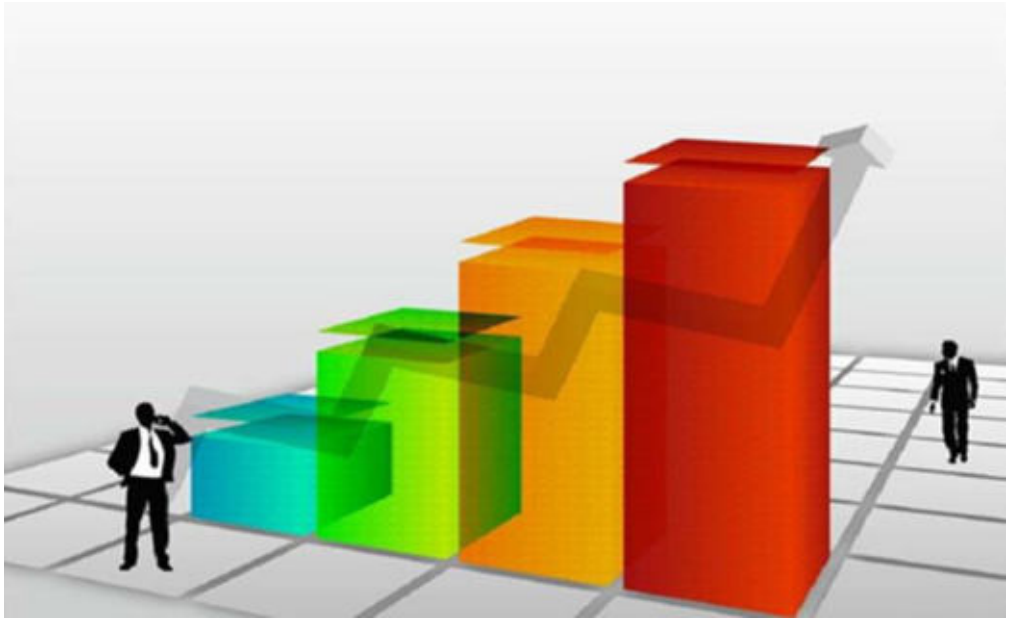
IBM (2011a). *IBM SPSS Statistics Base 20.* EUA. .Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Base.pdf

IBM (2011b).*Guía breve de IBM SPSS Statistics 20.* EUA.Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Brief_Guide.pdf

IBM (2011c). IBM SPSS Missing Values 20. EUA. .Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Missing_Values.pdf

Levin, R.,I.; Rubin, D.S. (2004). *Estadística para Administración y Economía*.7ª. Edición. México: Prentice-Hall

Capítulo 8. Análisis de La Varianza Univariante (ANOVA) y Multivariante (MANOVA)



8.1. Análisis de la Varianza: ¿Qué es?

El **análisis de la varianza** (**ANOVA**) fue introducido hace varias décadas según la formulación original de ***Wilks*** **[**Wilks,1932**]**. Sin embargo, hasta que se tuvieron desarrollados los contrastes apropiados a partir de distribuciones tabuladas y así como una amplia disponibilidad de software para cálculo de dichos contrastes en hardware personal y/o portátil con alta velocidad (computadoras), el **ANOVA** no era una herramienta práctica para los investigadores. Para saber más , vea: IBM,2011a; IBM, 2011b; IBM, 2011c.

El **ANOVA** es una **técnica de dependencia** que mide las **diferencias de dos o más variables métricas dependientes basadas en un conjunto de variables categóricas (no métricas) que actúan como predictores**, pudiendo ser (Hair et al., 1999):

1. De una sola variable dependiente:

Análisis de la varianza

$$Y_1 = X_1 + X_2 + X_3 + \cdots + X$$

(Métrica) (No métrica)

2. De más de una variable dependiente:

Análisis multivariante de la varianza

$$Y_1 + Y_2 + Y_3 \cdots + Y_n = X_1 + X_2 + X_3 + \cdots + X$$

(Métrica) (No métrica)

ANOVA, se refiere a **diferencias entre grupos (o tratamientos de experimentos)**. Sin embargo:

1.- El **ANOVA** se denomina como proceso **univariante** debido a que se emplea para valorar las **diferencias entre grupos utilizando una única variable dependiente métrica**.

2.-El **MANOVA** se denomina como **proceso multivariante** debido a que se usa para valorar **diferencias entre grupos a través de múltiples variables dependientes métricas de forma simultánea.** En el **MANOVA**, cada grupo de tratamiento es definido por **dos o más variables dependientes.**
La utilidad de **ANOVA** y **MANOVA** se refleja cuando son usados de manera conjunta con **diseños de experimentos**, es decir, en investigaciones donde se tiene el control y por lo tanto la **manipulación directa de una o más variables independientes** determinando el efecto sobre una (**ANOVA**) o más (**MANOVA**) **variables dependientes**. Ambas técnicas multivariantes, proporcionan las herramientas necesarias para **juzgar la fiabilidad de cualquier efecto observado,** sea el caso ejemplo, de si las diferencias observadas son producidas a un efecto del tratamiento o a la variabilidad aleatoria del muestreo.
El **MANOVA** es la extensión multivariante de las técnicas univariantes para **valorar las diferencias entre las medias de los grupos.** El procedimiento univariante, aplica **el contraste t** para situaciones con **2 grupos**, y el **ANOVA** en situaciones con **3 o más grupos definidos por dos o más variables independientes**. En este punto, se requiere recordar los principios básicos de las técnicas univariantes.

8.2. Procedimientos univariantes en la valoración de diferencias de grupo

Estos procedimientos se clasifican como **univariantes** no por el número de variables independientes, s**ino por el número de variables dependientes**. En la **regresión múltiple**, los términos **univariante y multivariante** se refieren al número de **variables independientes**, pero para **ANOVA** y **MANOVA**, se aplica la terminología al uso de **variables simples o múltiples**. La exposición mostrada a continuación aborda los **2 tipos** de procedimientos univariantes más comunes:

1. **El contraste t**, que compara una variable dependiente a través de **dos grupos** y
2. **ANOVA,** que se utiliza cuando el número de grupos es **tres o más.**

8.2.1. Contraste *t*

El **contraste t** valora la **significación estadística** de las **diferencias entre dos medias muestrales independientes**. Por ejemplo, Usted puede considerar una campaña de mercadotecnia basada en dos grupos de encuestados que perciben diferentes mensajes, uno **visual** y el otro **auditivo**, y consecuentemente pregunta a cada grupo sobre la pretensión del mensaje en una escala de diez puntos, donde **1= pobre y 10= excelente**. Los dos mensajes publicitarios distintos representan un **tratamiento** con **2 niveles** (**visual vs. auditivo**). Un **tratamiento,** es también conocido como un **factor** y es una **variable independiente no métrica**, manipulada u observada **experimentalmente**, que puede ser representada en varias **categorías o niveles**. En nuestro ejemplo, el tratamiento es el efecto de la pretensión **visual** frente a la **auditiva.**

Con el fin de determinar si los **2 mensajes** son percibidos de **manera diferente** (**significando que el tratamiento tiene un efecto**), se calcula **un estadístico *t***, el cual es **el ratio de las diferencias de las medias muestrales (μ_1-μ_2) y su error estándar.** Éste último, es una estimación de la diferencia entre las medias que se espera debido al error muestral, más que debido a diferencias reales entre las medias. Se puede mostrar esto con la ecuación

Estadístico t= [(μ_1-μ_2) /$SE_{\mu\ \mu_2}$

Donde:

μ_1=Media del grupo 1

μ_2=Media del grupo 2

$SE_{\mu\ \mu_2}$= Error estándar de la diferencia en las medias de grupo

Al construir el ratio entre la diferencia real entre las medias y la diferencia esperada debido al error muestral, se cuantifica la cantidad de impacto real del tratamiento que se debe al error muestral aleatorio. Es decir, **el valor *t* o estadístico *t*, representa la diferencia de grupo en términos de errores estándar.** Si el **valor *t* es suficientemente grande**, entonces podemos decir, **estadísticamente hablando**, que la **diferencia no se debe a la variabilidad del muestreo, sino que representa una diferencia real.**

Para comprobarlo, se lleva a cabo la **comparación del estadístico *t* con el valor crítico del estadístico ($t_{critica}$).** Si el valor absoluto del **estadístico *t* > al valor crítico**, del estadístico esto lleva al rechazo de la **hipótesis nula** de que **no existen diferencias en las pretensiones de los mensajes publicitarios entre los grupos**, lo que quiere decir que **la diferencia real debida a las pretensiones es mayor que la diferencia esperada a partir del error en el muestreo.**

Determinamos un **valor crítico del estadístico ($t_{critica}$)** para nuestro **estadístico *t*** y **contrastamos la significación estadística** de las diferencias observadas de la siguiente forma:

1. Calcule el **estadístico *t*** como el ratio de la diferencia entre la media de la muestra y su error estándar.
2. Especifique un **nivel de error de Tipo 1** (denotado por alfa-α- o nivel de **significación**), que indica el **nivel de probabilidad** que Usted aceptará para concluir que **las medias de los grupos son diferentes cuando realmente no lo son.**
3. Determine el valor crítico del estadístico **($t_{critica}$).** mediante la remisión a la distribución t con **N_1+N_2-2 grados de libertad** y especificación de **α. N y N_2** son los tamaños muestrales.
4. . Si el valor absoluto del **estadístico *t*** calculado es > valor crítico del estadístico ***($t_{critica}$)*** Usted puede concluir que los **dos mensajes publicitarios tienen diferentes niveles de pretensión** (μ_1< >μ_2), con una probabilidad de **error de Tipo I** de **α**. Usted puede entonces examinar los valores medios reales para determinar qué grupo es mayor en el valor dependiente.

8.2.2. ANOVA. Cómo entenderlo

En el ejemplo sobre el **contraste *t***, se consideró **2 grupos de encuestados** para diferenciar los mensajes publicitarios, y para ello se les pidió que clasificaran las pretensiones de los anunciantes en una escala de **10 puntos**.

Problema hipotético: Suponga estar interesado en evaluar 3 mensajes publicitarios en lugar de 2. Los encuestados serían asignados aleatoriamente a uno de los **3** grupos, y tendríamos **3 medias muestrales** para comparar.

Solución: Para analizar estos datos, una alternativa podría ser llevar a cabo **contrastes t separados** para **contrastar la diferencia entre cada par de medias** (por ejemplo, grupo 1 vs. grupo 2; grupo 1 vs. grupo 3; y grupo 2 vs. grupo 3).

Observaciones: Sin embargo, **los contrastes *t* múltiples sobre cuantifican el porcentaje del error Tipo I global** (más detalles en la siguiente sección). El **ANOVA evita este aumento del error de Tipo I** al comparar un conjunto de grupos de tratamiento, **determinando si el conjunto completo de medias muestrales indica que las muestras fueron tomadas de la misma población general**. Es decir, el **ANOVA** es empleado para **determinar la probabilidad de que las diferencias en las medias entre varios grupos sean debidas meramente al error muestral.**

El contraste **ANOVA** es muy sencillo y simple, ya que como su nombre lo indica en el análisis de la varianza**, se comparan 2 cálculos independientes de la varianza para la variable independiente:** uno que refleja la variabilidad general de los encuestados **entre los grupos** ($\boldsymbol{CM_I}$) y otro que representa las diferencias entre los grupos que se atribuyen a los **efectos del tratamiento** ($\boldsymbol{CM_E}$)**:**

l. El cálculo de la varianza dentro de los grupos ($\boldsymbol{CM_I}$cuadrado medio **intra grupos**): es una estimación de la variabilidad aleatoria de los encuestados sobre la variable dependiente dentro de un grupo de tratamiento, y está basado en desviaciones de puntuaciones individua les respecto de las medias de sus grupos respectivas, pero no incluye las diferencias entre las medias de los grupos. El cuadrado medio **intra grupos** es comparable **al error estándar entre dos medias calculado en el contraste t dado que representa la variabilidad dentro de los grupos**. Se denomina también **varianza del error**.

2. El cálculo de la varianza entre los grupos ($\boldsymbol{CM_E}$: cuadrado medio **entre** grupos): la segunda estimación de la varianza es la **variabilidad de las medias de los grupos de tratamiento sobre la variable dependiente**. Se basa en desviaciones de las medias de los grupos respecto de la media global de todas las puntuaciones. Bajo la $\mathbf{H_0}$ de que no hay efectos de tratamiento (por ejemplo, $\boldsymbol{\mu_1 = \mu_2 = \mu_3 \ldots = \mu_k}$), esta varianza, al contrario que el cuadrado medio intra grupos, refleja cualquier efecto de tratamiento que exista, es decir, las diferencias en el tratamiento implican un incremento en el valor esperado del cuadrado medio entre grupos.

Dado que la $\mathbf{H_0}$ = NO existencia de diferencias entre los grupos **es cierta**, los cuadrados medios **intra** y **entre grupos** representan **estimaciones independientes de la varianza poblacional**. Por tanto, **su ratio es una medida de cuánta varianza es atribuible a los diferentes tratamientos frente a la varianza esperada del muestreo aleatorio.** Este ratio nos proporciona un valor de un **estadístico *F***, cuyo cálculo es semejante al del **valor t**, y se calcula como:

$$\textit{Estadístico } F = CM_E / CM_I$$

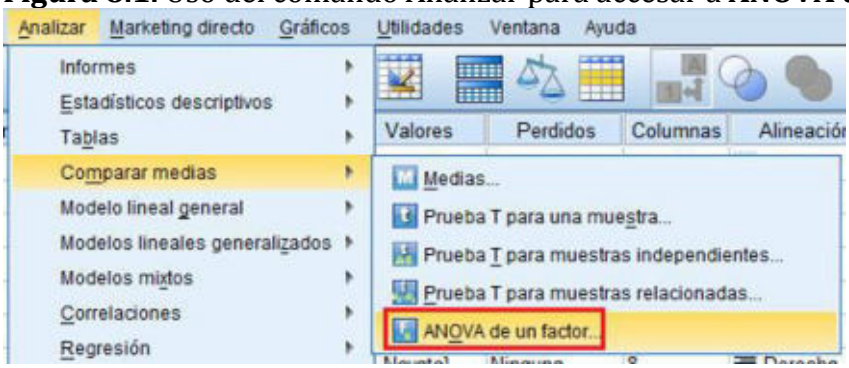
Dado que las diferencias entre los grupos tienden a aumentar el CM_E , valores grandes del **estadístico *F*** nos llevan **al rechazo** de la H_0 de que NO existen diferencias en las medias de los grupos. Si el análisis tiene varios tratamientos diferentes (**variables independientes**), entonces los valores del CM_E se calculan para cada tratamiento, con lo que se permite la valoración separada para cada tratamiento.

Para determinar si el **estadístico *F*** es suficientemente **grande** para apoyar el rechazo de la H_0, se sigue un **proceso similar al del contraste *t*:**

1. Determine el valor crítico para el **estadístico *F*** ($F_{crítico}$) atendiendo a la **distribución *F*** con (**k- 1**) y (**N- k**) grados de libertad para un nivel dado (donde **N**= N_1+... +N_k y ***k*** = número de grupos). Si el valor del **estadístico *F* calculado es** > $F_{crítico}$, concluiremos que las medias entre los grupos **NO** son iguales.
2. El examen de las medias de los grupos permite entonces valorar la **importancia relativa de cada grupo sobre la medida dependiente**. Aunque el **contraste *F* contrasta la hipótesis nula de igualdad de las medias**, **no resuelve la cuestión de qué medias son diferentes**. Por ejemplo, en una situación con **3 grupos**, los tres grupos pueden diferir significativamente, o dos pueden ser iguales pero diferir del tercero. Para valorar estas diferencias, Usted puede emplear **comparaciones planificadas o contrastes *post hoc*** (se explicarán más tarde)

Una de las facilidades más conocidas del **SPSS** es el modelo lineal general (**GLM.**-general linear model), tanto en el menú como en sus productos, y se encuentran disponibles varias versiones a través de submenús que el comando **Analizar** tiene disponibles. Ver **Figura 8.1 y 8.2.**

Figura 8.1. Uso del comando Analizar para accesar a **ANOVA de un Factor**



Fuente: SPSS 20 IBM

Figura 8.2. Uso del comando Analizar para accesar diferentes modalidades de **ANOVA**

Fuente: SPSS 20 IBM

Las diferentes modalidades de **ANOVA** son modelos que representan un problema en particular. El planteamiento de los modelos es muy importante para que los practicantes delas ciencias de la administración, sean capaces de plantear y resolver los problemas de manera que el pensamiento crítico sea capaz de identificar las diversas etapas por las cuales, se generaran las alternativas de solución. Una de estas etapas son los supuestos a los que cada modelo (en este caso de **ANOVA**), se deberán regir para la mejor toma de decisiones de los datos en análisis y determinar por ejemplo, si éstos son generados sistemáticamente o simplemente o son por el azar. Para explicarlo mejor, suponga que la empresa **MKT Digital** ha generado un juego de video en el que somete a sus 2 sucursales nacionales a probarlo para verificar las diferencias de su uso Un investigador cree que el supervisor de la sucursal del Norte favorece las pruebas a costa de los del Sur. La primer cuestión a la que el investigador se enfrenta es la verificar si esta situación existe o no. Suponga, que las 2 regiones son muy similares educacional y socioeconómicamente, y que los jugadores **NO** son diferentes en dichos términos a pesar de proceder de 2 regiones diferentes. Así también, no hay evidencia de que las instalaciones sean tan diferentes como para afectar los resultados. Así, en primera instancia estamos asumiendo que los jugadores realizan sus puntuaciones sin razones sistemáticas que los diferencien (ver **Figura 8.3**).

Figura 8.3. Tabla ejemplo

Puntuación Videojuego	**Sucursal**
52	Norte
49	Norte
51	Sur
55	Norte
48	Sur
52	Norte
49	Sur
45	Sur
50	Sur
54	Norte
50	Norte
47	Sur
53	Norte
46	Sur
48	Sur
51	Norte
50	**Media**

Fuente: propia

El investigador puede decir "*observo que los puntajes muestran el rango usual de jugadores muy jóvenes, siendo cierto que los jugadores del Norte lo hacen ocasionalmente. De todas formas algunos de la zona Sur lo hacen mejor que los del Norte – existen 51del Sur y 49 del Norte*." Apliquemos el modelo para confirmar si podemos decidir sobre si:

$\boldsymbol{H_0}$= No hay una diferencia entre los jugadores de las 2 sucursales en los puntajes del videojuego.

H_1= Si hay una diferencia entre los jugadores de las 2 sucursales en los puntajes del videojuego.
De los datos globales es posible afirmar que el jugador estándar tiene un **puntaje de 50** en la prueba (dado en el valor de la **media**). Así, la **media** de la población de los jugadores de la prueba, sugiere que no importa de dónde el jugador proviene, siempre consiguen un puntaje medio de **50**. Con esto, el primer paso en el modelo es el de argumentar que sin importar otros factores involucrados, cada jugador producirá un puntaje de **50**. Ahora, agrupemos a los jugadores por sucursal para ver si podemos hacer una comparación entre ellos. Ver **Figura 8.4.**

Figura 8.4. Tabla ejemplo agrupada

Puntuación Videojuego Sucursal Norte	Puntuación Videojuego Sucursal Sur
52	47
54	49
51	46
50	48
49	51
52	48
55	50
53	53
Media= 52	**Media=48**

Fuente: propia
Observe que las **medias** de los grupos, son diferentes: **52** y **48**, así, que es posible que las instalaciones de la sucursal **Norte** estén discriminando entre los jugadores (si las medias hubieran sido de **50 no habría evidencia de la diferencia**). Así, se encuentra que existe variación en los puntajes entre los grupos. Sin embargo, ¿se tienen diferencias producto de la oportunidad del azar, o debido a que existen condiciones en la sucursal que producen la discriminación? Aquí es donde la siguiente etapa del modelo. Cada jugador debe lograr el puntaje de **50** en la prueba, pero no lo logra por las instalaciones de la sucursal Norte. Entonces la media de **48** muestra un promedio de '**efecto supervisor de Sucursal Norte**' de **−2** para los jugadores de la **Sucursal Sur** y la media de **52** muestra un promedio '**efecto supervisor de Sucursal Norte** de **+2** para los jugadores de la **Sucursal Norte**. Para nuestro modelo tenemos que argumentar que el **efecto Sucursal Norte resulta en una variación consistente o sistemática** entre los grupos, así que cada jugador en el grupo de la **Sucursal Sur** debe tener el mismo efecto de **−2** y en cada jugador en el grupo de la **Sucursal Norte** debe tener el mismo efecto de **+2**, para ser consistente. Sin embargo, no todos los jugadores del grupo de la **Sucursal Sur** tienen puntaje de **48** y no todos de la del **Norte** tienen **52**. Para explicar esto por el modelo, podemos argumentar que existe variación dentro de los grupos debido a la **variación NO sistemática o azar** (tales como las diferencias individuales, oportunidad de efectos que ocurren en el día de la prueba, y otros factores). Nos referimos a esto como ***"error"*** por variación. Así la descripción final del modelo es:

Puntaje de jugador= Media dela población+ Efecto Supervisor Sucursal Norte+error

Ahora, presentamos la tabla de resultados en términos del modelo. Ver **Figura 8.5**

Figura 8.5. Tabla ejemplo reagrupada

Puntuación Videojuego Sucursal Norte	Puntuación Videojuego Sucursal Sur
52=50+2+0	47=50-2-1
54=50+2+2	49=50-2+1
51=50+2-1	46=50-2-2
50=50+2-2	48=50-2+0
49=50+2-3	51=50-2+3
52=50+2+0	48=50-2+0
55=50+2+3	50=50-2+2
53=50+2+1	45=50-2-3
Media= 52	**Media=48**

Fuente: propia

Por otro lado, se debe explicar algunas cuestiones adicionales. Recuerde que el modelo supone que los puntajes difieren debido al **efecto supervisor de la Sucursal Norte** y el **error**. El primero **es constante** dentro del grupo así solamente la razón por la cual los puntajes varían dentro del grupo es debido al **error**. Se espera (de acuerdo al modelo) que el error en cada grupo será el mismo, como los jugadores son ubicados al azar en los grupos (ya que como se dijo, no hay diferencias consistentes entre la gente que vive en torno a las 2 sucursales – es la oportunidad del azar que conduce a una persona vivir en cada una de las zonas seleccionadas). Si se tuviera una gran cantidad de **error** en uno de los grupos casi ninguno en el otro, esto sería muy extraño y socavaría el supuesto de nuestro modelo. Nosotros medimos **variabilidad** por la varianza estadística (**el cuadrado de la desviación estándar**). Formalmente, este supuesto es que cada grupo proviene de una población con homocedasticidad. Observando los errores, apreciamos que está bien. Finalmente, como estamos calculando estadísticos tales como la media, desviaciones estándar, y varianza, estos datos serán significativos si los puntajes provienen de escalas de intervalo.

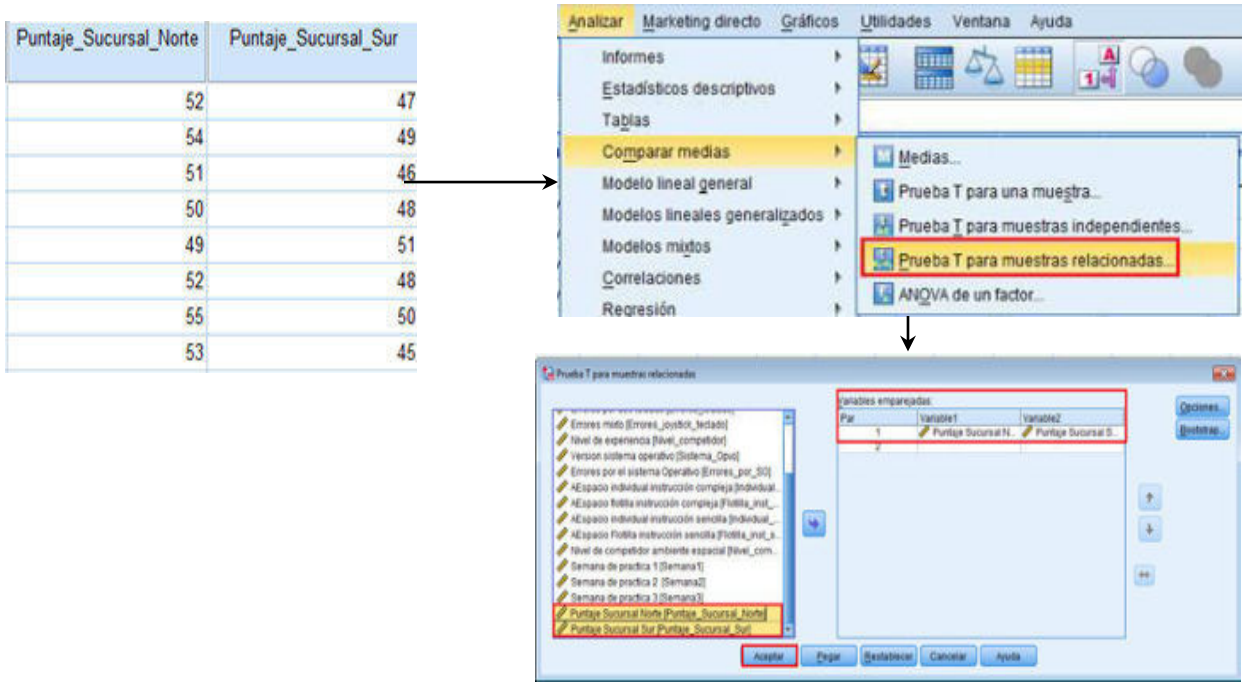
Una vez más, todos los supuestos del modelo se cumplen. Sin un modelo planteado, nuestros datos son simplemente datos: una colección de números. Aplicando un modelo, como lo hicimos anteriormente, damos orden a los datos y somos capaces de tomar decisiones estadísticas, como inferir el **efecto del supervisor de la Sucursal Norte**. Nuestro modelo es bastante simple, pero todavía requiere que hagamos un conjunto de suposiciones tales como "***todas las puntuaciones en un grupo serían las mismas excepto por error***" y "***la variabilidad de las puntuaciones en un grupo es la misma que la variabilidad de las puntuaciones en el otro grupo***", ya que el error debería, según el modelo, estar distribuido uniformemente entre los grupos.

Observe que esperamos que las puntuaciones en un grupo difieran sistemáticamente de las puntuaciones en el otro grupo debido al **efecto del supervisor de la Sucursal Norte**. A fin de determinar la solución, se debe realizar un **tratamiento al efecto** en cuanto se argumenta que los grupos no deben diferir si no los hacemos tratamiento alguno del **efecto Sucursal Norte**. Así, la descripción del modelo general es como sigue:

Puntuación= Media de la población + tratamiento del efecto + error

Este tipo de modelo se le conoce como **modelo lineal. Es un modelo aparentemente simple** dado que nos permite realizar varios análisis estadísticos muy potentes. Las pruebas estadísticas usan este modelo en la búsqueda de **variaciones sistemáticas** en los puntajes que pueden ser atribuidos a un **tratamiento al efecto y a una variación no sistemática o al azar** que pueda ser atribuido al **error**. Analizando las fuentes de variación en los puntajes, tenemos la posibilidad de emitir **decisiones estadísticas** (por ejemplo: ¿**existe un tratamiento al efecto o no**?); esto es lo que se quiere decir el **ANOVA**. **El modelo también tiene subyacente al modelo de regresión lineal.** El ejemplo anterior es el modelo de caso más simple (de hecho, también se le puede reconocer como una **prueba *t* independiente**), sin embargo, el **ANOVA** nos permite investigar la variación en los puntajes de los datos producidos por cualquier cantidad de variables independientes con cualquier número de grupos (**o condiciones**). Podemos incluso usarlo para ver más de una variable dependiente (**MANOVA**).
Justo en el interés de aplicar la **prueba *t* independiente** en los datos anteriores, tenemos:
-Teclear: Analizar-Comparar medias-> Prueba T para muestras relacionadas->Variables emparejadas: Puntaje_Sucursal_Norte; Puntaje_Sucursal_Sur->Aceptar. Ver Figura 8.6

Figura 8.6. - Proceso cálculo Prueba *t* para muestras relacionadas



Fuente: SPSS 20 IBM

Con una $t = 3.864$, $gl = 7$, $p=0.006 < 0.01$, significa que existe una **alta diferencia significativa entre los grupos en la dirección predicha**. (Una **prueba *t*** es un caso simple de **ANOVA con sólo 2 muestras**. Cuando este es el caso, tenemos la siguiente relación entre los 2: $t^2 = F$. Nuestro resultado es una **ANOVA** con valor $F = 14.9$, el cual indica que **la varianza entre los grupos**, **incluye el tratamiento al efecto**, el cual es **14.9 veces más grande** que la **varianza dentro de los grupos**, **el error de la varianza**, que muestra un gran efecto tratamiento). Ver **Figura 8.7**

Figura 8.7. Tabla de muestras relacionadas

Prueba de muestras relacionadas

		Diferencias relacionadas					t	gl	Sig. (bilateral)
		Media	Desviación típ.	Error típ. de la media	95% Intervalo de confianza para la diferencia Inferior	Superior			
Par 1	Puntaje Sucursal Norte - Puntaje Sucursal Sur	4.000	2.928	1.035	1.552	6.448	3.864	7	.006

Fuente: SPSS 20 IBM

Conclusión para nuestra decisión estadística: se piensa que el supervisor de la Sucursal Norte sesga los resultados en favor de los jugadores de su sucursal

El modelo general lineal, subyace en un gran número de pruebas, en particular **ANOVA** y **regresión lineal**. Esto se considera que permite al investigador una amplia flexibilidad de análisis:

1. Es posible tener tantas condiciones (grupos) como requiramos y de hecho, tantas variables como lo deseemos en el análisis y todavía conservar la misma estructura del modelo básico, subyacente, en todos los análisis.
2. El modelo puede tener una ligera complicación a nivel explicación en el caso de tener una gran cantidad de variables, por ejemplo: el uso de un modelo lineal al examinar la relación entre 2 variables donde el investigador está interesado en saber del comportamiento de 5 usuarios por entrar a las redes sociales más populares con equipos móviles (**cantidad de horas de uso de internet a la semana**) **vs.** El puntaje en cuanto a las **habilidades técnicas requeridas para ingresar a dichas redes sociales** (**calificadas de 1-20**). Ver **Figura 8.8**

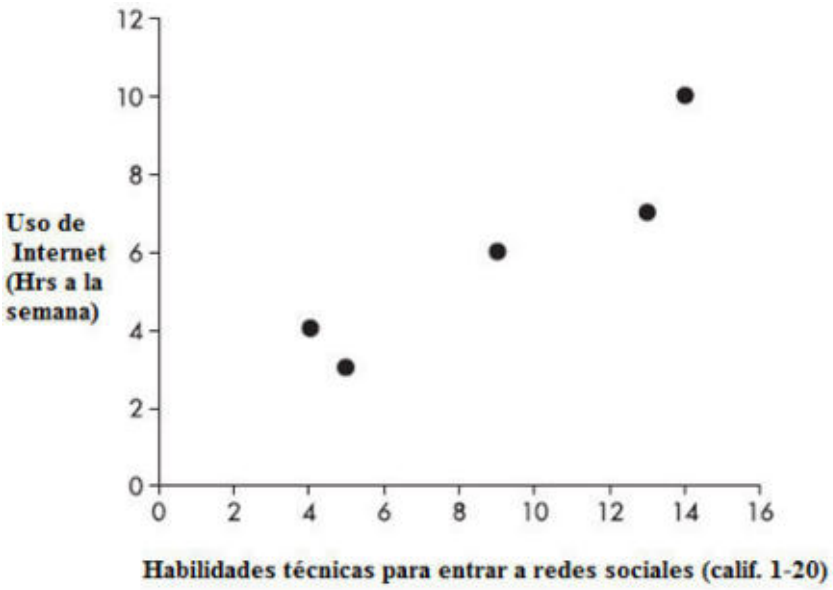
Figura 8.8. Tabla ejemplo

Participante	Uso de Internet móvil (Hrs)	Habilidades técnicas (calif. 1-20)
1	4	4
2	3	5
3	6	9
4	7	13
5	10	14

Fuente: propia

¿Cuál es la relación entre las variables? No está muy claro de la tabla así que deberá apoyarse de la representación gráfica de los datos de la **Figura 8.9**.

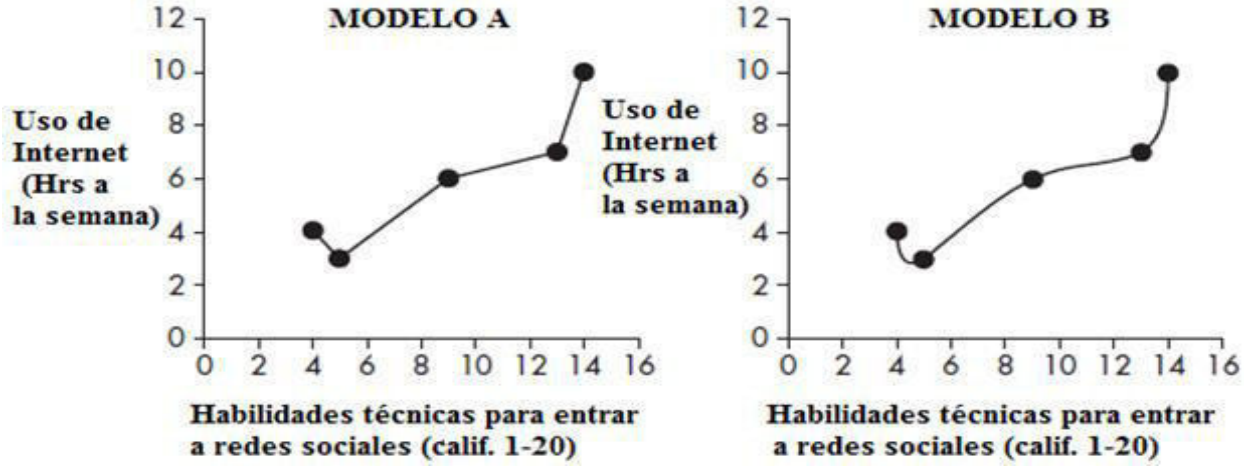
Figura 8.9. Gráfica Habilidades técnicas vs. Uso de Internet



Fuente: propia

Observe que la relación no es tan obvia, ¿cómo se puede modelar esta relación? Una opción simple es la de unir los puntos para determinar una **forma W** como el **modelo A** (**Figura 8.10**), o dibujar una línea ligeramente curveada a través de los puntos **modelo B** (**Figura 8.10**)

Figura 8.10. Gráfica comparativa de Modelos A y B

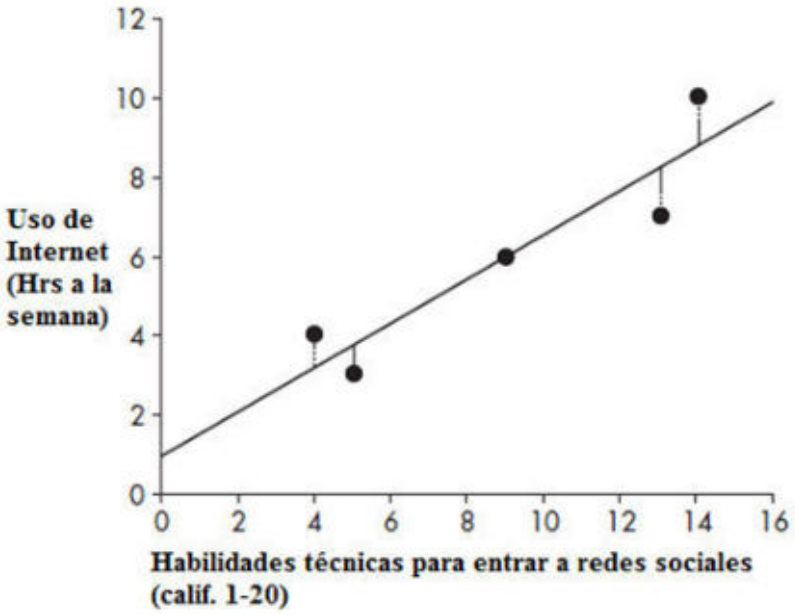


Fuente: propia

El problema con los modelos es que es complicado describirlos dado que no existe una simple descripción matemática para ello. Además, no hay garantía de que los datos nuevos se ajusten al modelo. Como alternativa podríamos argumentar que la relación entre las variables es simple, una relación lineal, pero la razón por la que los puntos en el gráfico no

siguen una línea recta se debe al ***"error"***. Si observamos el gráfico de dispersión nuevamente, podemos ver que los puntos caen dentro de una banda que va desde la parte baja izquierda la parte alta a la derecha, que parece seguir una línea. Ver **Figura 8.11**

Figura 8.11. Gráfica que ajusta los puntos a una línea recta



Fuente: propia

La figura incluye una línea recta en el gráfico para indicar la relación que creemos **habría ocurrido** si no hubiera habido **error** (**nuestro modelo**), con las líneas punteadas indicando el error - la distancia de cada punto de la línea. Matemáticamente definimos una recta por la fórmula:

$$Y = a + bX$$

Donde **'a'** es una constante llamada **intersección** ya que este es el valor de **Y** cuando **X = 0**. Si la línea pasa por el origen ***(X = 0, Y = 0***), entonces la Intersección será cero. La constante '**b**' se denomina **pendiente** de la recta. Ahora tenemos una situación en la que asumimos que nuestras puntuaciones son el resultado de una relación lineal (**un modelo lineal**) más el **error**. En este punto podríamos decidir si el puntaje de las **habilidades técnicas para entrar a redes sociales** son correctas **vs** las puntuaciones de **uso de Internet** contienen **variabilidad debido al error, o viceversa.**
En este ejemplo asumiremos los valores de X (**habilidades técnicas por entrar a redes sociales**) son correctos y los valores de Y (**uso de Internet**) **no se ajustan al modelo debido al error**, por lo que se tiene:

Valores observados de Y = a + b (valores observados de X) + error

Así que:

Uso de Internet = a + b * puntaje de habilidades técnicas + error

Podemos **calcular la línea recta que mejor se adapta a nuestros datos** mediante la búsqueda de la ecuación de la línea que nos **del menor error general**. **Esta línea se llama la línea de regresión Y**, que según el análisis de regresión, la mejor línea de ajuste para nuestros datos se define por la siguiente ecuación:

Uso de Internet = 0.951 + 0.561 * puntaje de habilidades técnicas

Lo que esto está diciendo es que **el modelo puede explicar cierta variabilidad en los datos** (**que sigue a la línea recta**) pero no otra variabilidad (**el error**). Por ejemplo, el participante 1 obtuvo 4 puntos en la prueba de habilidades técnicas. Poniendo **4** en la fórmula anterior para nuestro modelo lineal obtenemos una **predicción de que el participante pasó 3.195 horas en Internet** (Dado que **3.195 = 0.951 + 0.561 × 4**). De hecho pasaron **4 horas en Internet**, por lo que, de acuerdo con nuestro modelo, **3.195 explica la variabilidad y 0.805 es la variabilidad inexplicable o error. Ver Figura 8.12**

Figura 8.12. Tabla con datos del modelo lineal ajustado

Participante	Uso de Internet	Uso de Internet (explicado por el modelo)	Uso de Internet (error)	Habilidades técnicas (suponiendo que es correcto)
1	4	3.195	+0.805	4
2	3	3.756	-0.756	5
3	6	6	0	6
4	7	8.244	-1.244	9
5	10	8.805	+1.195	13

Ahora que hemos adaptado los datos a nuestro modelo lineal podemos tomar decisiones estadísticas sobre **la bondad de este ajuste.** Podemos examinar cuánto de **la variabilidad en las puntuaciones se explica por el modelo y cuánto es inexplicable o la variabilidad del error**. De nuevo, como se mencionó anteriormente, examinar las fuentes de la variabilidad en los datos nos permite tomar decisiones estadísticas. **¿Existe una variación sistemática en la variable dependiente que resulta a partir de una variación sistemática en la variable independiente?** Todo es cuestión de si los datos se ajustan bien al modelo o si hay simplemente demasiado **error** para que sea un buen ajuste. Si, al montar **el modelo, hay muy poco error entonces el modelo es una buena representación de los datos.** Tenemos que ser conscientes de los supuestos que estamos haciendo sobre nuestros datos al aplicar el modelo. Esperamos que haya más o menos la misma variabilidad en las puntuaciones Alrededor de la línea de regresión en cada punto de la línea, es decir, la variabilidad debida al error sería la misma en cada punto de la línea. Si sólo hubo un **pequeño error** en un punto **Y** con una gran cantidad de errores en otro lugar, **socavaría la validez del modelo**.

Esperamos que el error se distribuya uniformemente alrededor de la línea de regresión, ya que estamos asumiendo que **los factores aleatorios son la única razón del error.** Esta es la suposición de **"*homogeneidad*" u homoscedasticidad**.

En el ejemplo anterior hemos examinado **una sola variable dependiente, Y,** que en nuestro caso fue el **uso de Internet**, y una sola **variable independiente, X,** las habilidades técnicas puntuación. Pero este es el caso más simple del modelo lineal general. Podemos tener cualquier número con **variables independientes y** producir la siguiente ecuación de regresión múltiple:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 +. . b_nX_n$$

El punto clave es que la forma general de la función sigue siendo la misma y así se denomina **modelo lineal general**. Esto subyace a todos nuestros cálculos de **ANOVA** y **regresión**. Puede parecer sorprendente saber que tanto el **ANOVA** como el **análisis de regresión** son basados en el **modelo lineal general.** De hecho, son como dos caras de la misma moneda ya que ambos tipos de análisis están examinando cuánto de la **variación observada en la variable dependiente se puede atribuir a la variación en la variable independiente y cuánto se debe al error**. Es por eso que a veces verá tanto una regresión como un **ANOVA** en los resultados que genera **SPSS**.
Para que los datos sean apropiados para esta forma de modelo necesitamos hacer ciertos supuestos sobre el **"*error*"** que queda después de haber eliminado la variación explicada por el modelo:
• **Los errores se suman a cero.**
• **Los errores se distribuyen normalmente.**
• **Los errores son independientes entre sí.**
Estas suposiciones son necesarias ya que indican que **los errores son realmente aleatorios** y **No** hay una variación sistemática que se quede sin explicación en ellos después de aplicar el modelo. Sí existiera, es una variación sistemática en los términos de **error**, entonces indicará que **existe un mejor modelo para los datos y el que hemos encontrado no es el más apropiado.** Estas suposiciones subyacen en los supuestos de las **pruebas paramétricas.**
Cuando se analizan los resultados de un estudio, se observa que no todos los puntajes no son los mismos. Existe una **variabilidad de datos**, la cual se calcula mediante la **suma de los cuadrados**. En fórmula de la **desviación estándar** o la **varianza,** se observa que en la sección superior de la fórmula se tiene al término $\boldsymbol{\Sigma\ (X - Xmedia)^2}$ el cual, es la suma del cuadrado de las desviaciones de la media, o suma de cuadrados. Es posible **calcular el total de la variabilidad** de los datos al trabajar la variabilidad de todos los puntajes en el estudio. Sin embargo también se puede trabajar sumas de cuadrados por cada fuente de variabilidad, tal como la variabilidad debida a las variables independientes o a **fuentes de error**. Podemos trabajar cuanto del total de la suma de los cuadrados provienen de cada fuente de variación. Con esto, se necesitará trabajar en el **promedio de la variabilidad** debido a cada fuente. Esto es, que se tiene mayor posibilidad de obtener variación en 30 participantes que de 10 participantes, o de seis condiciones en lugar de tres condiciones, simplemente porque tenemos más de ellos. Así, se calcula una **variabilidad media** o una **media cuadrática dividiendo las sumas de cuadrados atribuidas a una fuente de variación por los grados de libertad para la misma fuente de variación**. La **media cuadrática** también se denomina **varianza (y es el cuadrado de la desviación estándar)**. Si ahora comparamos la **varianza** debido a una fuente particular, como una **variable independiente**, con una

varianza de error apropiada, el **radio de la varianza** resultante será **alta** si hay una gran variación sistemática en los datos debido a la variable independiente. Si es **pequeña** (**alrededor de 1**) entonces la variabilidad debida a la **variable independiente** no es diferente de la variabilidad que surge por el **error aleatorio**. Podemos examinar la probabilidad de producir una **radio de varianza** (**o valor *F***) de un tamaño particular cuando la **hipótesis nula es verdadera**. Podemos entonces utilizar el **valor calculado de *F*** para cada una de nuestras variables independientes para **decidir si indican diferencias estadísticamente significativas entre sus condiciones o no**. Un **ANOVA de un factor** calcula un **valor de *F* para la variable independiente** para examinar la **significación estadística**. Sin embargo, se produce una **varianza de error diferente para la medición de ANOVA de un factor independiente que la ANOVA con mediciones repetidas porque este último es capaz de eliminar la variación sistemática** debida a los participantes del término de error.

Un **ANOVA de dos factores produce tres valores de *F***, uno para cada una de las variables Independientes y una para la interacción. Si comparamos gerentes por género: hombres y mujeres en una tarea en la mañana y la tarde tendríamos **dos variables independientes: género y hora del día.** Una **interacción ocurre cuando el efecto de un factor es diferente para los diferentes niveles del segundo factor.** Por lo tanto, si las mujeres gerentes, eran mejores en la tarea en la mañana y los hombres gerentes eran mejores en la tarde, entonces **tendríamos una interacción.**

Si no encontramos una **interacción en un ANOVA de dos factores**, **observamos los efectos de las variables por separado.** Los medios marginales para el **"*género*"** darían la media para los hombres gerentes tanto por la mañana como por la tarde y también a las mujeres por la mañana y por la tarde. De manera similar, las **medias marginales** para la **"*hora del día*"** darían la **media de las puntuaciones** de la mañana y la media de las puntuaciones de la tarde promedio entre los hombres y las mujeres gerentes. Hay una serie de términos adicionales que necesitamos saber cuándo se utiliza el ANOVA en SPSS:

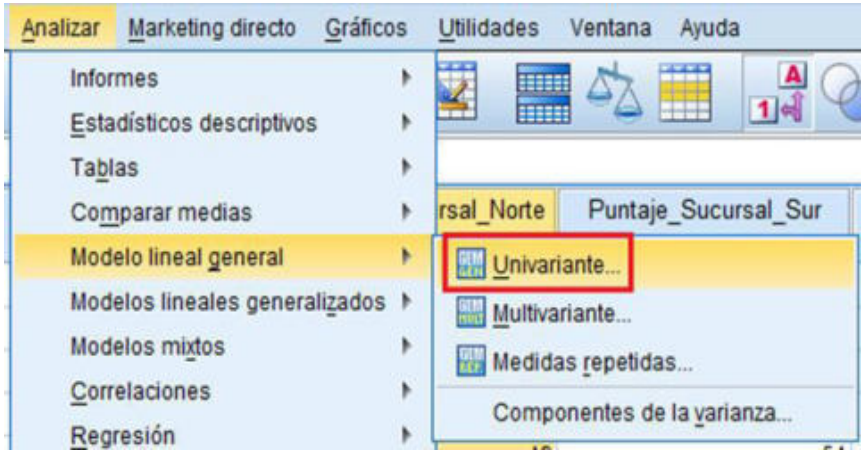
- Grupos, tratamientos, niveles y condiciones, todos referidos a las diferentes muestras de una **variable independiente.** Así que **"*hora del día*"** en el ejemplo anterior tiene dos condiciones, Niveles o grupos: **"*mañana*" y "*tarde*".**
- El término **"*sujetos*"** se utiliza para referirse a los participantes.
- Las tablas **de Efectos Inter-sujetos**, se refieren al resultado del análisis de **Factor de mediciones independiente.**
- Las tablas de **Efectos Intra-sujetos** se refieren al resultado del análisis **Factor de medidas repetidas.**

Por último, la **ANOVA realiza siempre una prueba de dos colas,** ya que mide la cantidad de **variabilidad pero no la dirección (es decir, qué condición tiene la media más alta o más baja).** Así que, a menos que tengamos solamente **2 condiciones**, tal vez desee realizar un análisis más donde las diferencias se encuentran entre las diferentes condiciones con **un estadísticamente significativo valor *F***. Las diferentes comparaciones que puede realizar se describen más adelante.

8.2.3. ANOVA Univariante

Esta opción, corresponde cuando tenemos una sola variable dependiente vs. Variables independientes que lo son de diseño original. **Ver Figura 8.13.**

Figura 8.13. Proceso de selección ANOVA univariante



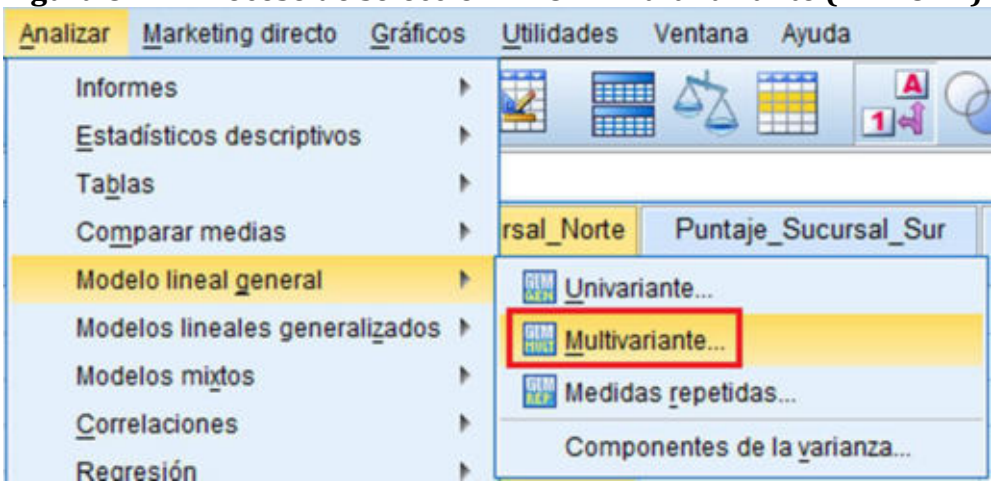
Fuente: SPSS 20 IBM

Cuando todas las variables independientes provienen de mediciones independientes, la **ANOVA** es relativamente simple como técnica que examina la variabilidad en los datos y los atributos de las fuentes que los generan planteando la pregunta: **¿cuánto es debido a una variabilidad sistemática entre las condiciones y cuánto lo es por cuestiones no sistemáticas o de error entre las mismas condiciones?**
Si se encontrara una cantidad relativamente grande de variabilidad sistemática comprada **vs** la variabilidad de error podremos afirmar que existe un **tratamiento de efecto genuino** que ocurre en los datos. En los libros de **ANOVA**, son generalmente descritos **ANOVA de un factor y de dos factores de medición independientes;** con estos 2 temas, los lectores son capaces de distinguir las características del análisis de **1** factor (**una variable independiente**) así como el de apreciar con una **interacción,** caso de análisis de **2** (2 **variables independientes**). **SPSS,** le permite ingresas cualquier cantidad de variables independientes con la misma lógica: habiendo realizado pruebas de **ANOVA de dos factores, las ANOVAS de 3 y 4 factores** son prácticamente lo mismo en su procedimiento.

8.2.4. ANOVA Multivariante (MANOVA)

Esta opción corresponde al caso de tener más de una variable dependiente así como varias variables de medición independientes. Por ejemplo, cuando estamos interesados en el efecto de la edad y el género de los empleados en la productividad de los diversos departamentos a través de un gran periodo de privación de recompensas por estímulos al personal. Edad y género, son nuestras variables independientes y productividad y estímulos son las variables dependientes. Ver **Figura 8.14**.

Figura 8.14. Proceso de selección ANOVA Multivariante (MANOVA)



Fuente: SPSS 20 IBM

8.2.5. ANOVA de mediciones repetidas

Esta opción se considera cuando una o más de las variables independientes requiere de mediciones repetidas. Ver **Figura 8.15.**

Figura 8.15. Proceso de selección ANOVA de mediciones repetidas

Fuente: SPSS 20 IBM

En este caso, el análisis es un tanto complejo que cuando se tiene solamente mediciones de variables independientes. De hecho, ésta opción produce una mayor cantidad de tablas de análisis por lo que a menudo se cuestiona desde lo más general: ¿por qué tantas tablas? hasta lo más particular: ¿cuál de las tablas tiene los **valores *F*** que requiere nuestro estudio?, para explicarlo, un ejemplo puede aplicarse como sigue:

Paso 1: Objetivos.

Problema 1: La empresa **MKT Digital** está interesada en conocer el efecto de su último videojuego de simulador de vuelo. Un investigador prepara las condiciones para que cada participante sea instruido en cómo jugarlo dados **2 minutos** para aprender de los controles y **10 minutos** para jugarlo. El investigador registra el número de errores que los jugadores

hacen durante los **10 minutos** de prueba. Los jugadores son probados una vez al día por **4 días** usando el mismo procedimiento. Así, existen **4 condiciones, prueba 1 a prueba 4**, de la variable independiente '**práctica**' con la variable dependiente de **número de errores**.

Paso 2: Diseño. Tamaño de muestrase supondrá en parámetros

Paso 3: Condiciones de aplicabilidad.
Así, podríamos realizar una **ANOVA** de factor de mediciones repetidas en estos datos. El problema del diseño de mediciones repetidas es la cuestión de la ***esfericidad***. Nuestro modelo subyacente requiere ciertos supuestos a cumplir, uno de ellos se refiere a la **homocedasticidad** (homogeneidad de la varianza) de las diferencias entre las muestras **(*esfericidad*)**. Esto también es referido como la **homogeneidad de la covarianza entre pares de condiciones**, o el supuesto de la **homogeneidad de la covarianza** Si no tuviéramos ***esfericidad*** entonces nuestros cálculos de la varianza pueden ser distorsionados y nuestro **valor *F* sería demasiado grande**, así la ***esfericidad* es un supuesto clave de la ANOVA de mediciones repetidas.** La ***esfericidad* NO** es un problema cuando tenemos **ANOVA de mediciones independientes** como resultado del hecho de que se tiene el mismo número de jugadores en cada una de las condiciones.
Tampoco es problema en el caso de **ANOVA de factores de mediciones repetidas** cuando se tienen solamente 2 condiciones (en cuanto esto siempre tiene ***esfericidad***). **Sin embargo, cuando tenemos 3 o más condiciones de factores de mediciones repetidas debemos realizar pruebas de *esfericidad***. Además, existe un gran riesgo de violar el supuesto en la medida que el número de condiciones de las mediciones repetidas se incrementa. Fundamental mente, **la *esfericidad* es violada** cuando:

1. Los participantes tienen diferentes responsabilidades para el **tratamiento de efectos** a través de las condiciones, o
2. Cuando existen diferentes portadores de efectos a través de las condiciones para los diferentes participantes. Esto se conoce como un **tratamiento de interacción por sujetos.**

Paso 4: Ejecución y ajuste, se supondrá calculado

Paso 5: Interpretación
Diferentes personas tienen diferentes susceptibilidades a ciertas temáticas de videojuegos, así que una persona puede ser inestable en digamos, al jugar cansado o con hambre, mientras que otra persona puede todavía hacerlo sin problemas. Esto causa un problema para el modelo de **ANOVA**, el cual está en la búsqueda de la consistencia en el tratamiento de efectos a través de los participantes y a través de las condiciones. Podemos ver los resultados simulados del ejemplo del simulador de vuelo para demostrarlo. Primero con ***esfericidad***. Ver **Figura 8.16**

Figura 8.16. Prueba 1 de varianza de mediciones repetidas

Jugador	Prueba 1	Prueba 2	Prueba 3	Prueba 4
A	12	10	8	6
B	9	7	5	3
C	6	4	2	0
Varianza	**9**	**9**	**9**	**9**

Fuente: propia

Observe que en cada condición la varianza es la misma, así que existe homocedasticidad. Ahora, observemos en las diferencias entre las condiciones por cada persona. Ver **Figura 8.17.**

Figura 8.17. Prueba 1 de diferencia entre grupos

Jugador	Prueba 1-Prueba 2	Prueba 1-Prueba 3	Prueba 1-Prueba 4	Prueba 2-Prueba 3	Prueba2-Prueba4	Prueba 3-Prueba4
A	2	4	6	2	4	2
B	2	4	6	2	4	2
C	2	4	6	2	4	2
Varianza	**0**	**0**	**0**	**0**	**0**	**0**

Fuente. propia

Aquí, tenemos **homocedasticidad** entre las diferencias de las condiciones. Todas las varianzas son las mismas (Son cero en nuestro ejemplo, circunstancia que no se obtendría en la vida real, sólo es para propósitos ilustrativos), así que tenemos ***esfericidad*, los efectos de los tratamientos** son consistentes a través de la gente y las condiciones. La evidencia mostrada es que cada participante está mejorando en la misma manera con mayor práctica en el simulador de vuelo.

Ahora, observemos lo que sucede cuando la *esfericidad* es violada. Ver Figura 8.18.

Figura 8.18. Prueba 2 de varianza de mediciones repetidas

Jugador	Prueba 1	Prueba 2	Prueba 3	Prueba 4
A	12	9	6	3
B	9	7	5	3
C	6	5	4	3
Varianza	**9**	**4**	**2**	**0**

Fuente: propia

Observe que ya no existe homocedasticidad. Ahora, observemos las diferencias entre las condiciones de cada persona. Ver **Figura 8.19.**

Figura 8.19. Prueba 2 de diferencia entre grupos

Jugador	Prueba 1-Prueba 2	Prueba 1-Prueba 3	Prueba 1-Prueba 4	Prueba 2-Prueba 3	Prueba2-Prueba4	Prueba 3-Prueba4
A	3	6	9	3	6	3
B	2	4	6	2	4	2
C	1	2	3	1	2	1
Varianza	**1**	**4**	**9**	**1**	**4**	**1**

Fuente: propia

Aquí, las diferencias no son consistentes entre las personas y las condiciones, por lo que **no se encuentra la *esfericidad***. No tenemos **homocedasticidad** de las diferencias de tratamientos entre pares: uno es nueve veces más grande que otro. El **jugador A** comienza mal con **12** errores y mejora mucho en cada prueba (**reduciendo sus errores en 3**), pero el **jugador C** empieza bien (**6 errores**) y sólo mejora un poco cada vez (**por 1**). El **jugador B** está en el medio. Si hicimos un **ANOVA de medidas repetidas** en este segundo conjunto de datos, **obtendríamos un valor *F* altamente significativo e incorrecto**. Por lo tanto, debemos **corregir la violación de *esfericidad*** para obtener un **valor más preciso de F.** Por eso hay tantas tablas de reporte para una **ANOVA de medidas repetidas** porque **SPSS** nos proporciona mucha información para **ayudarnos a decidir cómo elegir el valor *F* correcto,** tomando en cuenta que:

1. **SPSS** calcula **ANOVA Univariada** (en cuanto se tiene una **variable dependiente**) y reporta una '**valor *F* de *esfericidad* asumida** para nuestro **factor de medidas repetidas .**Un valor llamado ***Epsilon*** es reportado también y si tiene un valor de **1 entonces el supuesto de *esfericidad* se cumple** y podemos finalizar tomando directamente, el **valor de *F*.**
2. Si ***Epsilon* es menor que 1** entonces podemos tener un problema con *esfericidad*. **SPSS** reporta un **'límite inferior' o el peor caso para *Epsilon***. Para **3 condiciones**, la peor infracción de *esfericidad* dará un *Epsilon* de **0.5**. Cuantas más condiciones tengamos, más cerca el límite inferior llega a **cero**
3. **SPSS entrega los resultados de la prueba de *esfericidad* de *Mauchly*** para checar nuestros datos. **Si es significativo entonces la *esfericidad* es violada.** La dificultad con esto es que la **prueba de *Mauchly*** puede fallar una violación con muestras pequeñas e indicar violaciones con muestras grandes cuando no son lo suficientemente grandes como para preocuparse. Así que tenemos que ser sensibles al valor de ***Epsilon*** y lo que ***Mauchly*** genera. Ciertamente, un ***Epsilon* de 0.75 debe ser visto como bajo**. **Si se ha violado el supuesto de la *esfericidad*** debemos **corregir nuestro valor de *F***. La dificultad con este esquema es que la prueba de ***Mauchly* puede perder una violación con muestras pequeñas** e indicar violaciones con muestras grandes con los que cuando no lo es de manera suficiente para sentirlo. Así que necesitamos ser sensibles al valor de ***Epsilon*** a independientemente de lo que ***Mauchly*** nos reporte. Ciertamente con ***Epsilon* =0.75** debe ser visto como bajo. **Si el supuesto de la *esfericidad* y ha sido violado debemos corregir el valor de *F***
4. **SPSS** calcula un **valor *F* corregido** según **3 métodos de corrección diferentes**: - El método del **"*límite inferior*"** da un grado de **libertad corregido y el valor *F* para el "*peor escenario*"** de la violación de la *esfericidad*. No usamos esto ya que hay **2 correcciones menos severas:** -El ***Greenhouse-Geisser y el Huyn-Feldt.*** La primera es más conservadora que la segunda, sobre corrigiendo la *esfericidad*. A pesar de esto, recomendamos la corrección de ***Greenhouse-Geisser*** cuando se **viola la *esfericidad*,** ya que proporciona una posición media entre el límite inferior y los valores asumidos de *esfericidad*. Normalmente eso es todo lo que necesitamos hacer. Sin embargo, hay una alternativa a la corrección de una violación de *esfericidad*. Esto es **para evitar el problema realizando un análisis multivariado en lugar de un univariante.** En el ejemplo del simulador de vuelo, en lugar de considerar las

pruebas 1 a 4 como un factor de medidas repetidas, vemos los errores del prueba 1 como la primera variable dependiente, los errores del ensayo 2 como la segunda variable dependiente, los errores del ensayo 3 como la tercera variable dependiente Y el ensayo 4 e 4Errores como una cuarta variable dependiente. Podemos hacer esto ya que los mismos participantes han producido cada uno de los tres conjuntos de resultados. Así obtenemos resultados adicionales impresos.

5. **SPSS** calcula **ANOVA** o **MANOVA.** Esto puede ser calculado en un número de diferentes formas así que **SPSS** reporta un **valor *F* siguiendo 4 different methods** (***Trazo de Pillai, lambda de Wilks lambda, Trazo de Hotelling y raíz más grande de Roy***). Podemos tomar el **valor de F** que creemos sea el más apropiado, sin embargo el más popular utilizado es ***lambda de Wilks***.

Para evitar problemas ¿por qué no tomamos simplemente el **valor de *F* multivariado** e ignoramos el **análisis univariante** si la *esfericidad* es tal dolor de cabeza? La respuesta es que para la mayoría de los casos **el análisis univariado es más poderoso** (es mejor detectar un efecto cuando está ahí. Por lo tanto, normalmente preferimos la corrección univariante. Sin embargo, cuando tenemos una ***Epsilon* bajo** y **muestras grandes,** la **prueba multivariante puede ser más potente**. Esto puede sonar bastante confuso, particularmente cuando vemos primero toda la salida que obtenemos con un factor de medidas repetidas en un **ANOVA**. Prácticamente, con la mayoría de los datos no es un problema ya que todos los diferentes análisis dan el mismo resultado. Las diferentes formas de análisis indican lo mismo (significativo o no) y esto es muy tranquilizador. Si usted encuentra los diferentes análisis producir resultados diferentes, entonces vale la pena mirar los datos en más detalle para ver lo que realmente está pasando y esto puede necesitar un mayor nivel de conocimiento estadístico de lo que estamos asumiendo aquí. Usted tiene la opción de profundizar en el asunto usted mismo o consultar a un estadístico, dependiendo de lo seguro que se sienta en su propia comprensión de los datos.

8.2.5. ANOVA. Pruebas de contraste y las *post hoc* de comparación por pares múltiples

Tenemos que recordar que un **valor de *F*** estadísticamente **significativo en un ANOVA** nos permite **rechazan la hipótesis nula pero no nos dice qué hipótesis alternativa aceptar.** Si nosotros examinamos el efecto de los **estímulos económico a los empleados** en particular sobre el **rendimiento,** en **4 condiciones** (**sin estímulo, estímulo bajo, estímulo medio, estimulo alto**) en nuestro análisis, el **valor de *F* indicaría una diferencia entre las condiciones pero no donde se encuentra la diferencia**. Por ejemplo: ¿hay una diferencia en el **rendimiento** entre el **no estímulo** y **el estímulo medio**?, ¿es el efecto debido a la condición de **alto estímulo** que es diferente al resto? Sólo podemos encontrar esto haciendo comparaciones entre nuestras condiciones.

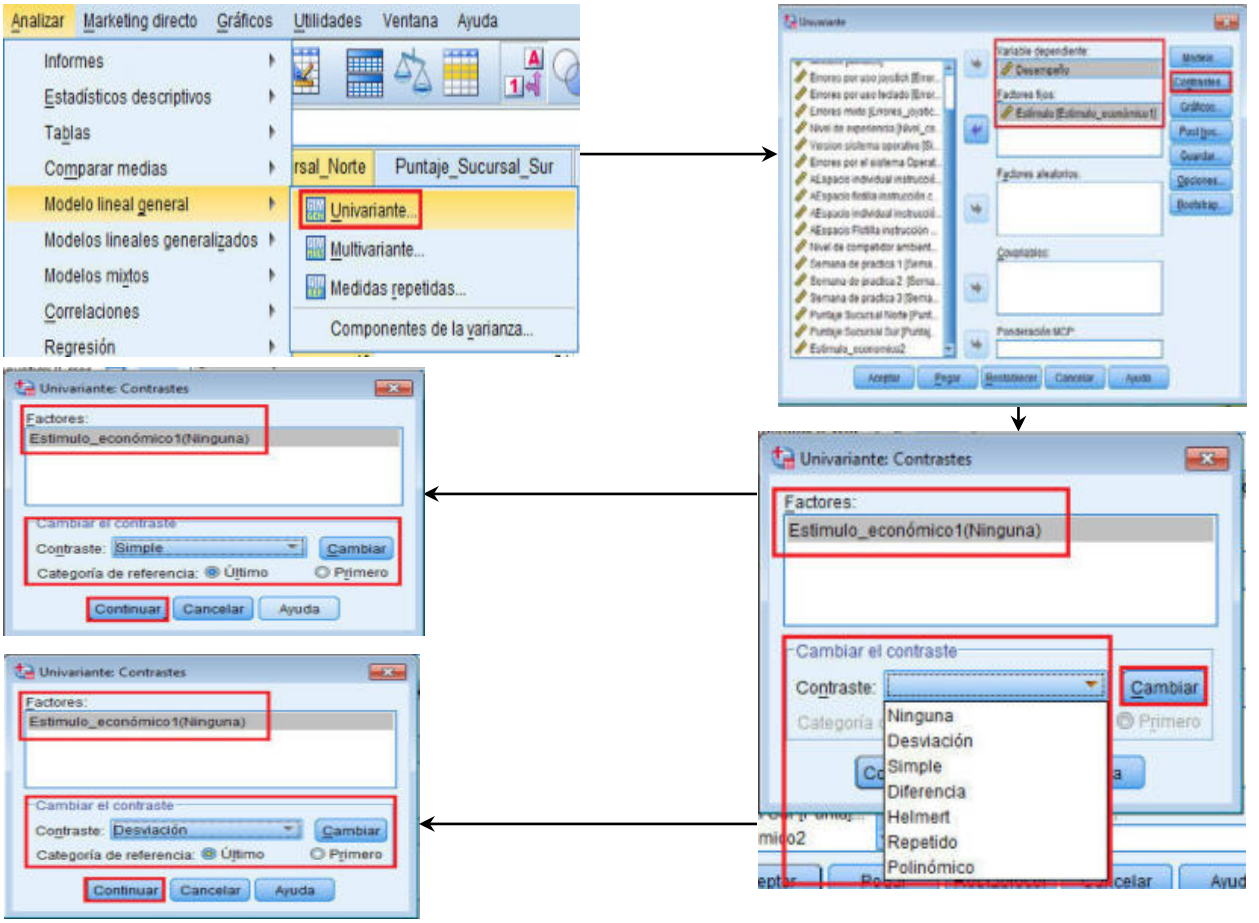
Podemos planificar nuestras comparaciones antes del análisis, sobre cuáles casos queremos hacer comparaciones específicas en lugar de comparar cada condición con todas las demás. En el ejemplo anterior podríamos estar interesados en comparar la condición de **sin estímulo** contra los demás para ver si existe un efecto sobre el **rendimiento**. **SPSS** se refiere a esta forma de comparación como de **contraste**. La planificación de sus comparaciones es a menudo recomendada ya que indica que hay una justificación para todo el estudio en lugar de emprenderlo con la intención de qué efectos encontrar que sean significativos.

Sin embargo, aunque pareciera despectivo el realizarlo sin planeación, podríamos haber decidido emprender un **estudio exploratorio y hacer una serie de comparaciones después del ANOVA**. Estos se denominan pruebas ***post hoc*** (que significa literalmente "***después de esto***"). Las pruebas ***post hoc*** normalmente nos permiten realizar **comparaciones entre pares**, es decir, comparar una condición con otra. Las pruebas hacen corrección por el aumento a **riesgo de errores de tipo I** para las comparaciones múltiples, lo que nos permite posibles comparaciones por parejas si así lo deseamos. **Desafortunadamente, SPSS NO permite realizar pruebas *post hoc* en ANOVA de variables de medidas repetidas** (existía debate en cuanto a los términos apropiados para usar en el análisis) **y también, es posible examinar la opción de efectos principales, que también nos proporcionan comparaciones entre pares.**
Los diferentes **contrastes y comparaciones de pares** que podemos utilizar para ayudarnos a comprender un **valor *F*** en un **ANOVA** se explican en las siguientes secciones.

8.2.6. ANOVA. Prueba de Contraste

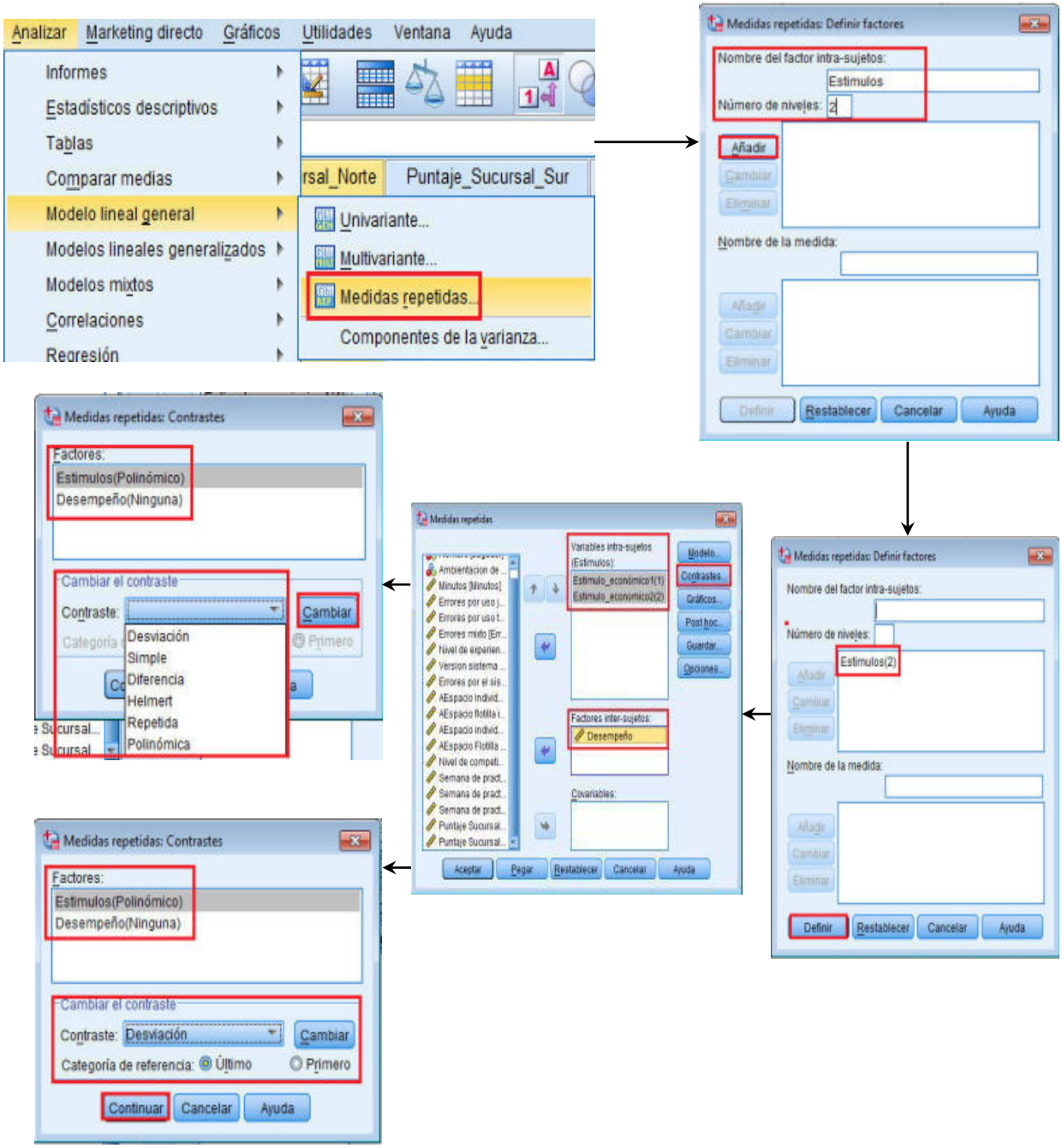
SPSS le permite emprender los mismos contrastes ya sea si aplica una **ANOVA** de medición independiente o repetida a través del comando Modo lineal general (**GLM**). Cuando hace ***click*** en el botón de **Contrastes. Ver Figura 8.20 y 8.21.**

Figura 8.20. Proceso de Prueba de Contrastes caso ANOVA Univariado



Fuente: SPSS 20 IBM

Figura 8.21. Proceso de Prueba de Contrastes caso ANOVA de mediciones repetidas



Fuente: SPSS 20 IBM

Observe que como valor predeterminado en la selección de **ANOVA de mediciones independientes no existen contrastes,** y para **ANOVA** de mediciones repetidas es **polinómica;** en ésta última podrá seleccionar el **contraste** requerido y presionar **Cambiar.**

Existen **6 contrastes estándar** que **SPSS p**ermite hacer (y hay dos versiones de los dos primeros contrastes dependiendo de si selecciona **Último** o **Primer**o como Categoría de referencia).

1. Desviación. La media de cada condición se contrasta con la media general excepto la última (cuando se selecciona **Ultimo**) o primero (cuando se selecciona **Primero**).

2. Simple. La media de cada condición se contrasta con la media de la última condición (cuando se selecciona **Último**) o la media de la primera condición (cuando se selecciona **Primero**).

3. Diferencia. La media de cada condición se compara con la media de las condiciones anteriores.

4. Helmert La media de cada condición se compara con la media de las condiciones subsecuentes.

5. Repetida La media de cada condición se contrasta con la media de la condición siguiente, así que con **3 condiciones** tenemos **condición 1 vs. Condición 2 y condición 2 vs. Condición 3.**

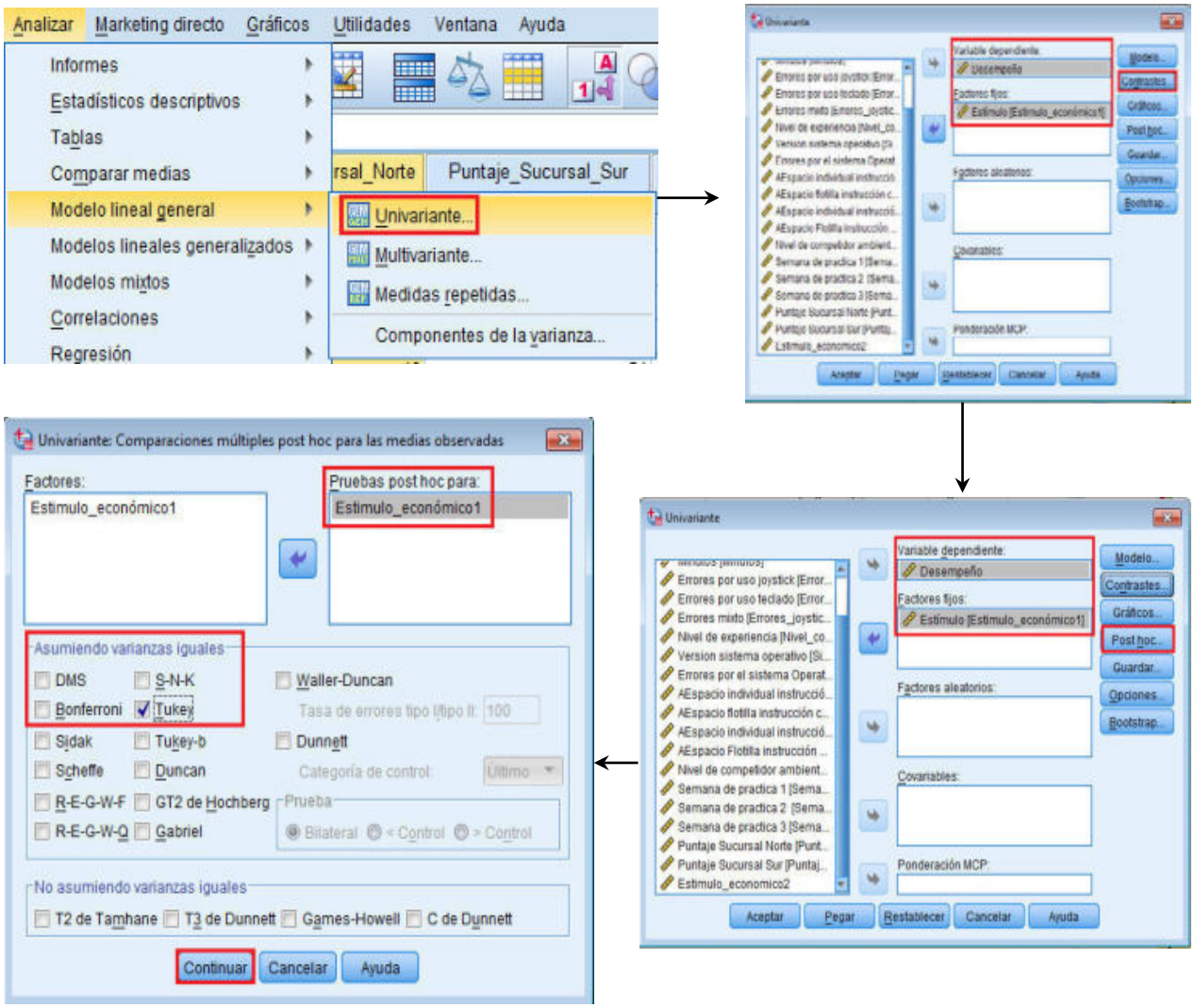
6. Polinómica. Las medias de las condiciones se contrastan en ´términos de si se ajustan a una tendencia. **Con 2** o más condiciones se prueba una **tendencia lineal**, **con 3** o más condiciones se prueban, tanto una **tendencia lineal como una tendencia cuadrática** y, **con 4** o más condiciones, se examinan una **tendencia lineal, cuadrática y una cúbica** y su importancia

Como observará, se tienen todas las opciones probables de contrastes planificados que desee realizar. Por ejemplo, en nuestro ejemplo del **rendimiento**, comparando la condición de **"*ningún estímulo económico*"** con todo el resto de opciones se lograría mediante un contraste de ***Helmert*.**

8.2.7. ANOVA. Prueba *post hoc* de comparación por pares múltiples

Si se ha detectado un **valor *F* significativo** durante el análisis, pero Usted **NO** tiene una hipótesis específica para probar, las **pruebas *post hoc* de comparación por pares múltiple pueden llevarlo a asegurar donde se encuentran las diferencias**. Esta es una opción clara y rápida **ANOVA de mediciones independientes,** como que existe botón de acceso a las pruebas ***post hoc*** en el cuadro de opción **Univariante. Ver Figura 8.22.**

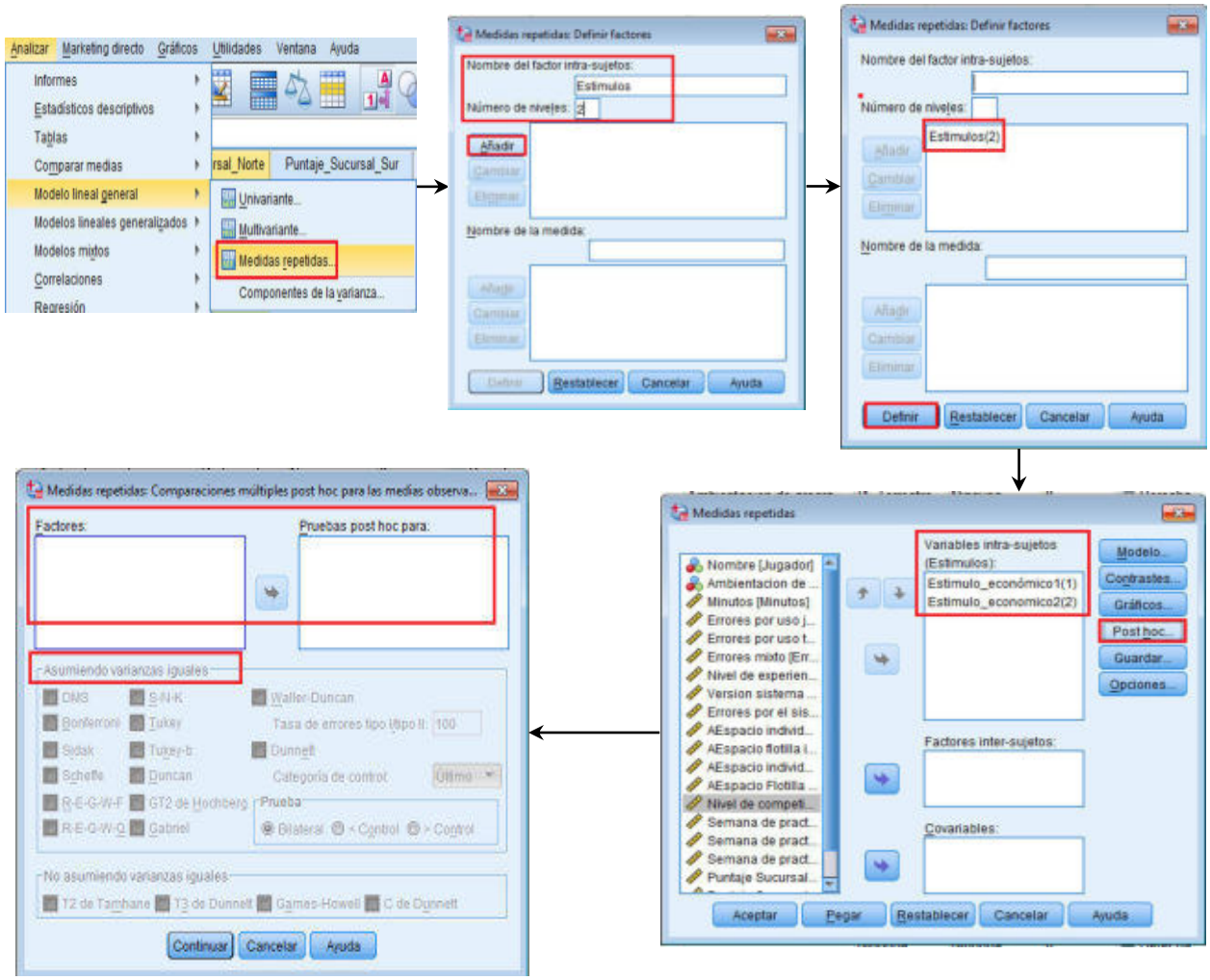
Figura 8.22. Proceso de pruebas con *post hoc* caso ANOVA Univariado



Fuente: SPSS 20 IBM

Hay un gran número de pruebas ***post hoc*** disponibles, siendo una de las más utilizadas la **Prueba de *Tukey***. Esta prueba utiliza un método similar a la **prueba *t*** al dividir la diferencia entre cualquiera de **2 medias** por el error estándar de la diferencia entre cualquiera de las 2 medias. La **prueba de *Tukey* utiliza un error estándar de "*propósito general*"** que puede utilizarse para cualquier par de medias. Cuando se ejecutan **múltiples pruebas *t*** siempre existe el problema de un mayor riesgo de **error Tipo I**. La **prueba de *Tukey*** supera este problema estableciendo un nivel general de significatividad, por ejemplo en **0.05**. Esto significa **que el riesgo de un error Tipo I** tiene una probabilidad de **0.05** cuando comparamos cada par de medias. Sin embargo, cuando tratamos de realizar una prueba ***post hoc*** en un **ANOVA de medidas repetidas en SPSS N0 es posible hacerlo. Ver Figura 8.23.**

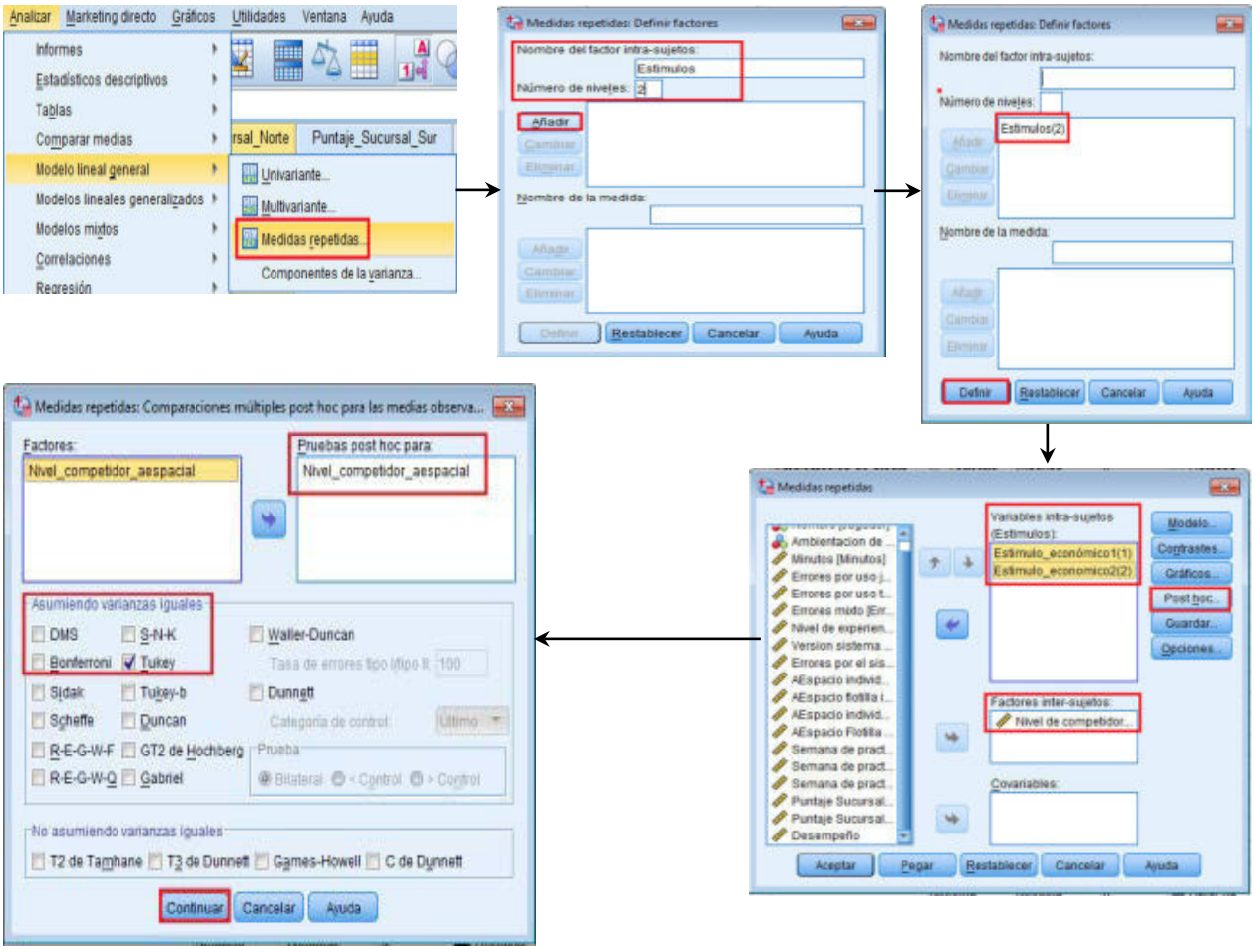
Figura 8.23. Proceso de pruebas con *post hoc* caso ANOVA de mediciones repetidas



Fuente: SPSS 20 IBM

Por último, es importante tener en cuenta que si tiene un **ANOVA por diseño combinado**, su factor de medidas independientes aparecerá en esta casilla y se puede elegir. Ver **Figura 8.24.**

Figura 8.24. Proceso de pruebas con *post hoc* caso ANOVA de por diseño combinado

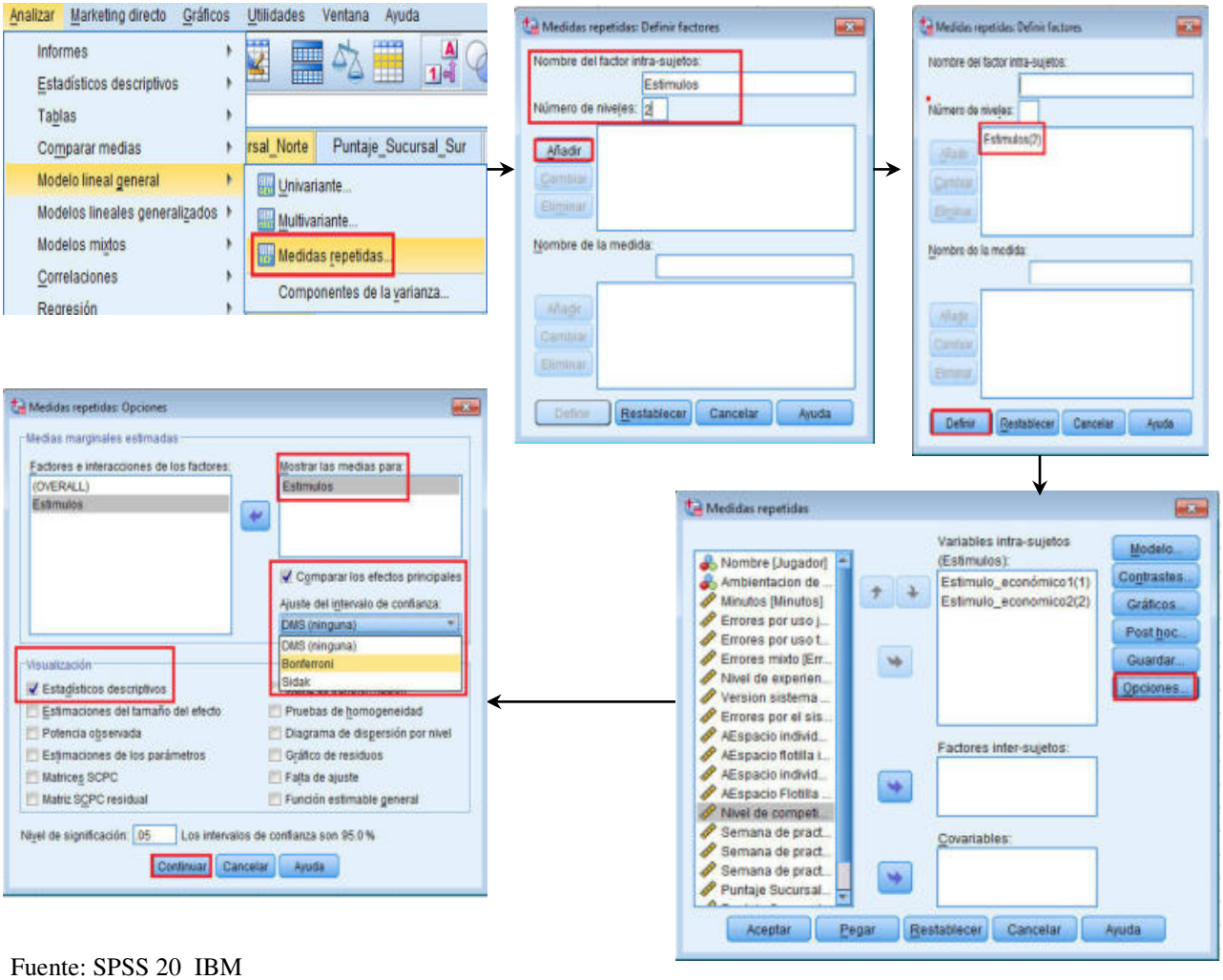


Fuente: SPSS 20 IBM

8.2.8. ANOVA. Efectos principales

Es posible realizar **la comparación de las condiciones de las medias de un factor de medidas repetidas** (o, si lo desea, de un factor de mediciones independientes) en una **ANOVA** a través del botón **Opciones**. Dentro del botón **Opciones** podemos seleccionar la opción **Comparar efectos principales**. Ver **Figura 8.25.**

Figura 8.25. Proceso de pruebas con *post hoc* caso ANOVA por diseño combinado



Fuente: SPSS 20 IBM

En el ejemplo, **como las comparaciones pares múltiples de un número de medias resultará en un aumento del riesgo de un error de Tipo I, esto debe ser controlado, y una corrección por *Bonferroni* se elige hacer esto.**

Para un **ANOVA de un factor**, esto producirá las comparaciones múltiples ***post hoc*** para la **ANOVA** de medidas repetidas.

Cuando tenemos un **ANOVA multifactor** (como un **ANOVA de dos factores**), la opción de **Comparar efectos principales sólo comparará las medias marginales para cada factor**. Así que **ANOVA de dos factores con 3 condiciones en cada factor** comparará los **3 factores marginales de la media del factor uno** y luego las **3 medias marginales del factor dos**. Solo deberíamos **Comparar los efectos principales** si tenemos un **valor *F* significativo para ese factor en el ANOVA y no una interacción significativa.**

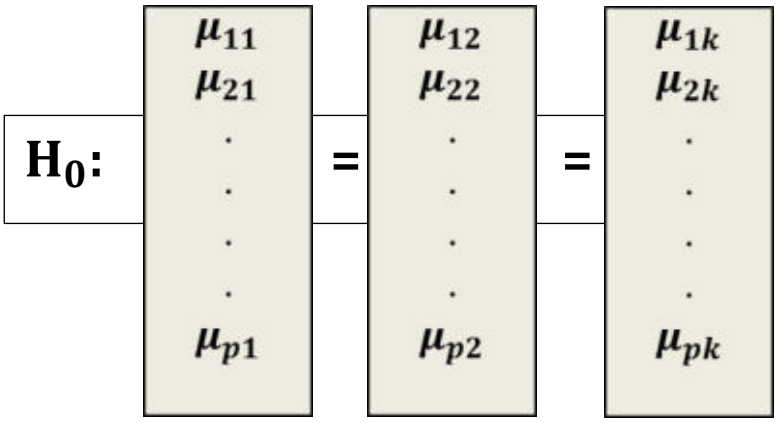
En conclusión, podemos aconsejar las siguientes reglas generales respecto a comparar la condición de las medias. Recuerde que sólo son válidos **si Usted tiene un, estadísticamente significativo valor *F* de ANOVA:**

• Si es posible, **seleccione previamente los contrastes planificados** para las comparaciones que le interesen.
• Sin embargo, si desea realizar pruebas ***post hoc*** (con ANOVA de un factor o ANOVA Multifactor sin interacciones significativas) entonces:
-Para medidas independientes, los **factores eligen una prueba *post hoc* de *Tukey.***
-Para medidas repetidas los factores eligen **Comparar efectos principales con una corrección de *Bonferroni.***

8.3. Análisis de la Varianza. (ANOVA/MANOVA)

Como técnicas de inferencia estadística, tanto las de tipo **univariantes** (**contraste t y ANOVA**) como el **MANOVA** se emplean para **contrastar la significación estadística de las diferencias entre los grupos**. En el **contraste t** y en el **ANOVA**, la $\mathbf{H_0}$ contrastada es la igualdad de las **medias de las variables dependientes** entre los grupos. En el **MANOVA**, la $_{\mathbf{0}}$ contrastada es la igualdad de vectores de medias de variables dependientes múltiples entre los grupos. **La diferencia entre las hipótesis contrastadas en el ANOVA y en el MANOVA** está representada en la **Figura 8.26.**

Figura 8.26. Contraste de la hipótesis nula del ANOVA y del MANOVA.H_0=$\mu_1 = \mu_2 = \mu_3 \ldots = \mu_k$; se plantea la hipótesis nula = las medias de todos los grupos, es decir, provienen de una misma población



Fuente: propia

Los contrastes a realizar, son:
1. Caso **univariante** se contrasta la **igualdad entre los grupos de una única variable dependiente.**
2. Caso multivariante, se contrasta la **igualdad de un valor teórico.** El concepto de un valor teórico es fundamental en el tratamiento multivariante.

En el **MANOVA**, el investigador tiene realmente **2 valores teóricos:**
-Uno construido a partir de las variables dependientes y el otro
-A partir las variables independientes.

El valor teórico con las **variables dependientes es más interesante**, ya que las medidas dependientes métricas pueden ser introducidas en una combinación lineal como ya se vio en la regresión múltiple y en el análisis discriminante. El aspecto propio del **MANOVA** es que **el valor teórico** combina óptimamente las medidas dependientes múltiples dentro de un valor único que maximiza las diferencias entre los grupos.

8.3.1. T^2 de *Hotelling*: el caso de dos grupos

En el ejemplo **univariante** anterior, Usted podría estar interesado en la pretensión de **2 mensajes publicitarios**. Pero, ¿qué ocurre si también quisieran conocer la intención de compra generada por los dos mensajes? Si se emplease solamente análisis univariante, los investigadores llevarían a cabo **contrastes *t* separados** tanto para la pretensión de los mensajes como para la intención de compra generada por los mensajes. Sin embargo, **las dos medidas no están relacionadas;** por ello**, lo deseable es un contraste de las diferencias entre los mensajes en ambas variables conjuntamente**. Aquí es donde se pueden usar la T^2 **de *Hotelling***, una **MANOVA** de forma especializada, es una **extensión directa del contraste t univariante**.

La T^2 **de *Hotelling* proporciona un contraste estadístico del valor teórico formado por las variables dependientes que produce la mayor diferencia entre los grupos**. También tiene en cuenta el problema de "***sobre cuantificar***" **la tasa del error de Tipo 1** que **aumenta** al realizar una **serie de contrastes *t*** para la comparación entre las medias de los grupos sobre las medidas dependientes. Este contraste controla el aumento de la tasa del error de **Tipo 1** proporcionando un único contraste general para contrastar las diferencias de los grupos entre todas las variables dependientes para un nivel de significación α dado.

¿Cómo logra la T^2 **de *Hotelling*** estos objetivos? Considere la siguiente ecuación de un valor teórico de las variables dependientes:

$$C = W_1 Y_1 + W_2 Y_2 + \ldots + W_n Y_n$$

Donde:

***C* =** Puntuación del valor teórico o de la combinación para un encuestado

W_n=Ponderación para la variable dependiente ***i***

Y_1= Variable dependiente ***i***

En el ejemplo, las clasificaciones de **las pretensiones de los mensajes** son **combinadas con las intenciones de compra** para formar la combinación. Para cualquier conjunto de ponderaciones, es posible:

1. Calcular puntuaciones de la **combinación para cada encuestado,**
2. Calcular un **estadístico *t* normal** para la diferencia entre los grupos sobre las puntuaciones de la combinación.

Sin embargo, si se encuentra un conjunto de ponderaciones que generan el valor máximo del **estadístico *t*** para este conjunto de datos, **estas ponderaciones serían las mismas que las de la función discriminante entre los dos grupos**. El **estadístico *t*** máximo que se obtiene con las puntuaciones de la combinación elaboradas a partir de la **función discriminante** pueden ser **elevadas al cuadrado** para obtener el valor de la T^2**de *Hotelling* [**Harris, 1975**]**. La expresión para el cálculo de la T^2 **de *Hotelling*** representa los

resultados de las derivaciones matemáticas empleadas para resolver un **estadístico *t*** máximo (e implícitamente, la combinación lineal más discriminante de las variables dependientes). Esto es equivalente a decir que si nosotros podemos **encontrar una función discriminante** para los grupos que producen una T^2 significativa, los dos grupos son considerados diferentes entre los vectores de las medias. Además, La T^2 proporciona **un contraste para la hipótesis** de que **no existen diferencias de los grupos sobre los vectores de puntuaciones medias**. Al igual que el **estadístico *t*** sigue una distribución conocida bajo la **hipótesis nula** de que no existen efectos del tratamiento sobre una única variable dependiente, la T^2 **de *Hotelling*** sigue una distribución conocida bajo la hipótesis nula de que no existen efectos del tratamiento sobre cualquiera de los conjuntos de las medida dependientes.
Esta distribución es una **distribución F** con **p** y N_1+N_2 -2 - 1 grados de libertad del ajuste (donde ***p* = el número de variables dependientes**). Para obtener el valor crítico para la T^2**de *Hotelling*,** encontramos el valor tabulado para $F_{critica}$ a un nivel de significación **α** dado y se calcula $T^2_{crítica}$ como sigue:

$$T^2_{crítica} = (p\,(N_1+N_2-2))/(N_1+N_2-p-1)$$

8.3.2 El caso de k grupos: MANOVA

El **MANOVA** puede ser considerado como una sencilla extensión del procedimiento de la T^2 **de *Hotelling*;** es decir, **Usted idea ponderaciones de las variables dependientes para elaborar una puntuación del valor teórico para cada encuestado**, como se describió anteriormente. Si requiere evaluar 3 mensajes publicitarios tanto en sus pretensiones como en las intenciones de compra que ellos generan, emplearíamos el **MANOVA.**

Con el **MANOVA** se desea encontrar ahora el **conjunto de ponderaciones que maximizan el valor de la F del ANOVA calculado sobre las puntuaciones** de los valores teóricos para todos los grupos. **MANOVA** también puede ser considerado como una **extensión** del **análisis discriminante** en donde pueden ser construidos los valores teóricos de las medidas dependientes si el número de grupos es tres o más. El primer valor teórico, denotado como **función discriminante**, establece un conjunto de **ponderaciones que maximizan las diferencias entre los grupos**, y por tanto **maximiza el valor de la *F*.** El **valor de la *F*** máximo nos permite calcular directamente lo que es conocido como **el estadístico de la mayor raíz característica** (**mrc**), que da cabida para el contraste estadístico de la primera función discriminante. Así, se calcula:
mrc= (k-1) F_{max} ***/ (N-k) [***Hubert y MmTis 1989***]***
Para obtener un único contraste de la hipótesis de que no existen diferencias de los grupos sobre los **vectores de las puntuaciones medias**, podemos emplear las **tablas de la distribución de la mrc**.
Al igual que el **estadístico *F*** seguía una distribución conocida bajo la **hipótesis nula** de que las medias de los grupos son equivalentes sobre una variable dependiente única, el estadístico ***mrc*** sigue una distribución conocida bajo la **hipótesis nula** de que los **vectores de las medias** de los grupos son equivalentes (por ejemplo, las medias de los grupos son equivalentes sobre un conjunto de medidas dependientes). La comparación de la ***mrc***

observada con la $\boldsymbol{mrc_{critica}}$ nos da una base para rechazar la hipótesis nula global de que los vectores de las medias de los grupos son equivalentes.

Cualquiera de las **funciones discriminantes** posteriores son **ortogonales**; éstas maximizan las diferencias entre los grupos basándose en la **varianza** que **permanece no explicada por la(s) función(es) anterior(es)**. Por ello, en muchos casos, **el contraste para las diferencias entre los grupos incluye no sólo un único valor teórico sino un conjunto de puntuaciones de valores teóricos que son evaluadas simultáneamente**. Se hallan disponibles un conjunto de contrastes multivariantes, cada uno de los cuales es el más apropiado para unas situaciones específicas para la contrastación de estos valores teóricos múltiples.

8.4. ANOVA/MANOVA vs. Análisis Discriminante

Ya comentados los elementos básicos tanto de los **contrastes univariantes** como los **multivariantes** para valorar las diferencias **entre los grupos en una o más variables dependientes**. De esta manera, observamos el cálculo de la función discriminante; en el caso de **MANOVA** es el **valor teórico de las variables dependientes que maximiza la diferencia entre grupos**. Así que, ¿cuál es la diferencia entre **MANOVA** y el análisis discriminante? En algunos aspectos, **MANOVA** y el análisis discriminante son ***"imágenes de espejo".*** Las variables dependientes en **MANOVA** (**variables métricas**) son las **variables independientes** en el **análisis discriminante** y una simple **variable dependiente no métrica** del **análisis discriminante** se convierte en la **variable independiente** en el **MANOVA**. Además, ambos utilizan los mismos métodos en la formación de valores teóricos y evalúan la significación estadística entre los grupos. Las diferencias, sin embargo, se centran alrededor de los objetivos de los análisis y el papel de la variable no métrica. El análisis discriminante emplea una variable no métrica simple como variable dependiente. Se supone que las categorías de la variable dependiente están dadas y que se utilizan las variables independientes para formar valores teóricos que son diferentes de manera máxima entre los grupos formados por las categorías de la variable dependiente.

En el **MANOVA**, la serie de variables métricas actúan ahora como variables dependientes y el objetivo es encontrar grupos de encuestados que exhiben diferencias sobre la serie de variables dependientes. Los grupos de encuestados no son especificados previamente; en su lugar, el investigador utiliza una o más variables independientes (variables no métricas) para formar grupos. El **MANOVA**, incluso mientras forma estos grupos, retiene todavía la capacidad de valorar el impacto de cada variable no métrica por separado.

8.5. ANOVA/ MANOVA cuando utilizar

Usted puede obtener diferentes ventajas con el empleo del **ANOVA/MANOVA**, ya que esta técnica permite el examen de **varias medidas dependientes simultáneamente. Desde** El punto de vista del uso **ANOVA/MANOVA,** éste se puede analizar desde la base de la **eficiencia y de control de la precisión estadística** mientras nos proporcione la manera apropiada para contrastar cuestiones multivariantes.

8.5.1. Control del porcentaje de errores experimentales

El uso de **ANOVAs univariantes separados o de contrastes *t*** puede crear un problema al controlar la **tasa de errores experimentales o globales [**Hubert y MmTis, 1989**]**. Por ejemplo, suponga que analizamos una serie de cinco variables dependientes utilizando **ANOVAs** separados, con **0.05** nivel de significación cada uno. **Si no existen diferencias reales** entre las variables dependientes, **esperaríamos que no se observase ningún efecto** de alguna variable dependiente el **5 %** de las veces. Sin embargo, en los **5 contrastes separados**, la probabilidad del **error de Tipo 1** se situaría entre el **5%** (si todas las variables dependientes están **perfectamente correlacionadas**) y (**1 – 0.955**) **23%** si todas las variables dependientes **no están correlacionadas**.

Así, **un conjunto de contrastes separados No nos permite ningún control de nuestro porcentaje efectivo del error de Tipo l**. Si Usted **quiere mantener el control sobre el porcentaje del error experimental** y existe algún **grado de correlación entre las variables dependientes,** entonces la técnica más apropiada es el **MANOVA.**

8.5.2. Diferencias entre una combinación de variables dependientes

Una serie de **contrastes univariantes ANOVA** tampoco tiene en cuenta la posibilidad de que alguna combinación (**combinación lineal**) de las **variables dependientes** pueda proporcionar **evidencia de la existencia de alguna diferencia global en los grupos**, **que puede no detectarse si se examina cada variable dependiente separadamente**. **Los contrastes individuales ignoran las correlaciones entre las variables dependientes y por ello no se emplea toda la información disponible para valorar diferencias globales en los grupos.** Si existe **multicolinealidad** entre las variables dependientes, el **MANOVA será más potente que las contrastes univariantes separados**. De esta manera, el **MANOVA** puede detectar diferencias combinadas que no se encuentran con los contrastes univariantes. Además, si se construyen valores teóricos múltiples, entonces estos pueden proporcionar dimensiones de diferencias que puedan distinguir entre los grupos mejor que las variables de forma independiente.

Sin embargo, en algunos casos donde existe un gran número de variables dependientes, **la potencia de los contrastes ANOVA excede lo que se obtiene empleando un único MANOVA**.

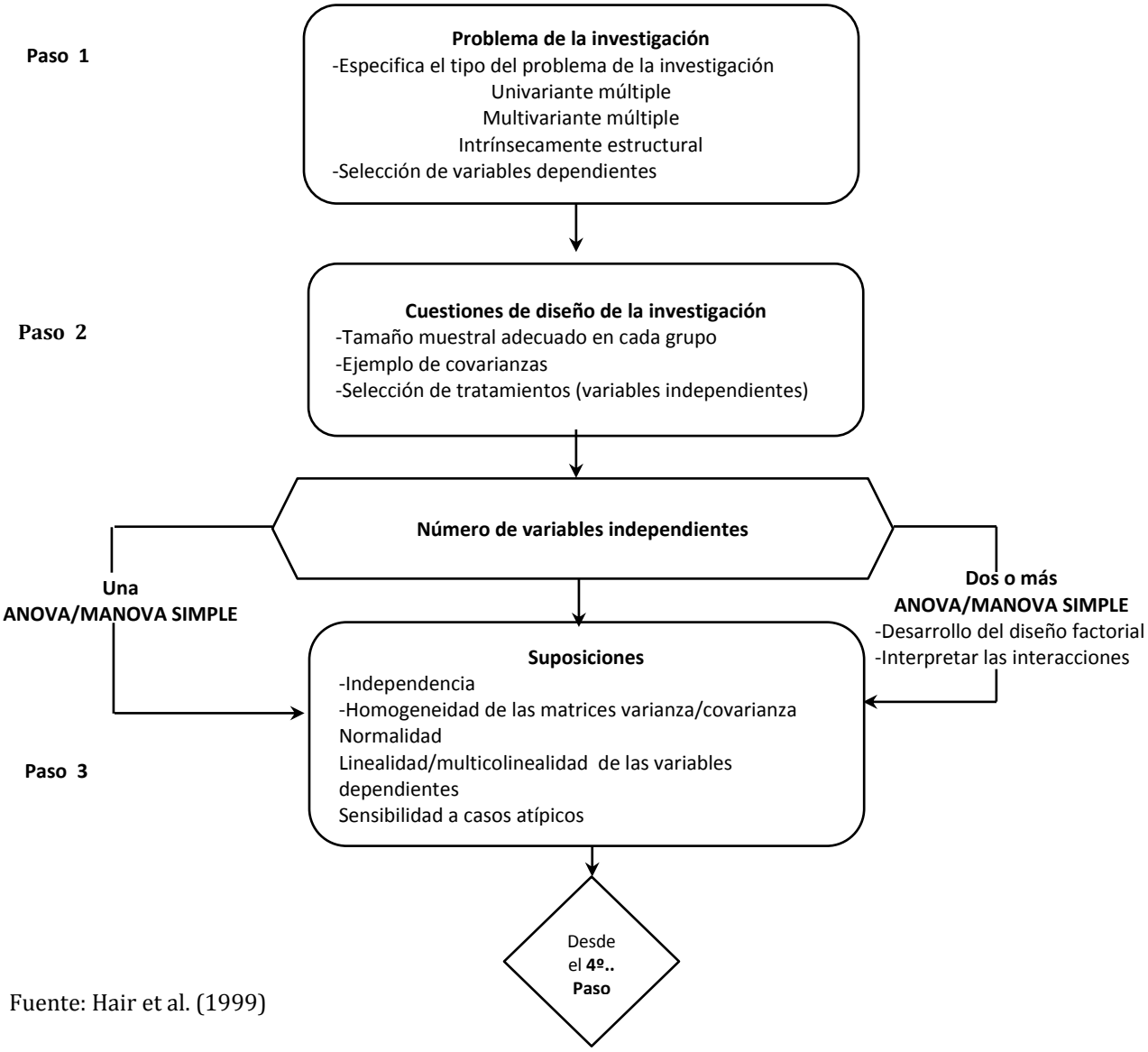
El proceso para llevar a cabo el análisis multivariante de la varianza es similar al existente en otras muchas técnicas multivariantes, por lo que puede ser descrito a través del proceso de construcción de un modelo en seis pasos descrito en el **Capítulo 2**. Este proceso comienza con la especificación de los objetivos de la investigación. Después continúa con un conjunto de cuestiones de diseño a las que se enfrente el análisis multivariante y con el estudio de los supuestos básicos del **MANOVA**. Con estas cuestiones ya tratadas, el proceso avanza con la estimación del modelo **MANOVA** y la valoración del ajuste global del modelo.

Cuando se encuentra un modelo **MANOVA** aceptable, entonces se deben interpretar los resultados con más detalle. El último paso comprende los intentos para validar los resultados y asegurar generalidad en la población.
Las **Figura 8.2** (pasos 1-3) y **8.3** (pasos 4-6) nos proporcionan un esquema gráfico del proceso, el cual se discute en detalle en las siguientes secciones.

8.6. ANOVA/MANOVA y el proceso de decisión

De acuerdo al modelo de seis pasos que se introdujo en el **Capítulo 2**, la **Figura 8.27.** muestra tres pasos iniciales de la aproximación estructurada para la construcción de modelos multivariantes.

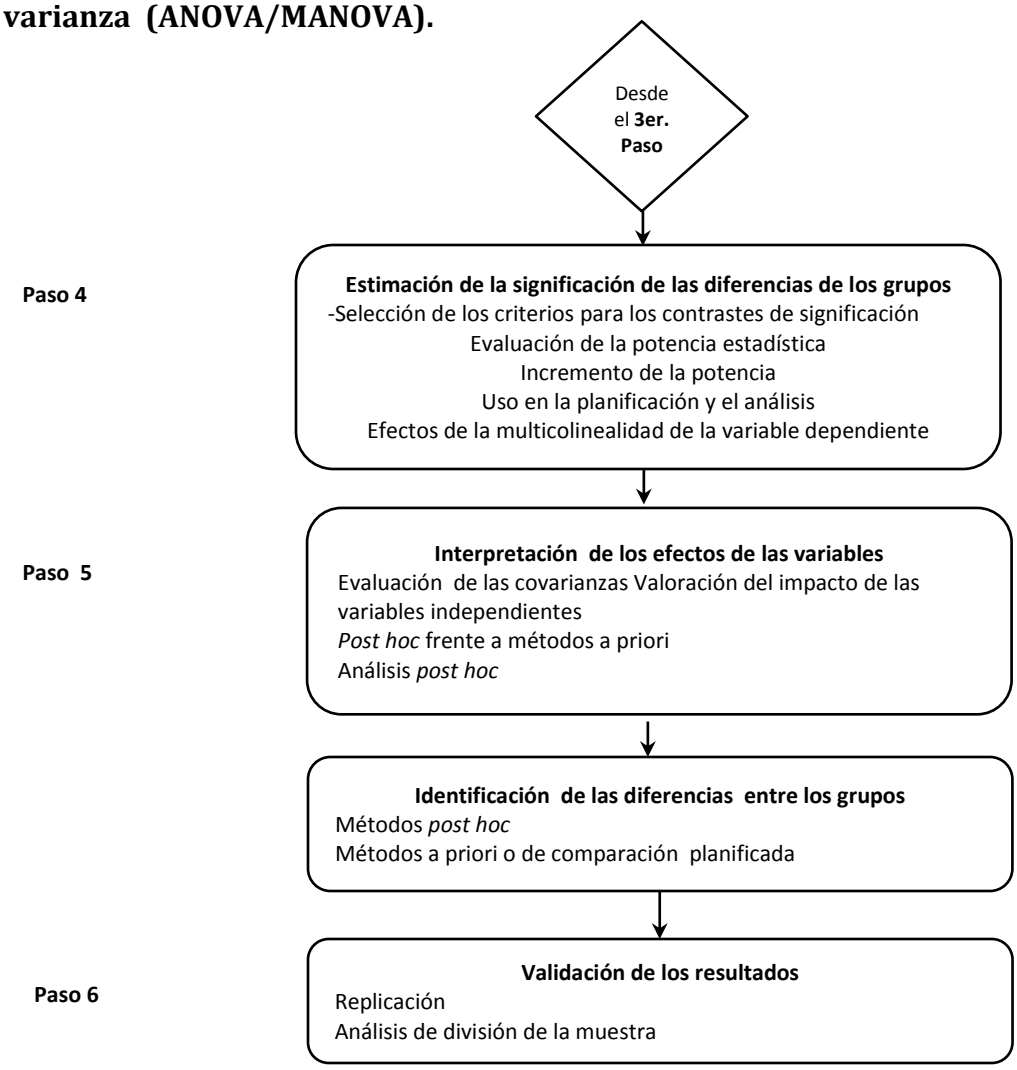
Figura 8. 27. Pasos 1 a 3 en el diagrama de decisión del análisis multivariante de la varianza (ANOVA/MANOVA).



Fuente: Hair et al. (1999)

La **Figura 8.28** muestra en detalle los últimos tres pasos, y uno adicional (el séptimo paso) más allá de la estimación, la interpretación y la validación de los modelos factoriales, que ayuda a la selección de las variables sustitutas, las puntuaciones de factores o la creación de las escalas aditivas para la utilización en otras técnicas multivariantes. A continuación se presenta un análisis de cada paso

Figura 8.28. Pasos 4 a 6 en el diagrama de decisión del análisis multivariante de la varianza (ANOVA/MANOVA).



Fuente: Hair et al. (1999)

8.7. ANOVA/MANOVA. Paso 1: Objetivos

Paso 1: Establecimiento de objetivos

La selección del **MANOVA** está basada en el requerimiento de analizar una relación de dependencia representada como las **diferencias en un conjunto de medidas dependientes a través de una serie de grupos formados por una o más medidas independientes categóricas.** Como tal, el **MANOVA** representa una poderosa herramienta analítica apropiada para una amplia colección de cuestiones de investigación, ya que se emplea en **situaciones cuasi-experimentales** o reales (tales como trabajos de campo o estudios de encuestas donde las **medidas independientes son categóricas**), puede proporcionar no sólo un **entendimiento de la naturaleza y del poder predictivo** de las medidas independientes, sino también de la **interrelaciones y diferencias** observadas en el conjunto de **medidas dependientes**.

Las ventajas del **MANOVA** frente a un conjunto de **ANOVAs** univariantes se dan en el campo estadístico discutido anteriormente, y también en su capacidad de **proporcionar un único método de contrastar un amplio conjunto de cuestiones multivariantes diferentes**. A lo largo del texto, nos centraremos en la naturaleza interdependiente del análisis multivariante. El **ANOVA/MANOVA** tiene un carácter interdependiente que le permite la flexibilidad a que Usted **seleccione los contrastes estadísticos** más apropiados para las cuestiones que le afectan. Se tiene clasificación una clasificación de los problemas multivariantes en **3 categorías**, cada una de las cuales emplea diferentes aspectos del **ANOVA/MANOVA** en su resolución, a través de formular **3 preguntas de:**

1. **ANOVA/MANOVA univariante múltiple:** Cuando Usted lo aborda, identifica un número de **variables dependientes separadas** (por ejemplo: renta, edad, educación de los consumidores, etc.) que son **analizadas de forma separada** pero en donde se necesita algún **control** sobre el **porcentaje de error experimental.** En este caso, se emplea el **MANOVA** para valorar si se encuentra **alguna diferencia global entre los grupos,** y después se llevan a cabo **contrastes univariantes separados** para dar respuestas individuales a cada variable dependiente**.**
2. **ANOVA/MANOVA estructurado:** Cuando Usted lo practica, reúne dos o más medidas dependientes que tienen unas determinadas relaciones entre ellas. Una situación habitual en esta categoría son las **medidas repetidas**, donde las respuestas múltiples son reunidas para cada sujeto, quizá a lo largo del tiempo o en una orientación **precontraste-postcontraste** para algún estímulo, como por ejemplo un anuncio publicitario. En este caso, el ANOVA/ MANOVA proporciona un método estructurado para especificar las comparaciones de las diferencias de los grupos sobre un conjunto de medidas dependientes mientras se mantiene la eficiencia estadística.
3. **ANOVA/MANOVA intrínseco:** Cuando Usted lo practica, se enrola en un conjunto de medidas dependientes en donde la **principal cuestión** es cómo se diferencian **como un total entre los grupos**. Las diferencias en las medidas dependientes individuales son de menor interés que su efecto conjunto. Un ejemplo es la **contrastación de medidas** de respuesta múltiple que deben ser **consistentes**, tales como **opinión**, **preferencia e intención de compra**, todas ellas relacionadas con diferentes campañas publicitarias. La potencia completa del **ANOVA/MANOVA** se utiliza en este caso para valorar no solamente las diferencias globales sino también las diferencias entre las combinaciones de medidas dependientes que de otro modo no serían evidentes. Este tipo de cuestión

se aborda de forma correcta con el **ANOVA/MANOVA** debido a su capacidad para detectar diferencias multivariantes, incluso cuando ningún contraste univariante muestra diferencias.

8.8. ANOVA/MANOVA.Paso 2: Diseño

Paso 2: Diseño

Aunque el **MANOVA** contrasta los supuestos básicos de la misma forma que el **ANOVA** y comparte los mismos principios, algunos aspectos de su aplicación son únicos. Estas cuestiones afectan tanto al diseño como a la contrastación estadística del modelo **ANOVA/MANOVA**

8.8.1. El tamaño muestral

MANOVA requiere **tamaños muestrales más grandes** que los **ANOVAs univariantes** y el tamaño muestral debe exceder umbrales específicos en cada grupo del análisis. **Un tamaño de grupo mínimo es de 20 observaciones**, aunque es posible que se necesiten tamaños de grupo mayores para una **potencia estadística** aceptable. Como **mínimo**, el **tamaño en cada grupo debe ser más grande que el número de variables dependientes incluidas**. Aunque este aspecto puede parecer menor, la introducción de un número pequeño de **variables dependientes** (**5 a 10**) en el análisis pone una restricción algunas veces en el **campo de la experimentación o del estudio de investigación**, donde el investigador tiene menor control sobre la muestra obtenida.

8.8.2.-Diseños factoriales

Muchas veces, Usted deseará **examinar los efectos de varias variables independientes o tratamientos en vez de usar solamente un tratamiento único en las pruebas de ANOVA o MANOVA**. Un análisis con más de dos tratamientos se denomina **diseño factorial**. En general, un diseño **con *n* tratamientos** se denomina **diseño factorial de n-vías.**

Así, la selección procederá como:

1. Los diseños factoriales más habitualmente empleados **son esas preguntas de investigación que relacionan dos o más variables independientes no-métricas con un conjunto de variables dependientes.** En estos casos, las **variables independientes** se han especificado en el **diseño del experimento de campo en un cuestionario**.
2. En algunos casos, **los tratamientos son añadidos después de que el análisis ha sido diseñado**. Los tratamientos adicionales de uso más común se introducen como **factor en bloques**, que es una característica **no-métrica** empleada de forma posterior **para dividir a los encuestados con el fin de obtener una homogeneidad dentro del grupo más grande y reducir la fuente de diferencias intra-grupo**. Si se realiza esto, la capacidad de los contrastes estadísticos para identificar las diferencias **aumenta**. Como ejemplo, supongamos que en nuestro ejemplo de los anuncios publicitarios anterior, se descubriera que los hombres reaccionan de forma diferente que las mujeres frente a los anuncios. Si el **género** es entonces empleado como un **factor en bloques**, las diferencias entre los mensajes pueden llegar a ser más evidentes, mientras se tenía que las **diferencias estaban ocultas** cuando se consideraba que los **hombres y las mujeres reaccionaban de forma similar y no estaban separados**. Los efectos del tipo

de mensaje y el sexo son entonces **evaluados de forma separada**, proporcionado un contraste más preciso de sus efectos individuales.

3. Un caso de ejemplo de un diseño factorial simple de **2 tratamientos:** Supongamos que un productor de **Smartphone** desea examinar el impacto de tres posibles colores (negro, blanco y oro) y tres formas diferentes de adquisición (contrato pre-pago, contrato post-pago, de contado) sobre la evaluación global de un consumidor. Podríamos examinar el impacto de estas variables independientes simultáneamente empleando un **diseño factorial 3 X 3**. Los encuestados serían asignados aleatoriamente para **evaluar una de las nueve posibles combinaciones** de color y forma de adquisición (en una escala de valoración general de **diez puntos**). En el análisis de este diseño, se pueden contrastar tres efectos diferentes globales con el **ANOVA:**

   l. El efecto principal del color: ¿existen algunas diferencias entre las clasificaciones **medias dadas al negro** (por ejemplo, incluyendo todas las clasificaciones de pre-pago negro, post-pago negro y pago directo negro), con respecto al blanco y al oro?

   2. El efecto principal de la forma de adquisición: ¿existen algunas diferencias entre las clasificaciones medias dadas al pre-pago (por ejemplo, incluyendo todas las clasificaciones de prepago negros, pre-pago blancas, y prepago oro), con respecto a los post-pago y de contado?

   3. El efecto interacción del color y la forma: al igual que la diferencia global entre los colores, ¿es esta diferencia la misma cuando se examine de forma separada de las pre-pago, post-pago, de contado? Por ejemplo, si el negro fue clasificado muy alto pero recibió una clasificación muy baja cuando fue clasificado como de contado (en relación al blanco y al oro), **este resultado daría evidencia de un efecto interacción**; **es decir, el efecto del color depende de con qué forma de adquisición sea considerado.** Podríamos plantear esta cuestión de la interacción de una forma equivalente **preguntando si el efecto de la forma de adquisición depende de con qué color es considerado**.

   4. En los diseños factoriales del **ANOVA**, cada uno de estos **3 efectos sería contrastado con un estadístico *F***. Los diseños factoriales del **MANOVA** son una sencilla extensión del **ANOVA**; es decir, **para cada estadístico *F* en el ANOVA que evalúa un efecto sobre una única variable dependiente, existe un correspondiente estadístico multivariante** (por ejemplo, ***mrc*** o ***lambda de Wilks***) que evalúa el mismo efecto sobre un **conjunto (vector) de medias de las variables dependientes.**

8.8.3. Interpretación de los términos de la interacción

Interacción es **el efecto conjunto de dos tratamientos y es el efecto que debe ser examinado en primer lugar.**

1. **Si la interacción no es estadísticamente significativo**, entonces los efectos de **los tratamientos serán independientes** esto es, que **el efecto de un tratamiento es el mismo para cada nivel del (los) otro(s) tratamiento(s), y que los efectos principales pueden ser interpretados directamente.**
2. **Si la interacción es significativa,** entonces se debe determinar el tipo de interacción, esto es: **ordinal o disordinal**:

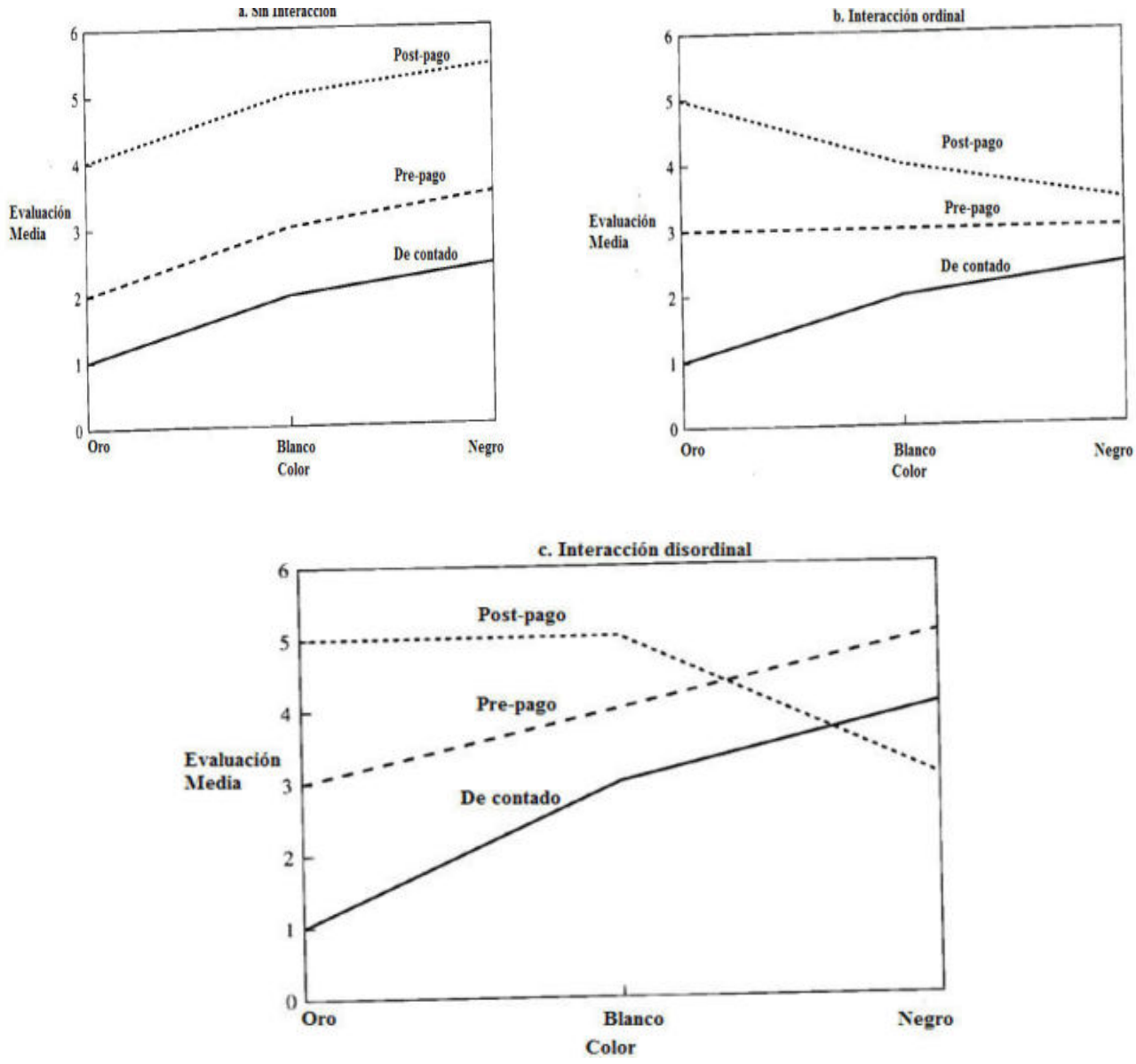
-Una **interacción ordinal** se produce cuando **los efectos de un tratamiento no son iguales para todos los niveles del otro tratamiento,** pero la magnitud es siempre de la misma dirección.

-En una **interacción disordinal,** los efectos de un tratamiento son positivos para algunos niveles y negativos para los otros niveles del otro tratamiento.

Las diferencias entre las interacciones son mejor descritas gráficamente. En la **Figura 8.29.** del caso **Smartphone**, el eje x re presenta las tres categorías de color (negro, blanco y oro). Las líneas conectan la media de la categoría para cada forma a través de los tres colores. Por ejemplo en el gráfico superior, el valor para las **post-pago oro** es aproximadamente **4.0**, el valor para la **post-pago blanca** es aproximadamente **5.0** y el valor se incrementa ligeramente hasta aproximadamente **5.5** para el **post-pago negro**. En cuanto a las interacciones:

1. C**aso a.**, **no existe interacción**. **La falta de interacción se puede ver** cuando las líneas que representan las diferencias de las distintas formas entre los niveles del color son **paralelas** (se observaría el mismo efecto si se representasen las diferencias en el color entre los tres tipos de formas). En el caso donde no existe interacción, **los efectos de cada tratamiento son constantes en cada nivel y las líneas son aproximadamente paralelas**.
2. **Caso b.**, vemos que los efectos de cada tratamiento **no son constantes y por ello las líneas no son paralelas.** Las diferencias para el **oro** son grandes, pero disminuyen ligeramente para el **Smartphone blanco** e incluso más para el **negro**. Por ello, las diferencias en color varían entre las formas. Pero la ordenación relativa entre los niveles de la forma de adquisición es la misma, con las **post-pago** siempre como la más alta, seguidas de los de pre-pago y después de las **de-contado**.
3. **Caso c.** las diferencias en el color no sólo varían en magnitud sino también en dirección. Este efecto se observa cuando **las líneas no son paralelas** y se cortan entre los niveles. Por ejemplo, los **pre-pago y de contado** tienen una valoración mayor que las de **post-pago** cuando el color es el negro, pero la valoración es menor para los colores blanco y oro.

Figura 8.29. Efectos de interacción en los diseños factoriales……..



Fuente: propia

Si las interacciones significativas **son ordinales**, el investigador **debe interpretar la interacción y asegurar** que sus resultados son **conceptualmente aceptables, con:**

1. Si es así, entonces se pueden describir los efectos de cada tratamiento.
2. Pero si la interacción significativa es **disordinal**, entonces **los efectos principales de los tratamientos no pueden ser interpretados** y el **estudio debe ser rediseñado,** ya que con la interacción **disordinal** los efectos varían, no solamente entre los niveles del tratamiento, sino también en la dirección (**positiva o negativa**). Por ello, los tratamientos no representan un efecto consistente.

Si las interacciones significativas son **ordinales**, Usted deberá interpretar la interacción y asegurar que sus resultados son conceptualmente aceptables. Si es así, entonces se pueden describir los efectos de cada tratamiento. Pero si la interacción significativa es **disordinal**, entonces los efectos principales de los tratamientos no pueden ser interpretados y el estudio debe ser rediseñado, ya que con la interacción **disordinal** los efectos varían, no solamente entre los niveles del tratamiento, sino también en la dirección **(positiva o negativa**). Por ello, los tratamientos no representan un efecto consistente.

8.8.4. Uso de covariaciones- ANCOVA y MANCOVA

En cualquier diseño del **ANOVA** se pueden incluir variables **independientes métricas**, denominadas **covarianzas**. El diseño es entonces calificado como **diseño del análisis de la covarianza (ANCOVA)**. Las **covarianzas métricas** son incluidas normalmente en un diseño **experimental** para **eliminar influencias extrañas de la variable dependiente**, **que incrementen la varianza dentro de los grupos** ($\mathbf{CM_I}$. Esta situación es similar al uso de un factor de bloqueo, solamente que esta vez la **variable es métrica.** Se han empleado procedimientos similares a la **regresión lineal** para **eliminar la variación de la variable dependiente asociada con una o más covarianzas**. Después se lleva a cabo un análisis **ANOVA** convencional sobre la variable dependiente ajustada. El análisis multivariante de covarianza (**MANCOVA**) es una extensión simple de los principios de **ANCOVA** para el análisis multivariante (**variables dependientes múltiples**); es decir, se puede considerar **MANCOVA** como **MANOVA** de los residuos de regresión, o varianza en las variables dependientes que no son justificados por las covarianzas.

8.8.5. Objetivos del análisis de la covarianza

El **análisis de covarianzas** es apropiado para lograr 2 objetivos específicos:

1. **Eliminar cualquier error sistemático** fuera del control del investigador que pueda sesgar los resultados, y
2. **Tener en cuenta las diferencias en las respuestas debidas a características propias de los encuestados**. Un sesgo sistemático puede ser eliminado por medio de la asignación aleatoria de los encuestados a varios tratamientos. Sin embargo, **en estudios no experimentales, estos controles no son posibles.**

Por ejemplo, al contrastar los anuncios publicitarios, los efectos pueden diferir dependiendo del momento del día o de la composición de la audiencia y de sus reacciones. **El objetivo de la covarianza es eliminar cualquiera de los efectos que (a) influyen solamente a una parte de los encuestados, (b) varían entre los encuestados.** Por ejemplo, las diferencias personales, tales como **actitud u opiniones**, pueden afectar a las respuestas, pero el experimento **no las incluye como un factor de tratamiento**. El investigador utiliza una **covarianza** para extraer cualquiera de las diferencias debidas a estos factores antes de que los efectos del experimento sean calculados. **Este es el segundo papel del análisis de la covarianza**.

8.8.6. Selección de las covarianzas.

Una covarianza efectiva en el **ANCOVA** es aquella que está **altamente correlacionada con la variable dependiente pero no está correlacionada con las variables independientes**. La varianza de **la variable dependiente** forma la base de nuestro

término de error en el ANOVA. Si nuestra **covarianza** está correlacionada con la **variable dependiente**, podemos **explicar algo de la varianza** (por medio de la **regresión lineal**), y nos quedamos solamente con la **varianza residual** de la **variable dependiente** que no hubiera sido explicado por la **variable independiente** de todas formas (**porque la covarianza no está correlacionada con la variable independiente**). Esta varianza residual proporciona un término de error más pequeño (**CM_I** para el **estadístico *F*** y por ello un contraste más eficiente para los efectos del tratamiento. No obstante, **si se correlaciona la covarianza con la variable independiente**, **entonces la covarianza "*explicará*" parte de la varianza que podría haber sido "*explicada*" por la variable independiente y reducir sus efectos.** Dado que se extrae primero la covarianza**, cualquier variación que se asocia con la covarianza no está disponible para las variables independientes.**

Una cuestión habitual es determinar cuántas covarianzas se añaden al análisis. Aunque Usted quiera tener en cuenta tantos efectos extraños como sea posible, **un número demasiado grande reducirá la eficiencia estadística de los procedimientos,** proponiéndose:

El número de covarianzas < (**0.10* tamaño muestra**)- (**número de grupos -1**) [Huitema, 1980]

Por ejemplo, para un tamaño muestral de **100 encuestados y 5 grupos**, el número de covarianzas debe ser **< 6**, esto es, **0.10*100-(5-1)**.

Usted siempre debe intentar minimizar el número de covarianzas, mientras también **asegura las que no van a ser eliminadas**, ya que en muchos casos, particularmente cuando hay tamaños muestrales pequeños, estas covarianzas pueden mejorar de manera importante la sensibilidad de los contrastes estadísticos.

Existen dos requisitos para el empleo del análisis de la covarianza:

1. Las covarianzas deben tener algún tipo de relación con las medidas dependientes, y
2. Las covarianzas deben presentar una homogeneidad del efecto de la regresión, significando esto que la(s) covarianza(s) tiene(n) **efectos iguales sobre la variable dependiente entre los grupos**. Existen contrastes estadísticos para valorar si este supuesto básico es cierto para cada covarianza empleada. Si alguno de estos requisitos no se da, **entonces el empleo de la covarianza es inapropiado**.

8.8.7. Caso especial del MANOVA: Medidas repetidas

Hemos tratado un conjunto de situaciones en las que deseamos examinar las diferencias sobre varias medidas dependientes. Una situación especial de este tipo se produce cuando el mismo encuestado proporciona varias medidas, tales como puntuaciones de un contraste a lo largo del tiempo, y nosotros deseamos examinarlas para comprobar si existe alguna tendencia. Sin embargo, sin un tratamiento especial, estaríamos violando el supuesto básico más importante, **la independencia**. Existen modelos **MANOVA** especiales, **denominados medidas repetidas**, que pueden tener en cuenta esta dependencia y sirven para **averiguar si existen algunas diferencias entre los in dividuos del conjunto de variables dependientes. La perspectiva de una persona-de-dentro** es importante para que cada persona sea considerada en **"*igualdad*";** por ejemplo, supongamos que estuviéramos valorando la mejora en las puntuaciones de un contraste a lo largo de un semestre. Debemos tener en cuenta las puntuaciones de los contrastes anteriores **y cómo**

éstas se relacionan con las puntuaciones posteriores, y podríamos esperar que se observaran diferentes tendencias para aquellos con **puntuaciones iniciales bajas frente a los que las tienen altas.** Por ello, debemos "***emparejar***" las puntuaciones de cada encuestado cuando se lleva a cabo el análisis. Mayor información en [Cattell, 1966, Cooley y Lohnes 1971, Orcen, 1978, Oreen, y Tull 1979, Harris, 1975, Morrison. D. F. 1967, Tatsuoka,1971].

8.9. ANOVA/MANOVA. Paso 3: Supuestos

Paso 3: Supuestos de aplicabilidad.

Los procedimientos de los contrastes univariantes del **ANOVA** descritos en este capítulo son válidos (en un sentido formal) solamente si se **supone que la variable dependiente está distribuida normalmente y que las varianzas son iguales para todos los grupos de tratamiento.** Sin embargo, existe evidencia **[**Meyers, 1975**,** Winer, 1962**]** de que los **contrastes F** en el **ANOVA** son robustos respecto a estos supuestos excepto en algunos casos extremos. Para que los procedimientos de los **contrastes multivariantes del MANOVA sean válidos**, se deben cumplir tres supuestos:

1. **Las observaciones deben ser independientes,**
2. **Las matrices de varianzas-covarianzas deben ser iguales** para todos los grupos de tratamiento y
3. El conjunto de las **p-variables dependientes** debe seguir una **distribución normal multivariante** (por ejemplo, **cualquier combinación lineal de las variables dependientes debe seguir una distribución normal**) **[**Harris, R. J. 1975**].** Además de los estrictos supuestos estadísticos, el investigador también debe considerar varios aspectos que afectan a los posibles efectos a saber, **la linealidad y la multicolinealidad** del valor teórico de las **variables dependientes.**

8.9.1. Independencia

Es el incumplimiento más básico, aunque más importante, de un supuesto, es decir, cuando existe una falta de independencia entre las observaciones. Ya sea en situaciones experimentales o no, este supuesto puede ser violado fácilmente en casos como:

1. Producir un efecto de ordenación en el tiempo (**correlación serial**) si las medidas son tomadas a lo largo del tiempo, incluso de diferentes encuestados.
2. Por **obtención de información en la composición de los grupos**, por lo que una experiencia común (como una habitación ruidosa o un conjunto confuso de instrucciones) produciría un **subconjunto de individuos** (aquellos con experiencias comunes) que tendrían respuestas que estarían un tanto correlacionadas.
3. **Los efectos extraños y no medidos** pueden afectar a los resultados creando una dependencia entre los encuestados.

Aunque **no existen contrastes** con una certeza absoluta para detectar todas las formas de dependencia, Usted debe **explorar todos los posibles efectos y corregirlos de encontrarlos**, mediante**:**

1. **Combinar a éstos dentro de los grupos y analizar la puntuación media del grupo** en lugar de las puntuaciones de los diferentes encuestados.
2. **Emplear alguna forma de análisis de la covariación** para tener en cuenta la dependencia.

En cualquier caso, o cuando se sospecha que existe dependencia, Usted **debe emplear un menor nivel de significación (0.01 o incluso menor).**

8.9.2. Igualdad de las matrices de varianzas-covarianzas entre grupos

Es segundo supuesto básico del **MANOVA** En este caso, al igual que con el problema de la heteroscedasticidad vista en la regresión múltiple, estamos interesados en las **diferencias sustanciales en la cantidad de varianza de un grupo en comparación con la de otro grupo** para las mismas variables. Sin embargo, en el **MANOVA** el interés se centra en las **matrices de varianzas-covarianzas** de las medidas dependientes para cada grupo. El requisito de igualdad es un contraste muy riguroso, ya que en lugar de contrastar igualdad de varianzas para una única variable como en el **ANOVA**, el contraste del **MANOVA examina todos los elementos de la matriz de covarianzas**. Por ejemplo, para cinco variables dependientes, se contrasta la igualdad de las cinco correlaciones y de las diez covarianzas entre los grupos. Afortunadamente, **la violación de este supuesto tiene un mínimo impacto si los grupos son aproximadamente de igual tamaño** (si el tamaño del grupo más grande dividido por el tamaño del grupo más pequeño **es menor de 1.5**). Si los tamaños difieren más que esta medida, entonces Usted, si es posible, **debe contrastar y corregir la posible existencia de varianzas distintas**. Los programas del **MANOVA** proporcionan el **contraste para la igualdad de las matrices de covarianzas** (**contraste de *Box***) y aportan los niveles de significación para el contraste. Si Usted encuentra una diferencia significativa que requiere algún tipo de tratamiento, se utiliza una de las **muchas transformaciones de estabilización de la varianza disponibles** (**vea Capítulo 3**). Aun así, **el contraste de *Box* es muy sensible a la falta de normalidad [**Harris, 1975**,** Stevens, 1972**]**. Por ello se debe siempre comprobar la **normalidad univariante** de todas las medidas dependientes antes de llevar a cabo este contraste.

Si persisten las diferencias entre las varianzas después de la transformación y los tamaños delos grupos difieren de forma importante, Usted debe **realizar una serie de ajustes que tengan en cuenta estos efectos:**

1. Debe determinar qué grupo presenta la varianza mayor. Esta determinación se realiza fácilmente, bien:
   -Examinando la matriz de varianzas-covarianzas o bien
   -Usando el determinante de esta matriz, que es calculado por todos los paquetes estadísticos. Si las mayores varianzas pertenecen a los **grupos de mayor tamaño, el nivel alfa de significación tiende a ser más grande.** Este resultado significa que las diferencias deben ser realmente tenidas en cuenta **empleando para ello algún valor más pequeño** (por ejemplo, **emplear 0.03 en lugar de 0.05**).
2. **Si las mayores varianzas se dan en los grupos de tamaño más pequeño**, entonces se producirá lo **contrario**. Es decir, la potencia del contraste se vería reducida, y por ello el investigador **debería incrementar el nivel de significación**

8.9.3. Normalidad

Es el último supuesto básico para considerar en las medidas dependientes. En estricto sentido, el supuesto es que todas las variables siguen una **distribución normal multivariante,** la cual supone que **el efecto conjunto de 2 variables se distribuye normalmente.** Aunque este supuesto subyace en la mayoría de las técnicas multivariantes, **no existe un contraste preciso para la normalidad multivariante**, por lo que **la mayoría de los investigadores contrastan la normalidad univariante para cada variable.** Pero aunque la **normalidad univariante no garantiza la normalidad multivariante,** si todas las variables cumplen ese requisito, entonces **cualquier posible incumplimiento de este supuesto es generalmente insignificante.** Las violaciones de este su puesto tienen:

1. Una **pequeña influencia si los tamaños muestrales son grandes**, al igual que ocurre en el **ANOVA**.
2. La violación de este supuesto crea principalmente problemas para utilizar **el contraste de *Box***, pero las transformaciones pueden corregir estos problemas en la mayoría de las situaciones. (Para transformación de variables, vea **Capítulo 3**).
3. Cuando los tamaños muestrales son **medianos**, los **incumplimientos pequeños** pueden ser suavizados siempre y cuando las diferencias sean debidas a **asimetría** y no a datos **anómalos.**

8.9.4. Linealidad y multicolinealidad entre las variables dependientes

Aunque el **MANOVA** valora las diferencias entre **combinaciones de medidas dependientes, solamente se puede construir una relación lineal entre las medidas dependientes (y cualquier covariación, si está incluida)**. Así que se sugiere:

1. **Examinar los datos, valorando la posible presencia de relaciones no lineales.** Si existen, entonces se debe tomar la decisión de si se deben incluir dentro del conjunto de variables dependientes, que tiene como coste aumentar la complejidad pero también se incrementa la representatividad. El **Capítulo 3** aborda estas pruebas.
2. Además del requisito de la linealidad, **las variables dependientes no deben presentar alta multicolinealidad entre ellas,** porque esto sólo indica medidas dependientes redundantes y tiende a **disminuir la eficiencia estadística.** Recuerde el impacto de la multicolinealidad sobre la **potencia estadística del MANOVA** en la próxima sección.

8.10. ANOVA/MANOVA. Paso 4: Estimación y Ajuste

Paso 4: Estimación y ajuste

Una vez que el análisis del **MANOVA** ha sido formulado y los supuestos básicos contrastados, se de realizar la valoración de las diferencias significativas entre los grupos formados por el(los) tratamiento(s) (vea **Figura 8.3**). Con esta valoración, Usted debe seleccionar los contrastes estadísticos más apropiados para los objetivos que se persiguen. Además, en cualquier situación, pero especialmente cuando el análisis llega a ser más complejo, Usted debe evaluar la potencia de los contrastes estadísticos para proporcionar la mejor representación de los resultados obtenidos.

8.10.1. Criterios para la contrastación de la significación

Los criterios **MANOVA** con los que valora las diferencias multivariantes los grupos, son:

1. ***Raíz mayor de Roy***
2. ***El lambda de (también conocido como el estadístico-U)***,
3. ***La traza de Hotelling*** *y el*
4. ***El criterio de Pillai.***

Del ana lisis discriminante en el **Capítulo 6**, estos criterios valoran las diferencias entre "***dimensiones***" de las variables dependientes. **La Raíz mayor de Roy** como su nombre indica, **mide las diferencias solamente sobre la primera raíz canónica (o función discriminante) entre las variables dependientes.** Este criterio proporciona alguna ventaja sobre **potencia y especificidad de los contrastes**, pero los hace menos útiles en ciertas situaciones donde todas las dimensiones deben ser consideradas. El contraste de la **Raíz mayor de Roy** es el más adecuado cuando las **variables dependientes** están **fuertemente interrelacionadas** en una sola dimensión, pero también **es la medida que es más probable que se vea afectada** gravemente por incumplimientos de los supuestos básicos. Las otras **3 medidas** valoran todas las **posibles fuentes de diferencia entre los grupos.** El contraste más comúnmente empleado para la **significación global del MANOVA es la *lambda de Wilks*.** En realidad, **existen p o (k- 1) (cualquiera que sea el menor)** raíces características o funciones discriminantes, donde **p es el número de variables dependientes y k es el número de grupos.** A diferencia del **estadístico mrc,** que está basado en **la primera (mayor) raíz característica, el *lambda de Wilks* considera todas las raíces características.**

Compara **si los grupos son de algún modo diferentes sin estar afectados por el hecho de que los grupos difieran en al menos una combinación lineal de las variables dependientes.** El ***lambda de Wilks*** es mucho **más fácil de calcular que el estadístico mrc: |W| / |W+A|** donde **|W|** es el determinante (un sólo número) de la matriz de dispersión multivariante dentro de los grupos, y **|W+A|** es el determinante de la suma de **W** y **A** donde **A** es la matriz de dispersión multivariante entre los grupos. **Cuanto mayor es la dispersión entre los grupos, más pequeño es el valor del lambda de *Wilks* y mayor la significación.** Aunque la distribución de la ***lambda de Wilks*** es compleja, se tienen buenas aproximaciones para contrastar la significación, transformándolo en un **estadístico *F* [**Rao,1978**].**

¿Qué estadístico se prefiere? Usted tiene estos **2 estadísticos** más un conjunto de posibles medidas que también pueden elegir. Otras medidas ampliamente empleadas son el ***criterio de Pillai y la traza de Hotelling***, ambas similares a la ***lambda de Wilks***, dado que consideran todas las raíces características y pueden ser aproximadas por un **estadístico *F.***

La medida a utilizar será la más invulnerable a las violaciones de los supuestos básicos del **MANOVA**, y la que aún mantenga la mayor potencia. Existe un acuerdo entre los expertos en cuanto a que el ***criterio de Pillai*** *o* ***el lambda de Wilks*** son los que mejor cumplen estas requisitos, aunque la evidencia sugiere que el **criterio de Pillai es más robusto** y debe ser empleado si los tamaños muestrales **disminuyen**, si se dan **tamaños de grupos distintos** o si se **incumple la homogeneidad de las covarianzas**.

Sin embargo, si Usted está convencido de que todos los supuestos se cumplen estrictamente, y que las medidas dependientes son representativas de un sola dimensión de los efectos, **entonces la Raíz mayor de Roy es el estadístico más potente.** La mayoría de los

paquetes estadísticos **incluyen las 3 medidas,** y se puede realizar una comparación entre ellas**.**

8.10.2. Potencia estadística de los contrastes multivariantes

La potencia estadística es la probabilidad de que el contraste estadístico identifique un efecto del tratamiento si este realmente existe. La potencia puede ser definida como uno menos la probabilidad del **error de Tipo II** (**β**). Como tal, la potencia está relacionada con el nivel de significación alfa (**α**), que define el **error de Tipo I** aceptable (Vea **Capítulo 2**).

El nivel de potencia para cualquiera de los cuatro criterios estadísticos, como el: ***Raíz mayor de Roy, el lambda de Wilks, la traza de Hotelling o el criterio de Pillai***, se basa en:

1. **El nivel de significación alfa,**
2. **El efecto tamaño del tratamiento** y el
3. **Tamaño muestral de los grupos**.

La **potencia está inversamente relacionada con el nivel de significación alfa seleccionado.** A medida que alfa se incrementa (es más cauteloso, pasando de **0.05 a 0.01**), la potencia disminuye. Por tanto, si el investigador reduce el **nivel de significación alfa** para **reducir el error de Tipo I**, como en el caso de una posible dependencia entre las observaciones o un ajuste para comparaciones múltiples, la potencia disminuye. Usted siempre debe conocer las implicaciones del ajuste del **nivel de significación alfa**, ya que el objetivo del análisis que no se tiene en cuenta, no es solamente evitar los **errores de Tipo I**, sino también **identificar los efectos del tratamiento** si en verdad existen. Si se impone un **nivel de significación alfa** demasiado riguroso, entonces la potencia puede ser **demasiado baja** para identificar resultados válidos. Usted debe considerar, no sólo el **nivel de significación alfa**, sino también la **potencia resultante**, y que intente mantener un **nivel de significación alfa aceptable** con una potencia cerca de **0.80.**

8.10.3. El incremento de potencia en ANOVA y MANOVA

Con un nivel de significación alfa especificado, ¿cómo incrementar la potencia? :

1. Se sugiere primero partir del **tamaño muestral de los grupos**, por lo que se necesita comprender el impacto de este efecto, que es una **medida estandarizada de las diferencias de los grupos**, generalmente **definida como las diferencias en las medias de los grupos divididas por sus desviaciones estándar**. La magnitud del efecto tamaño tiene un efecto directo sobre la potencia del contraste estadístico. Para cualquier tamaño muestral dado**, la potencia del contraste estadístico será mayor cuanto mayor sea el efecto tamaño**. Contrariamente, si el tratamiento tiene un efecto tamaño esperado más pequeño, se debe conseguir un tamaño muestral mayor para lograr la misma potencia que un tratamiento con un mayor efecto tamaño.
2. Una vez especificado el nivel de **significación alfa** e **identificado el efecto tamaño**, el último elemento que afecta a la potencia es el **tamaño muestral**. En muchos casos, este es el elemento sobre el que Usted tiene menos control. Como se discutió anteriormente, incrementos en el tamaño muestral generalmente reducen el error muestral e incrementan la sensibilidad (potencia) del contraste.
3. **En trabajos con tamaños de grupos menores a 50 miembros**, obtener niveles de potencia adecuados puede ser bastante problemático. El aumento de los **tamaños**

muestrales en cada grupo tiene efectos importantes, hasta que se alcanzan unos tamaños de aproximadamente **150 miembros**, y entonces los incrementos en la potencia disminuyen notablemente. También se debe tener cierta precaución cuando los tamaños muestrales son grandes.

4. En muchos contrastes estadísticos, **tamaños muestrales grandes reducen el componente del error muestral a un nivel tan pequeño que cualquier diferencia se considera estadísticamente significativa**. Cuando los tamaños muestrales llegan a ser grandes y la significación estadística es la indicada, Usted debe examinar la potencia y los efectos tamaños para asegurarse, no solamente significación estadística, sino también significación práctica.

8.10.4. La utilización de potencia en la planificación y el análisis

La estimación de la potencia debe ser empleada tanto para:

1. **Planificar el análisis.** En la etapa de planificación, Usted determina el tamaño muestral necesario para identificar el efecto tamaño estimado. En muchos casos, el efecto tamaño puede ser estimado a partir de una investigación previa a partir de juicios razonados o incluso ponerlo a un nivel mínimo de significación práctica. En cada caso, se puede determinar el tamaño muestral necesario para lograr un determinado nivel de potencia con un nivel de significación específico.
2. **La valoración de los resultados,** de los contrastes después de que el análisis ha sido realizado, se está proporcionando un contexto para interpretar los resultados, especialmente si no se han encontrado diferencias significativas. Usted debe determinar si la **potencia obtenida fue suficiente (0.80 o más)**. Si no fuese así, **¿el análisis podría ser reformulado para lograr más potencia?** Las posibles soluciones, son:

-Realizar alguna forma de tratamiento en bloques o

-Realizar un análisis de la covarianza que hará que el contraste sea más eficiente haciendo énfasis en el efecto tamaño**. Si la potencia fuese la adecuada y no se encontrase significación estadística para el efecto del tratamiento, entonces lo más probable es que el efecto tamaño para el tratamiento fuese demasiado pequeño para considerar significación estadística o práctica.**

8.10.5. El cálculo de los niveles de potencia

Para calcular la potencia en los análisis **ANOVA**, se encuentran disponibles **tanto fuentes publicadas [**Cohen,1977**,** Stevens,1980**]** como programas de computador. Sin embargo, **los métodos de cálculo de la potencia en los MANOVA están mucho más limitados**. Afortunadamente, la mayoría de los programas informáticos proporcionan una valoración de la potencia para los **contrastes de significación** y le permitan a Usted determinar si la potencia debe jugar un papel en la interpretación de los resultados. Con respecto al material publicado con objetivos de planificación, existe poco para el **MANOVA** dado que muchos elementos afectan a la potencia de un análisis **MANOVA**. Una fuente **[**Lauter, 1978**]** con tablas publicadas refleja la potencia de un conjunto de situaciones habituales donde se ha aplicado el **MANOVA**. La **Figura 8.5** proporciona un panorama de los tamaños muestrales necesarios para varios niveles de complejidad del análisis. Revisando la tabla se observa:

1. Un incremento del número de variables dependientes requiere un incremento del **tamaño muestral** para mantener un nivel de potencia dado. El **tamaño muestral** adicional que se necesitaría es más pronunciado para los efectos tamaño más pequeños.
2. Si se espera que los **efectos tamaño sean pequeños**, Usted debe estar preparado para llevar a cabo un esfuerzo de investigación sustancial para lograr niveles aceptables de potencia. Por ejemplo, para lograr una potencia propuesta de **0.80** cuando se valoran los efectos tamaño en un estudio con **4 grupos**, se requieren **115 individuos** si se emplean 2 medidas dependientes. El **tamaño muestral** requerido **aumenta a 185 por grupo** si se consideran **8 variables dependientes**. Los **beneficios de la parsimonia** en el conjunto de variables dependientes aparecen no solamente en la interpretación sino también en los contrastes estadísticos para las diferencias de los grupos. Ver **Figura 8.30.**

Figura 8.30. Requisitos de tamaño muestral para obtener una potencia estadística de 0.80 en el MANOVA

	Número de grupos											
	3				4				5			
	Número de variables dependientes				Número de variables dependientes				Número de variables dependientes			
Efecto Tamaño	2	4	6	8	2	4	6	8	2	4	6	8
Muy Grande	13	16	18	21	14	18	21	23	16	21	24	27
Grande	26	33	38	42	29	37	44	48	34	44	52	58
Medio	44	56	66	72	50	64	74	84	60	76	90	100
Pequeño	98	125	145	160	115	145	165	185	135	170	200	230

Fuente: Liiuter, J. (1978), "Sample Size Requirements for the f2 Test of MANOVA (Tables for One-Way Classification) ", Biometrical Journal 20: 389-406.

8.10.6. Los efectos de la multicolinealidad de la variable dependiente sobre la potencia

Hasta el momento, la potencia desde una perspectiva aplicable tanto al **ANOVA** como **MANOVA**. Sin embargo, en **MANOVA**, Usted también tiene que considerar los efectos de **multicolinealidad** de las **variables dependientes** sobre la **potencia de los contrastes estadísticos**. El investigador, bien en la **etapa de planificación** o bien en la **etapa de análisis**, tiene que considerar **la fuerza y dirección de las correlaciones** además de los

efectos tamaño de las variables dependientes. Si clasificamos las variables por sus tamaños de efecto como **fuertes o débiles**, entonces surgen pautas **[**Cole et al. 1994**]:**

1. Si la pareja de variables correlacionadas está compuesta de variables **fuertes-fuertes o débiles-débiles,** entonces se obtiene la máxima potencia cuando la correlación entre las variables es altamente negativa.
2. Lo anterior, sugiere que el **MANOVA está en las mejores condiciones con la adición de las variables dependientes que tienen correlaciones altamente negativas**. Por ejemplo, en vez de incluir dos medidas de satisfacción redundantes, Usted podría reemplazarlas con las medidas de **correlación de satisfacción y desatisfacción** para incrementar la potencia.
3. Cuando la pareja de variables correlacionada es una **mezcla** (**fuerte-débil**), entonces la **potencia está maximizada** cuando la **correlación es alta, bien positiva o negativa**.
4. **Una excepción** es el resultado de que el uso de los ítems múltiples para incrementar la fiabilidad tiene como resultado una ganancia neta de potencia, incluso si los ítems son redundantes y positivamente correlacionados.

8.11. ANOVA/MANOVA. Paso 5: Interpretación

Paso 5: Interpretación

Una vez que se ha evaluado la **significación estadística de los tratamientos**, Usted puede desear examinar los resultados por medio de:

1. **Interpretar los efectos de las covarianzas** si éstas han sido empleadas,
2. **Valorar qué variable(s) dependiente(s) presenta(n) diferencias entre los grupos,** o
3. **Identificar qué grupos difieren en una sola variable dependiente o en el valor teórico dependiente completo.**

8.11.1. Evaluación de las covarianzas

Habiéndose dado los supuestos para la aplicación de las **covarianzas**, Usted puede **reinterpretar su efecto real sobre el valor teórico dependiente** y su impacto sobre los verdaderos **contrastes estadísticos de los tratamientos**. Dado que el **ANCOVA y MANCOVA** son aplicaciones de los procedimientos de regresión dentro del **análisis del método de la varianza**, la valoración del impacto de las **covarianzas** sobre las variables dependientes es bastante similar a examinar las **ecuaciones de regresión**. Para cada **covarianza** se forma una ecuación de regresión que señala la **validez de la relación predictiva**. Si las **covarianzas** representan teóricamente efectos despreciables, entonces estos resultados proporcionan una base objetiva para aceptar o rechazar las relaciones propuestas. En la práctica, Usted puede examinar el **impacto de las covarianzas y eliminar aquéllas con poco o ningún efecto.**

También se debe examinar el impacto global de la introducción de la(s) covarianza(s) en un contraste estadístico de los tratamientos. **El enfoque más directo es realizar el análisis con y sin las covarianzas,** como sigue:

1. Aquellas que sean efectivas mejorarán la potencia estadística de los contrastes y reducirán la varianza dentro del grupo.

2. Si no se observa ninguna mejora sustancial, entonces pueden ser eliminadas, ya que reducen los grados de libertad disponibles para los contrastes de los efectos del tratamiento.
3. Este enfoque también puede servir para identificar aquellos casos en los que la **covarianza es "*demasiado potente*"** y reduce la varianza de tal manera que los tratamientos no son significativos; hecho que se produce muchas veces cuando se incluye una covarianza que está correlacionada con una de las variables independientes y de esta manera "***elimina***" esta varianza, reduciendo por ello la capacidad explicativa de la variable independiente.

8.11.2. Evaluación del valor teórico dependiente

El siguiente paso es el **análisis del valor teórico dependiente** para evaluar cuál de las variables dependientes contribuye a las diferencias globales señaladas en los **contrastes estadísticos**. Este paso es esencial ya que se puede identificar un conjunto de variables que, o bien acentúa las diferencias mientras otras variables no son significativas, o bien oculta los efectos significativos de las restantes. Los procedimientos descritos en esta sección se denominan **contrastes *post hoc*** (contrastes que se realizan después de examinar el modelo de los datos. Otro enfoque es el empleo de **contrastes a priori**) contrastes que se planean con anterioridad a observar los datos desde un punto de vista de toma de decisión teórico o práctico. Desde un punto de vista pragmático, las situaciones surgen cuando una variable dependiente clave debe ser aislada y contrastada con máxima potencia. Recomendamos que se lleve a cabo un contraste a priori en tales situaciones. El enfoque más común consiste en realizar los contrastes univariantes para las variables seleccionadas. Por ejemplo, en el caso de dos grupos, un **contraste t ordinario** es un contraste a priori para una **variable dependiente** dada. Sin embargo, Usted debe saber que a medida que se incrementa el número de estos contrastes a priori, **se invalida una de las principales ventajas del enfoque multivariante para contrastar la significación** (el control del porcentaje del **error de Tipo I**), a menos que se realicen ajustes específicos que controlen el aumento del **error de Tipo l.**

Un análisis **MANOVA** de dos grupos implica un ajuste del estadístico $\mathbf{T^2}$. Cuando el estadístico $\boldsymbol{T^2 > T^2_{critica}}$ para un nivel alfa (**α**) dado, concluimos que los **vectores de las puntuaciones medias son diferentes**. La función discriminante nos informa de qué combinación lineal de las variables dependientes produce la diferencia entre los grupos más fiable, pero otras comparaciones también pueden ser de interés. Si quisiéramos contrastar las diferencias de los grupos individualmente para cada una de las variables dependientes, calcularíamos un **estadístico *t* estándar** y lo compararíamos con la raíz cuadrada de $\boldsymbol{T^2_{critica}}$, para juzgar su significación. Este procedimiento aseguraría que la probabilidad de cualquier **error de Tipo 1** entre todos los contrastes debería mantenerse en alfa (**α**) (donde alfa (**α**) se especificó en el cálculo de $\boldsymbol{T^2_{critica}}$) **[**Harris, 1975**]**. Podríamos realizar contrastes similares en situaciones con k grupos ajustando el nivel alfa (**α**) con la **desigualdad de *Bonferroni***, la cual establece que **el nivel alfa (α) debe ser ajustado según el número de contrastes que se están realizando.** El nivel **alfa (α) ajustado** empleado en cualquier contraste separado se define como **nivel alfa total dividido por el número de contrastes:**

Alfa ajustado= alfa total / número de contrastes

También se puede emplear un procedimiento conocido como **análisis de reducción [**Koslowsky. y Caspy,1991**,** Stevens, 1972**]** para **valorar individualmente las diferencias entre las variables dependientes**. Este procedimiento comprende el **cálculo de un estadístico univariante F** para una variable dependiente después de eliminar los efectos de otras variables dependientes que la preceden en el análisis.

El procedimiento es de alguna forma similar a la **regresión por etapas**, pero aquí se examina si una determinada **variable dependiente** proporciona **información única (incorrelacionada) sobre las diferencias de los grupos**. Los resultados del análisis de reducción serían exactamente los mismos que los que se obtendrían si se realizase un **análisis de covarianzas** con las otras variables dependientes **precedentes empleadas como covarianzas**. Un supuesto decisivo del **análisis de reducción** es que **Usted conoce el orden en que se deben introducir las variables dependientes, ya que la interpretación puede variar fuertemente según diferentes órdenes para entrar.** Si la ordenación tiene un apoyo teórico, entonces el contraste del análisis de reducción es válido. Las variables identificadas como **no significativas** son "***redundantes***" con las anteriores variables significativas, con lo que añaden más información referente a las diferencias entre los grupos. **Se puede cambiar el orden de las variables dependientes para comprobar si los efectos de las variables son redundantes o únicos, pero el proceso es bastante complicado conforme se incrementa el número de variables**. Otros procedimientos incluyen más **análisis de las funciones discriminantes**, en particular sobre la primera función discriminante, para obtener información adicional sobre qué variables diferencian mejor entre los grupos. Todos estos análisis están pensados para verificar las variables dependientes contribuye a las diferencias en el valor teórico dependiente entre el(los) tratamiento(s).

8.11.3. Identificación de las diferencias entre los distintos grupos

Aunque los contrastes univariantes y multivariantes del **ANOVA** y del **MANOVA** nos permiten probar la **hipótesis nula** de que **las medias de los grupos son todas iguales**, **no nos señalan dónde se establecen esas diferencias significativas**. Los **contrastes *t* múltiples no son apropiados para contrastar la significación de las diferencias entre las medias de pares de grupos dado que la probabilidad del error de Tipo I se incrementa con el número de comparaciones realizadas entre grupos** (igual que el problema al emplear múltiples **ANOVAs** univariantes frente al **MANOVA**). Se encuentran disponibles muchos procedimientos para investigaciones posteriores de **diferencias de las medias de grupos específicos de interés**, que pueden ser clasificados como a **priori o *post hoc***. Estos procedimientos emplean enfoques diferentes para controlar el porcentaje del **error de Tipo I** entre los múltiples contrastes.

8.11.4. Métodos *post hoc*

Entre los métodos ***post hoc*** más habituales consideramos:

1. **Contraste de *Scheffer***,
2. **Método de la diferencia de verdad significativa (DVS) de *Tukey*,**
3. **Extensión del enfoque de la diferencia menos significativa (DMS) de *Tukey*,**
4. **Contraste de rango múltiple de *Duncan*,** y

5. **El contraste de *Newman-Kuels*.**

Cada método identifica qué comparaciones entre los grupos (por ejemplo, grupo 1 frente a grupo 2 y 3) presentan **diferencias significativas**. Proporcionan al investigador los contrastes para cada combinación de grupos, simplificando con ello el proceso interpretativo. Aunque estos métodos simplifican la identificación de las diferencias de los grupos, todos ellos comparten el problema de tener niveles bastante bajos de potencia. Dado que los **contrastes *post hoc*** deben examinar todas las posibles combinaciones, la potencia de cualquier contraste individual es bastante baja. A estos 5 contrastes de significación post hoc o de comparación múltiple se les ha contrastado su potencia.**Las conclusiones son que el contraste de Scheffer es el más prudente con respecto al error de Tipo l.** Los restantes contrastes son clasificados en este orden: **DVS de *Tukey*, DMS de *Tukey, Newman-Kuels y Duncan***. Si los efectos tamaño son grandes o el número de grupos es pequeño, los métodos ***post hoc*** pueden identificar las diferencias de los grupos. Pero Usted debe tener en cuenta las limitaciones de estos métodos y emplear otros métodos **si se pueden identificar comparaciones más específicas.** Se pueden encontrar explicaciones y tratamientos excelentes de estos procedimientos en varios textos **[**Huitema. , 1980**,** Winer, 1962**].**

8.11.5. Contrastes a priori o comparaciones planificadas

Usted también puede realizar comparaciones específicas entre los grupos empleando un contraste a priori o **comparaciones planificadas**. Este método **es similar a los contrastes *post hoc*** descritos anteriormente pero se diferencia en que el investigador establece qué comparaciones de los grupos se van a realizar en las comparaciones planificadas frente a contrastar el conjunto completo, como se hacía en los **contrastes *post hoc***. Las comparaciones planificadas son más potentes porque el número de comparaciones es menor, pero tener más potencia resulta de poca utilidad si el investigador no contrasta específicamente las comparaciones de grupos "***correctas***". Las comparaciones planificadas son más apropiadas cuando los fundamentos conceptuales pueden apoyar las comparaciones específicas que se van a realizar. **Las comparaciones planificadas no deben emplearse de una manera explicativa porque no tienen controles efectivos frente al aumento de los niveles del error de Tipo I total.** Usted establece los grupos que se van a comparar a través de un **contraste**, que es justamente una combinación de las medias de los grupos que representa una comparación planificada específica. Los contrastes pueden ser construidos generalmente como:

$$C=W_1G_1+W_2G_2+...W_kG_k$$

Donde:

***C*= valor del contraste**

***W* = ponderaciones**

***G* = medias de los grupos**

El contraste se realiza asignando ponderaciones positivas y negativas para especificar los grupos que se van a comparar, mientras se asegura que la **suma de las ponderaciones es cero**. Por ejemplo, suponga que tenemos **3 medias de grupos**. Para contrastar la diferencia entre $\mathbf{G_1}$ y $\mathbf{G_2}$, $\mathbf{C= (1)\ G_1 + (-1)\ G_2 + (0)\ G_3}$. Para contrastar si la media de $\mathbf{G_1}$ y $\mathbf{G_2}$ es diferente de la de $\mathbf{G_3}$, el contraste se construye como:

$C= (0.5)\ G_1 + (0.5)\ G_2 + (-1)\ G_3$. Se calcula un ***estadístico F*** separado para cada grupo. De esta manera, Usted puede definir cualquier comparación deseada y contrastarla directamente, pero la probabilidad del **error de Tipo 1** para cada comparación a priori es igual a alfa (**α**). De esta manera, varias comparaciones planificadas incrementarán **el nivel total del error de Tipo l.** Todos los paquetes estadísticos pueden realizar tanto **contrastes a priori como *post hoc*** para variables dependientes únicas. Si Usted desea llevar a cabo comparaciones del valor teórico dependiente completo, están disponibles extensiones de estos métodos. Después de concluir que los vectores de las **medias** de los grupos no son iguales, Usted podría estar interesado en ver **si existen algunas diferencias de los grupos sobre el valor teórico dependiente compuesto.** Se puede calcular **un estadístico *F* ANOVA estándar** y compararlo con :

$F_{critica} = (N - k)\ mrc_{critica} / (k - 1)$, donde el valor de $\mathbf{mrc_{critica}}$ está tomado de la distribución de ***mrc*** con grados de libertad adecuados.

8.12. ANOVA/MANOVA. Paso 6: Validación

Paso 6: validación

ANOVA y MANOVA se han desarrollado según la tradición de la experimentación, donde la replicación es el principal medio de validación. La especificidad de los tratamientos experimentales permite un amplio empleo del mismo experimento en múltiples poblaciones para evaluar la generalidad de los resultados. **Este es el principal dogma del método científico**. Sin embargo, en las ciencias sociales y en la investigación económica, la experimentación verdadera es muchas veces sustituida por contrastes estadísticos en situaciones no experimentales como un estudio con encuesta. La capacidad para validar los resultados en estas situaciones se basa en la replicación de los tratamientos. En muchos casos, se usan como tratamientos características demográficas tales como edad, sexo, renta y los gustos. Estos tratamientos puede parecer que cumplen el requisito de comparabilidad, pero el investigador debe asegurarse que el elemento adicional de la asignación aleatoria a las celdas también se cumple; sin embargo, muchas veces en un estudio con encuesta esto no es cierto. Por ejemplo, el uso de la edad y el género como las variables independientes es un ejemplo común del empleo del **ANOVA o MANOVA** en un estudio con encuesta. Pero en términos de validación, el investigador debe ser precavido al analizar múltiples poblaciones y comparar los resultados como una mera prueba de validez. Dado que los encuestados en un sentido simple "***Se seleccionan ellos mismos***", los tratamientos en este caso no pueden ser asignados por el investigador y con ello la asignación aleatoria no se puede realizar. Por tanto, el investigador debe considerar de manera importante el empleo de covarianzas para controlar otras características que podrían estar relacionadas con los grupos edad/género y que podrían afectar a las variables dependientes, pero que no están incluidas en el análisis. Otra cuestión es la exigencia de causalidad cuando se emplean métodos o técnicas experimentales. Los principios de la causalidad son examinados con más detalle en ecuaciones estructurales. Usted debe recordar que en todos los campos de investigación, incluyendo el experimental, ciertos criterios conceptuales (por ejemplo, **ordenación temporal de los efectos y resultados**) deben ser establecidos antes que la causalidad pueda ser respaldada. La aplicación simple de una técnica en particular en un ámbito experimental no asegura la causalidad.

8.13. ANOVA. Resumen

- Es un método estadístico para determinar si diversos conjuntos de muestras aleatorias de una determinada variable, proceden de una misma población o bien de poblaciones distintas. En general, cada conjunto de muestras se caracteriza por estar afectado por un tratamiento específico, que eventualmente puede influir en los valores que tome la variable que es objeto de estudio.
- Se denomina **factor** a la variable que supuestamente ejerce una influencia sobre la variable estudiada o analizada, a la que se le denomina **variable dependiente.**
- En el Análisis de la Varianza, el factor cuya influencia se requiere corroborar se introduce de forma discreta, independientemente de que sea de naturaleza continua o no. Cuando el factor sea de naturaleza discreta, que será la situación más frecuente, se utilizará de forma equivalente los términos grupo o nivel.
- A menudo deseamos comparar más de 2 condiciones de una variable **independiente**. Cuando este es el caso no es posible usar más las **pruebas *t*** pero sí el uso de **ANOVA**. La ventaja es que nos permite incluir tantas condiciones como lo planeemos en una sola prueba.
- En el **ANOVA** de un factor obtenemos un estadístico basado en el radio de varianzas conocido como ***F***, el cual permite observar la **variabilidad** de las puntuaciones entre las condiciones comparadas a la **variabilidad** en las puntuaciones debidas a factores del **azar** o el **error.**
- Si existiera un efecto de la **variable independiente** sobre la **variable dependiente** esto generará una **variabilidad apreciable** en las puntuaciones de las condiciones y así producirá también, un valor ***F*** apreciable. Note que un valor de ***F*** nos informa que hay un efecto de la variable independiente sobre la variable dependiente **pero no reporta qué condiciones la producen**. Considere un programa de videojuegos ambientado en 3 diferentes escenarios: terrestre, acuático (ambos con tips de ayudas) y aéreo (sin ayuda). Los resultados de los jugadores pueden mostrar un valor significante de ***F*** en una **ANOVA**. Sin embargo, sin más investigación podría ser que las 3 ambientaciones produzcan puntuaciones muy diferentes, o que 2 de ellas produzcan resultados de puntuación entre ellos similares vs la tercera con resultados de puntuación diferentes. El **ANOVA** solamente nos informa que existe una diferencia en algún grupo y que será necesario aplicar por lo tanto, una **prueba de comparación de pares múltiple *post hoc***, para determinar exactamente cuales condiciones son la causa de los efectos.
- Existen 2 formas de **ANOVA de un factor**, justo como existen **2 formas de las pruebas *t***:
  1. El diseño de **medición de un factor independiente** es para las situaciones donde las puntuaciones en cada condición provienen de deferentes participantes
  2. El diseño de **mediciones repetidas** es para situaciones donde las puntuaciones de cada condición provienen de los mismos participantes.

8.14. ANOVA de un factor independiente. Ejemplos

Paso 1: Objetivos

-Problema 1: La empresa **MKT Digital** está interesada en detectar indicios en el pre lanzamiento de su videojuego en 3 versiones de ambientación (terrestre, acuática con tips

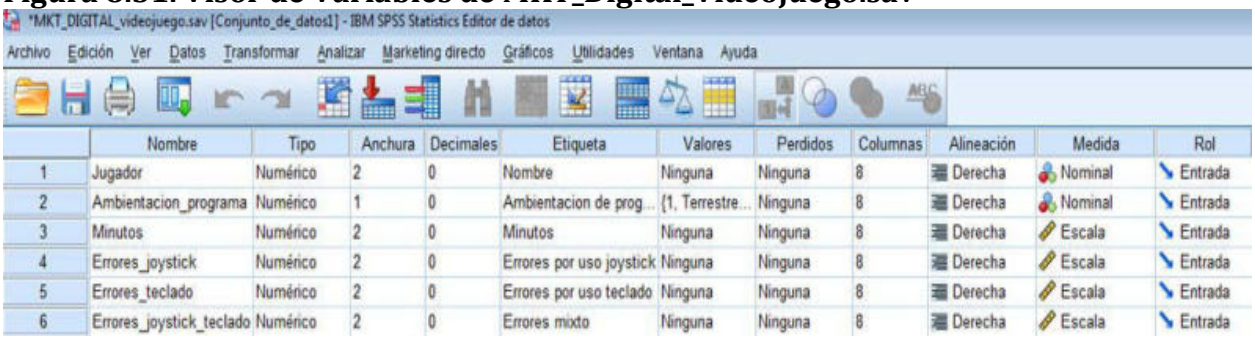
de ayuda y aérea sin tips de ayuda), con los tiempos en los cuales 30 sujetos ganaron en el juego para subir de nivel. La base de datos es **MKT_Digital_Videojuego.sav.**

H_0= El tiempo para ganar la competencia de los jugadores, en los ambientes terrestre y acuático que reciben tips de ayuda, son iguales que el aéreo que no recibe tips de ayuda.

H_1= El tiempo para ganar la competencia de los jugadores, en los ambientes terrestre y acuático que reciben tips de ayuda, son diferentes que el aéreo que no recibe tips de ayuda.

Ver Figuras 8.31 y 8.32.

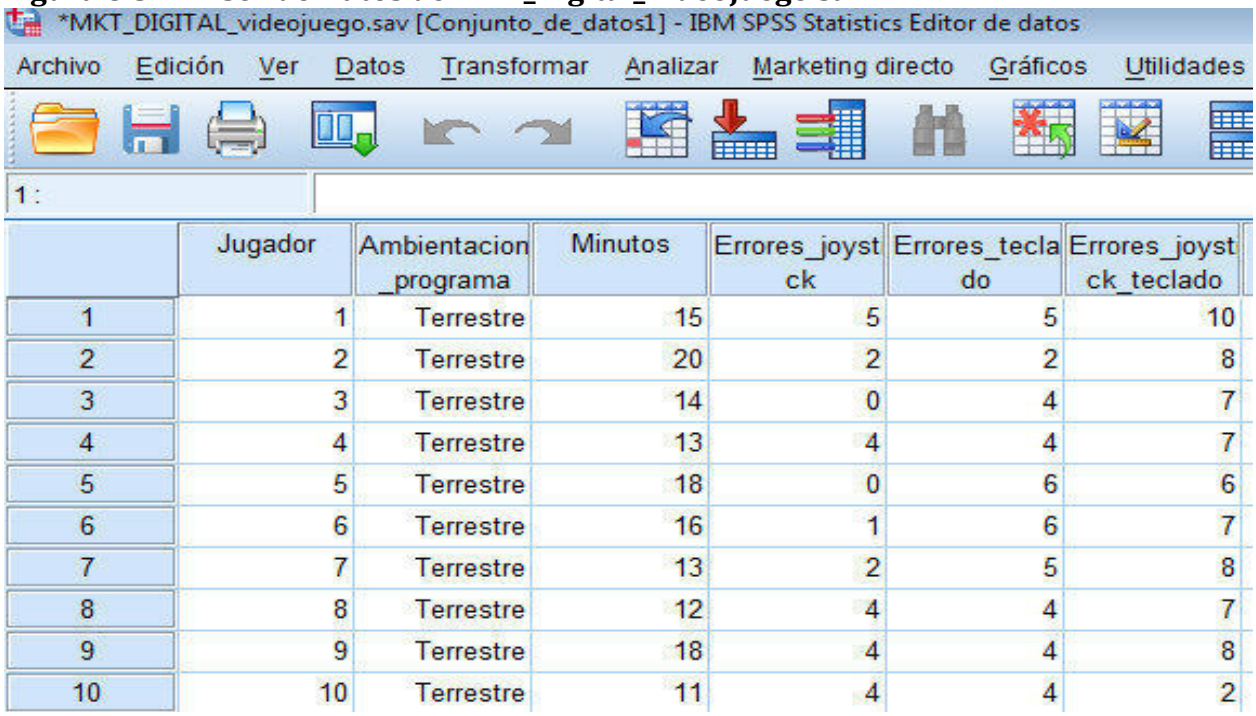
Figura 8.31. Visor de Variables de MKT_Digital_Videojuego.sav



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	Jugador	Numérico	2	0	Nombre	Ninguna	Ninguna	8	Derecha	Nominal	Entrada
2	Ambientacion_programa	Numérico	1	0	Ambientacion de prog...	{1, Terrestre...	Ninguna	8	Derecha	Nominal	Entrada
3	Minutos	Numérico	2	0	Minutos	Ninguna	Ninguna	8	Derecha	Escala	Entrada
4	Errores_joystick	Numérico	2	0	Errores por uso joystick	Ninguna	Ninguna	8	Derecha	Escala	Entrada
5	Errores_teclado	Numérico	2	0	Errores por uso teclado	Ninguna	Ninguna	8	Derecha	Escala	Entrada
6	Errores_joystick_teclado	Numérico	2	0	Errores mixto	Ninguna	Ninguna	8	Derecha	Escala	Entrada

Fuente: SPSS 20 IBM

Figura 8.32. Visor de Datos de MKT_Digital_Videojuego.sav



	Jugador	Ambientacion_programa	Minutos	Errores_joystick	Errores_teclado	Errores_joystick_teclado
1	1	Terrestre	15	5	5	10
2	2	Terrestre	20	2	2	8
3	3	Terrestre	14	0	4	7
4	4	Terrestre	13	4	4	7
5	5	Terrestre	18	0	6	6
6	6	Terrestre	16	1	6	7
7	7	Terrestre	13	2	5	8
8	8	Terrestre	12	4	4	7
9	9	Terrestre	18	4	4	8
10	10	Terrestre	11	4	4	2

Fuente: SPSS 20 IBM

Paso 2: Diseño

- En adelante para efectos de aprendizaje, sólo se tomarán de forma directa los datos con el fin de aprender a utilizar la técnica.

Paso 3: Condiciones de Aplicabilidad

- No olvidar realizar los análisis previos de inspección visual de la base de datos para detectar ausentes y/o atípicos (corregir en su defecto), confiabilidad, normalidad, homocedasticidad y linealidad
- Existen **2 procedimientos** para realizar la medición de **ANOVA** de un factor, las cuales producen diferentes productos, con los mismos resultados, pero con diferentes beneficios en cada uno.
- El **método 1 (vía: Anova de un factor)** es más rápido para generar tablas que son más fáciles de interpretar y le permite realizar **contrastes *post hoc*** muy fácilmente. Sin embargo, este procedimiento sólo puede llevarse a cabo con mediciones independientes de **ANOVA.** Muy solicitado por los principiantes en **SPSS**
- El **método 2 (modelo lineal general vía: Univariado)** es el método general de **contrastes** para todos los tipos de **ANOVA** y es el método más preferido por los investigadores quienes aplican diferentes **ANOVAS** con el **SPSS.**

Comúnmente, no hay problema de llevar a cabo tanto las pruebas de **comparación por pares múltiples vía *post hoc*** como la de **contrastes planeados**. Sin embargo, se explicarán **ambas técnicas en ambos métodos** para su mayor comprensión a fin de que Usted decida cuál es el más apropiado para su investigación. Una regla general, sería:

1. Si **NO** está seguro del sentido de su investigación, emplee el **de comparación por pares múltiple vía *post hoc.***
2. Si Usted está probando una hipótesis, emplee el de especificar los **contrastes planeados,** como la más apropiada.

En un caso práctico, No es necesario más que hacer una de las pruebas.

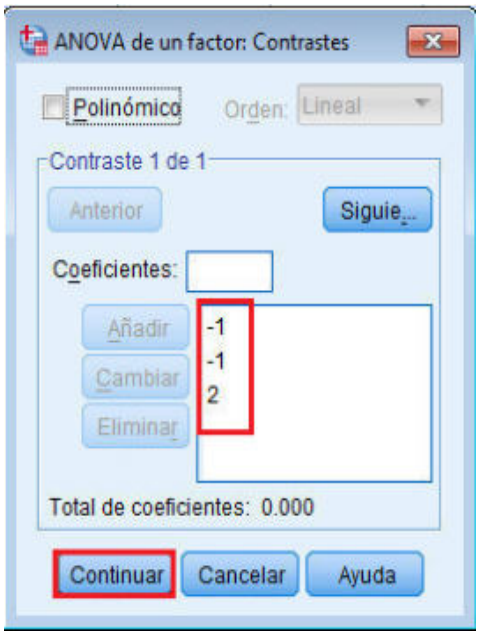
Paso 4: Estimación y ajuste

8.14.1. Contrastes planeados

Si Usted desea hacer uso de pruebas de **contrastes planeados con el método 1 (vía Anova de un factor)**, deberá realizar lo que se sugiere:

1. Es necesario **ponderar los grupos usando coeficientes que totalicen cero (0).** Los contrastes planeados le indican al **SPSS cuáles grupos combinar y cuáles grupos comparar.**
2. Usted elije los valores de los coeficientes a ponderar los grupos de acuerdo a las siguientes reglas:
   a. Coeficientes para grupos de un lado de la comparación tienen coeficientes positivos (**+**) y coeficientes en el otro lado tienen coeficientes negativos (**-**)
   b. Es necesario ponderar los grupos usando coeficientes que **totalicen cero.**
   c. En ambos lados de los coeficientes son ponderados igualmente, así que si 2 grupos son marcados con **+1** en un lado de la comparación, el otro lado debe tener una ponderación de **−2. Ver Figura 8.33.**

Figura 8.33. Cuadro de diálogo para contrastes planeados



Fuente: SPSS 20 IBM

- Al teclear en el botón de **Contrastes** se comparan los efectos de varios grupos de puntuaciones
- Nuestra comparación seleccionada es para contrastar los efectos combinados de los **grupos 1 y 2 vs. 3** (**versión de ambientación terrestre v acuático (sí recibieron ayuda, -1-1) VS ambientación aéreo** (**no recibieron, +2**).
- Para ésta combinación se necesita asignar coeficiente de **−1** a cada uno de los 2 primeros grupos y combina sus efectos vs. el tercer al cual se le asigna el coeficiente de **+2**
- Teclear: **Continuar.** Ver **Figura 8.33.**

Si deseara realizar comparaciones por pares entonces debe seleccionar el comando ***post hoc*.**
La opción del botón: **Contrastes** también da la opción para llevar a cabo un **análisis de tendencias**. Al llevar a término esta prueba, le dará más información en cuanto al modelo subyacente que mejor ajusta los datos. Esto se logra al seleccionar el botón **Polinómico** el tipo de tendencia requerida, siendo los más comunes:

- **El modelo de análisis lineal** que ajusta la tendencia a una **relación lineal**. Esto puede ser llevado a término en **2 o más condiciones**
- **El modelo de análisis cuadrático**, el cual analiza el ajuste de la tendencia a una **línea curva.** Esto puede realizarse **en 3 o más condiciones**
- **El modelo de análisis cúbico** que analiza el ajuste de la tendencia a una **línea ondulada.** Esto puede llevarse a cabo en **4 o más condiciones**

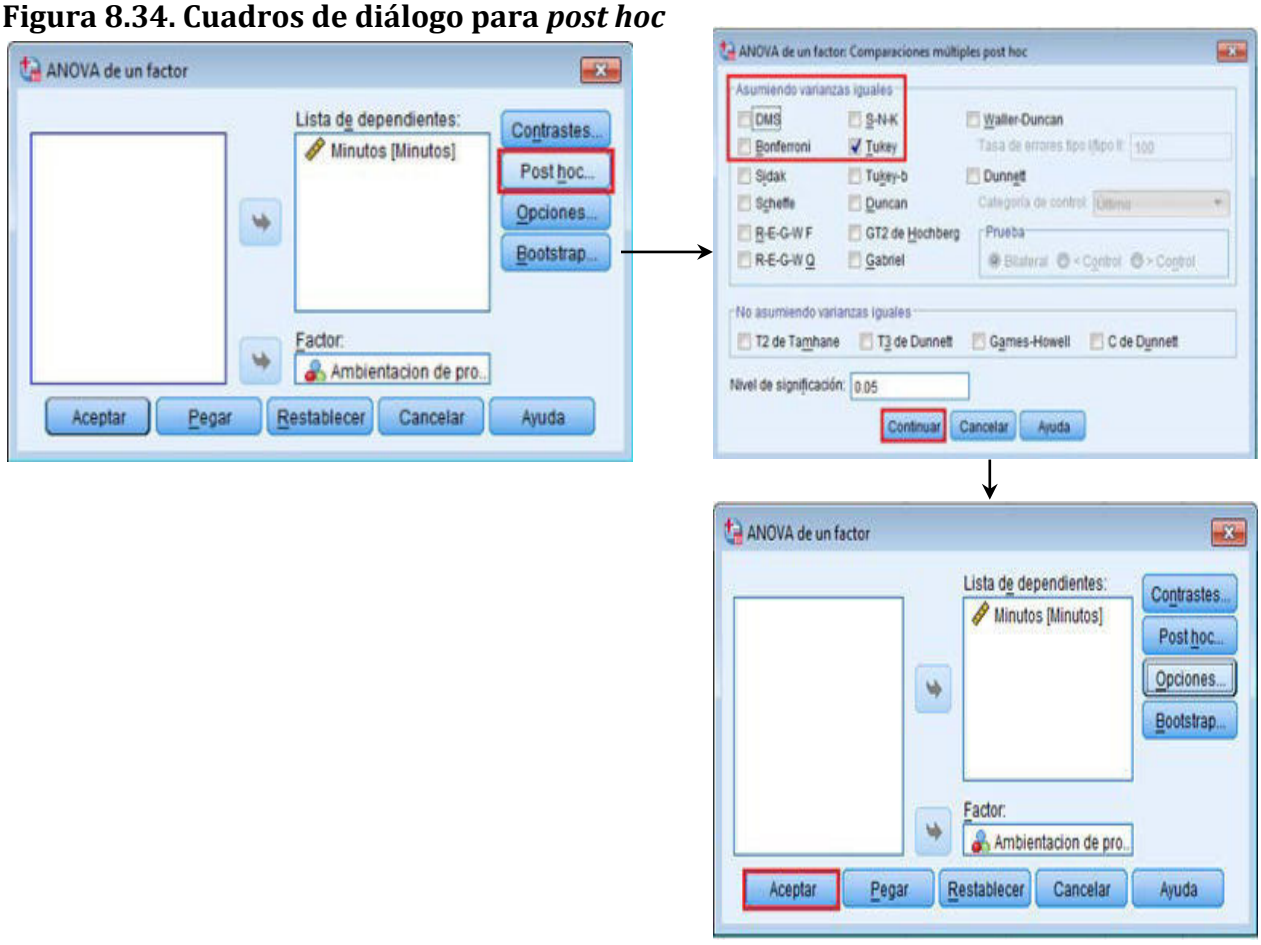
- El **SPSS** puede calcular las tendencias relevantes de sus datos pero deberá asegurarse de que tengan las suficientes condiciones, para que sean apropiadas a realizar conclusiones claras y sensibles.

8.14.2. Pruebas *post hoc*

Si su análisis se puede realizar mejor por las pruebas de comparaciones múltiples, entonces la opción ***post hoc*** es la forma más eficiente.

• Teclear-> botón *post hoc* y marcar en la prueba post hoc la prueba *post hoc* por comparación múltiple de: Tukey->Continuar->Aceptar. Ver. Ver Figura 8.34.

Figura 8.34. Cuadros de diálogo para *post hoc*



Fuente: SPSS 20 IBM

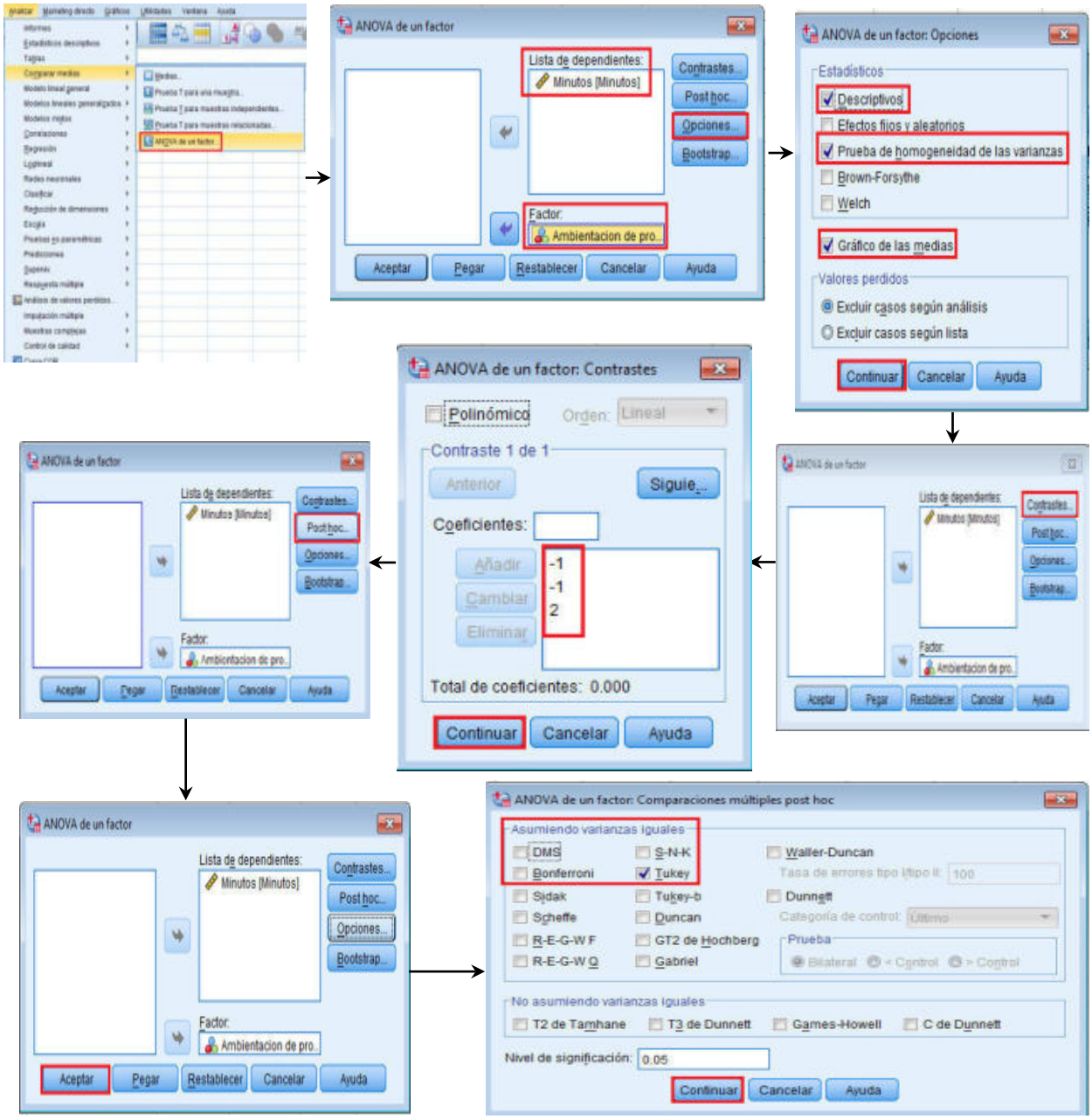
Existen ventajas y desventajas de las diferentes pruebas ***post hoc.*** Se recomienda el **la prueba de *Tukey*** ya que **controla el nivel global de error de tipo 1 y es razonablemente poderoso.**

Método 1 (vía: ANOVA de un Factor)

Paso 4: Ejecución y ajuste

Teclear: Analizar->Comparar medias->ANOVA de un factor-> en Lista de dependientes: variable métrica (minutos); Factor: variable (no métrica): ambientación de programa->Opciones->Estadísticos: Descriptivos; Prueba de homogeneidad de varianzas->Gráfico de medias->Continuar.->Contrastes->Añadir:-1-1+2->Continuar-> *post hoc* -> Asumiendo varianzas iguales: Tukey->Continuar->Aceptar. Aceptar. Ver Figura 8.35.

Figura 8.35. Proceso ANOVA de un factor



Fuente: SPSS 20 IBM

Paso 5: Interpretación

8.14.3. Resultados generales

La primera tabla que se produce es la de Estadísticos Descriptivos. Ver la **Figura 8.36**

Figura 8.36. Resultados descriptivos de ANOVA de un factor

ANOVA de un factor

[Conjunto_de_datos1] C:\Users\Juan\Desktop\MKT_DIGITAL_videojuego.sav

Variable Independiente

Variable dependiente

Descriptivos

Minutos

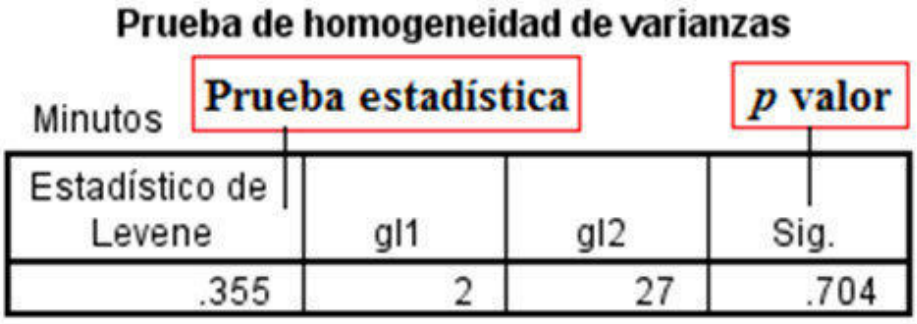
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Terrestre	10	15.00	2.944	.931	12.89	17.11	11	20
Acuatico	10	24.00	3.464	1.095	21.52	26.48	18	29
Aereo	10	28.00	4.397	1.390	24.85	31.15	20	36
Total	30	22.33	6.557	1.197	19.89	24.78	11	36

Fuente: SPSS 20 IBM

- La primera columna de la **tabla Descriptivos** detalla el número de participantes (N) en cada grupo
- La tabla despliega el tiempo promedio en completar los jugadores las distintas ambientaciones de programa (terrestre, acuático, aéreo). Puede verse que los participantes tuvieron más problemas con la interface de juego aérea (tiempo promedio = 28 minutos) ya que esta no presenta ayudas para el jugador. Las otras interfaces, por su diseño de prestar ayuda al jugador de ambientación terrestre y acuática. Cuando los jugadores reciben ayuda para terminar exitosos los promedios se mejoran (acuática=24 minutos), aunque no tan rápido como cuando se aplica a la versión terrestre (15 minutos). Estas diferencias soportarán nuestra hipótesis a fin de afirmar si los resultados son significantes o debido a las oportunidades a la tabla de **ANOVA**
- La **desviación estándar** indica la dispersión de puntuaciones en las 3 condiciones de programa. La más grande se obtuvo en la versión aéreo (**4.3970 minutos**).
- La tabla también despliega la media total y las desviaciones estándar para las 3 condiciones.
- El **error estándar** (**error típico**) es un estimado de la desviación estándar de la distribución de la muestra de la media.

- **El 95% del intervalo de confianza** para la media indica de que estamos seguros al 95% de confianza que la verdadera (población) media estará entre los límites superior e inferior. El muestreo de las medias cae entre estos 2 valores.
- SPSS produce la tabla **Prueba de homogeneidad de varianzas**, la cual nos dice si ya encontramos nuestro segundo supuesto (los grupos tienen aproximadamente la igual varianza en la variable dependiente). Ver **Figura 8.37**

Figura 8.37. Prueba de homogeneidad de varianzas



Prueba de homogeneidad de varianzas

Minutos

Estadístico de Levene	gl1	gl2	Sig.
.355	2	27	.704

Prueba estadística

p valor

Fuente: SPSS 20 IBM

Si el **resultado de la prueba de *Levene*** es:

- **No significante** (***p* > 0.05**), entonces las **varianzas son aproximadamente iguales.**

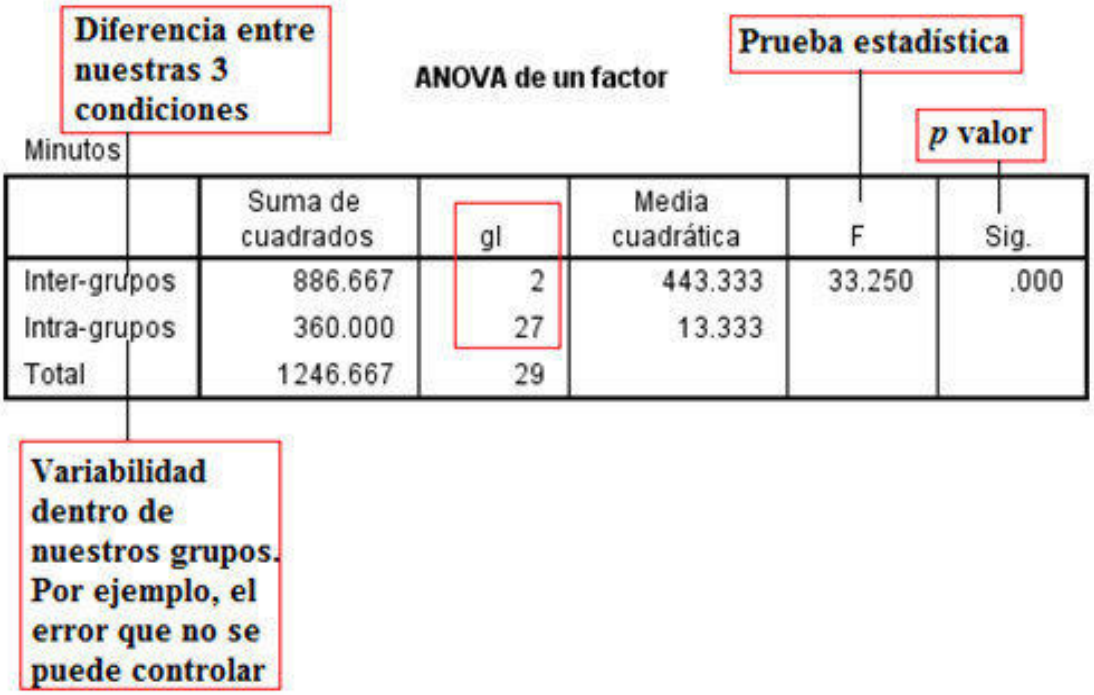
Aquí, el valor **Sig.** es de **0.704**, que es **>0.5**, por lo que se puede afirmar que las varianzas son aproximadamente iguales.

- **Si significante** (***p* < 0.05**) entonces las varianzas son significativamente diferentes.

Si este es el caso entonces necesitará considerar la transformación para hacer sus varianzas más homogéneas.

- La tabla **ANOVA de un factor** se encuentra desplegada a continuación. Esta tabla contiene información clave acerca de nuestro **estadístico *F* calculado. Ver Figura 8.38.**

Figura 8.38. ANOVA de un factor



Fuente: SPSS 20 IBM

- Los **grados de libertad** (**gl**) se reportan. En **ANOVA se encuentran 2 valores,** uno para el **factor** (Inter-grupos) y otro para el **error** (**Intra-grupos**), así que **gl = (2, 27).**
- Si el **SPSS** establece que la probabilidad (**Sig.**) = **0.000**, significa que el **SPSS** lo ha redondeado o abajo de la cantidad o al número más cercano **a 3 lugares decimales**. Sin embargo es deseable siempre que lo realice desde el último 0 a 1, así que ***p* < 0.001.**
- La forma convencional de reportar los hallazgos es la de establecer la prueba estadística **(F)**, los grados de libertad (**gl**), y la probabilidad (**Sig.**), como sigue:

F(2,27) = 33.250; p < 0.001

- Como ***p* < 0.001**, esto indica que hay una alta diferencia significativa entre los **3 grupos. Sin embargo,** no establece en donde se encuentra.
- La **Suma de los cuadrados** nos reporta una medida de la variabilidad en las puntuaciones debido a una fuente particular.
- La **Media cuadrática es la varianza** (suma de los cuadrados divididos por los grados de libertad). Observe que hay una gran cantidad de variabilidad debido a nuestro factor y mucho menos debido al error

8.14.4. Resultados generales. Contrastes planeados

Dados **los contrastes planeados** a través de la opción **Contrastes**, se generan diversas tablas como las que indican sobre las que se han programado y llevado a cabo. **Ver Figura 8.39.**

Figura 8.39. Coeficiente de los contrastes

Coeficientes de los contrastes

Contraste	Ambientacion de programa		
	Terrestre	Acuatico	Aereo
1	-1	-1	2

Fuente: SPSS 20 IBM

El contraste seleccionado en el ejemplo, fue el comparar el grupo de **programación aéreo que no recibió ayuda** con los otros 2 grupos Nuestro análisis se aplicó a la homogeneidad de las varianzas y, como se discutió previamente, **no encontramos diferencias significativas en ellas**, por lo que asumimos **que son iguales.** En nuestro ejemplo, tomamos la parte más alta de información del renglón de contraste. Si estuviéramos menos seguros sobre nuestras varianzas, y las encontraremos significativamente diferentes en los 3 grupos, entonces seleccionaríamos los resultados más conservadores, como los mostrados en el **segundo renglón**.
Ver **Figura 8.40**

Figura 8.40. Prueba para los contrastes

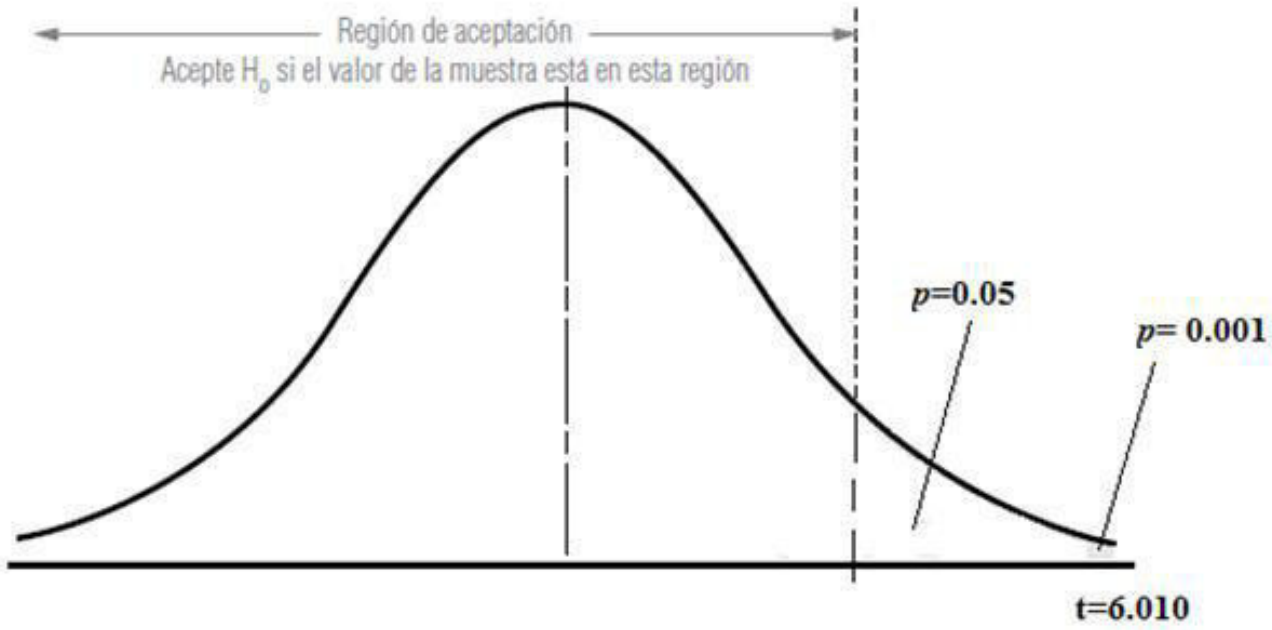
Pruebas para los contrastes

		Contraste	Valor del contraste	Error típico	t	gl	Sig. (bilateral)
Minutos	Asumiendo igualdad de varianzas	1	17.00	2.828	6.010	27	.000
	No asumiendo igualdad de varianzas	1	17.00	3.130	5.430	13.942	.000

Fuente: SPSS 20 IBM

- La tabla de la **Figura 8.40 Pruebas para los contrastes** debe ser leída de manera similar a la tabla de **Prueba *t*. Ver Figura 8.41.**

Figura 8.41. Prueba *t* de una cola

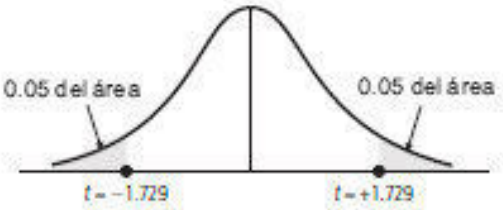


Fuente: propia

- El Valor del contraste de **17.00** al ser analizado por las **pruebas *t,*** con un valor *t*= 6.010 **en 27 grados de libertad** (**gl**), se encuentra significativa en ***p* < 0.001**.
  Ver Figura 8.42.

Figura 8.42. Tabla Prueba *t* de una cola

Áreas en los dos extremos combinados para la distribución *t* de Student.*



EJEMPLO: Para encontrar el valor de t que corresponde a un área de 0.10 en los dos extremos combinados de la distribución, cuando existen 19 grados de libertad, busque en la columna del 0.10 hacia abajo hasta el renglón correspondiente a 19 grados de libertad, el valor *t* apropiado es 1.729

Grados de libertad	Área en los dos extremos combinados			
	0.10	0.05	0.02	0.01
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
16	1.746	2.120	2.583	2.291
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.725	2.086	2.528	2.845
21	1.721	2.080	2.518	2.831
22	1.717	2.074	2.508	2.819
23	1.714	2.069	2.500	2.807
24	1.711	2.064	2.492	2.797
25	1.708	2.060	2.485	2.787
26	1.706	2.056	2.479	2.779
27	1.703	2.052	2.473	2.771
28	1.701	2.048	2.467	2.763
29	1.699	2.045	2.462	2.756

Fuente: Levin y Rubin (2004)

- **Conclusión: Se rechaza la H_0 y se acepta H_1 donde:**

H_1 = El tiempo para ganar la competencia de los jugadores, en los ambientes terrestre y acuático que reciben tips de ayuda, **son diferentes** que el aéreo que no recibe tips de ayuda.

8.14.5. Resultados generales. Comparaciones por pares múltiples *post-hoc*

Dada las comparaciones por pares múltiples entre los grupos a través del uso de la prueba ***post hoc*** de ***Tukey post hoc*** test, se tienen los resultados mostrados en la **Figura 8.43.**

Figura 8.43. Resultados comparaciones por pares múltiples con prueba *post hoc*

Pruebas post hoc

Comparaciones múltiples

Variable dependiente: Minutos

HSD de Tukey

(I) Ambientacion de programa	(J) Ambientacion de programa	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
Terrestre	Acuatico	-9.000*	1.633	.000	-13.05	-4.95
	Aereo	-13.000*	1.633	.000	-17.05	-8.95
Acuatico	Terrestre	9.000*	1.633	.000	4.95	13.05
	Aereo	-4.000	1.633	.053	-8.05	.05
Aereo	Terrestre	13.000*	1.633	.000	8.95	17.05
	Acuatico	4.000	1.633	.053	-.05	8.05

*. La diferencia de medias es significativa al nivel 0.05.

Fuente: SPSS 20 IBM

- La **Figura 8.43** muestra todas las comparaciones por parejas para nuestros grupos de participantes
- En cada comparación, existe un grupo con el identificador **'I'** y el segundo con **'J'**. Esto se hace evidente en la columna de las **Diferencia de medias (I-J),** el cual reporta la cifra resultante de la media de un grupo **(J)** que ha sido sustraído de la media de otro grupo (**I**).
- En nuestro ejemplo, la media del grupo 1 (**ambientación terrestre**) se muestra en **15.00 minutos (ver Tabla 8.36)** de los cálculos de estadística descriptiva, y la media del segundo grupo (ambientación acuática) **24.00 minutos. Así : 15.0000 (I) – 24.0000 (J) = –9.000**
- La columna **Sig**. nos permite evaluar si las diferencias de las medias entre los grupos son significativas.
- Podemos observar de nuestro ejemplo que la diferencia entre los grupos de la **ambientación terrestre y la ambientación acuática es significativa, como lo es la diferencia entre los grupos de ambientación terrestre y la ambientación aérea** ya que los **valores *p* son < 0.05.**
- Por otro lado, **no se encuentran diferencias significativas entre los grupos de ambientación acuática y ambientación aérea ya que p > 0.05**. Sin embargo, esto sólo es justo fuera del ámbito de reclamar una diferencia significativa, por lo **que el examen de los intervalos de confianza** puede dar más información acerca de la fuerza de esta diferencia.
- El **Intervalo de confianza al 95%** nos permite realizar un método diferente de evaluación de las diferencias de nuestros grupos. De observar los niveles de

significancia concluimos que no había una diferencia significativa entre nuestros grupos **'ambientación aérea' and 'ambientación acuática'** (**p > 0.05**). Sin embargo, el intervalo de confianza calculado sugiere que puede haber una diferencia.

- **Los límites superiores e inferiores de los intervalos de confianza, son: -0.05 a 8.05 (Figura 8.18),** es decir, un amplio rango, **sin embargo incluye el cero justo en un extremo del rango,** a pesar de que la **Diferencia de medias (I-J)** sea **4.**
- Los intervalos de confianza son por lo tanto una adecuada manera de complementar los niveles de significancia, particularmente si las cifras se encuentran al filo de la misma.
- SPSS también calcula lo denominado Subconjuntos homogéneos, como la **Figura 8.44** (**Minutos**) que combina las comparaciones por pares múltiples que no fueron encontradas a ser significativamente diferentes unas de otras.

Figura 8.44. Subconjuntos homogéneos

Subconjuntos homogéneos

Minutos

HSD de Tukey[a]

Ambientacion de programa	N	Subconjunto para alfa = 0.05	
		1	2
Terrestre	10	15.00	
Acuatico	10		24.00
Aereo	10		28.00
Sig.		1.000	.053

Se muestran las medias para los grupos en los subconjuntos homogéneos.

a. Usa el tamaño muestral de la media armónica = 10.000.

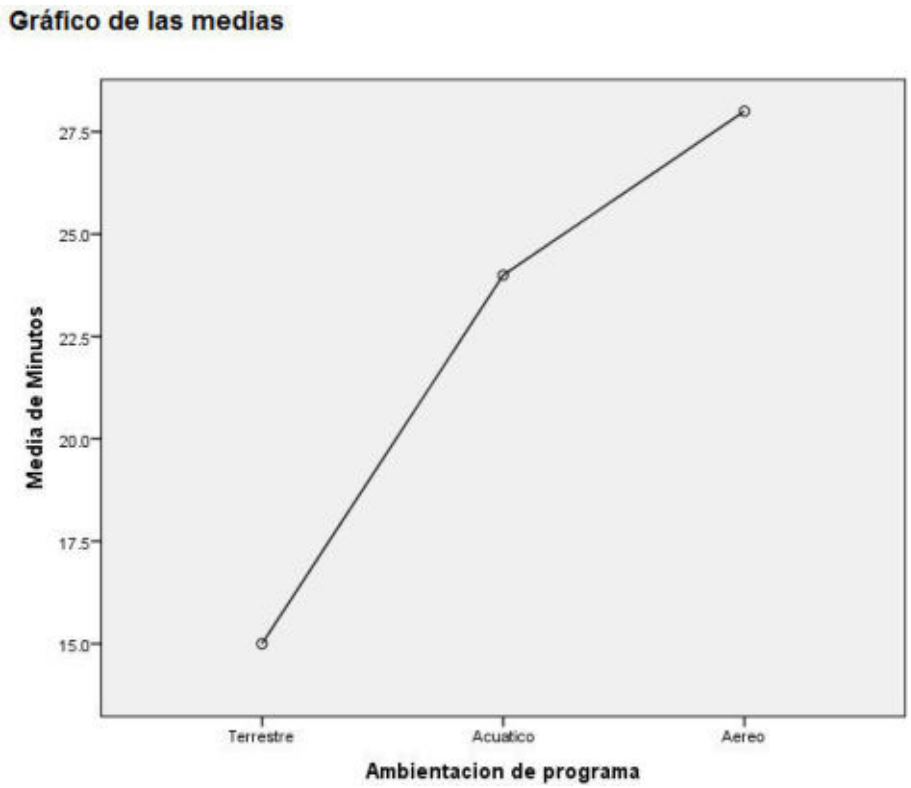
Fuente: SPSS 20 IBM

- Como se observó en la **Figura 8.43 de Comparaciones múltiples**, al grupo que se le asignó la **ambientación terrestre lo llevó a cabo significativamente diferente** de los otros 2 grupos restantes (ambientación acuática y ambientación aérea). **Sin embargo, estos 2 últimos grupos no lo llevaron a cabo significativamente diferente el uno del otro.**
- **SPSS**, por lo tanto ha creado 2 subconjuntos de los datos. Debido a que el primer grupo de ambientación terrestre fue determinado como haberlo llevado a cabo de forma diferente a los otros 2 grupos , aparece en un subconjunto por sí mismo
- Los **2 grupos restantes** fueron determinados como diferentes desde el grupo de ambientación terrestre, pero **NO** diferentes entre ellos mismos, y por lo tanto aparecen en el mismo subconjunto.

- Si todos los **3 grupos fueran determinados a ser significativamente diferentes entre ellos, 3 subconjuntos separados deberían haber sido creados,** uno para cada grupo en el estudio.

Al seleccionar la opción de **SPSS** para producir un gráfico de las medias de nuestros grupos de participantes, se crea (Ver **Figura 8.45**)

Figura 8.45. Gráfico de las medias



Fuente: SPSS 20 IBM

- Se observa del gráfico de las medias, los patrones previamente discutidos.
- Aquellos participantes en los grupos que fueron asignados como ambientación terrestre con tips de ayuda, ganaron los juegos en el tiempo más corto
- El grupo que lo llevó a cabo en el peor de los tiempos fue el que no recibió tips de ayuda como el de ambientación aérea.
- **Conclusión: Se rechaza la H_0 y se acepta H_1 donde:**

H_1 = El tiempo para ganar la competencia de los jugadores, en los ambientes terrestre y acuático que reciben tips de ayuda, **son diferentes** que el aéreo que no recibe tips de ayuda….siendo el aéreo el que más tiempo utilizó y que sus diferencias con el acuático no son tan marcadas como con el terrestre

Método 2 (modelo lineal general vía: Univariante)

Paso 4: Estimación y ajuste

8.14.6. Las pruebas de contrastes planeados y comparación múltiple de pares *post hoc*

Como el **método 1 ANOVA** de generación de factores independientes en el **SPSS**, ambas versiones de las pruebas **planeado y *post hoc*** pueden llevarse a cabo usando el **modelo lineal general.** Nuevamente, solamente es necesario realizar una de las pruebas, ya sea los **contrastes planeados** o el ***post hoc*** a sus datos, dependiendo del diseño de las preguntas de investigación de su estudio. A continuación se explicarán cada uno de los métodos.

8.14.7. Contrastes planeados

Calculando **ANOVA independiente** a través del método descrito anteriormente, le permite al usuario especificar la combinación del contraste que mejor ajuste al diseño de su estudio. Sin embargo, cuando se calcula **ANOVA** a través del **modelo lineal general,** contrastes pre-creados son reportados a Usted dentro de la opción **Contrastes** de la sección **Univariada.**

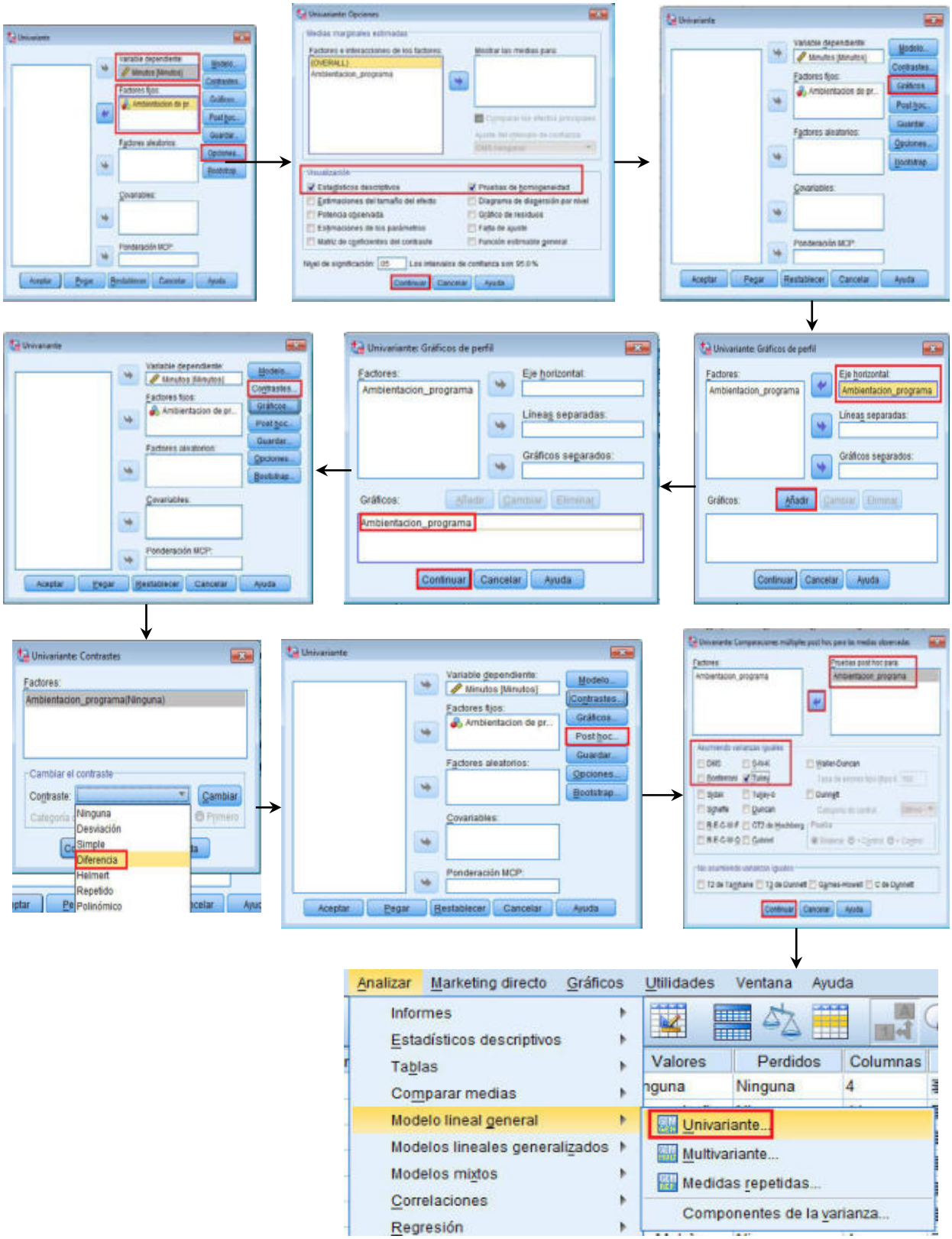
8.14.8. Pruebas de comparación por pares múltiple

Si su análisis es mejor realizad a través del uso de comparación múltiple de pares, las pruebas de ***post hoc*** deben ser la opción. Seleccione la prueba ***post hoc*** y coloque la variable independiente Señale en loa caja de opciones el método apropiado ***post hoc***, por ejemplo: ***Tukey*.**

Oprima Continuar y entonces **OK** para completar el proceso de **ANOVA.** Existen ventajas y desventajas de cada una de las pruebas ***post hoc***; se recomienda usar ***Tukey*** ya que tiene un mejor control de la **tasa de errores de tipo** y es razonablemente preciso**.**

Recuerde que toda la estadística inferencial se encuentra bajo el comando de **SPSS Analizar Teclear: Modo linear general->Univariante-> Variable dependiente métrica : Minutos; Factores fijos (variable no métrica): Ambientación_programa->Opciones->Visualización, selecciones: Estadísticos descriptivos; Pruebas de homogeneidad->Continuar->Gráficos->Factores (variable no métrica); pulsar a Eje horizontal; Añadir->Continuar->Contrastes->Cambiar el contraste: Diferencia->Continuar->*post hoc*->Factor (variable no métrica): Ambientación_programa; Asumiendo variables iguales: seleccionar Tukey->Continuar->Aceptar.** Ver **Figura 8.46.**

Figura 8.46. Método Univariante



Fuente: SPSS 20 IBM

-Resultados. La primera tabla que **SPSS** produce es la **Factores inter-sujetos**. Esto lista el número de participantes en cada grupo, y confirma cuantos grupos han sido usados en el cálculo. **Ver Figura 8.47**

Figura 8.47. Tabla Factores Inter-sujetos

Análisis de varianza univariante

[Conjunto de datos2] C:\Users\Juan\Desktop\MKT_

Variable Independiente

Factores inter-sujetos

		Etiqueta del valor	N
Ambientacion de programa	1	Terrestre	10
	2	Acuatico	10
	3	Aereo	10

Fuente: SPSS 20 IBM

- La siguiente tabla que el **SPSS** produce es la de **Estadísticos descriptivos**. **Ver Figura 8.48.**

Figura 8.48. Tabla Estadísticos descriptivos

Variable Independiente

Estadísticos descriptivos

Variable dependiente: Minutos

Ambientacion de programa	Media	Desviación típica	N
Terrestre	15.00	2.944	10
Acuatico	24.00	3.464	10
Aereo	28.00	4.397	10
Total	22.33	6.557	30

Fuente: SPSS 20 IBM

- La tabla de **Estadísticos descriptivos** las **Media** veces tomadas completar los videojuegos en sus 3 ambientaciones. Se puede observar que cuando los participantes

no recibieron ayuda (ambientación aérea), les tomó más tiempo de juego para ganar. En la ambientación acuática, los participantes ganaron el juego más rápidamente que cuando era de ambientación aérea, pero no tanto como los de ambientación terrestre. Estas diferencias soportan nuestras hipótesis, y para asegurar que si son resultados sistemáticos o productos del azar, se debe examinar la tabla de **Pruebas de los efectos intersujetos**

- **La Desviación típica** (deviación estándar) muestra que la dispersión de las puntuaciones en la condición donde no se dio ayuda (ambientación aérea) fue más grande (**4.3970**), que con las puntuaciones de las condiciones de ambientación terrestre (**2.9439**).
- La tabla también despliega la **Media Total** y las **Desviaciones típicas Totales (desviación estándar)** de las 3 condiciones de ambientación de programa, siendo **N**, el número de participantes.
- La tabla que sigue despliega es el **Contraste de *Levene* sobre la igualdad de las varianzas de error** que nos dice si hemos encontrado el supuesto de homogeneidad de varianza (los grupos tienen aproximadamente igual varianza sobre la variable dependiente).**Ver Figura 8.49.**

Figura 8.49. Contraste de *Levene*

Contraste de Levene sobre la igualdad de las varianzas error[a]

Prueba estadística

p Valor

Variable dependiente: Minutos

F	gl1	gl2	Sig.
.355	2	27	.704

Contrasta la hipótesis nula de que la varianza error de la variable dependiente es igual a lo largo de todos los grupos.

a. Diseño: Intersección + Ambientacion_programa

Fuente: SPSS 20 IBM

- Si la prueba de ***Levene*** **NO** es significativa (**$p > 0.05$**), indica que las son aproximadamente iguales.
- En nuestro caso, la probabilidad es **$p=0.704>0.05$**, por lo que asumimos que **las varianzas son aproximadamente iguales.**
- Si la prueba de ***Levene*** **SI** es significativa (**$p < 0.05$**) entonces las varianzas son significativamente diferentes
- A continuación, pudiera desear revisar más de cerca a sus datos para ver si hay algún **resultado anómalo**, o considerar una transformación para hacer sus varianzas más homogéneas.
- La tabla resumen del modelo **ANOVA** es etiquetada como **Pruebas de los efectos inter-sujetos.** Esta tabla contiene información clave respecto a nuestra prueba calculada del **estadístico *F*. Ver Figura 8.50.**

Figura 8.50. Pruebas de los efectos inter-sujetos

Variable independiente (lo que SI podemos explicar con nuestro modelo)

Pruebas de los efectos inter-sujetos

Prueba estadística

***p* Valor**

Variable dependiente: Minutos

Origen	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
Modelo corregido	886.667[a]	2	443.333	33.250	.000
Intersección	14963.333	1	14963.333	1122.250	.000
Ambientacion_programa	886.667	2	443.333	33.250	.000
Error	360.000	27	13.333		
Total	16210.000	30			
Total corregida	1246.667	29			

a. R cuadrado = .711 (R cuadrado corregida = .690)

Variablidad dentro del grupo (lo que NO podemos explicar con nuestro modelo)

Fuente: SPSS 20 IBM

- En la tabla **Figura 8.50** de **ANOVA** hay 2 renglones de máximo interés: nuestro **factor** (**Ambientación_programa**) y el **Error**. El valor ***F*** del renglón **Ambientación_programa** muestra la **significancia** de nuestro factor (***p* = 0.000**).
- Si el **SPSS** establece que la probabilidad es **0.000,** significa que el **SPSS** ha redondeado la cantidad al número más cercano en tres lugares decimales. Sin embargo, con **0.000** siempre redondeamos del último 0 a 1, así que ***p* < 0.001.**
- En un **ANOVA** existen dos valores de grados de libertad a reportar. Uno para el **factor (Ambientación_programa)** el cual tiene **2 grados de libertad** y el otro para el **Error**, con **27 grados de libertad.**
- La forma convencional de reportar los hallazgos es establecer la **prueba estadística (F)**, los grados de libertad (**gl**), y el valor de probabilidad (**Sig**.) como sigue:

***F* (2, 27) = 33.250; *p* < 0.001**

- Como ***p* < 0.001**, esto indica que existe una alta diferencia significativa entre los 3 grupos, pero sin establecer dónde se ubica la significancia.
- El **Modelo corregido** muestra cuanta variabilidad en los datos se explican por nuestra variable independiente. Observe que la **Suma de cuadrados** son las mismas que las **Sumas de cuadrados de ambientación_programa** (**886.667**).
- La **ANOVA** trabaja las cantidades de variabilidad de los datos en torno del **valor de la media**. Si esta **media global** es **0** entonces la **Intersección** será también cero. Sin embargo, cuando la **media no es cero** no estaremos interesados en cómo las puntuaciones individuales difieren de cero, estaremos interesados en cómo ellos **difieren de la media global**. En este caso las sumas de **Intersección** de los cuadrados simplemente nos indicarán cuánta variabilidad es debida a la **media global** siendo

- diferente de cero. Así, se puede **remover** esta variabilidad (ya que no es relevante para el cálculo de nuestra **ANOVA**) para producir las **Suma de cuadrados total corregida.**
- En nuestro ejemplo, el renglón de **Intersección** nos muestra que nuestra **media global** es **significativamente diferente de cero**.
- Las **Sumas de cuadrados** reportan una medida de la variabilidad en las puntuaciones debido a una fuente particular de variabilidad. La **Suma de cuadrados total corregida** (**1246.667**) es la **Suma de cuadrados** de nuestro **factor Ambientación_programa** (**886.667**) y las sumas de **Error de cuadrados** (**360.000**). Observe que hay una gran cantidad de variabilidad debido más a nuestro **factor** que al error.
- La **Media cuadrática** es la cantidad de varianza (**Sumas de cuadrados** divididos por los **grados de libertad**).

Los valores de **R cuadrada y R ajustada** nos reportan un indicativo dela cantidad de variabilidad en las puntuaciones que pueden ser explicadas por nuestra **variable independiente**. Se calcula al dividir las **Sumas del Modelo corregido** (**886.667**) por **Suma de cuadrados total corregida** (**1246.667**), reportando el valor de **0.711**. Por lo tanto, nuestro modelo puede explicar el **71.1 %** de la variabilidad.

- La tabla de **Comparaciones múltiples *post hoc*** el ***Tukey* HSD (*Honestly Significant Difference***); este compara cada par de condiciones para verificar si sus diferencias diferencia es significante. Las comparaciones múltiples son emprendidas cuando **No tenemos planeado un contraste. Ver Figura 8.51.**

Figura 8.51. Comparaciones múltiples

Comparaciones múltiples

Variable dependiente: Minutos

DHS de Tukey

(I)Ambientacion de programa	(J)Ambientacion de programa	Diferencia de medias (I-J)	Error típ.	Sig.	Intervalo de confianza 95%	
					Límite inferior	Límite superior
Terrestre	Acuatico	-9.00*	1.633	.000	-13.05	-4.95
	Aereo	-13.00*	1.633	.000	-17.05	-8.95
Acuatico	Terrestre	9.00*	1.633	.000	4.95	13.05
	Aereo	-4.00	1.633	.053	-8.05	.05
Aereo	Terrestre	13.00*	1.633	.000	8.95	17.05
	Acuatico	4.00	1.633	.053	-.05	8.05

Basadas en las medias observadas.

El término de error es la media cuadrática(Error) = 13.333.

*. La diferencia de medias es significativa al nivel .05.

Fuente: SPSS 20 IBM

De la **Figura 8.51**, las columnas importantes, son: **Diferencias de medias (I-J)** y **Sig.**
La tabla muestra **todas las posibles comparaciones** de nuestros **3 grupos** de participantes.

En cada comparación, un grupo se le da el identificador '**I**' y al segundo '**J**'. Esto es evidente en la columna de la **Diferencia de medias**, la cual indica las cifras resultantes cuando la media de un grupo (**J**) ha sido sustraída de la media del otro grupo (**I**).
En nuestro ejemplo, la media del grupo 1(ambientación terrestre) se muestra como **15.00** en la estadística descriptiva (ver **Figura 8.48**), y la media del segundo grupo (ambientación acuática) **24.00** minutos. Así que:

15.0000 (I) – 24.0000 (J) = –9.0000

La columna **Sig.** nos permite evaluar si las diferencias de las medias entre los grupos son significantes.
Se observa de nuestro ejemplo que la diferencia entre la ambientación terrestre y la acuática es significante, como loes la diferencia entre la ambientación terrestre y la ambientación aérea ya que los ***p* valores** son **< 0.05.**
No se ha encontrado una diferencia significativa entre los grupos de ambientación acuática y la ambientación aérea ya que **p > 0.05**. Sin embargo, es apenas notoria por lo que se sugiere examinar los intervalos de confianza que puedan reportar más información con mayor información en cuanto a la fuerza de esta diferencia.
La sección del **Intervalo de confianza 95%** nos reporta un diferente método para evaluar las diferencias en nuestros grupos. Al observar los niveles de significancia, se concluye que NO hubo una diferencia significativa entre el grupo de ambientación aéreo y el grupo de ambientación acuático. Sin embargo, el nivel de confianza calculado sugiere que es posible exista una diferencia.
Los límites superior e inferior de los intervalos de confianza son: -**0.05** a **8.05**. Testo es un claro rango amplio, sin embargo incluye al **cero** justo al límite del rango, a pesar de que la media es **4.**
Los intervalos de confianza son por lo tanto una buena manera de completar los niveles de significancia, particularmente si las cifras encontradas están al borde de los límites.
Los hallazgos de lo citado hasta ahora, se encuentran resumidos en la **Tabla de subconjuntos homogéneos** (**Minutos**), mostrada en la **Figura 8.52.**

Figura 8.52. Subconjuntos homogéneos

Subconjuntos homogéneos

Minutos

DHS de Tukey[a,b]

Ambientacion de programa	N	Subconjunto 1	Subconjunto 2
Terrestre	10	15.00	
Acuatico	10		24.00
Aereo	10		28.00
Sig.		1.000	.053

Se muestran las medias de los grupos de subconjuntos homogéneos.
Basadas en las medias observadas.
El término de error es la media cuadrática(Error) = 13.333.

a. Usa el tamaño muestral de la media armónica = 10.000

b. Alfa = .05.

Fuente: SPSS 20 IBM

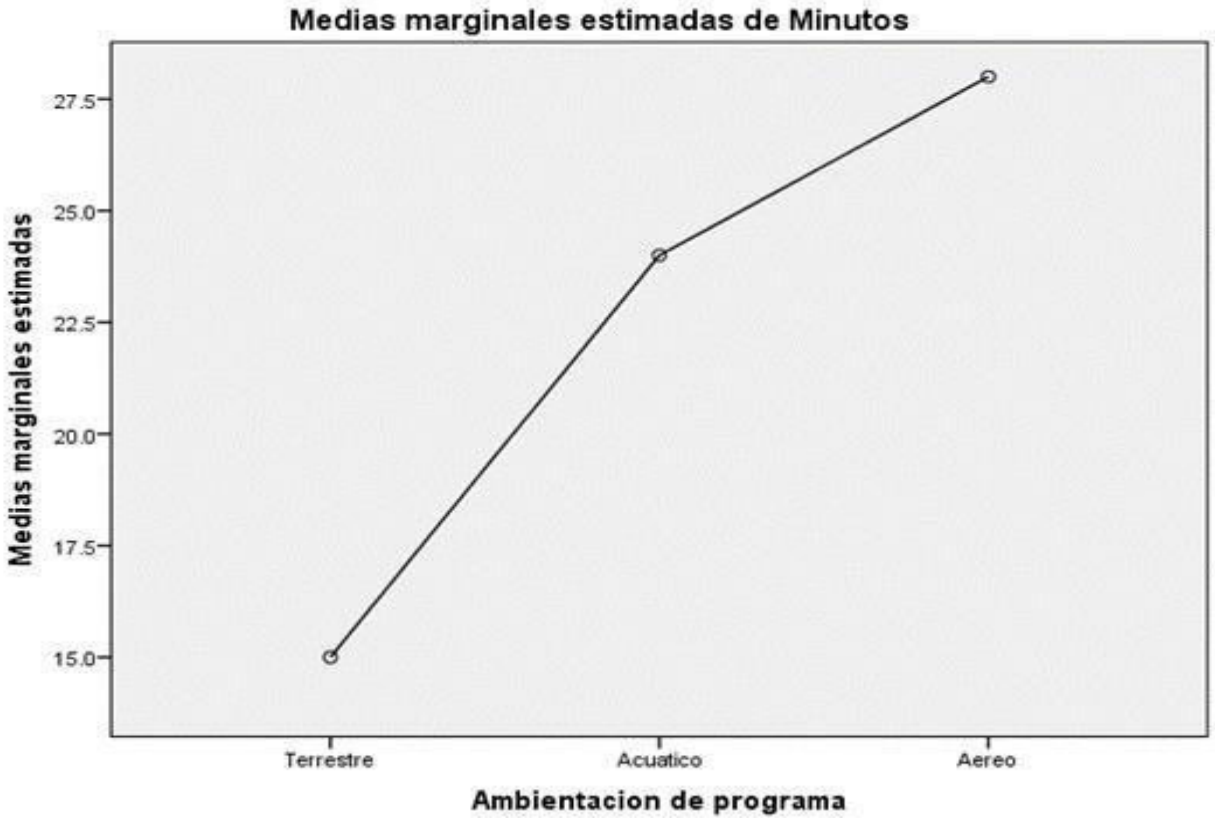
Esta tabla combina aquellas **comparaciones de pares** las cuales NO se determinaron como significativamente diferentes entre ellas.

- Como se observa de la tabla de **Comparaciones múltipl**e (**Figura 8.51**) el grupo de ambientación terrestre **SI** actuó significativamente diferente de los otros 2 grupos (ambientación acuática y ambientación aérea). Sin embargo, estos 2 últimos grupos **NO** actuaron significativamente diferente entre ellos.
- SPSS, tiene ha creado 2 subconjuntos de los datos. Dado que el primer grupo fue determinado que actuó diferente de los otros 2, aparece en un subconjunto aparte.
- Los 2 grupos que restan se determinaron diferentes del primer grupo de ambientación terrestre, pero no entre cada uno de ellos, y por lo tanto, aparecen en el mismo subconjunto.
- Si todos los 3 grupos fueran determinados como significativamente diferentes entre ellos, entonces 3 subconjuntos separados habrían sido creados y reportados, uno para cada grupo de estudio.

Fuente: SPSS 20 IBM

En nuestro proceso de **ANOVA** seleccionamos la opción para **SPSS** para generar un gráfico de las medias de nuestros grupos de participantes. Este nos permite obtener una imagen visual del desempeño de los jugadores de los 3 grupos de ambientación. Ver **Figura 8.53.**

Figura 8.53. Gráfico de las Medias (Comparar con Figura 8.45)



Fuente: SPSS 20 IBM

- Se puede observar los patrones previamente discutidos
- Aquellos participantes en el grupo que fue asignado con ambientación terrestre y con tips de ayuda completó el juego en el menor tiempo posible.
- El grupo que peor actuó fue aquel cuyos miembros no tuvieron los tips de ayuda para completar el juego (Ambientación aérea)
- **Conclusión: Se rechaza la H_0 y se acepta H_1 donde:**

$_1$ = El tiempo para ganar la competencia de los jugadores, en los ambientes terrestre y acuático que reciben tips de ayuda, **son diferentes** que el aéreo que no recibe tips de ayuda….siendo el aéreo el que más tiempo utilizó y que sus diferencias con el acuático no son tan marcadas como con el terrestre.

8.15. ANOVA de un factor de mediciones repetidas. Ejemplos.

Paso 1: Objetivos

Problema 2: Problema: la empresa **MKT_Digital** a través de la base de datos **MKT_Digital_videojuegos.sav,** tiene registrados a 30 jugadores y la cantidad de errores generados antes de ganar el nivel de un videojuego próximo a lanzar en 3 ambientaciones: terrestre, acuático (ambos con tips de ayuda) y aéreo (sin tips de ayuda). Las pruebas de error en el videojuego se basaron en 3 interfaces en desarrollo para el videojuego: joystick, teclado y joystick-teclado (mixto) por lo que desea conocer cuáles son las 2 mejores interfaces que tiene mayores posibilidades de éxito.

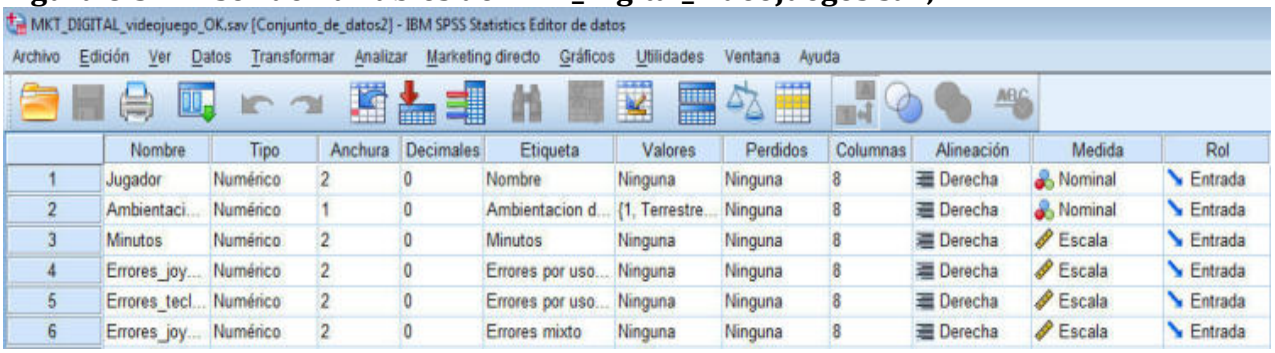
$_1$= La interface joystick (1) y tecclado (2) tienen más posibilidades de éxito que la interface joystick-teclado (mixta 3)

$_2$= La interface joystick (1) y joystick-teclado (mixta 3) tienen más posibilidades de éxito que la interface tecclado (2)

$_3$= La interface teclado (2)) y joystick-teclado (mixta 3) tienen más posibilidades de éxito que la interface joystick (1)

Ver Figuras 8.54 y 8.55

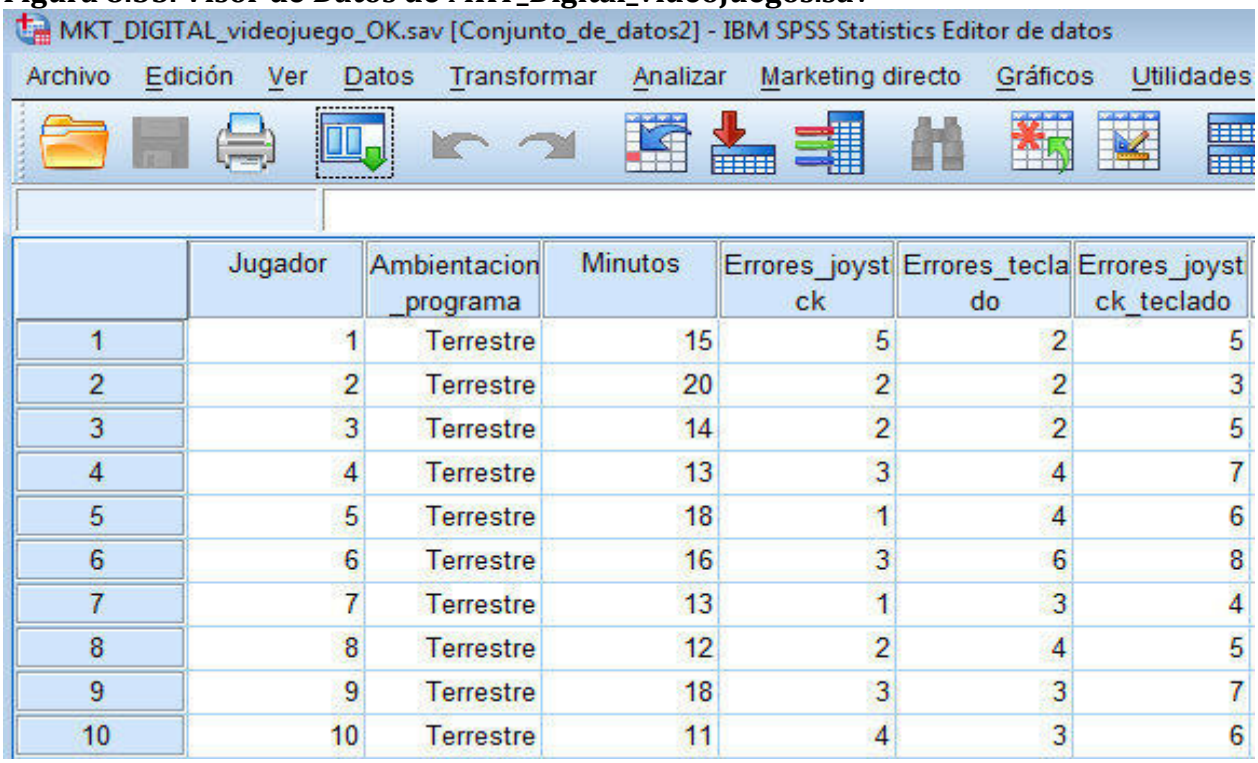
Figura 8.54. Visor de Variables de MKT_Digital_videojuegos.sav,



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	Jugador	Numérico	2	0	Nombre	Ninguna	Ninguna	8	Derecha	Nominal	Entrada
2	Ambientaci...	Numérico	1	0	Ambientacion d...	{1, Terrestre...	Ninguna	8	Derecha	Nominal	Entrada
3	Minutos	Numérico	2	0	Minutos	Ninguna	Ninguna	8	Derecha	Escala	Entrada
4	Errores_joy...	Numérico	2	0	Errores por uso...	Ninguna	Ninguna	8	Derecha	Escala	Entrada
5	Errores_tecl...	Numérico	2	0	Errores por uso...	Ninguna	Ninguna	8	Derecha	Escala	Entrada
6	Errores_joy...	Numérico	2	0	Errores mixto	Ninguna	Ninguna	8	Derecha	Escala	Entrada

Fuente: SPSS 20 IBM

Figura 8.55. Visor de Datos de MKT_Digital_videojuegos.sav



	Jugador	Ambientacion_programa	Minutos	Errores_joystck	Errores_teclado	Errores_joystck_teclado
1	1	Terrestre	15	5	2	5
2	2	Terrestre	20	2	2	3
3	3	Terrestre	14	2	2	5
4	4	Terrestre	13	3	4	7
5	5	Terrestre	18	1	4	6
6	6	Terrestre	16	3	6	8
7	7	Terrestre	13	1	3	4
8	8	Terrestre	12	2	4	5
9	9	Terrestre	18	3	3	7
10	10	Terrestre	11	4	3	6

Fuente: SPSS 20 IBM

Paso 2: Diseño

- En adelante para efectos de aprendizaje, sólo se tomarán de forma directa los datos con el fin de aprender a utilizar la técnica.
- Esta técnica de **ANOVA** es usada cuando se tiene a los mismos participantes en cada una de las condiciones de la **variable independiente**, tales como una prueba de los mismos jugadores de un videojuego en 3 ambientaciones diferentes y **tiempo de prueba diferentes** a fin de verificar si existe algún efecto de ***memoria*** en su desempeño. Es muy posible que se requiera implementar medidas que balanceen esta situación a fin de probar el control del ***efecto memoria*** en la práctica. Existen ventajas en las medidas repetidas sobre el método de mediciones de **factores independientes** de la **ANOVA**. Una de ellas es que cuando se usa a los **mismos participantes en todas las condiciones** de la **variable independiente,** estamos en posición de remover a las ***diferencias individuales*** del análisis antes de entrar al proceso estadístico. Lo anterior produce una inserción pequeña del término de error y en una alta probabilidad de encontrar un efecto significativo cuando se encuentra presente.

Por ejemplo, si probamos al mismo grupo de jugadores en su comprensión de las 3 ambientaciones del videojuego para verificar si existe alguna diferencia en su entendimiento entre las mismas ambientaciones del guion del juego, es posible que encontremos que e1 jugador 5 entiende de todas las ambientaciones del juego y consiga altas puntuaciones en su comprensión, pero muy bajas respecto a su desempeño en el juego, por ejemplo de ambientación acuática; incluso comparado con el jugador 28 que consigue

bajos registros de comprensión de la temática de las ambientaciones del videojuego, pero que logra bajos registros en la ambientación acuática. Con esto, aunque el jugador 5 está logrando registros más altos que el jugador 28 el patrón de resultados a través de los guiones de las ambientaciones es la misma, con la ambientación terrestre dando la puntuación más baja para ambos (incluso cuando el puntaje por comprensión, del jugador 5>jugador 28).

Paso 3: Condiciones de Aplicabilidad

- No olvidar realizar los análisis previos de inspección visual de la base de datos para detectar ausentes y/o atípicos (corregir en su defecto), confiabilidad, normalidad, homocedasticidad y linealidad.
- Las mediciones repetidas de **ANOVA** remueven las diferencias individuales de los participantes (o **dentro de los sujetos de la varianza**), los cuales en el caso del jugador 28 generalmente puntean mucho más que el jugador 5, y entonces las diferencias del análisis entre las condiciones, ambos jugadores puntean a bajo nivel en la ambientación terrestre que en las otras 2y 3. El resultado de esto es que el **análisis del ANOVA de un factor de mediciones repetidas es más complicado para realizar** y el **SPSS** genera una gran cantidad de reportes, con las siguientes características:
  1. Al igual que los supuestos del método de medición de **ANOVA de un factor independiente,** requiere de intervalos de datos en cada condición, que partan d**e distribuciones normales**, con **homocedasticidad** y o **linealmente relacionadas**, las mediciones de **ANOVA de un factor de mediciones repetidas** tiene el supuesto adicional de la ***esfericidad***.
  2. Lo anterior a menudo complica más que explica a las pruebas. Esencialmente **ANOVA** supone que las diferencias entre puntuaciones en cada par de condiciones tiene homocedasticidad. Necesita partir de esto para realizar el **ANOVA** propiamente, de otra forma trabajaría la homocedasticidad de manera inadecuada y el análisis no sería tan significativo.
  3. En el diseño de mediciones repetidas, con los mismos participantes actuando en cada condición, corremos el riesgo de violar el supuesto anterior y causar el cálculo incorrecto de un **valor del estadístico *F*,** así que **SPSS** aplica la aproximación de la ***esfericidad***, la cual se logra:

Paso 4: Estimación y ajuste

- Con el aseguramiento de pruebas como el de ***Mauchly*** y ***Epsilon***, para verificar si la ***esfericidad*** se encuentra en parámetros. Si es así, entonces puede usarse el valor de la '***esfericidad*** **supuesta de F**.
- Sin embargo, si existe algún problema, **SPSS** nos reporta **valores seleccionados de *F* corregida** dada la violación en ***esfericidad*** (por ejemplo: ***uso*** **de la técnica de *Greenhouse– Geisser***).

5. SPSS va más allá y para aquellos afectados por el incumplimiento de ***esfericidad***, se tiene la alternativa la **prueba original univariada**. Esto es el '***análisis multivariado***' que **no requiere el supuesto de *esfericidad*** del todo (por ejemplo, usa ***lambda de Wilks***) el cual se puede usar en su lugar. La recomendación aquí es que debe checar la **prueba de *Mauchly*** y la **prueba de *Epsilon*** así, utilizar el valor de la '***esfericidad*** **supuesta F'**, si la

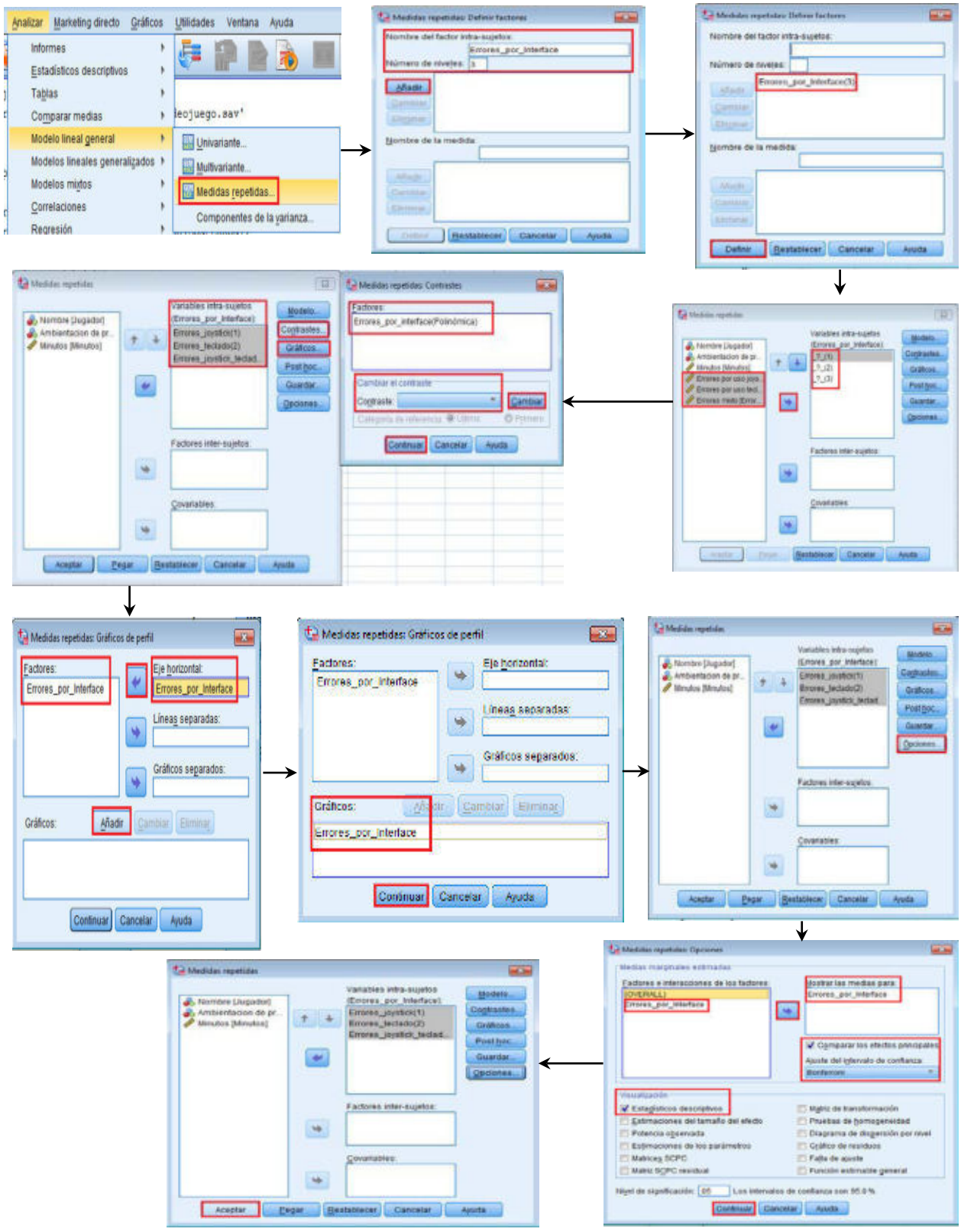
prueba de *Mauchly* no es significativa y *Epsilon* está cerca de 1. Sin embargo, puede tener un rápido vistazo tanto a las pruebas de ***Greenhouse–Geisser y lambda de Wilks'*** para checar si se están dando los mismos resultados. Si las pruebas d**e *Mauchly*** indican una violación a la *esfericidad* o Usted tiene preocupación del valor de ***Epsilon***, Usted debe escoger una corrección tal como la ***Greenhouse–Geisser*** o la ligeramente conservadora **Huynh–Feldt.**

6. Por otro lado, con muestras muy grandes puede seleccionar una prueba multivariante, tal como ***lambda de Wilks*** (la más popular), donde la *esfericidad* no es un problema, sin embargo, **la prueba univariada es a menudo la prueba más poderosa.**

7. Para muchos análisis, la *esfericidad* estará bien (la prueba de ***Mauchly*** será no significativa y el valor de la ***esfericidad F*** estará bien (con las pruebas como la de ***lambda de Wilks*** o la de ***Greenhouse–Geisser*** indicando el mismo resultado). Pero es de simple sentido observar las diferentes tablas para checar que esto sea así y no es entonces si no es así, tomar más tiempo para decidir sobre la mejor estadística a usar.

Teclear: Analizar->Modelo lineal general->Medidas repetidas (las mediciones repetidas de la variables llamada Nombre del factor intra-sujetos, en nuestro caso: Errores_por_Interface) ->Número de niveles: 3->Añadir->Definir->Variables intra-sujetos: ingresar las variables de estudio: (en nuestro caso Errores por uso joystick, teclado y mixto correspondientes)->Contrastes->Cambiar el contraste; contraste: Polinómico->Cambiar->Continuar->Gráficos->Factores: Errores_por_Interface->EjeHorizontal->Añadir->Continuar->Opciones-> Mostrar medias para: Errores_por Interface->Comparar los efectos principales->Ajuste del intervalo de confianza->Bonferroni->Visualización->Estadísticos descriptivos->Continuar-->Aceptar. Ver Figura 8.56.

Figura 8.56. Método ANOVA de un factor de mediciones repetidas



Fuente: SPSS 20 IBM

Paso 5: Interpretación

La primera tabla que produce el **SPSS** es **Factores Intra-sujetos** que reporta una descripción de los factores ingresados dentro de la ecuación de la **ANOVA**. Esto confirma que nuestro factor intra-sujetos (nuestros factores de medición repetidas) tiene 3 condiciones: Errores _Joystick, Errores_Teclado, Errores_Joystick_Teclado. Ver **Figura 8.57.**

Figura 8.57. Tabla Factores-intrasujetos

Factores intra-sujetos

Medida: MEASURE_1

Errores_por_Interface	Variable dependiente
1	Errores_joystick
2	Errores_teclado
3	Errores_joystick_teclado

Fuente: SPSS 20 IBM

La siguiente es la de los descriptivos Ver **Figura 8.58.**

Figura 8.58. Tabla Estadísticos descriptivos

Estadísticos descriptivos

	Media	Desviación típica	N
Errores por uso joystick	3.13	1.548	30
Errores por uso teclado	3.50	1.106	30
Errores mixto	5.73	1.660	30

Fuente: SPSS 20 IBM

- La **Figura 8.58** muestra la **Media de errores** generados por los participantes en cada una de las interfaces
- Por observación de las **medias**, se tiene que el número de errores incrementa del joystick (1), al teclado (2) y así hasta el uso mixto (3) (3.13, 3.50 y 5.73 respectivamente).
- La **Desviación típica** (desviación estándar), indica la dispersión de las puntuaciones dentro de las pruebas de las interfaces. Por observación delas mismas, podemos determinar que la interface mixta tiene más variabilidad en su desempeño (1.660),

- mientras que la 2 (teclado) produjo la menor variabilidad de errores (1.106) entre los participantes de la prueba del videojuego.
- N representa el número de participantes quienes realizaron la prueba de cada interface.
- La siguiente tabla a analizar es **Contrastes multivariados**, la cual se genera automáticamente por el **SPSS** durante el proceso de **ANOVA** de un factor de mediciones repetidas. Solamente se utiliza si la *esfericidad* muestra algún problema con nuestros datos. Ver **Figura 8.59.**

Figura 8.59. Tabla Contrastes multivariados

Contrastes multivariados[a]

Efecto		Valor	F	Gl de la hipótesis	Gl del error	Sig.
Errores_por_Interface	Traza de Pillai	.817	62.327[b]	2.000	28.000	.000
	Lambda de Wilks	.183	62.327[b]	2.000	28.000	.000
	Traza de Hotelling	4.452	62.327[b]	2.000	28.000	.000
	Raíz mayor de Roy	4.452	62.327[b]	2.000	28.000	.000

a. Diseño: Intersección
Diseño intra-sujetos: Errores_por_Interface

b. Estadístico exacto

Fuente: SPSS 20 IBM

- Las pruebas anteriores hacen algunos supuestos acerca de los datos y pueden ser apropiados si la **prueba de *Mauchly* de *esfericidad* es significativa o el valor de *Epsilon* NO es suficientemente grande** (ver **Figura 8.60**).

Figura 8.60. Tabla Prueba de *esfericidad* de *Mauchly*

Prueba de esfericidad de Mauchly[a]

Medida: MEASURE_1

Efecto intra-sujetos	W de Mauchly	Chi-cuadrado aprox.	gl	Sig.	Epsilon[b]		
					Greenhouse-Geisser	Huynh-Feldt	Límite-inferior
Errores_por_Interface	.995	.126	2	.939	.996	1.000	.500

Contrasta la hipótesis nula de que la matriz de covarianza error de las variables dependientes transformadas es proporcional a una matriz identidad.

a. Diseño: Intersección
Diseño intra-sujetos: Errores_por_Interface

b. Puede usarse para corregir los grados de libertad en las pruebas de significación promediadas. Las pruebas corregidas se muestran en la tabla Pruebas de los efectos inter-sujetos.

Fuente: SPSS 20 IBM

- La prueba más popular es la ***Lambda de Wilks.*** **(Ver Figura 8.59)**

- De nuestro ejemplo, usando ***Lambda de Wilks*** se concluye que **SI HAY** una diferencia significativa entre los desempeños de las interfaces:

$$F(2,28) = 62.327;\ p < 0.05$$

- En este ejemplo en particular, las pruebas de contraste multivariado producen un resultado significativo. Esto coincide con la prueba univariada mostrada más adelante. En el caso de que fuera lo contrario (que el resultado fuera no significativo), es posible que una prueba univariada indique que es significativa, dado que las **pruebas multivariadas son menos poderosas que las pruebas univariadas en pequeñas muestras de datos**
- Cuando se tienen más que 2 condiciones de las mediciones de variables repetidas, procedemos a verificar el supuesto de ***esfericidad*** antes de calcular los resultados de los valores ***F***
- Si el **supuesto de *esfericidad*** está presente, entonces procedemos a reportar el valor de **F** a partir de **la línea de *Esfericidad* asumida** (ver **Figura 8.60**) de la tabla **Prueba de efectos intra-sujetos**, abajo anexa.
- Si el **supuesto de *esfericidad* no está presente** y **la prueba de *esfericidad* de *Mauchly* es significativa,** entonces **NO** podemos tomar **la línea de *Esfericidad* asumida de la Prueba de efectos intra-sujetos** y necesitaremos realizar correcciones. **SPSS** nos reporta varios modelos de corrección, el de ***Greehouse–Geisser*** es el que usualmente se reporta.
- La **Figura 8.60** reporta una prueba **W de *Mauchly* estadística** de ***0.995, gl = 2; p=0.939 > 0.05***. Con esto es posible concluir que **el supuesto de *esfericidad* ha sido determinado** podemos usar los resultados desde el **modelo univariado SIN corrección.**
- Existe, sin embargo, algunos debates de cómo la **sensibilidad de la prueba de *Mauchly*, en sus alcances y habilidades detecta la *esfericidad***.
- Dada esta circunstancia, existe la alternativa de apoyarse en la consulta del **valor de *Epsilon*** anotada en la columna de ***Greenhouse-Geisser.*** **Esta cifra debe ser tan cercana a 1 como sea posible a fin de evitar problemas de no *esfericidad***
- Nuestro valor es cercano a **1** (**0.996**) así que podemos confiar que los problemas de *esfericidad* no afectan nuestros cálculos.
- Si el supuesto de la *esfericidad* se encuentra, como en el ejemplo, se pueden tomar los valores desde los renglones de la ***Esfericidad* asumida de la tabla de Pruebas de efectos intra-sujetos, Ver Figura 8.61**

Figura 8.61. Tabla Pruebas de efectos intra-sujetos

Pruebas de efectos intra-sujetos.

Medida: MEASURE_1

Origen		Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
Errores_por_Interface	Esfericidad asumida	118.822	2	59.411	66.465	.000
	Greenhouse-Geisser	118.822	1.991	59.679	66.465	.000
	Huynh-Feldt	118.822	2.000	59.411	66.465	.000
	Límite-inferior	118.822	1.000	118.822	66.465	.000
Error (Errores_por_Interface)	Esfericidad asumida	51.844	58	.894		
	Greenhouse-Geisser	51.844	57.740	.898		
	Huynh-Feldt	51.844	58.000	.894		
	Límite-inferior	51.844	29.000	1.788		

Fuente: SPSS 20 IBM

- Los renglones importantes son los de ***Esfericidad* asumida** en esta tabla y han sido encuadrados en color rojo
- Los grados de libertad (**gl**) para ambos, la variable y el error deben ser reportados, **gl** = (**2,58**).
- La forma convencional de reportar los hallazgos es establecer la prueba estadística (**F**), los grados de libertad (**gl**), y el ***p* valor** (**Sig.**):

$$F(2,58) = 66.465,\ p < 0.01$$

- **Como $p < 0.01$, indica que se ha encontrado una diferencia significativa** en el desempeño de nuestras interfaces. No se sabe sin embargo en donde se encuentran las diferencias y por lo tanto debe consultarse los resultados de las pruebas ***post hoc*** para esta información
- La **tabla de Pruebas de contrastes intra-sujetos** es generada automáticamente por **SPSS** durante el proceso de **ANOVA de un factor de mediciones repetidas** y es una dirección de análisis. **Ver Figura 8.62.**

Figura 8.62. Tabla Pruebas de contrastes intra-sujetos

Pruebas de contrastes intra-sujetos

Medida: MEASURE_1

Origen	Errores_por_Interface	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
Errores_por_Interface	Lineal	101.400	1	101.400	106.543	.000
	Cuadrático	17.422	1	17.422	20.840	.000
Error (Errores_por_Interface)	Lineal	27.600	29	.952		
	Cuadrático	24.244	29	.836		

Fuente: SPSS 20 IBM

- **La tabla de Pruebas de contrastes intra-sujetos** examina las tendencias desplegadas de nuestros datos y reporta información en cuanto al mejor ajuste de datos de nuestro modelo subyacente Fuente: SPSS 20 IBM
- Como se tienen 3 interfaces de los que se están probando los números de errores, las 2 posibles tendencias son : **el modelo lineal o cuadrático** de contraste
- En nuestro ejemplo, **se tiene hemos encontrado una tendencia significante lineal** en nuestros datos***: $F(1,29) = 106.543; p < 0.05$*.** Sin embargo, TAMBIEN se ha identificado una tendencia cuadrática: ***$F(1,29) = 20.840; p<0.05$*.**
- La **tabla de pruebas de los efectos inter-sujetos** es generada automáticamente por **SPSS** durante el proceso de **ANOVA de un factor de mediciones repetidas.** Como estamos calculando un **ANOVA** para un modelo de un factor y por lo tanto, no tiene un factor entre sujetos, la información que esta tabla nos reporta es con referencia a la **intersección.** Sin embargo, si tuviéramos una segunda variable, que fuera de una medición independiente, como en la **ANOVA** de 2 factores por diseño combinados, el efecto del factor independiente debería mostrarse aquí. **Ver Figura 8.63.**

Figura 8.63. Tabla Pruebas de los efectos intra-sujetos

Pruebas de los efectos inter-sujetos

Medida: MEASURE_1
Variable transformada: Promedio

Origen	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
Intersección	1529.344	1	1529.344	333.494	.000
Error	132.989	29	4.586		

Fuente: SPSS 20 IBM

En esta ocasión, como no se tienen variables independientes, sólo existe la intersección, el cual nos indica que nuestra **media global, es significativamente diferente de cero** Fuente: SPSS 20 IBM

- Las siguientes tablas se refieren al análisis de comparación por pares múltiple requerido para nuestras mediciones de factor repetidas, las cual fue una comparación de los principales efectos con el ajuste de ***Bonferroni.*** **Ver Figura 8.64.**

Figura 8.64. Tabla Pruebas de los efectos intra-sujetos

Estimaciones

Medida: MEASURE_1

Errores_por_Interface	Media	Error típ.	Intervalo de confianza 95%	
			Límite inferior	Límite superior
1	3.133	.283	2.555	3.711
2	3.500	.202	3.087	3.913
3	5.733	.303	5.114	6.353

Fuente: SPSS 20 IBM

- La **Media** indica las medias de puntuación de errores de cada una de las 3 interfaces. Esto es similar al caso de las medias discutidas en la **Figura 8.58** de **Estadística descriptiva**
- El **Error típico** es un estimado de la **Desviación típica** (desviación estándar) de la distribución de la media de las muestras. Un valor pequeño nos dice que deberíamos esperar una media similar si hacemos la prueba de nuevo, pero un gran valor indica una gran variabilidad pronosticada en las medias.
- El **Error típico** de la media es una cifra útil ya que es usada en el cómputo de pruebas de significancia comparando las medias y en el cálculo de intervalos de confianza
- El **Intervalo de confianza 95%** indica que estamos al **95%** de confianza de que la media de la población verdadera mean se encuentre entre el límite superior e inferior mostrados. La media de la muestra cae entre estos **2 valores.**

La **tabla de Comparaciones por pares múltiples** nos reporta la comparación de las medias de todas las combinaciones pareadas de las 3 condiciones de medición repetidas, que en nuestro ejemplo es titulado como **Errores por Interface**. Todas las comparaciones son ajustadas usando el **método de *Bonferroni***. Esta tabla debe inspeccionarse para asegurar donde se encuentran las diferencias significativas, que son evidentes del cálculo de nuestro modelo **ANOVA**. **Figura 8.65.**

Figura 8.65. Tabla Pruebas de los efectos intra-sujetos

Comparaciones por pares

Medida: MEASURE_1

(I)Errores por Interface	(J)Errores por Interface	Diferencia de medias (I-J)	Error típ.	Sig.[b]	Intervalo de confianza al 95 % para la diferencia[b]	
					Límite inferior	Límite superior
1	2	-.367	.242	.422	-.982	.248
	3	-2.600*	.252	.000	-3.240	-1.960
2	1	.367	.242	.422	-.248	.982
	3	-2.233*	.238	.000	-2.839	-1.628
3	1	2.600*	.252	.000	1.960	3.240
	2	2.233*	.238	.000	1.628	2.839

Basadas en las medias marginales estimadas.

*. La diferencia de medias es significativa al nivel .05.

b. Ajuste para comparaciones múltiples: Bonferroni.

Fuente: SPSS 20 IBM

- La **tabla de Comparaciones por pares múltiples** muestra todas las posibles comparaciones para los 3 niveles de nuestras variables bajo medición repetida.
- En cada comparación a un nivel se le da el identificador '**I**' y al segundo '**J**'. Esto se ubica en la columna **Diferencia de medias (I-J),** el cual indica la cifra resultante cuando la media del nivel de la **variable (J) ha sido restada de la otra (I)**.
- En el ejemplo, la media global de nuestra primer interface **(I) Error por Interface= 0.367** como cálculo de nuestra estadística descriptiva y que la media global de la segunda interface **(J) Error por Interface=2.600**, así que:

  0.367 (I) – 2.600 (J) = –2.233
- La columna (**Sig.**) nos permite evaluar si las diferencias de la media entre los niveles de la variable son significativos
- Se puede observar de nuestro ejemplo de comparaciones por pares múltiples que **son significativamente diferentes las comparaciones de: interface 1 con interface 3 y de la interface 2 con interface 3. Las diferencias de la media aquí son 2.666 y 2.233** respectivamente, con una significancia en ***p* < 0.05**. La **otra comparación de interface 1 con interface 2, no se encontró significativa ya que p > 0.05.**
- Los valores de los **Errores tip**. (errores estándar) se consideran pequeños, indican baja variabilidad en las diferencias de las medias pronosticadas
- **El Intervalo de confianza al 95% para la diferencia** nos indica que estamos 95% seguros que la diferencia de la media de la población se encontrarán entre el límite superior y el límite inferior.
- Como se generaron **comparaciones por pares múltiples vía el método *Bonferroni***, la **ANOVA** también las **tablas de Contrastes multivariadas** también como las de ***Bonferroni***.
- **La tabla no es de especial interés en cuanto sigamos el método de análisis univariado,** llevando a cabo primero los chequeos sugeridos en los datos (***Mauchly* y *Epsilon***). Ver **Figura 8.66 (similar a Figura 8.59)**

Figura 8.66. Tabla Contrastes multivariados

Contrastes multivariados

	Valor	F	Gl de la hipótesis	Gl del error	Sig.
Traza de Pillai	.817	62.327[a]	2.000	28.000	.000
Lambda de Wilks	.183	62.327[a]	2.000	28.000	.000
Traza de Hotelling	4.452	62.327[a]	2.000	28.000	.000
Raíz mayor de Roy	4.452	62.327[a]	2.000	28.000	.000

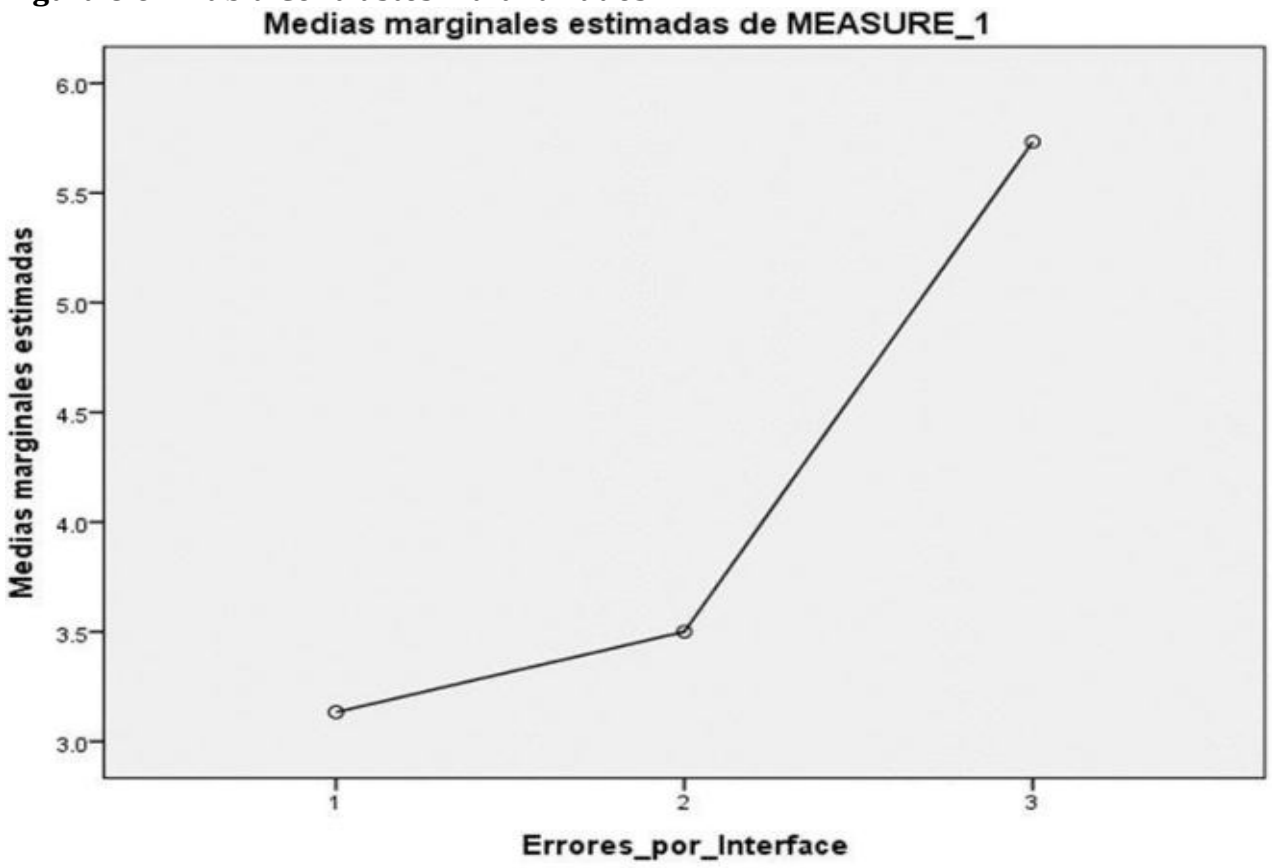
Cada prueba F contrasta el efecto multivariado de Errores_por_Interface. Estos contrastes se basan en las comparaciones por pares, linealmente independientes, entre las medias marginales estimadas.

a. Estadístico exacto

Fuente: SPSS 20 IBM

La parte final de los resultados es la gráfica de las medias de los errores producto del uso de las 3 interfaces. Ver **Figura 8.67.**

Figura 8.67. Tabla Contrastes multivariados



Fuente: SPSS 20 IBM

Conclusión: Como se puede ver de la gráfica**, los participantes en nuestra muestra hicieron más errores en cuanto cambiaron de** la **interface 1 a interface 3 y de la interface 2 a la interface 3, por lo que las interfaces con mejores posibilidades de éxito son las interfaces 1 y 2 con respecto a la 3. Se rechazan las hipótesis 2 y 3, aceptándose la hipótesis 1.**

H_1= La interface joystick (1) y tecclado (2) tienen más posibilidades de éxito que la interface joystick-teclado (mixta 3)…Se acepta

H_2= La interface joystick (1) y joystick-teclado (mixta 3) tienen más posibilidades de éxito que la interface teclado (2)…..Se rechaza

H_3= La interface tecclado (2)) y joystick-teclado (mixta 3) tienen más posibilidades de éxito que la interface joystick (1)…..Se rechaza

• Es claro también que en el despliegue de los datos se encuentra una tendencia de tipo cuadrático a lineal.

8.16. ANOVA de dos Factores. Resumen

Se lleva a cabo **ANOVA de dos factores** cuando deseamos analizar el efecto de **2 variables independientes** sobre una variable **dependiente, y los supuestos de una prueba paramétrica deban cumplirse**. **ANOVA de dos factores** es una de las pruebas más populares, debido en parte, a que se modela una interacción. Además de analizar el efecto de las variables independientes por separado (referido esto como los principales efectos) también es capaz de analizar sus efectos combinados (referido como **interacción**). Suponga que está investigando el **efecto de la música y el tipo de producto sobre cierto tipo de consumidor de artículos de entretenimiento**. Selecciona a los participantes tanto del ámbito juvenil urbano y suburbano y analiza la forma en que gasta en sus necesidades básicas como las de lujo. Es posible que se encontrara con efectos propios de la educación del consumidor, su nivel socioeconómico, o la geografía en la que se encuentra. Se podría encontrar un efecto principal en el tipo de producto, motivado por la moda, el precio y el valor que percibe el consumidor. También es posible encontrar una interacción, tal como que los jóvenes con más poder adquisitivo adquieren los productos de entretenimiento de mayor precio y en mayor cantidad que los artículos primarios que compran los jóvenes de menor poder adquisitivo. **Una interacción muestra que el efecto de una variable independiente no es la misma en cada condición de la otra variable independiente**. Así, que cuando lleve a cabo la **ANOVA de dos factores,** es posible se obtengan **3 valores de F**: Uno para el principal efecto de cada variable independiente y otra para la **interacción**. **Existen 3 tipos de ANOVA de dos factores** y necesitamos de estar seguros de seleccionar la correcta para nuestros datos. Así, cuando ambas variables independientes son mediciones independientes (es decir, que existen diferentes participantes en cada condición) llevamos a cabo una medición de **ANOVA de dos factores independientes**. Cuando ambas **variables independientes** han repetido mediciones a través de las distintas condiciones (es decir, cada participante contribuye a un puntaje para cada condición de la variable) emprenderá una medición de dos factores repetidos de **ANOVA**. Finalmente, cuando una **variable independiente** es independiente en mediciones y la otra variable también entonces llevamos a cabo un diseño de dos factores por diseño combinados de **ANOVA**. En el ejemplo anterior, hemos combinado el diseño tanto como haya mediciones independientes de nivel socioeconómico y con mediciones repetidas **sobre el efecto de la**

música y el tipo de producto sobre cierto tipo de consumidor de artículos de entretenimiento

8.17. ANOVA de dos factores. Ejemplos.

Paso 1: Objetivos

Problema 3: la empresa **MKT_Digital** requiere determinar cómo introducir un innovador sistema operativo que soportará los procesos de fabricación de su hardware, en sus instalaciones. El cuestionamiento es el de resolver si requerirá sólo cambiar su personal activo, que hoy día conoce y opera el sistema operativo antiguo o debe inclinarse por contratar nuevo personal quien jamás ha tenido contacto con los sistemas operativos. Para resolver, Usted tomará 12 personas del staff quienes tienen la experiencia de manejar el sistema operativo actual vs. 12 personas quienes nunca han tenido experiencia. Así también, la mitad del personal de ambas opciones se instala tanto en el sistema operativo innovador como actual, en los que son anotados los errores producidos

$_0$= La introducción del innovador sistema operativo que soportará los proceso de fabricación de software SI presenta efecto principal respecto del software de sistema operativo antiguo

$_1$= La introducción del innovador sistema operativo que soportará los procesos de fabricación de software NO presenta efecto principal respecto del software de sistema operativo antiguo

Ver Figuras 8.68 y 8.69

Figura 8.68. Visor de Variables de MKT_Digital_videojuego.sav

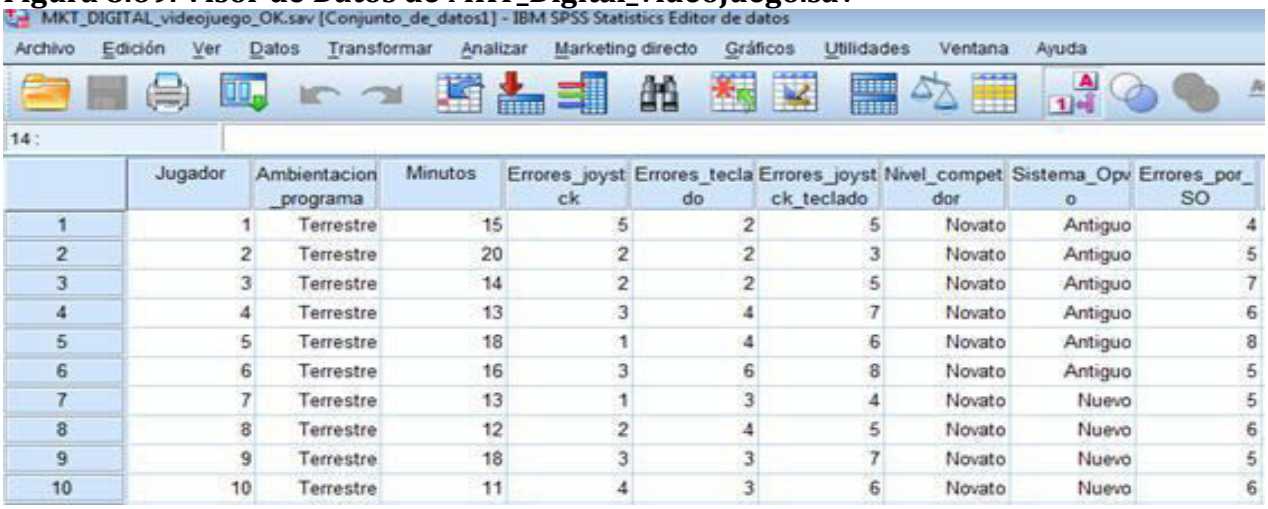
MKT_DIGITAL_videojuego_OK.sav [Conjunto_de_datos1] - IBM SPSS Statistics Editor de datos

Archivo Edición Ver Datos Transformar Analizar Marketing directo Gráficos Utilidades Ventana Ayuda

	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	Jugador	Numérico	2	0	Nombre	Ninguna	Ninguna	8	Derecha	Nominal	Entrada
2	Ambientacion_pro...	Numérico	1	0	Ambientacion de ...	{1, Terrestre...	Ninguna	8	Derecha	Nominal	Entrada
3	Minutos	Numérico	2	0	Minutos	Ninguna	Ninguna	8	Derecha	Escala	Entrada
4	Errores_joystick	Numérico	2	0	Errores por uso jo...	Ninguna	Ninguna	8	Derecha	Escala	Entrada
5	Errores_teclado	Numérico	2	0	Errores por uso t...	Ninguna	Ninguna	8	Derecha	Escala	Entrada
6	Errores_joystick_...	Numérico	2	0	Errores mixto	Ninguna	Ninguna	8	Derecha	Escala	Entrada
7	Nivel_competidor	Numérico	8	0	Nivel de experiencia	{0, Novato}...	Ninguna	8	Derecha	Escala	Entrada
8	Sistema_Opvo	Numérico	8	0	Version sistema ...	{0, Antiguo}...	Ninguna	8	Derecha	Escala	Entrada
9	Errores_por_SO	Numérico	8	0	Errores por el sist...	Ninguna	Ninguna	8	Derecha	Escala	Entrada

Fuente: SPSS 20 IBM

Figura 8.69. Visor de Datos de MKT_Digital_videojuego.sav



	Jugador	Ambientacion_programa	Minutos	Errores_joystck	Errores_teclado	Errores_joystck_teclado	Nivel_competidor	Sistema_Opvo	Errores_por_SO
1	1	Terrestre	15	5	2	5	Novato	Antiguo	4
2	2	Terrestre	20	2	2	3	Novato	Antiguo	5
3	3	Terrestre	14	2	2	5	Novato	Antiguo	7
4	4	Terrestre	13	3	4	7	Novato	Antiguo	6
5	5	Terrestre	18	1	4	6	Novato	Antiguo	8
6	6	Terrestre	16	3	6	8	Novato	Antiguo	5
7	7	Terrestre	13	1	3	4	Novato	Nuevo	5
8	8	Terrestre	12	2	4	5	Novato	Nuevo	6
9	9	Terrestre	18	3	3	7	Novato	Nuevo	5
10	10	Terrestre	11	4	3	6	Novato	Nuevo	6

Fuente: SPSS 20 IBM

Paso 2: Diseño

En adelante para efectos de aprendizaje, sólo se tomarán de forma directa los datos con el fin de aprender a utilizar la técnica.

Las mediciones **ANOVA** de dos factores independientes es la más simples forma d**e ANOVA** y produce la mínima cantidad de productos de **SPSS**. Esto es debido a que ambas **variables independientes son mediciones independientes**, que aportan puntaje en cada condición de los diferentes participantes. Por ejemplo, si se comparan los efectos de **género** de los gerentes de una empresa de telecomunicaciones, con su **posición** dentro de las distintas áreas de la empresa (producción, ingeniería, mercadotecnia) en función a su habilidad negociadora para **incrementar los ingresos**, **ambas variables son de mediciones independientes.** El aspecto importante de las **ANOVAS** de dos factores de mediciones independientes es que analiza los dos factores juntos, así que se produce una interacción también como los principales efectos de género y posición dentro de la empresa, para incrementar los ingresos de la compañía. Si por ejemplo, los gerentes del área de mercadotecnia tienen mejores resultados en sus habilidades de negociación que el resto de los gerentes, para obtener mayores ingresos para la compañía y se asocia a que las gerentes se presentan en mayor número como gerentes mujeres entonces, Usted habría encontrado una interacción de los dos factores para aumentar los ingresos: género y posicionamiento de trabajo. Existen supuestos que necesitarán ser ubicados para resolverse a través de **ANOVA**, y que deberán cumplir: a nuestro modelo de datos:

•Los datos son data tomados al azar.

•Los puntajes son medidos en una escala de intervalo y son provienen de poblaciones normalmente distribuidas

• Las muestras en cada condición son tomadas de poblaciones con homocedasticidad

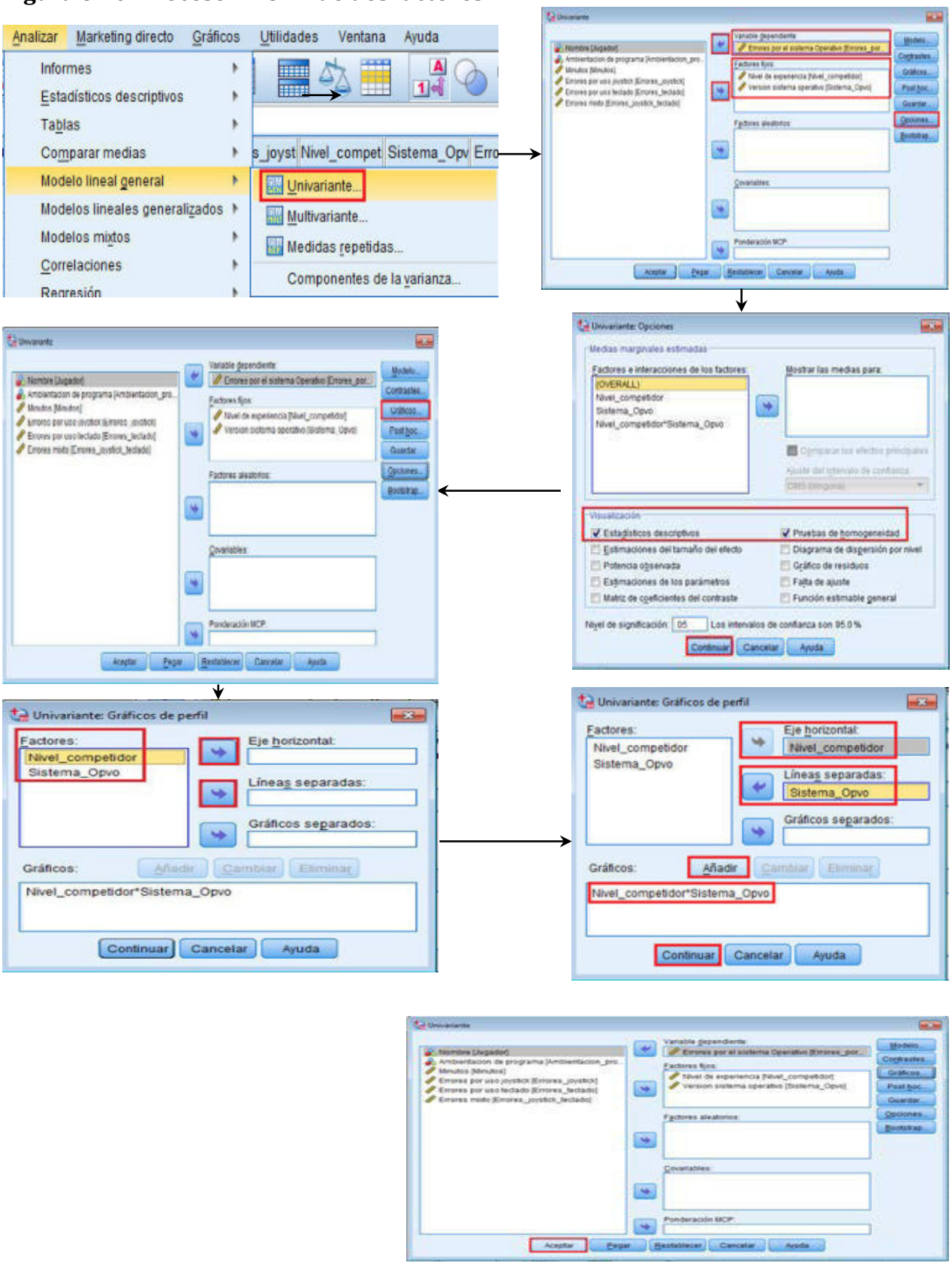
Paso 3: Condiciones de Aplicabilidad

No olvidar realizar los análisis previos de inspección visual de la base de datos para detectar ausentes y/o atípicos (corregir en su defecto), confiabilidad, normalidad, homocedasticidad y linealidad.

Paso 4: Estimación y ajuste

Teclear: Analizar->Modelo lineal general->Univariante->Variable dependiente: Errores por el sistema operativo->Factores fijos: Nivel de experiencia y Versión sistema operativo->Opciones->Visualización: Estadísticos descriptivos; Pruebas de homogeneidad->Continuar->Gráficos->Eje horizontal: Nivel_Competidor->Líneas separadas: Sistema_Opvo->Añadir->Continuar->Aceptar. Ver Figura 8.70.

Figura 8.70. Proceso ANOVA de dos factores……



Fuente: SPSS 20 IBM

Nota:

- Observe en este ejemplo **NO** se utilizó ninguna prueba ***post hoc*** **ya que ambos factores independientes solamente tuvieron 2 niveles.**
- Para **2 factores independientes de medición de ANOVA, donde uno o ambos factores tienen 3 o más niveles**, haga ***click*** en el botón ***post hoc*** y seleccione la prueba apropiada.

PASO 5: Interpretación

La primera tabla que SPSS produce es la **Factores Inter-sujetos**, la cual reporta cuántos participantes se encuentran en cada grupo y cuantas condiciones hay de cada una de las variables independientes. Ver **Figura 8.71.**

Figura 8.71. Tabla Factores-Intersujetos

Factores inter-sujetos

		Etiqueta del valor	N
Nivel de experiencia	0	Novato	12
	1	Experto	12
Version sistema operativo	0	Antiguo	12
	1	Nuevo	12

Variables independientes

Nivel de variables

Fuente: SPSS 20 IBM

La siguiente tabla que el **SPSS** nos muestra es la de **Estadísticos descriptivos** la cual describe la **media, desviación típica (estándar) y el número de participantes de cada grupo. Ver Figura 8.72.**

Figura 8.72. . Tabla Estadísticos descriptivos

Estadísticos descriptivos

Variable dependiente: Errores por el sistema Operativo

Nivel de experiencia	Version sistema operativo	Media	Desviación típica	N
Novato	Antiguo	5.83	1.472	6
	Nuevo	5.50	.548	6
	Total	5.67	1.073	12
Experto	Antiguo	2.17	.753	6
	Nuevo	8.17	.753	6
	Total	5.17	3.215	12
Total	Antiguo	4.00	2.216	12
	Nuevo	6.83	1.528	12
	Total	5.42	2.358	24

Fuente: SPSS 20 IBM

La tabla de **Estadísticos descriptivos** despliega la **Media** como el número de errores hechos por los trabajadores novatos y experimentados en los sistemas operativos antiguos y nuevos.

- Puede verse del renglón **Total** que el global general, no parece ser una gran diferencia en el número de errores cometidos por los principiantes y los trabajadores experimentados (una media de **5.67** para los trabajadores novatos en comparación con **5.17** para los trabajadores experimentados).
- Sin embargo, cuando se considera el **factor: tipo de sistema operativo** se puede ver claramente que surgen diferencias. Los trabajadores novatos hicieron casi el mismo número de errores en el viejo antiguo operativo (**5,83**) como en el nuevo tipo de sistema operativo (**5.5**). Los trabajadores con experiencia, sin embargo, estaban haciendo un menor número de errores en la el sistema operativo antiguo (**2.17**) que en la nueva máquina (**8.17**).
- Al observar el **TOTAL** en la fila inferior, se puede ver que los errores totales por tipo de sistema operativo, independientemente de si el trabajador es o no experto, SÍ muestra algunas diferencias, con más errores que se realizan cuando se trabaja con el nuevo sistema operativo.
- La **Desviación típica** (desviación estándar) muestra que al comparar las puntuaciones generales de los dos grupos, el grupo de trabajador **experto** tiene la mayor difusión de las puntuaciones (**3.215**), con las puntuaciones más estrechamente relacionados que se encuentran en el grupo de novatos (**1.0731**). Incluso aunque la **Desviación típica** en general es más grande para el grupo con experiencia, cuando consideramos el tipo de

- sistema operativo es el de novatos, que trabajan con sistemas operativos antiguos los que muestran la mayor difusión de las puntuaciones (**1.472**).

Con nuestras mediciones de **ANOVA** de dos factores independientes, estamos listos para buscar los principales efectos significativos de nuestros **2 factores**: **nivel de experiencia** y tipo de **sistema operativo así como una posible interacción entre los 2.** Toda la información está contenida en las **pruebas de los efectos inter-sujetos**. Ver **Figura 8.73.**

Figura 8.73. Tabla Pruebas de los efectos inter-sujetos

Pruebas de los efectos inter-sujetos

Variable dependiente: Errores por el sistema Operativo

Origen	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
Modelo corregido	109.833[a]	3	36.611	40.679	.000
Intersección	704.167	1	704.167	782.407	.000
Nivel_competidor	1.500	1	1.500	1.667	.211
Sistema_Opvo	48.167	1	48.167	53.519	.000
Nivel_competidor * Sistema_Opvo	60.167	1	60.167	66.852	.000
Error	18.000	20	.900		
Total	832.000	24			
Total corregida	127.833	23			

a. R cuadrado = .859 (R cuadrado corregida = .838)

Fuente: SPSS 20 IBM

Podemos ver que:

-De nuestro factor **Nivel_competidor NO se ha encontrado un efecto principal significativo,** o sea: ***F* (1,20) = 1.667, *p*=0.211> 0.05.** Esto se esperaba de las discusiones anteriores de la estadística descriptiva.

-De nuestro factor **Tipo de sistema operativo SÍ hemos encontrado un efecto principal significativo para el factor,** es decir, ya sea que los trabajadores estén usando una antigua o nuevo sistema operativo, resultando: ***F* (1,20) = 53.519, *p* <0.001.**

- Recuerde que si SPSS establece que la probabilidad (**Sig.**) Es 0.000, significa que SPSS ha redondeado la cantidad al número más cercano a tres cifras decimales. Sin embargo, se prefiere considerar el redondeo del último 0 a 1, por lo que ***p* <0.001.**
  Nuestros resultados indican que SI existe una interacción significante de nuestros dos factores *F (1, 20) = 66.852, p < 0.001.*

El **Modelo corregido** muestra la cantidad de variabilidad en los datos que podemos explicar por nuestras **variables independientes**. Tenga en cuenta que estas **Sumas de cuadrados se** componen de las **sumas de cuadrados del *"nivel de experiencia***", "***tipo de sistema operativo***" **y su interacción.**

• En nuestro ejemplo, la fila de **intersección** nos muestra que nuestra **media global es significativamente diferente de cero.**

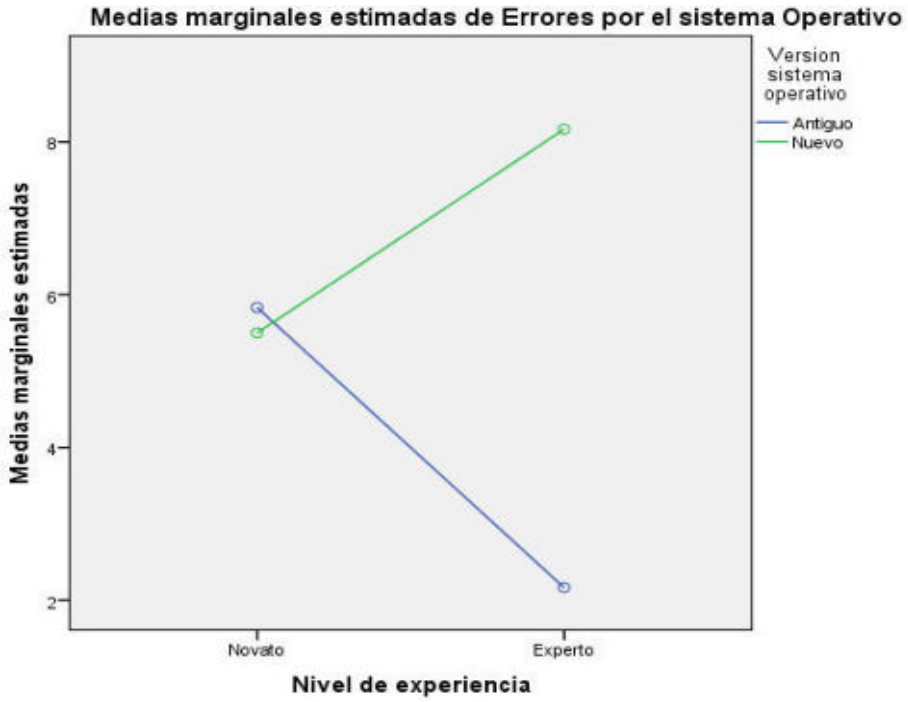
• Las **Sumas de cuadrados** dan una medida de la variabilidad en las puntuaciones debido a una en particular fuente de variabilidad. **La Suma de cuadrados Total Corregida** (**127.833**) está formado por las **Sumas** de nuestro factor **'nivel de experiencia**'(**1.500**), "***tipo de sistema operativo*** " (**48.167**) y la interacción entre ellos (**60.167**), además de las sumas de cuadrados (**18.000**). Tenga en cuenta que hay una gran cantidad de variabilidad debido al error de nuestro factor **nivel de experiencia**

• La **Media de cuadrática** es la varianza (**sumas de cuadrados divididos por grados de libertad**).

• El **R al cuadrado** y los valores de **R cuadrada ajustada** nos da una indicación de la cantidad de la variabilidad en las puntuaciones que se pueden explicar por nuestras variables independientes. Esto es se calcula dividiendo las **sumas de cuadrados del modelo corregido (109.833) por la Suma de cuadrados total corregida por las sumas totales de cuadrados** (**127.833**), dando un valor de **0.859**.

La parte final de **SPSS** es el gráfico de interacción, el cual permite analizar los patrones discutidos previamente. Ver **Tabla 8.74.**

Tabla 8.74. Gráfico de Interacción



- Fuente: SPSS 20 IBM

- El gráfico **representa las medias** y confirma nuestros hallazgos discutidos previamente con una significativa de interacción entre las dos variables.
- Recuerde, cada vez que las líneas en un gráfico de interacción **no son paralelas**, **esta indica que existe una interacción, aunque esto puede no ser estadísticamente** una interacción significativa.
- Podemos ver en la gráfica de interacción que los trabajadores con **nivel de experiencia,** como es lógico, hacen menos errores en el sistema operativo antiguo, pero hacen más en el sistema operativo nuevo
- Esto parece un caso de **transferencia negativa**, **donde las habilidades aprendidas anteriormente pueden ser un obstáculo más que una ayuda. Los trabajadores novatos parecen desempeñar con igual exactitud en ambos sistemas operativos.**
- Es posible que desee llevar a cabo más pruebas con los datos que le permitan comprender más plenamente los patrones emergentes en nuestro análisis. Por ejemplo, **podemos estar bastante seguros que nuestros operadores experimentados están haciendo más errores con el software de sistema operativo nuevo en comparación al software de sistema operativo antiguo, es decir, hemos identificado un simple efecto principal**. Sin embargo, estamos menos seguros acerca de las diferencias entre las máquinas para el principiante operadores.
- Una prueba para tales efectos principales simples por lo tanto, se puede calcular a través de la sintaxis mando dentro de SPSS.

H_0= La introducción del innovador sistema operativo que soportará los procesos de fabricación de software SI presenta efecto principal respecto del software de sistema operativo antiguo….dado que los empleados con mayor experiencia al interactuar con el sistema operativo nuevo generan más errores.

8.18. ANOVA de dos factores de medidas repetidas. Resumen

Usted realizará un ANOVA de dos factores de medidas repetidas cuando tengamos que realizar mediciones en ambas variables. En un estudio sobre la percepción, los investigadores estaban interesados en el tiempo que tardó una persona para encontrar una forma oculta (un circulo, un cubo o un cuadrado) en los patrones visuales. También estaban interesados en saber si había diferencia en los tiempos de detección para el ojo dominante en comparación con el ojo no dominante, por lo que cada persona vio la mitad de los patrones con un ojo y la mitad con la otra. En la medida que cada participante toma parte de cada condición de ambas variables independientes, tenemos mediciones repetidas en ambos factores. Los tiempos para la detección de las formas con cada ojo se registraron para el análisis. (Usted puede averiguar cuál de sus ojos es el dominante mediante colocar y observar un dedo con el brazo extendido. Alinéelo con un objeto en el otro lado de la habitación con los dos ojos abiertos. Con un ojo el dedo permanece en frente del objeto y así se determina su ojo dominante; con el otro ojo el dedo cambia y se "***reduce***", siendo este el ojo no dominante). **Como se tienen dos factores (dos variables independientes) habrá tres valores *F* generadas por ANOVA; una por el principal efecto de cada uno de las dos variables independientes y una por la interacción**. En el ejemplo anterior el **principal efecto del "*ojo*"** nos diría si hubo una diferencia en la detección de las veces del ojo dominante y del no dominante, y el efecto principal de la **"*forma*"** nos diría si hubo diferencia en las veces de detección de las formas diferentes. Una **interacción** ocurre

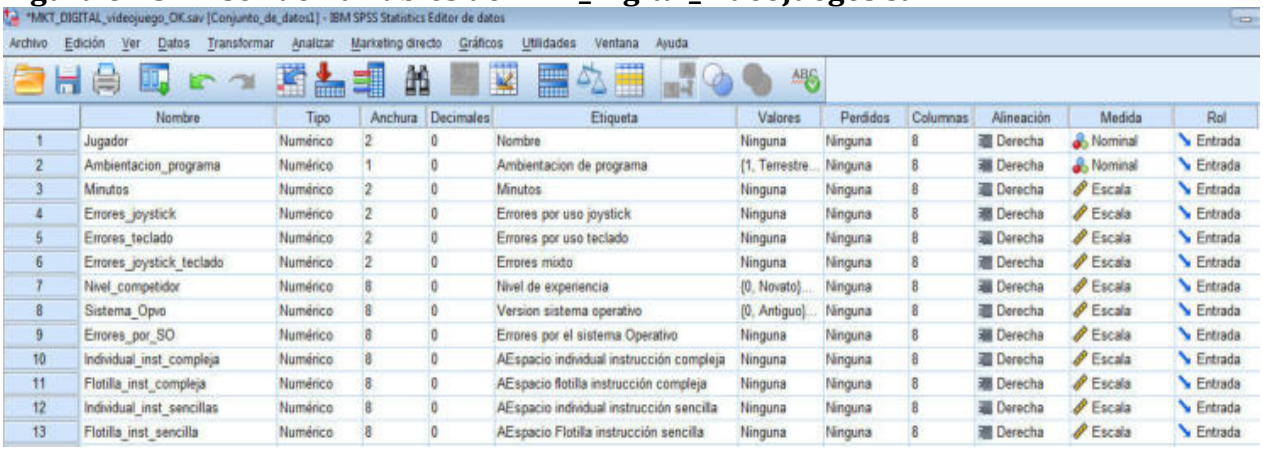
cuando el efecto de un factor es diferente en los niveles diferentes del segundo sector. Por ejemplo, si el ojo dominante fue mejor al detectar círculos y el ojo no-dominante lo es para detectar cubos, sin ninguna diferencia para los cuadrados.

Paso 1: Objetivos

Problema 4: la empresa **MKT_Digital** tiene dos videojuegos de serie de colección con ambiente en el espacio que opera como sigue: uno requiere que el jugador siga un set de instrucciones complejas y el otro, es más simple de operar. En la ambientación del juego, existen a su vez 2 variantes de combate: individual y en flotilla. El administrador desea que ambos juegos, se desarrollen con un mínimo de errores por parte de los usuarios (retraso, falta de precisión, atascos). Así, un investigador decide estudiar tanto el efecto de **variantes de combate** (individual vs. flotilla) así como el de **instrucción** (instrucciones complejas vs. sencillas) presentes en los errores producidos por los usuarios, a los que se les colocará en un **sistema de rotación de ambientes** (individual y flotilla). De esta forma se escogerán como prueba, 6 jugadores en un concurso de forma aleatoria y sus errores de desempeño serán medidos durante el cambio de modo de individual a flotilla. Un balanceo apropiado es emprendido de forma tal que los 3 jugadores de prueba inician primero con el modo individual, y los otros 3 en modo flotilla, no se tenga influencia de uno a otro por el cambio de juego.

- **H_0= Los jugadores producen más errores en la ambientación espacial, con instrucciones complejas y en variante de combate flotilla, que con el resto de combinaciones.**
- **H_1= Los jugadores producen menos errores en la ambientación espacial, con instrucciones complejas y en variante de combate flotilla, que con el resto de combinaciones. Ver Figuras 8.75 y 8.76**

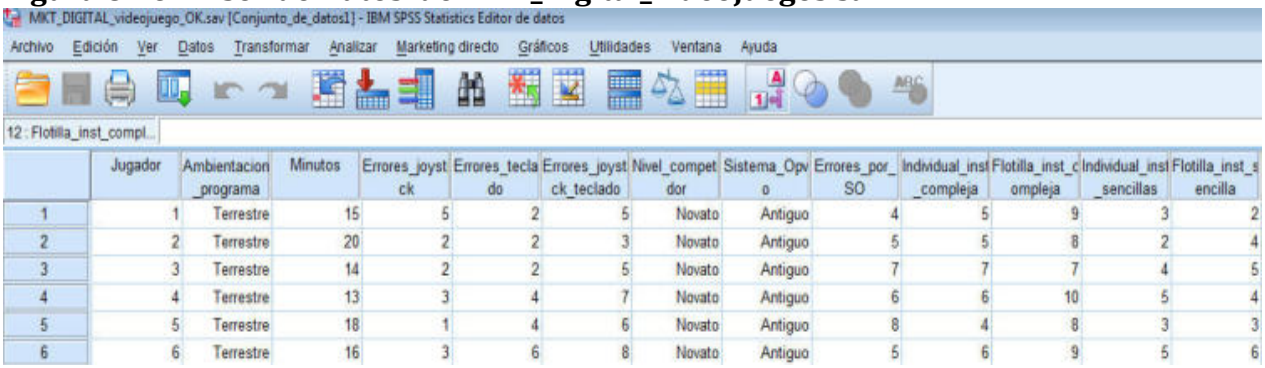
Figura 8.75. Visor de Variables de MKT_Digital_videojuegos.sav



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	Jugador	Numérico	2	0	Nombre	Ninguna	Ninguna	8	Derecha	Nominal	Entrada
2	Ambientacion_programa	Numérico	1	0	Ambientacion de programa	{1, Terrestre...	Ninguna	8	Derecha	Nominal	Entrada
3	Minutos	Numérico	2	0	Minutos	Ninguna	Ninguna	8	Derecha	Escala	Entrada
4	Errores_joystick	Numérico	2	0	Errores por uso joystick	Ninguna	Ninguna	8	Derecha	Escala	Entrada
5	Errores_teclado	Numérico	2	0	Errores por uso teclado	Ninguna	Ninguna	8	Derecha	Escala	Entrada
6	Errores_joystick_teclado	Numérico	2	0	Errores mixto	Ninguna	Ninguna	8	Derecha	Escala	Entrada
7	Nivel_competidor	Numérico	8	0	Nivel de experiencia	{0, Novato}...	Ninguna	8	Derecha	Escala	Entrada
8	Sistema_Opvo	Numérico	8	0	Version sistema operativo	{0, Antiguo}...	Ninguna	8	Derecha	Escala	Entrada
9	Errores_por_SO	Numérico	8	0	Errores por el sistema Operativo	Ninguna	Ninguna	8	Derecha	Escala	Entrada
10	Individual_inst_compleja	Numérico	8	0	AEspacio individual instrucción compleja	Ninguna	Ninguna	8	Derecha	Escala	Entrada
11	Flotilla_inst_compleja	Numérico	8	0	AEspacio flotilla instrucción compleja	Ninguna	Ninguna	8	Derecha	Escala	Entrada
12	Individual_inst_sencillas	Numérico	8	0	AEspacio individual instrucción sencilla	Ninguna	Ninguna	8	Derecha	Escala	Entrada
13	Flotilla_inst_sencilla	Numérico	8	0	AEspacio Flotilla instrucción sencilla	Ninguna	Ninguna	8	Derecha	Escala	Entrada

Fuente: SPSS 20 IBM

Figura 8.76. Visor de Datos de MKT_Digital_videojuegos.sav



	Jugador	Ambientacion_programa	Minutos	Errores_joystick	Errores_teclado	Errores_joystck_teclado	Nivel_competidor	Sistema_Opvo	Errores_por_SO	Individual_inst_compleja	Flotilla_inst_compleja	Individual_inst_sencillas	Flotilla_inst_sencilla
1	1	Terrestre	15	5	2	5	Novato	Antiguo	4	5	9	3	2
2	2	Terrestre	20	2	2	3	Novato	Antiguo	5	5	8	2	4
3	3	Terrestre	14	2	2	5	Novato	Antiguo	7	7	7	4	5
4	4	Terrestre	13	3	4	7	Novato	Antiguo	6	6	10	5	4
5	5	Terrestre	18	1	4	6	Novato	Antiguo	8	4	8	3	3
6	6	Terrestre	16	3	6	8	Novato	Antiguo	5	6	9	5	6

Fuente: SPSS 20 IBM

Paso 2: Diseño

- En adelante para efectos de aprendizaje, sólo se tomarán de forma directa los datos con el fin de aprender a utilizar la técnica.
- Para todo **ANOVA** de dos factores necesitamos hacer los supuestos:
  - Los datos están aleatoriamente seleccionados de la población.
  - Las puntuaciones son medidas en una escala de intervalo y provienen de una población normalmente distribuida
  - Las muestras en cada condición son tomadas de poblaciones con homocedasticidad

  Sin embargo, como los factores son de mediciones repetidas también debemos cumplir el supuesto de "***esfericidad***" y por lo tanto habrá más tablas de reporte generadas ´por **SPSS** que las realizadas con una medición independiente de **ANOVA** que nos permitirá analizar la "***esfericidad***" de los datos y corregir por alguna violación del supuesto si es necesario.

Paso 3: Condiciones de Aplicabilidad

- No olvidar realizar los análisis previos de inspección visual de la base de datos para detectar ausentes y/o atípicos (corregir en su defecto), confiabilidad, normalidad, homocedasticidad y linealidad

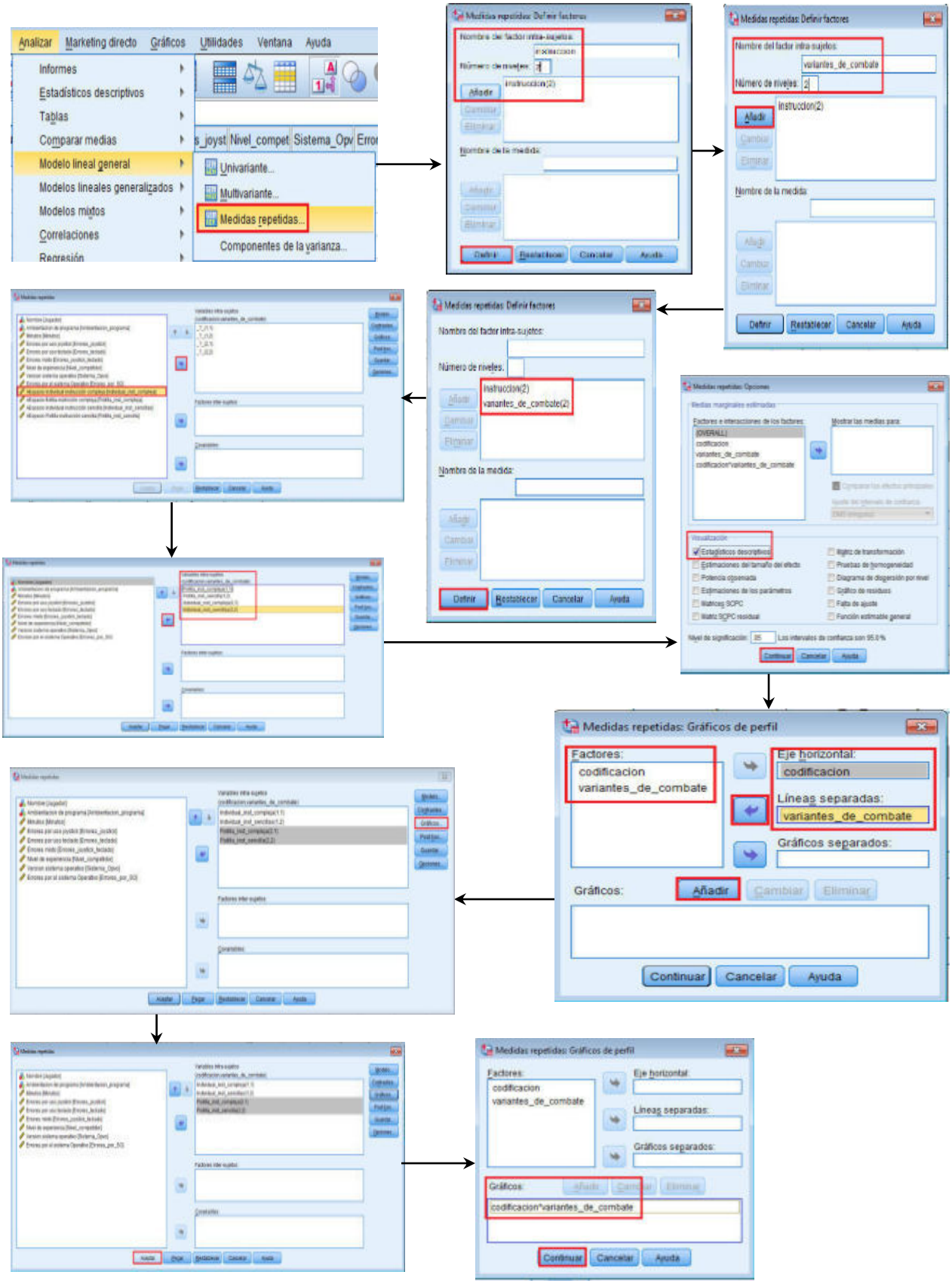
Paso 4: Estimación y Ajuste

-Teclear: Analizar->Modo lineal general->Medidas repetidas->Nombre del factor intra-sujetos: instrucción->Número de niveles: 2->Añadir-> Nombre del factor intra-sujetos: variantes_de_ combate->Número de niveles: 2->Añadir->Definir->Variables intra-sujetos:
individual_inst_compleja(1,1) ;individual_inst_sencilla(1,2) ;flotilla_inst_compleja(2,1); flotilla_inst_sencilla((2,2)*->Opciones->Visualización: Estadísticos descriptivos->Continuar->Gráficos-> Factores->Eje horizontal: instrucción->Lineas separadas: variantes_de_combate->Añadir->Continuar->Aceptar. Ver Figura 8.77.**

* **.-**Cada variable se asigna de acuerdo a la combinación de niveles a que corresponde a un factor

****.-**Si tuviéramos más de **2 niveles** para nuestros factores podríamos necesitar la **prueba *post hoc* de *Bonferroni*** a fin de asegurar dónde se detectan diferencias significativas.

Figura 8.77 Proceso ANOVA de dos factores de medidas repetidas……



Fuente: SPSS 20 IBM

Paso 5: Interpretación

La primera tabla generada por **SPSS** llamada **Factores intra-sujetos** nos reporta una descripción de los 2 factores ingresados dentro del cálculo e **ANOVA.** La primera columna muestra ambas variables, **instrucción y variantes de combate**. La segunda columna muestra el nombre de la **variable dependiente** formada por la combinación de los diferentes niveles de los factores. Ver **Figura 8.78.**

Figura 8.78. Tabla Factores intra-sujetos

Factores intra-sujetos

Medida: MEASURE_1

instruccion	variantes de combate	Variable dependiente
1	1	Flotilla_inst_compleja
	2	Flotilla_inst_sencilla
2	1	Individual_inst_compleja
	2	Individual_inst_sencillas

Fuente: SPSS 20 IBM

- Como se puede observar la combinación del primer nivel de la variable **instrucción** y el primer nivel de la variable **variantes de combate** se le etiqueto como **Flotilla_inst_compleja**
- Nivel 1 de **instrucción** y nivel 2 de **variantes de combate** es etiquetada como **Flotilla_inst_sencilla**
- El resto de las combinaciones de los niveles de los factores son asignados de igual manera.
- La siguiente tabla generada por el **SPSS** es la de **Estadísticos descriptivos**. Es qué que podemos analizar nuestros datos para potenciales diferencias de tasas de error. Ver **Figura 8.79.**

Figura 8.79. Tabla Estadísticos descriptivos

Estadísticos descriptivos

	Media	Desviación típica	N
AEspacio flotilla instrucción compleja	8.50	1.049	6
AEspacio Flotilla instrucción sencilla	4.00	1.414	6
AEspacio individual instrucción compleja	5.50	1.049	6
AEspacio individual instrucción sencilla	3.67	1.211	6

Fuente: SPSS 20 IBM

- Por observación de los errores de la **Media** podemos observar que haciendo la **codificación** del programa más complejo los jugadores producen números más grandes de errores (**A Espacio individual instrucción compleja: 5.50 vs. A Espacio individual instrucción sencilla: 3.67** y **A Espacio flotilla l instrucción compleja: 8.50 vs. A Espacio individual instrucción sencilla: 4.00.**
- Cuando se comparan las 2 **combinaciones de instrucción simple** y las 2 **combinaciones de instrucción compleja**, se observa que en ambos casos, los de instrucción compleja son los que producen más errores, así como la variante de combate en **flotilla.**
- La combinación que produjeron más errores fueron los de instrucciones complejas.
- La **Desviación típica** (desviaciones estándar) no dan indicios de una gran dispersión de los puntajes en cualquiera de las condiciones.

Fuente: SPSS 20 IBM

- En suma, se ha visto que se ha encontrado una posible diferencia entre **AEspacio flotilla e individual** y entre instrucciones **compleja y sencilla**, con efectos posibles de interacción de nuestras variables como se indica por el gran número de errores hechos por las **instrucciones complejas.**
- La tabla de pruebas de **Contrastes multivariados** es generada por el **SPSS** durante el proceso de **ANOVA** de dos factores de medidas repetidas. Solamente usamos esta tabla si la ***esfericidad*** mostrada tiene problemas de nuestros datos. Sin embargo, como solamente tenemos 2 niveles por cada uno de los valores de los factores de medición repetida, la ***esfericidad*** **no es un problema aquí**. Ver **Figura 8.80.**

Figura 8.80. Tabla Contrastes multivariados

Contrastes multivariados[a]

Efecto		Valor	F	Gl de la hipótesis	Gl del error	Sig.
instruccion	Traza de Pillai	.877	35.714[b]	1.000	5.000	.002
	Lambda de Wilks	.123	35.714[b]	1.000	5.000	.002
	Traza de Hotelling	7.143	35.714[b]	1.000	5.000	.002
	Raíz mayor de Roy	7.143	35.714[b]	1.000	5.000	.002
variantes_de_combate	Traza de Pillai	.940	78.478[b]	1.000	5.000	.000
	Lambda de Wilks	.060	78.478[b]	1.000	5.000	.000
	Traza de Hotelling	15.696	78.478[b]	1.000	5.000	.000
	Raíz mayor de Roy	15.696	78.478[b]	1.000	5.000	.000
instruccion * variantes_de_combate	Traza de Pillai	.593	7.273[b]	1.000	5.000	.043
	Lambda de Wilks	.407	7.273[b]	1.000	5.000	.043
	Traza de Hotelling	1.455	7.273[b]	1.000	5.000	.043
	Raíz mayor de Roy	1.455	7.273[b]	1.000	5.000	.043

a. Diseño: Intersección
Diseño intra-sujetos: instruccion + variantes_de_combate + instruccion * variantes_de_combate

b. Estadístico exacto

Fuente: SPSS 20 IBM

- Como ambos factores son de mediciones repetidas, **3 pruebas multivariadas** son generadas; una por cada factor, y otra para las interacciones entre los dos
- La prueba más popular es la ***Lambda de Wilks.*** De ésta se puede observar que hay un efecto principal significativo para nuestro factor ***instrucción,*** así que:

F(1,5) = 35.714; p =0.002< 0.01.

- Así también, se ha encontrado un efecto principal significativo en **variantes de combate, así que:**

F(1,5) = 78.478; p < 0.01.

- Cuando analizamos la tabla para determinar una posible **interacción** entre las 2 variables podemos consultar el reporte de lo que esto significa, como:

F(1,5) = 7.273; p < 0.05.

- La siguiente tabla generada por **SPSS** muestra las diversas pruebas de *esfericidad*. Nuevamente, estas pruebas son generadas al procesar datos durante el proceso ANOVA de mediciones repetidas. Ver **Figura 8.81.**

Figura 8.81. Tabla Prueba de *esfericidad* de *Mauchly*

Prueba de esfericidad de Mauchly[a]

Medida: MEASURE_1

Efecto intra-sujetos	W de Mauchly	Chi-cuadrado aprox.	gl	Sig.	Epsilon[b]		
					Greenhouse-Geisser	Huynh-Feldt	Límite-inferior
instruccion	1.000	.000	0	.	1.000	1.000	1.000
variantes_de_combate	1.000	.000	0	.	1.000	1.000	1.000
instruccion * variantes_de_combate	1.000	.000	0	.	1.000	1.000	1.000

Contrasta la hipótesis nula de que la matriz de covarianza error de las variables dependientes transformadas es proporcional a una matriz identidad.

a. Diseño: Intersección
Diseño intra-sujetos: instruccion + variantes_de_combate + instruccion * variantes_de_combate

b. Puede usarse para corregir los grados de libertad en las pruebas de significación promediadas. Las pruebas corregidas se muestran en la tabla Pruebas de los efectos inter-sujetos.

Fuente: SPSS 20 IBM

- Como se observa la columna **Sig.** (probabilidad) está en blanco, sin grados de libertad (**gl**) reportados. Esto es así, debido a que la ***esfericidad***, solamente es un problema si Usted tuviera **más de 2 condiciones** en sus factores de mediciones repetidas. Ambos de nuestros factores solamente tienen **2 niveles** y por lo tanto, la ***esfericidad NO será un problema de datos.***
- La tabla, por lo tanto puede permanecer ignorada en este ejemplo. Sin embargo, si una o más de sus variables **tuvieran más de 2 niveles, entonces necesitaría checar esto de la misma manera para ANOVA de mediciones repetidas de un factor.**
- Como la ***esfericidad*** **NO** es un problema en nuestros datos, podemos tomar los valores desde los renglones de ***Esfericidad*** **asumida de la tabla Prueba de efectos intra-sujetos**.Ver **Figura 8.82.**

Figura 8.82. Tabla Prueba de efectos intra-sujetos

Origen		Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
instruccion	Esfericidad asumida	16.667	1	16.667	35.714	.002
	Greenhouse-Geisser	16.667	1.000	16.667	35.714	.002
	Huynh-Feldt	16.667	1.000	16.667	35.714	.002
	Límite-inferior	16.667	1.000	16.667	35.714	.002
Error(instruccion)	Esfericidad asumida	2.333	5	.467		
	Greenhouse-Geisser	2.333	5.000	.467		
	Huynh-Feldt	2.333	5.000	.467		
	Límite-inferior	2.333	5.000	.467		
variantes_de_combate	Esfericidad asumida	60.167	1	60.167	78.478	.000
	Greenhouse-Geisser	60.167	1.000	60.167	78.478	.000
	Huynh-Feldt	60.167	1.000	60.167	78.478	.000
	Límite-inferior	60.167	1.000	60.167	78.478	.000
Error (variantes_de_combate)	Esfericidad asumida	3.833	5	.767		
	Greenhouse-Geisser	3.833	5.000	.767		
	Huynh-Feldt	3.833	5.000	.767		
	Límite-inferior	3.833	5.000	.767		
instruccion * variantes_de_combate	Esfericidad asumida	10.667	1	10.667	7.273	.043
	Greenhouse-Geisser	10.667	1.000	10.667	7.273	.043
	Huynh-Feldt	10.667	1.000	10.667	7.273	.043
	Límite-inferior	10.667	1.000	10.667	7.273	.043
Error (instruccion*variantes_de _combate)	Esfericidad asumida	7.333	5	1.467		
	Greenhouse-Geisser	7.333	5.000	1.467		
	Huynh-Feldt	7.333	5.000	1.467		
	Límite-inferior	7.333	5.000	1.467		

Fuente: SPSS 20 IBM

- Los renglones más importantes de la tabla son los enmarcados como ***Esfericidad* asumida.**
- Así, se deberá estar en búsqueda de un efecto principal de forma significativa para las variables: **instrucción, variantes_de_combate** y su posible **interacción.**
- Así, se tiene un efecto principal significativo para el factor **instrucción**, como

F(1,5) = 35.714, p =0.002< 0.01.

- Otro efecto principal significativo para el factor **variantes_de_combate,** como

F(1,5) = 78.478, p =0.000< 0.01.

- Una **interacción** evidente entre las variables, como:

F(1,5) = 7.273, p < 0.05.

- **Hallazgo:** podemos concluir que, como se indicó en los resultados descriptivos, ambos factores **instrucción** y **variantes_de_combate** tienen un efecto sobre el número de errores producidos por los jugadores y esto resulta en una interacción significativa entre los 2 factores.
- La tabla de pruebas de **Contrastes intra-sujetos** se genera por **SPSS** durante el proceso de cálculo de **ANOVA** de mediciones repetidas y es un análisis de tendencia. **Figura 8.83.**

Figura 8.83. Tabla Prueba de contrastes intra-sujetos

Pruebas de contrastes intra-sujetos

Medida: MEASURE_1

Origen	instruccion	variantes_de_combate	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
instruccion	Lineal		16.667	1	16.667	35.714	.002
Error(instruccion)	Lineal		2.333	5	.467		
variantes_de_combate		Lineal	60.167	1	60.167	78.478	.000
Error (variantes_de_combate)		Lineal	3.833	5	.767		
instruccion * variantes_de_combate	Lineal	Lineal	10.667	1	10.667	7.273	.043
Error (instruccion*variantes_de_combate)	Lineal	Lineal	7.333	5	1.467		

Fuente: SPSS 20 IBM

- Esta tabla analiza las tendencias desplegadas en nuestros datos. **Esta reporta información de cómo el modelo subyacente mejora el ajuste de datos.**
- Como solamente teneos 2 niveles para cada uno de los factores con mediciones repetidas la única tendencia posible a seguir es la del **modelo lineal.**
- En nuestro ejemplo, podemos ver que ambos factores siguen una tendencia lineal, así como la interacción entre ambos. Esto es esperado, dado que cada factor tiene solamente **2** niveles.
- La tabla de **Pruebas de los efectos inter-sujetos** se genera por **SPSS** durante el proceso de cálculo de **ANOVA** de mediciones repetidas. Con ambos factores de medidas repetidas, NO tenemos sujetos entre los factores, por lo que la información que se reporta está con referencia a la **intersección**. Si, por otro lado, tuviéramos mediciones de variables independientes, como lo que ocurre con el **ANOVA** de dos factores **combinados,** el efecto del factor independiente se mostraría también. **Ver Figura 8.84.**

Figura 8.84. Tabla Prueba de los efectos inter-sujetos

Pruebas de los efectos inter-sujetos

Medida: MEASURE_1

Variable transformada: Promedio

Origen	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
Intersección	704.167	1	704.167	237.360	.000
Error	14.833	5	2.967		

Fuente: SPSS 20 IBM

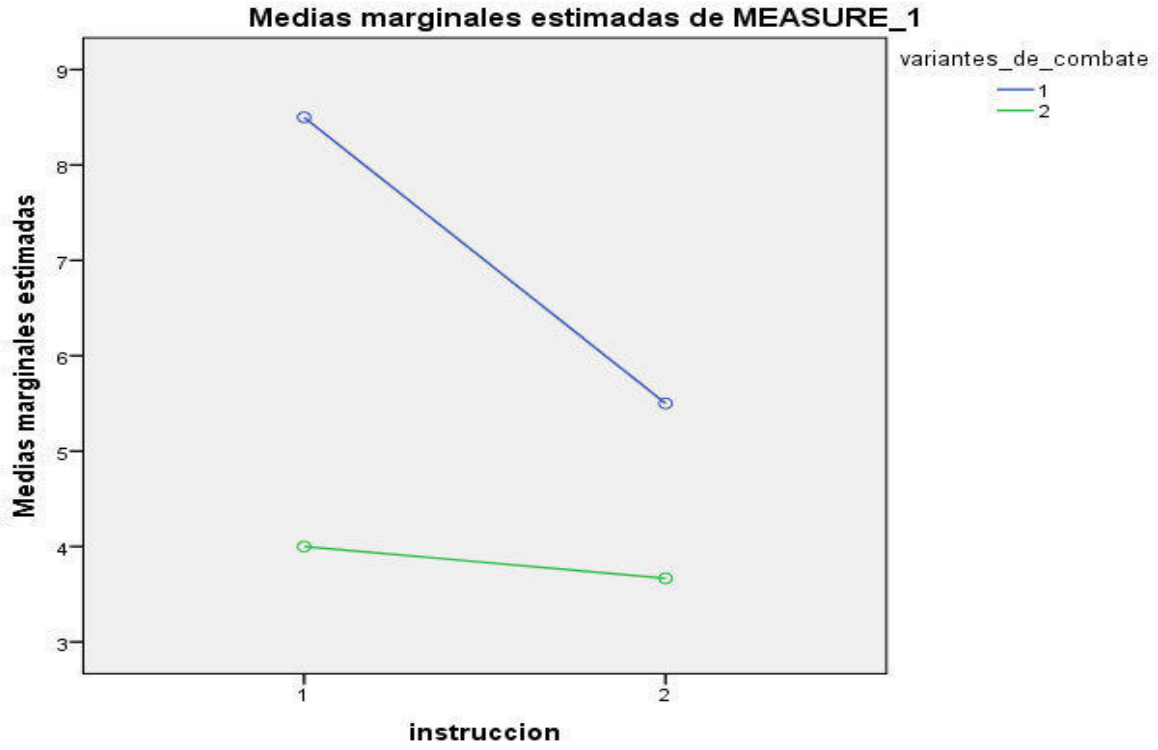
- En esta tabla, al no tener mediciones de variables independientes, sólo se podrá explicar a través de la **intersección** la cual nos indica que nuestra **media global es significativamente diferente de cero**.

Fuente: SPSS 20 IBM

- La parte final de la salida es el diagrama de interacción de las medias de las cuatro condiciones.
- Tenga en cuenta que hemos utilizado el Editor de gráficos para agregar etiquetas. Ver **Figura 8.85.**

Figura 8.85. Tabla Gráfico de interacción

Gráficos de perfil



Fuente: SPSS 20 IBM

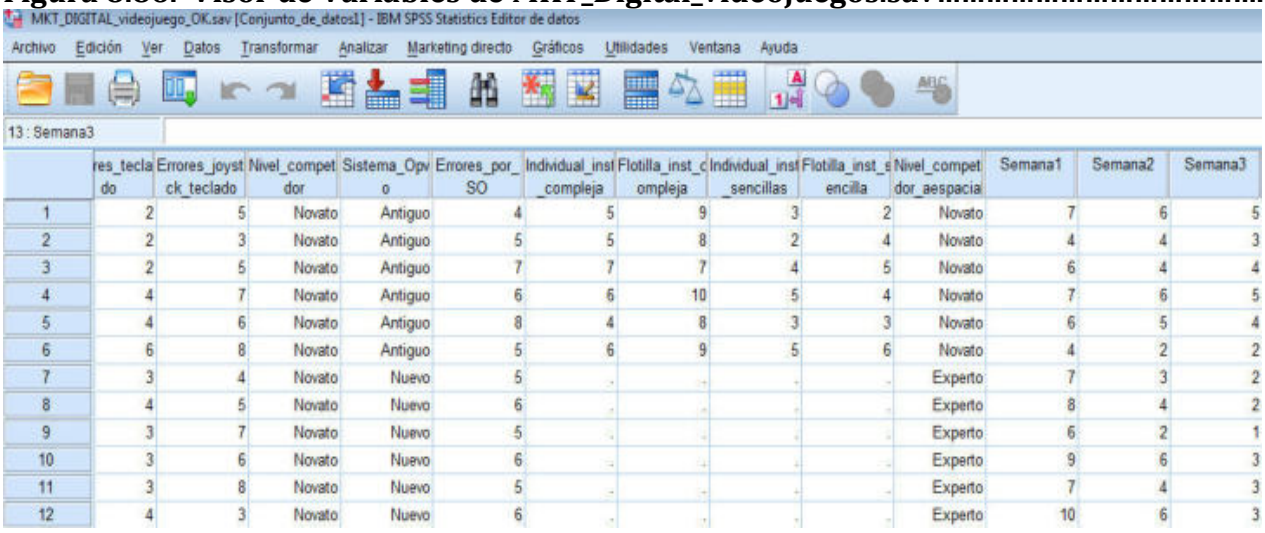
- Podemos observar del patrón de gráficos de interacción, previamente discutido.
- **Los jugadores hicieron más errores en con instrucciones complejas que con las sencillas por lo que tiene una interacción significativa en el efecto identificado.**
- Es posible llevar a cabo más pruebas para verificar si somos capaces de interpretar los patrones emergentes en nuestro análisis.
- Así, Se acepta H_0: **los jugadores producen más errores en la ambientación espacial, con instrucciones complejas y en variante de combate flotilla, que con el resto de combinaciones. Sin embargo, se tiene menos confianza acerca de las diferencias entre la variante de combate cuando las instrucciones son sencillas**
- Una prueba de dichos efectos principales sencillos puede ser calculado a través del comando Syntax ubicado dentro de **SPSS.**

8.19. ANOVA de dos factores por diseño combinado. Ejemplos

Este diseño es emprendido cuando se tienen mediciones independientes en uno de nuestros factores y mediciones repetidas en el segundo de nuestros factores. Por ejemplo, un investigador Un investigador está interesado en la habilidad gerencial de recordar eventos de impacto a nivelo mundial y ha diseñado un cuestionario sobre los eventos de los últimos 20 años, con igual número de reactivos (a nivel dificultad) porcada uno de esos años. El investigador está también interesado en comparar a los gerentes jóvenes adultos (24 a 44 años de edad) con un grupo más experimentado (45 a 54 años de edad), con el mismo instrumento. Las **2** variables independientes (**factores**) son: **años**, con **20** condiciones, el cual es un f**actor de medición repetida** y **edad**, con **2** condiciones, el cual es un **factor de medición independiente**.
Con esto, habrá **2** efectos principales de **año** y **edad** así como una **interacción** entre ambos. Una **interacción** cuando el efecto de uno de los factores es diferente en las condiciones diferentes del segundo factor. En el ejemplo, si los gerentes maduros conocieron más sobre varios de los años que los gerentes jóvenes pero éstos conocieron más que otros años, entonces existiría una interacción.

Paso 1: Objetivos
Problema 5: la empresa **MKT_Digital** ha terminado de rediseñar y actualizar su videojuego en ambiente espacial y requiere verificar que los jugadores incrementen sus habilidades en el mismo. Se tiene el particular interés de comparar el desempeño de jugadores a nivel **experto** con los que recién se introduzcan en el mismo (**novatos**). Un investigador selecciona al azar a **6** jugadores de nivel **experto y 6 de nivel novato** para monitorear sus errores en el videojuego rediseñado y actualizado en un periodo de 3 semanas para observar si existen diferencias entre los 2 grupos.
H_0= Los jugadores expertos y novatos NO muestran diferencias en el incremento de habilidades del videojuego rediseñado y actualizado en ambientación espacial.
H_1= Los jugadores expertos y novatos SI muestran diferencias en el incremento de habilidades del videojuego rediseñado y actualizado en ambientación espacial.
Ver Figuras 8.86. y 8.87.

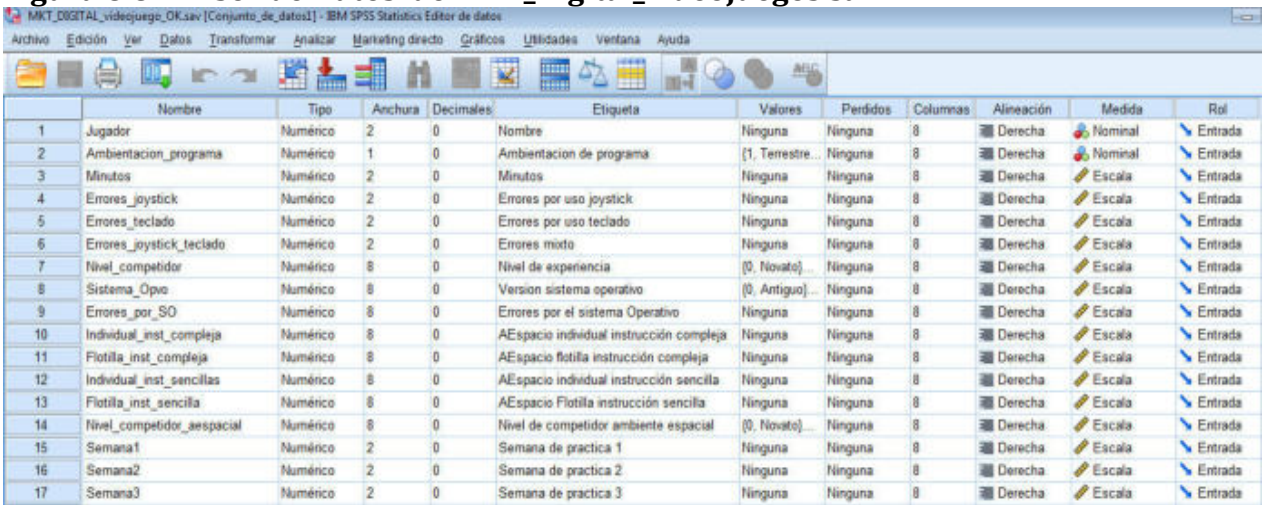
Figura 8.86. Visor de Variables de MKT_Digital_videojuegos.sav……



	res_tecla do	Errores_joyst ck_teclado	Nivel_competi dor	Sistema_Opv o	Errores_por_ SO	Individual_inst _compleja	Flotilla_inst_c ompleja	Individual_inst _sencillas	Flotilla_inst_e encilla	Nivel_competi dor_aespacia	Semana1	Semana2	Semana3
1	2	5	Novato	Antiguo	4	5	9	3	2	Novato	7	6	5
2	2	3	Novato	Antiguo	5	5	8	2	4	Novato	4	4	3
3	2	5	Novato	Antiguo	7	7	7	4	5	Novato	6	4	4
4	4	7	Novato	Antiguo	6	6	10	5	4	Novato	7	6	5
5	4	6	Novato	Antiguo	8	4	8	3	3	Novato	6	5	4
6	6	8	Novato	Antiguo	5	6	9	5	6	Novato	4	2	2
7	3	4	Novato	Nuevo	5	.	.	.	.	Experto	7	3	2
8	4	5	Novato	Nuevo	6	.	.	.	.	Experto	8	4	2
9	3	7	Novato	Nuevo	5	.	.	.	.	Experto	6	2	1
10	3	6	Novato	Nuevo	6	.	.	.	.	Experto	9	6	3
11	3	8	Novato	Nuevo	5	.	.	.	.	Experto	7	4	3
12	4	3	Novato	Nuevo	6	.	.	.	.	Experto	10	6	3

Fuente: SPSS 20 IBM

Figura 8.87. Visor de Datos de MKT_Digital_videojuegos.sav



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	Jugador	Numérico	2	0	Nombre	Ninguna	Ninguna	8	Derecha	Nominal	Entrada
2	Ambientacion_programa	Numérico	1	0	Ambientacion de programa	{1, Terrestre...	Ninguna	8	Derecha	Nominal	Entrada
3	Minutos	Numérico	2	0	Minutos	Ninguna	Ninguna	8	Derecha	Escala	Entrada
4	Errores_joystick	Numérico	2	0	Errores por uso joystick	Ninguna	Ninguna	8	Derecha	Escala	Entrada
5	Errores_teclado	Numérico	2	0	Errores por uso teclado	Ninguna	Ninguna	8	Derecha	Escala	Entrada
6	Errores_joystick_teclado	Numérico	2	0	Errores mixto	Ninguna	Ninguna	8	Derecha	Escala	Entrada
7	Nivel_competidor	Numérico	8	0	Nivel de experiencia	{0, Novato}...	Ninguna	8	Derecha	Escala	Entrada
8	Sistema_Opvo	Numérico	8	0	Version sistema operativo	{0, Antiguo}...	Ninguna	8	Derecha	Escala	Entrada
9	Errores_por_SO	Numérico	8	0	Errores por el sistema Operativo	Ninguna	Ninguna	8	Derecha	Escala	Entrada
10	Individual_inst_compleja	Numérico	8	0	AEspacio individual instrucción compleja	Ninguna	Ninguna	8	Derecha	Escala	Entrada
11	Flotilla_inst_compleja	Numérico	8	0	AEspacio flotilla instrucción compleja	Ninguna	Ninguna	8	Derecha	Escala	Entrada
12	Individual_inst_sencillas	Numérico	8	0	AEspacio individual instrucción sencilla	Ninguna	Ninguna	8	Derecha	Escala	Entrada
13	Flotilla_inst_sencilla	Numérico	8	0	AEspacio Flotilla instrucción sencilla	Ninguna	Ninguna	8	Derecha	Escala	Entrada
14	Nivel_competidor_aespacial	Numérico	8	0	Nivel de competidor ambiente espacial	{0, Novato}...	Ninguna	8	Derecha	Escala	Entrada
15	Semana1	Numérico	2	0	Semana de practica 1	Ninguna	Ninguna	8	Derecha	Escala	Entrada
16	Semana2	Numérico	2	0	Semana de practica 2	Ninguna	Ninguna	8	Derecha	Escala	Entrada
17	Semana3	Numérico	2	0	Semana de practica 3	Ninguna	Ninguna	8	Derecha	Escala	Entrada

Fuente: SPSS 20 IBM

Paso 2: Diseño

En adelante para efectos de aprendizaje, sólo se tomarán de forma directa los datos con el fin de aprender a utilizar la técnica.

Paso 3: Condiciones de Aplicabilidad

- No olvidar realizar los análisis previos de inspección visual de la base de datos para detectar ausentes y/o atípicos (corregir en su defecto), confiabilidad, normalidad, homocedasticidad y linealidad
- Para toda prueba paramétrica se debe suponer que:
  -Los datos provienen de una población al azar

-Las puntuaciones son medidas en escala de intervalo y provienen de poblaciones normalmente distribuidas

-Las muestras en cada condición provienen de poblaciones con homoscedasticidad

- Como hay factores de mediciones repetidas debemos también suponer que existe ***esfericidad***. Incluso, aunque podamos tener diferentes números de participantes en los factores de medición independientes (por ejemplo, más gerentes maduros que jóvenes) **es recomendable el tener igual número de participantes por razones de mantener la *esfericidad***

Paso 4: Estimación y Ajuste

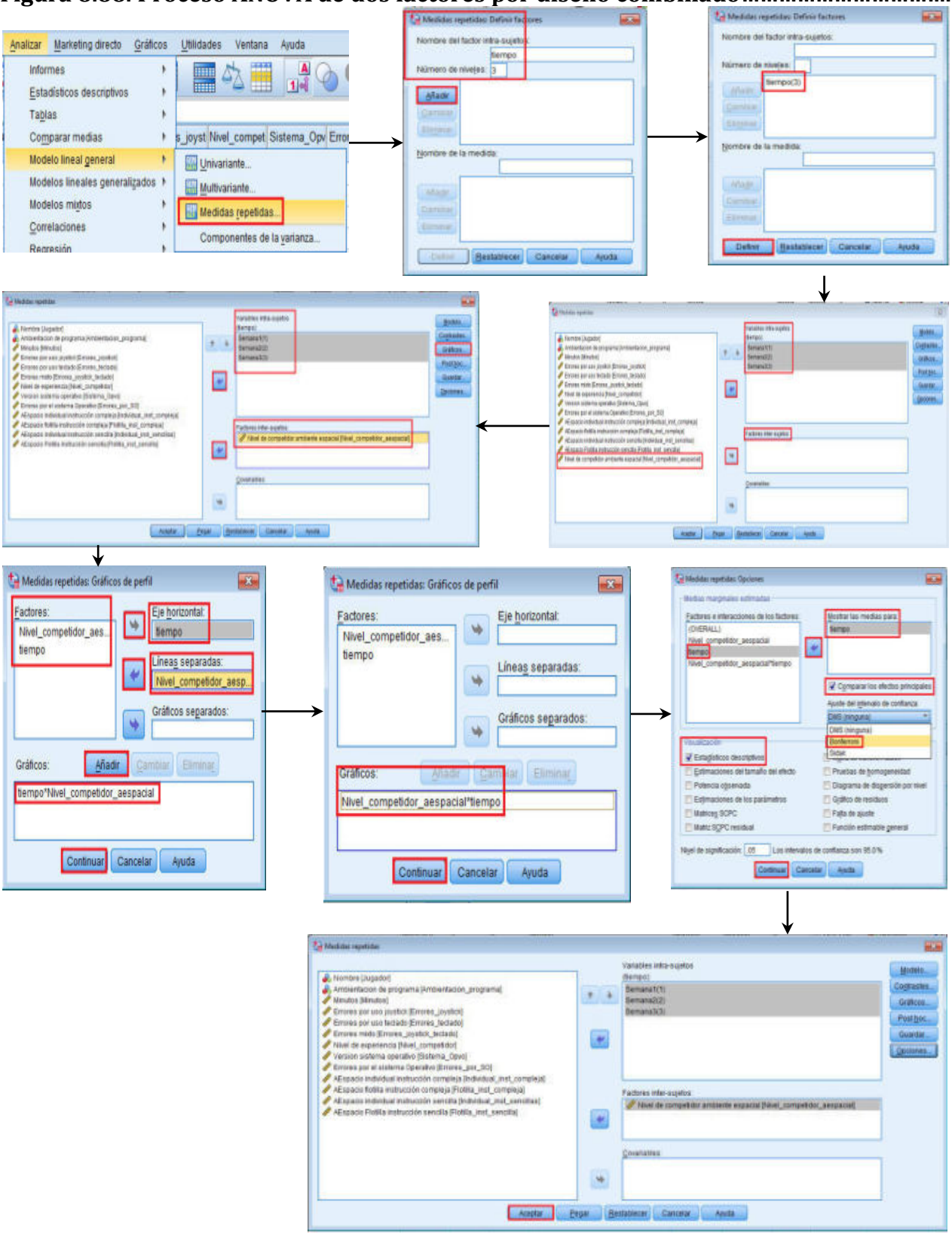
-Teclear: Analizar->Modo lineal general->Medidas repetidas-*>Nombre del factor intra-sujetos: tiempo->Número de niveles: 3->Añadir-> Definir->Variables intra-sujetos: Semana1(1); Semana2(2) ; Semana3(3) ->Factores inter-sujetos: Nivel de competidor ambiente espacial (Nivel_competidor_aespacial)->Gráficos->Factores->Eje horizontal: tiempo ->Líneas separadas: Nivel_competidor_aespacial->Añadir->Continuar->Opciones->Visualización: Estadísticos descriptivos-> Mostrar las medias para: tiempo->Comparar los efectos principales: Bonferroni-> Continuar->Aceptar ***. Ver Figura 8.88.**

* **.-**Todos los cálculos de **ANOVA** se hacen con este procedimiento. El tipo de **ANOVA** seleccionado dependerá del diseño del estudio. En este ejemplo tenemos una **ANOVA de dos factores por diseño combinado,** con un factor: **Nivel_competidor_aespacial,** con mediciones independientes y otro factor (subyacente) **tiempo,** con mediciones repetidas.
Así, se deberá enviar la variable en la que se realizarán las mediciones repetidas (**tiempo**) al cuadro de diálogo **Mostrar las medias para.**

****.-**Una prueba de comparación por parejas para los factores de medición repetidas necesita realizarse en cuanto se tienen **más 2 niveles.**

*****.-**Es posible hacer pruebas de homocedasticidad para nuestro factor de medidas repetidas al seleccionar el cuadro **Pruebas de homogeneidad.** En el ejemplo solamente necesitamos de una prueba ***post hoc*** para el factor de medidas repetidas: **tiempo.** Nuestras mediciones de la variable independiente **Nivel_competidor_aespacial,** solamente tiene **2 niveles** así que la prueba ***post hoc*** **NO** es necesaria Si nuestra variable independiente no fuera de 2 niveles se optaría por la prueba post hoc apropiada

Figura 8.88. Proceso ANOVA de dos factores por diseño combinado……



Fuente: SPSS 20 IBM

Paso 5: Interpretación

Las primeras 2 tablas que genera **SPSS** nos reportan una descripción de los 2 factores que ingresan al **ANOVA**. Esto confirma que las tablas **Factores intra-sujetos** (**con factor de mediciones repetidas**) tienen **3 niveles**, y la tabla **Factores inter-sujetos** (**factor de mediciones independientes**) tiene **2 niveles**. **Figura 8.89.**

Figura 8.89. Tablas Factores intra-sujeto y Factores inter-sujetos

Factores intra-sujetos

Medida: MEASURE_1

tiempo	Variable dependiente
1	Semana1
2	Semana2
3	Semana3

Factores inter-sujetos

		Etiqueta del valor	N
Nivel de competidor ambiente espacial	0	Novato	6
	1	Experto	6

Fuente: SPSS 20 IBM

La siguiente tabla que reporta los primeros resultados para analizar es la de Estadísticos descriptivos. **Figura 8.90.**

Figura 8.90. Tabla de Estadísticos descriptivos

Estadísticos descriptivos

	Nivel de competidor ambiente espacial	Media	Desviación típica	N
Semana de practica 1	Novato	5.67	1.366	6
	Experto	7.83	1.472	6
	Total	6.75	1.765	12
Semana de practica 2	Novato	4.50	1.517	6
	Experto	4.17	1.602	6
	Total	4.33	1.497	12
Semana de practica 3	Novato	3.83	1.169	6
	Experto	2.33	.816	6
	Total	3.08	1.240	12

Fuente: SPSS 20 IBM

- Esta tabla despliega como la **Media** el número de errores cometidos por los **2** grupos en **3 períodos de tiempo. La Media Total** es el global de los errores cometidos en la semana, sin tomar en cuenta de si los jugadores fueran expertos o novatos
- Al observar las medias se aprecia que hay 3 con un aparente patrón en el número de errores producidos por los 2 grupos a lo largo del tiempo. En la **Semana de práctica 1**

- los jugadores de **nivel Experto** hicieron más errores que los **Novatos** (**7.83** vs. **5.67**) pero la diferencia se redujo en la **Semana de práctica 2** (**4.17** vs. **4.50**). En la **Semana de práctica 3** final, la tendencia se revirtió, con los **novatos** ahora cometiendo más errores que los **Expertos** (**2.33** compare con **3.83**).
- El número total de errores se reduce de forma continua en el tiempo.
- La Desviación típica (estándar) muestra que la dispersión de los puntajes entre los grupos en cada período es similar. A través de las 3 semanas, el total de las desviaciones típicas (estándar) se reduce, de **1.765** en la **Semana de práctica 1 a 1.24** en la **Semana de práctica 3**. Esto nos indica que al final del periodo hay menos variación en el desempeño de los jugadores
- **N** representa el número de jugadores en cada grupo y periodo.

La siguiente tabla es la de **Contrastes multivariados**. Esta es generada por el **SPSS** en el proceso de cálculo la **ANOVA** de dos factores por diseño combinado y es usualmente consultada si el supuesto de ***esfericidad*** **es violado** Ver **Figura 8.91.**

Figura 8.91. Tabla de Contrastes multivariados

Contrastes multivariados[a]

Efecto		Valor	F	Gl de la hipótesis	Gl del error	Sig.
tiempo	Traza de Pillai	.968	135.928[b]	2.000	9.000	.000
	Lambda de Wilks	.032	135.928[b]	2.000	9.000	.000
	Traza de Hotelling	30.206	135.928[b]	2.000	9.000	.000
	Raíz mayor de Roy	30.206	135.928[b]	2.000	9.000	.000
tiempo * Nivel_competidor_aespacial	Traza de Pillai	.885	34.773[b]	2.000	9.000	.000
	Lambda de Wilks	.115	34.773[b]	2.000	9.000	.000
	Traza de Hotelling	7.727	34.773[b]	2.000	9.000	.000
	Raíz mayor de Roy	7.727	34.773[b]	2.000	9.000	.000

a. Diseño: Intersección + Nivel_competidor_aespacial
Diseño intra-sujetos: tiempo

b. Estadístico exacto

Fuente: SPSS 20 IBM

- **SPSS** genera una prueba multivariada para el factor con mediciones repetidas y su interacción.
- Podemos escoger los resultados de la prueba de **Contrastes multivariados** si estamos en el cometido de que los supuestos de las pruebas univariadas **NO** son encontradas. Por ejemplo si la **prueba de *esfericidad de Mauchly* es significativa o el valor de Epsilon es bajo.**
- La prueba más popular, ***Lambda de Wilks***. Desde esta prueba podemos observar que existe un efecto principal para nuestro factor de medidas repetidas **tiempo**, asi que:

F(2,9) = 135.928; p =.000< 0.01.

- Existe también un efecto principal significante por la **interacción de nuestros 2 factores,** así que**:**

F(2,9) = 34.773; p =.000< 0.01.

La siguiente tabla muestra las pruebas de la ***esfericidad***. Esta tabla es generada por **SPSS** durante el proceso de cálculo de **ANOVA** de dos factores por diseño combinada.Ver **Figura 8.92.**

Figura 8.92. Tabla de Prueba de *Esfericidad* de *Mauchly*

Prueba de esfericidad de Mauchly[a]

Medida: MEASURE_1

Efecto intra-sujetos	W de Mauchly	Chi-cuadrado aprox.	gl	Sig.	Epsilon[b]		
					Greenhouse-Geisser	Huynh-Feldt	Límite-inferior
tiempo	.937	.584	2	.747	.941	1.000	.500

Contrasta la hipótesis nula de que la matriz de covarianza error de las variables dependientes transformadas es proporcional a una matriz identidad.

a. Diseño: Intersección + Nivel_competidor_aespacial
Diseño intra-sujetos: tiempo

b. Puede usarse para corregir los grados de libertad en las pruebas de significación promediadas. Las pruebas corregidas se muestran en la tabla Pruebas de los efectos inter-sujetos.

Fuente: SPSS 20 IBM

- Cuando existen **más de 2 condiciones** de la variable de mediciones repetidas podemos revisar el supuesto de *esfericidad* antes de calcular los valores del **estadístico *F***.
- Si el supuesto de ***esfericidad*** **SÍ** se encuentra, entonces procedemos a inspeccionar la línea de ***Esfericidad* asumida de la tabla de Prueba de efectos intra-sujetos** (**Modelo Univariado**).
- Si el supuesto de ***esfericidad*** **NO** se encuentra y la **Prueba de *Esfericidad* de *Mauchly*** es **significativa NO** tomamos la **línea de *Esfericidad* asumida de la tabla de Prueba de efectos intra-sujetos (Modelo Univariado)** y requeriremos realizar una **corrección**. **SPSS** tiene varios modelos de corrección, el más utilizado es el de ***Greehouse–Geisser.***
- **La Prueba de *Esfericidad* de *Mauchly*** del ejemplo, reporta una prueba estadística **W de *Mauchly*** = **0.937**, **df = 2; p > 0.05.** Por lo tanto, podemos concluir que el supuesto de ***esfericidad*** se ha encontrado y que podemos utilizar el **Modelo Univariado SIN corrección.**
- Existe sin embargo cierto debate en cuanto a la sensibilidad de la **prueba de *Mauchly*** en su habilidad de detectar la *esfericidad*. Por lo tanto, la alternativa de consultar el **valor de Epsilon** en la columna de ***Greenhouse-Geisser***. Esta figura debe ser lo más cercana a **1.00** como sea posible a fin de conseguir que no existan problemas de *esfericidad*. Nuestro valor es de **0.941** as que podemos confiar claramente en que la ***esfericidad* NO** afecta nuestros ca lculos. Si el supuesto de la *esfericidad* se encuentra, como en el ejemplo, podemos tomar los valores de los **renglones de la Esfercidad asumida** de las **Pruebas de Efectos intra-sujetos**. Ver **Figura 8.93.** Esta tabla nos permite decidir si hemos encontrado un efecto principal significativo para nuestra medida repetida **'tiempo'**, y una **interacción significaiva entre** los factores: **tiempo vs. Nivel_competidor_aespacial.**

Figura 8.93. Tabla de Pruebas de Efectos intra-sujetos

Pruebas de efectos intra-sujetos.

Medida: MEASURE_1

Origen		Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
tiempo	Esfericidad asumida	83.389	2	41.694	150.100	.000
	Greenhouse-Geisser	83.389	1.882	44.313	150.100	.000
	Huynh-Feldt	83.389	2.000	41.694	150.100	.000
	Límite-inferior	83.389	1.000	83.389	150.100	.000
tiempo * Nivel_competidor_aespacial	Esfericidad asumida	21.056	2	10.528	37.900	.000
	Greenhouse-Geisser	21.056	1.882	11.189	37.900	.000
	Huynh-Feldt	21.056	2.000	10.528	37.900	.000
	Límite-inferior	21.056	1.000	21.056	37.900	.000
Error(tiempo)	Esfericidad asumida	5.556	20	.278		
	Greenhouse-Geisser	5.556	18.818	.295		
	Huynh-Feldt	5.556	20.000	.278		
	Límite-inferior	5.556	10.000	.556		

Fuente: SPSS 20 IBM

Fuente: SPSS 20 IBM

- Esta tabla nos habilita para decidir si hemos encotrado un efecto principal de nuestro factor de medidas repetidas **tiempo**, y una **interacción** significativa entre nuestros 2 factores: **tiempo** vs. **Nivel_competidor_aespacial.**
- Los datos más importantes de la tabla, han sido enmarcados y son del renglón ***Esfericidad* asumida.**
- Recuerde que debido a que es una **ANOVA de dos factores** el factor de mediciones repetidas y la **información de la interacción aparecen en esta tabla**. Las cifras para posibles efectos principalesde nuestra variable independiente se mostrarán en la tabla de **Pruebas de los efectos inter-sujetos.**
- El efecto principal detectado para la variable **tiempo, es:** ***F(2,20) = 150.100; p =.000< 0.001.***
- Recuerde que si el **SPSS es**tablece que el resultado es **0.000**, significa que **SPSS** ha redondeado la cantidad lo más cercano al número en **3 lugares decimales**. Sin embargo, **podemos afirmar que redondea el último 0 a 1**, así que ***p* < 0.001.**
- Como ***p* < 0.001**, esto indica que hemos encontrado un efecto principal significativo para nuestras mediciones repetidas en el factor **tiempo**. Sin embargo, no conocemos dónde se ubican las diferencias, y por lo tanto debemos consultar los resultados de la prueba ***post hoc*** para obtener esta información.
- También, hemos encontrado una **interacción significativa** entre los 2 factores, así que:

  F(2,20) = 37.900, p =.000< 0.001
- Podemos tomar con confianza las cifras del modelo de ***Esfericidad* asumida** debido a que al observar la **Prueba de *Esfericidad* de *Mauchly*** fue encontrado que **NO es significativo**, y que el **valor de Epsilon=0.941**. Si esto no es el caso entonces se deberá aplicar uno de los modelos de corrección sugeridos **SPSS** tiene varios modelos de corrección como el ***Greehouse–Geisser*,** usualmente utilizado.

- La **Suma de cuadrados** nos entrega una medición de la variabilidad de los puntajes debido a una fuente en particular de variabilidad. La **Media cuadrática** es la varianza (**Suma de los cuadrados** dividida por los grados de libertad). **Observe que existe una gran variabilidad debido a nuestros factores y mucho menos debido al error indicando un gran efecto.**
- Las **pruebas de Contrastes intra-sujetos** y es generada por **SPSS** durante el proceso de cálculo de una **ANOVA** de dos factores por diseño combinado y las mediciones repetidas y es una tendencia de análisis. Ver Figura **8.94.**

Figura 8.94. Tabla de Pruebas de Contrastes intra-sujetos

Pruebas de contrastes intra-sujetos

Medida: MEASURE_1

Origen	tiempo	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
tiempo	Lineal	80.667	1	80.667	254.737	.000
	Cuadrático	2.722	1	2.722	11.395	.007
tiempo * Nivel_competidor_aespacial	Lineal	20.167	1	20.167	63.684	.000
	Cuadrático	.889	1	.889	3.721	.083
Error(tiempo)	Lineal	3.167	10	.317		
	Cuadrático	2.389	10	.239		

Fuente: SPSS 20 IBM

- Esta tabla analiza las tendencias desplegadas por los datos. Nos reporta información en cuanto al **modelo subyacente que mejor ajusta los datos.**
- Como tenemos **3 periodos**, las **2 posibles tendencias** son aquellas que siguen una tendencia **lineal o cuadrática.**
- En nuestro ejemplo podemos observar que **hemos encontrado una tendencia lineal** en nuestros datos del factor de mediciones repetidas **tiempo**, así que: ***F(1,10) = 254.737; p=.000 < 0.001***. Sin embargo, nuestros datos también evidencian una tendencia cuadrática: ***F(1,10) = 11.395; p=.007 < 0.01*****.**
- La interacción entre los factores tiene un modelo subyacente linealmente significativo: ***F(1,10) = 63.684; p=.000 < 0.001*****.** La tendencia cuadrática **NO** es significaiva: ***F(1,10) = 3.721; p > 0.05*****.**
- La tabla de **Pruebas de los efectos inter-sujetos** nos habilita para decidir si tenemos un efecto principal significativo para nuestro factor de medidas independientes **Nivel_competidor_aespacial** (novatos/expertos).Ver Figura **8.95.**

Figura 8.95. Tabla de Pruebas de Contrastes intra-sujetos

Pruebas de los efectos inter-sujetos

Medida: MEASURE_1
Variable transformada: Promedio

Origen	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
Intersección	802.778	1	802.778	163.462	.000
Nivel_competidor_aespacial	.111	1	.111	.023	.883
Error	49.111	10	4.911		

Fuente: SPSS 20 IBM

- El efecto de nuestra variable de medición independiente **Nivel_competidor_aespacial NO** es estadísticamente significativa: ***F*(1,10) = 0.023, *p* = .883> 0.05.**
- Como ***p*>0.05** podemos concluir que NO hemos encontrado un efecto principal significativo para **Nivel_competidor_aespacial**, esto es, que NO hay diferencias significativas globales entre los jugadores **Novatos** vs **Expertos** sobre el número de errores que cometieron en las 3semanas.
- El renglón de Intersección muestra nuestra **media global es significativamente diferente de cero.**
- La siguiente tabla, la de **Estimaciones** apoya a la Estadística descriptiva**.** Ver Figura **8.96.**

Figura 8.96. Tabla de Estimaciones

Estimaciones

Medida: MEASURE_1

tiempo	Media	Error típ.	Intervalo de confianza 95%	
			Límite inferior	Límite superior
1	6.750	.410	5.837	7.663
2	4.333	.450	3.330	5.337
3	3.083	.291	2.435	3.732

Fuente: SPSS 20 IBM

La **Media** indica el número medio de errores producidos en 3 semanas.

El **Error típico nos reporta** un estimado de la **desviación estándar** de la **distribución de muestreo de la media**. Esta es una cifra útil cuanto si es usada en el cómputo de las pruebas de comparación de significancia sde la media y en los cálculos de los intervalos de confianza.

- El **intervalo de confianza 95%** de la diferencia nos reporta un estimado de la media de la población. Por ejemplo en el tiempo 1 (**Semana de práctica 1)** estamos al 95% de confianza de que la media de la población se ubica entre **5.837 y 7.663.**
- La tabla de **Comparaciones por pares** nos entrega una comparación de las medias para todas las combinaciones pareadas de los niveles de nuestro factor de mediciones repetidas (**tiempo**). Todas las comparaciones son ajustada usando el método de ***Bonferroni***. Esta tabla debe ser inspeccionada para asegurar donde, las diferencias significativas fueron evidentes del cálculo de nuestro efecto principal para el factor **tiempo** se encuentra. Ver Figura **8.97.**

Figura 8.97. Tabla de Comparaciones por pares

Comparaciones por pares

Medida: MEASURE_1

(I)tiempo	(J)tiempo	Diferencia de medias (I-J)	Error típ.	Sig.[b]	Intervalo de confianza al 95 % para la diferencia[b]	
					Límite inferior	Límite superior
1	2	2.417*	.186	.000	2.001	2.832
	3	3.667*	.230	.000	3.155	4.179
2	1	-2.417*	.186	.000	-2.832	-2.001
	3	1.250*	.227	.000	.745	1.755
3	1	-3.667*	.230	.000	-4.179	-3.155
	2	-1.250*	.227	.000	-1.755	-.745

Basadas en las medias marginales estimadas.

*. La diferencia de medias es significativa al nivel .05.

b. Ajuste para comparaciones múltiples: Diferencia menos significativa (equivalente a la ausencia de ajuste).

Fuente: SPSS 20 IBM

- Lo importante de la tabla se ha encuadrado
- La tabla muestra todas las posibles comparaciones para los 3 niveles de nuestra variable de mediciones repetidas
- En cada comparación un nivel es dado con el identificador 'I' y al segundo 'J'. Es evidente en la **columna de Diferencia de medias** la cual indica la cifra resultado de cuando la media de un nivel de la variable (J) ha sido sustraída del segundo nivel (I).
- En el ejemplo, la **media global** de nuestro primer nivel (**Semana de práctica 1)** =**6.750** en nuestra estad stica descriptiva y la media del **nivel 2** (**Semana de práctica 2**) = **4.333**. Por lo tanto: ***6.750 (I) – 4.333 (J) = 2.417***
- La columna **Sig.** nos permite evaluar **si las diferencias de las medias** entre los niveles de las variables son significativas Se puede ver , del ejemplo, que todas las posibles comparaciones por pares son significativas , en cuanto a los valores ***p*<0.01**.
- Los valores de **Error típ.** Son pequeños, lo cual da indicios de baja variabilidad en las diferencias de las medias proyectadas
- El **Intervalo de confianza al 95%** para las diferencias nos reporta un estimado de la diferencia de las medias de la población. Por ejemplo, estamos al 95% de confianza que la diferencia de las medias entre **tiempo 1 y tiempo 2 se ubica en tre 2.001 and 2.832**

- Como se produjeron comparaciones de pares a través del método de ***Bonferroni,*** la **ANOVA** también produce la siguiente tabla de pruebas de **Contrastes multivariados también como la de *Bonferroni*** Esta tabla no es de interés como hemos seguido a la del **método Univariado** de análisis, teniendo primero lo sugerido a revisar con ***Mauchly*** y **Epsilon**.Ver Figura **8.98**

Figura 8.98. Tabla de Comparaciones por pares

Contrastes multivariados

	Valor	F	Gl de la hipótesis	Gl del error	Sig.
Traza de Pillai	.968	135.928[a]	2.000	9.000	.000
Lambda de Wilks	.032	135.928[a]	2.000	9.000	.000
Traza de Hotelling	30.206	135.928[a]	2.000	9.000	.000
Raíz mayor de Roy	30.206	135.928[a]	2.000	9.000	.000

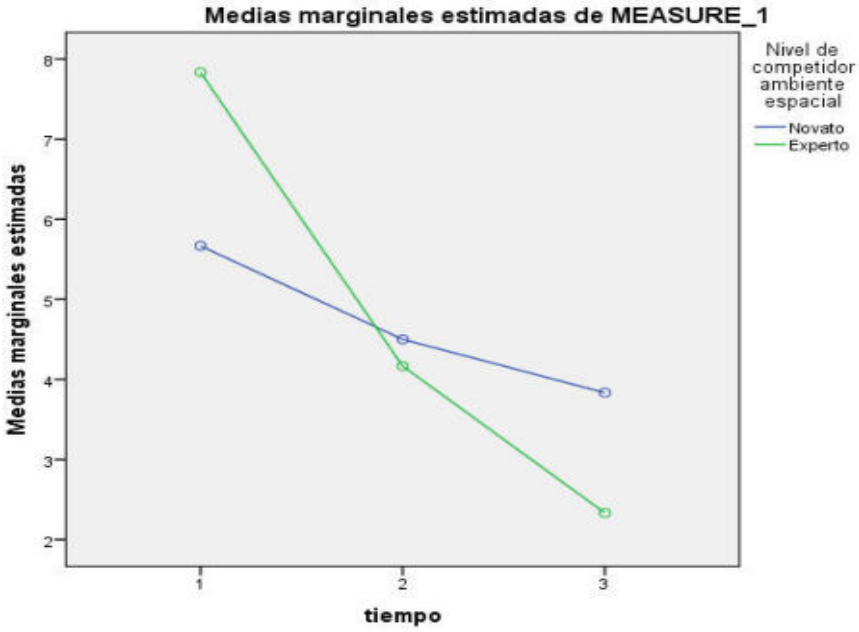
Cada prueba F contrasta el efecto multivariado de tiempo. Estos contrastes se basan en las comparaciones por pares, linealmente independientes, entre las medias marginales estimadas.

a. Estadístico exacto

Fuente: SPSS 20 IBM

La última parte que emite **SPSS** es el gráfico de interacción, el cual nos permite analizar los patrones previamente discutidos.Ver Figura **8.99**

Figura 8.99. Gráfico de interacción



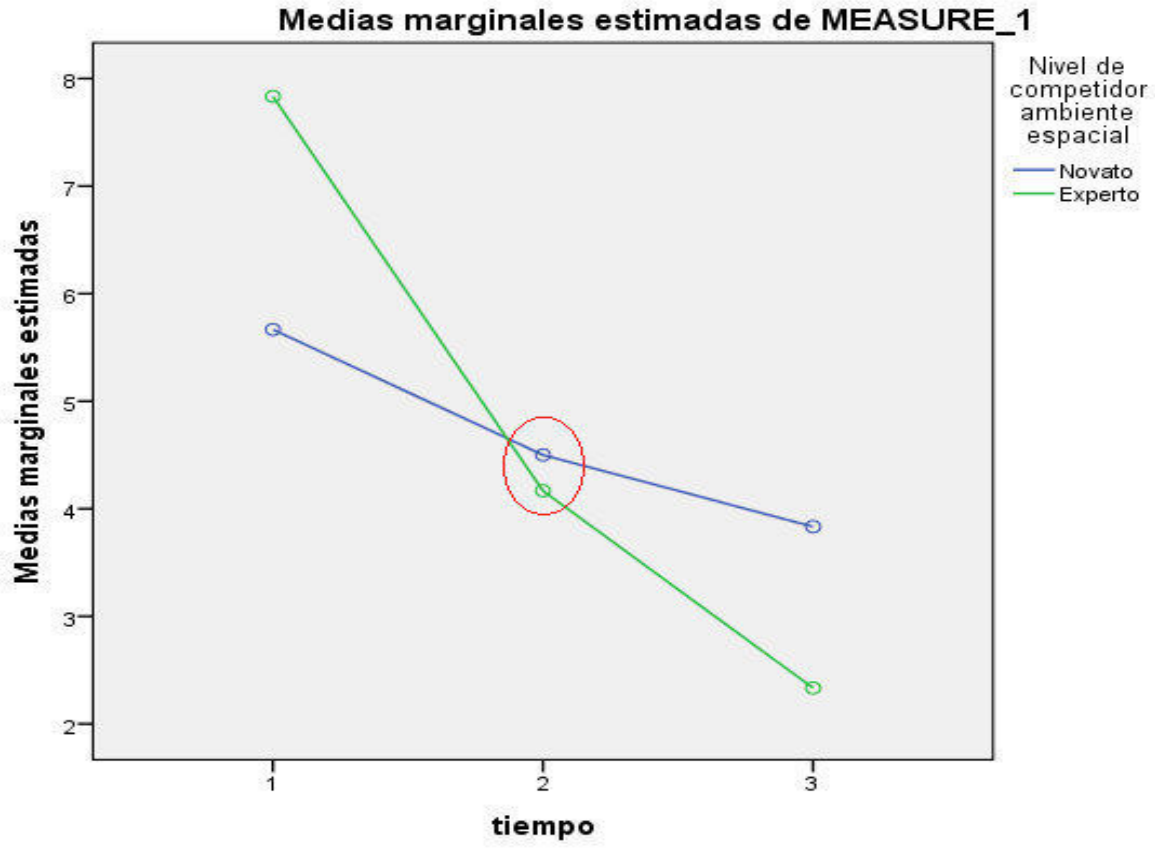
Fuente: SPSS 20 IBM

- El gráfico muestra los efectos principales y las interacciones que han sido identificadas como parte del **ANOVA.**
- El efecto principal significativo de nuestra variable de mediciones repetidas puede ser visto al analizar las tendencias a través de las **3 semanas**. Puede ser apreciado del gráfico en que ambos grupos de jugadores hicieron pocos errores sobre el periodo de pruebas.
- El gráfico indica que hubieron algunas diferencias entre nuestros grupos de jugadores, particularmente en el **tiempo 1** y **3**. Sin embargo, encontramos un NO-significativo efecto principal entre el grupo de jugadores. **Por lo tanto, NO existe evidencia de diferencias significativas entre los 2 grupos de jugadores a nivel global cuando usan cuando usan el videojuego rediselado y actualizado** .
- Se descubre una interacción significativa entre nuestras variables , que se puede observar directamente en el gráfico de interacción:
- En la **semana 1**los jugadores **Expertos** estuvieron haciendo más errores que los **Novatos**. Sin embargo, esta tendencia fue revertida en la **semana 3** cuando los jugadores **Expertos** hicieron menos errores que los **Novatos**.
- Los gráficos generados en esta forma pueden requerir editarse ya que la técnica de etiquetado no siempre es clara.
- Podríamos llevar más allá las pruebas de análisis de datos en la generación y comprensión de los modelos gráficos resultantes. Por ejemplo, en la explicación del porqué los jugadores **Expertos** inicialmente hicieron más errores que los **Novatos** y el porqué se revierte con el tiempo
- Una prueba para tales efectos principales puede por lo tanto ser calculada a través de Sintaxis de comando con **SPSS** para mostrar en la siguiente sección.

8.20. ANOVA de dos factores por diseño combinado con efecto simple principal. Resumen

Paso 1: Objetivo: Paso 2: Diseño; Paso 3: Condiciones de aplicabilidad Problema 6:Siguiendo el problema de la sección **8.16**, en algunos casos cuando realizamos **ANOVA de dos factores** necesitaremos evaluar los efectos simples al igual que los efectos principales por factor. Esto es, analizar el efecto de un factor por ejemplo: **Nivel de competidor ambiente espacial**, en cada nivel de nuestro segundo factor, **tiempo**, separadamente. Este proceso avanzado es posible **usando la función de sintaxis the SPSS**. El gráfico de interacción de la **Figura 8.69** al observarlo, podemos decir, con un razonado grado de confiabilidad que **existe una diferencia significativa entre nuestros 2 grupos** de jugadores en la **semana 1** y en la **semana 3**, con un cambio de dirección entre estos 2 tiempos. Sin embargo, podríamos NO estar seguros, en cuanto al efecto de los jugadores en la semana 2.
Es en este caso en el que podemos llevar a cabo un análisis estadístico adicional examinando el **efecto simple principal** en ese periodo, el cual es circulado en el gráfico de la **Figura 8.100**.

Figura 8.100. Gráfico de interacción



Fuente: SPSS 20 IBM

Paso 4: Estimación y ajuste

Las pruebas de los efectos simples principales son logradas a través de la opción de Sintaxis ubicada dentro de los procedimientos de **ANOVA** de dos factores. El siguiente ejemplo, se realizará basado en la **ANOVA de dos factores por diseño combinados,** aunque el procedimiento para calcular los **efectos simples principales** para ambas **ANOVAS** de dos factores de mediciones repetidas y de dos factores independientes, sea similar. Seguiremos el procedimiento de **ANOVA** de dos factores por diseño combinado, hasta ahora descrito. Es importante que antes de llevar a cabo este comando, tanto los **factores** como la **interacción** hayan sido enviados al cuadro de **Medias,** ubicado dentro de la pantalla de **Opciones.**

Si se falla en este supuesto, provocará que el cálculo de su efecto simple principal más difícil, en tanto más comando necesite incorporar manualmente. Así, cuando se tenga el comando de **Opciones** , realice:

Teclear: Analizar->Modo lineal general->Medidas repetidas->Nombre del factor intra-sujetos: tiempo->Número de niveles: 3->Añadir-> Definir->Variables intra-sujetos: Semana1(1); Semana2(2) ; Semana3(3) ->Factores inter-sujetos: Nivel de competidor ambiente espacial (Nivel_competidor_aespacial)->Gráficos->Factores->Eje horizontal: tiempo ->Lineas separadas: Nivel_competidor_aespacial->Añadir-

>Continuar->Opciones->Visualización: Estadísticos descriptivos-> Mostrar las medias para: Nivel_competidor_aespacial; tiempo; Nivel_competidor_aespacial * tiempo;->Comparar los efectos principales: Bonferroni-> Continuar->Pegar*->En la 7ª. Línea, complementar la instrucción con: COMPARE (Nivel_competidor_aespacial) ADJ (BONFERRONI)** ->Ejecutar***->Hasta el final**

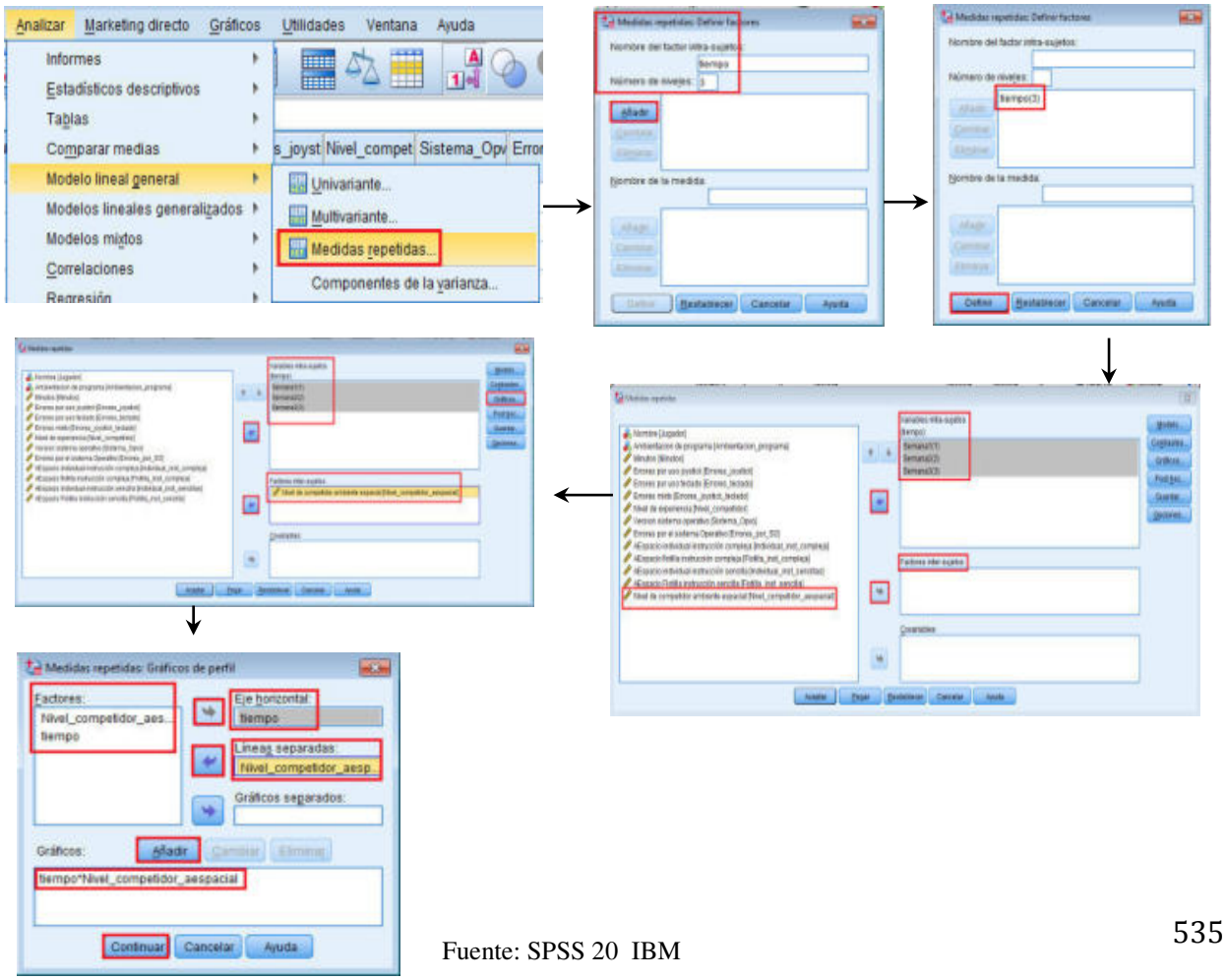
.- Aparece el cuadro de Syntaxis .** Este se presenta con líneas de sintaxis que confirman el análisis que se da a lugar; como se podido ver, los efectos principales de los 2 factores están siendo comparados por medio de la **prueba de *Bonferroni*.** Así, se necesita cambiar la syntaxis para asegrurarnos que, en adición a esto, la prueba **de *Bonferroni se está llevando a cabo en la intreracción de:

/EMMEANS=TABLES(Nivel_competidor_aespacial*tiempo). Con esto se asegura que los efectos simples principales sean producidos.

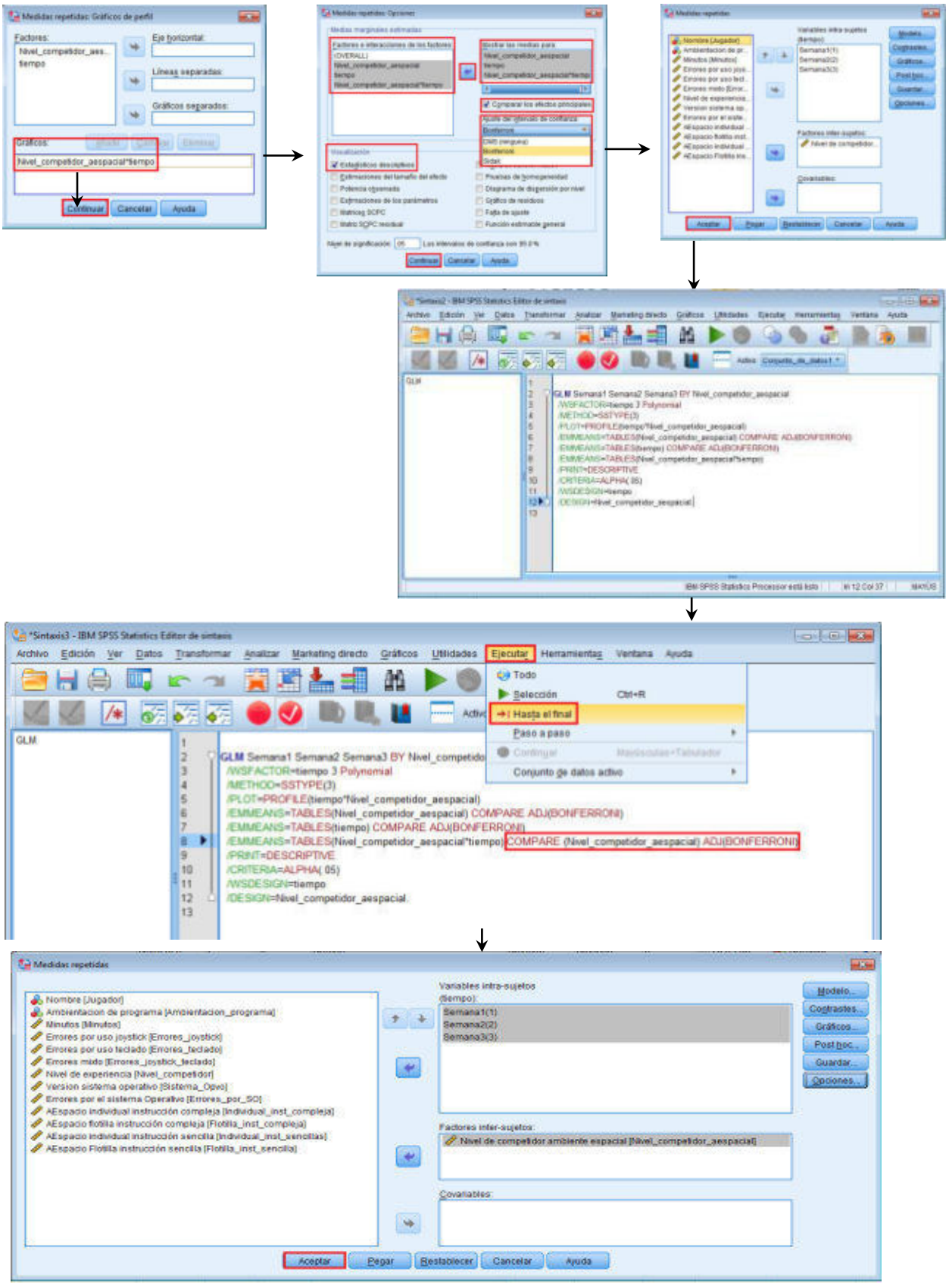
****.-** Deseamos comparar los 2 grupos de jugadores dentro de cada medición de tiempo, por lo tanto, escogemos la variable **Nivel_competidor_aespacial** para dl comando **COMPARE**.

*****.-** Seleccione **Ejecutar** del menú emergente y seleccione **Hasta el final**. Así, el **ANOVA** de dos factores por diseño combinado correrá en la misma forma que el anterior, pero esta vez., generará tablas adicionales para evaluación de la significancia de los efectos simples principales. **Aceptar Ver Figura 8.101.**

Figura 8.101. Proceso ANOVA de dos factores por diseño combinado con efecto simple principal.



Fuente: SPSS 20 IBM



Fuente: SPSS 20 IBM

Paso 5: Interpretación

Los productos de este proceso serán idénticos a los discutidos previamente con la excepción de **2 tablas adicionales: una tabla a mayor detalle de Comparaciones por pares y una tabla de Contrastes univariada**. Estas tablas serán encontradas siguiendo el orden original en las que aparecen sin cambios previos en la syntaxis. Como se observa, la tabla de Comparaciones por pares se genera dentro de cada tiempo de medición entre los 2 grupos de jugadores. Debe leerse de una forma similar a la tabla original. **Ver Figura 8.102.**

Figura 8.102. Tabla de Comparaciones por pares con Syntaxis modificada

Comparaciones por pares

Medida: MEASURE_1

tiempo	(I)Nivel de competidor ambiente espacial	(J)Nivel de competidor ambiente espacial	Diferencia de medias (I-J)	Error típ.	Sig.[b]	Intervalo de confianza al 95 % para la diferencia[b]	
						Límite inferior	Límite superior
1	Novato	Experto	-2.167*	.820	.025	-3.993	-.340
	Experto	Novato	2.167*	.820	.025	.340	3.993
2	Novato	Experto	.333	.901	.719	-1.673	2.340
	Experto	Novato	-.333	.901	.719	-2.340	1.673
3	Novato	Experto	1.500*	.582	.028	.203	2.797
	Experto	Novato	-1.500*	.582	.028	-2.797	-.203

Basadas en las medias marginales estimadas.

*. La diferencia de medias es significativa al nivel .05.

b. Ajuste para comparaciones múltiples: Bonferroni.

Fuente: SPSS 20 IBM

- Tomando el primer bloque de la tabla, **tiempo 1**, se observa que los 2 grupos de jugadores están llevando a cabo el juego de manera diferente, y de la columna **Sig.** columna se aprecia que **la diferencia es significativa** ya que ***p*=0.025 < 0.05**). Así, los jugadores **Expertos** están produciendo más errores en la **semana 1**.
- En la columna de **tiempo 3** las diferencias son también significativas, con ***p*= 0.028< 0.05.** Los jugadores **Novatos** están haciendo más errores en la **semana 3.**
- El periodo de medición en el que estamos enfocados es el de la **semana 2.** Podemos observar que en este periodo, la diferencia entre los 2 grupos **NO es significativa** ya que ***p*=0.719 > 0.05.**
- **Con lo anterior, podemos concluir, que NO existe un efecto simple principal en la semana 2**
  - La segunda tabla generada es la que se muestra en la **Figura 8.103.**

Figura 8.103. Tabla de Contrastes univariada por Syntaxis modificada

Contrastes univariados

Medida: MEASURE_1

tiempo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Contraste	14.083	1	14.083	6.983	.025
	Error	20.167	10	2.017		
2	Contraste	.333	1	.333	.137	.719
	Error	24.333	10	2.433		
3	Contraste	6.750	1	6.750	6.639	.028
	Error	10.167	10	1.017		

Cada prueba F contrasta el efecto de Nivel de competidor ambiente espacial. Estos contrastes se basan en las comparaciones por pares, linealmente independientes, entre las medias marginales estimadas. Estas pruebas se basan en las comparaciones por pares linealmente independientes entre las medias marginales estimadas.

Fuente: SPSS 20 IBM

- Esto produce una tabla de ANOVA que muestra el efecto de los contrastes de los que estamos interesados y nos indica si existe algún efecto dentro del mismo periodo que estamos estudiando.
- Podemos observar que las diferencias significativas entre los jugadores, ocurren entre la **semana 1** y la **semana 3**, con los detalles mostrados en la tabla de la **Figura 8.103.**

8.21. MANOVA.Resumen

La **ANOVA** es típicamente referida como una prueba univariada en cuanto a que el análisis involucra sólo una **variable dependiente**, a pesar de permitir más de una variable independiente. Sin embargo, existen casos en los que requerimos analizar datos **con más de una variable dependiente**. Es así así, que en este caso llevamos acabo el análisis multivariado de varianza (**MANOVA**). Por ejemplo, podemos tener datos recolectados sobre la satisfacción de servicios de acceso a internet en cuanto a la relación de precio y valor recibidos por los clientes de una compañia de telecomunicaciones. Podríamos realizar análisis por grupo de edades de usuarios (**variable independiente**) tanto en satisfacción en usuarios residenciales como empresariales **(2 variables dependientes**). Más que realizar 2 mediciones de un factor independiente de **ANOVA** separadas, es posible realizar una sola **MANOVA** en los datos. El punto clave es observar que **MANOVA** ba sicamente analiza el efecto de las variables independientes sobre la variable dependiente. En nuestro ejemplo, **MANOVA** nos reportará si hay un efecto de la edad sobre la combinación de variables dependientes de satisfacción: residencial y empresarial.

8.22. MANOVA de mediciones independientes. Ejemplos

La **MANOVA** de mediciones independientes es aquella en la que se realizan mediciones independientes sobre las variables independientes. Por ejemplo, podemos realizar análisis de la diferencia entre los gerentes de las tecnologías de información de una gran compañía de telecomunicaciones, como grupos de liderazgo que lo practican a nivel: transformacional

y transaccional, para un gran número de tareas como :toma de decisiones, productividad, procesos de innovación que generan, involucrando personal tanto interno como externo, así como él mismo. En este caso, el liderazgo transformacional-transaccional es una variable de medición independiente. La hipótesis nula: Existe un efecto de liderazgo transformacional-transaccional en las tareas de toma de decisiones entre los gerentes de las tecnologías de información de una gran compañía de telecomunicaciones, (**como una variable dependiente compuesta**). Usualmente la **ANOVA univariada** reportará suficiente informaci ón para mostrar que grado las variables dependientes están contribuyendo a una **MANOVA** significante. Así que podríamos analizar el efecto del liderazgo transformacional-transaccional de cada una de las tareas de toma de decisiones por separado. Sin embargo, si Usted estuviera interesado en las relaciones subyacentes entre las variables dependientes en combinacio n, con respecto a una variable independiente, entonces deber a llevarse a cabo un análisis de función discriminante para investigar dichas relaciones. En este caso los análisis deberían generar una función de las variables dependientes que serian capaces de clasificar a un gerente ya sea transformacional o transaccional.

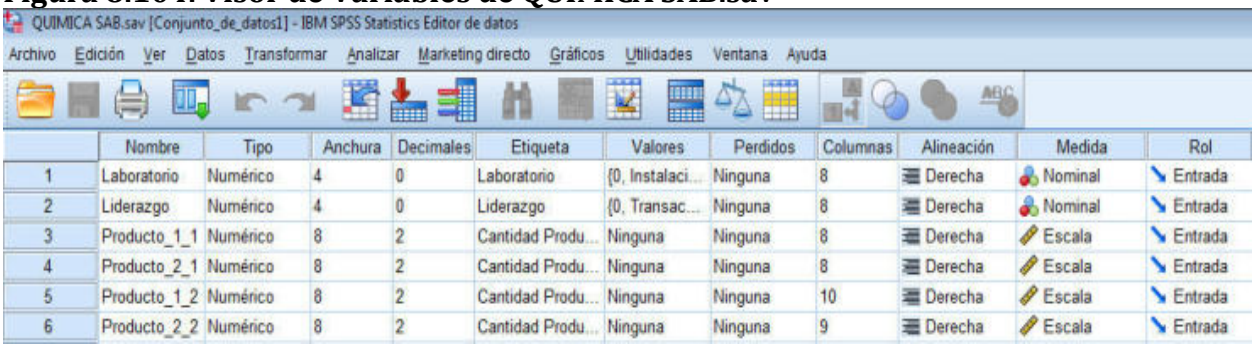
Paso 1: Objetivos

-Problema 7: El consejo de dirección de la empresa de Química SAB quiere evaluar si la estrategia de implementación de una innovación organizacional en sus 10 nuevas instalaciones, tienen mayores niveles de productividad en los dos productos estrella de la empresa, al compararlos con los procesos de las 10 instalaciones actuales, sin dicha implementación de innovación organizacional dispersos a nivel nacional. Los datos se **calificaron de 10 a 100** y encuentran en **QUIMICA SAB.sav**

$_0$= La introducción de la innovación organizacional en las nuevas instalaciones tienen mayores niveles de productividad en los dos productos estrella de la empresa, que las que tienen los procesos actuales

$_1$= La introducción de la innovación organizacional en las nuevas instalaciones NO tienen mayores niveles de productividad en los dos productos estrella de la empresa que las que tienen los procesos actuales Ver Figuras 8.104 y 8.105

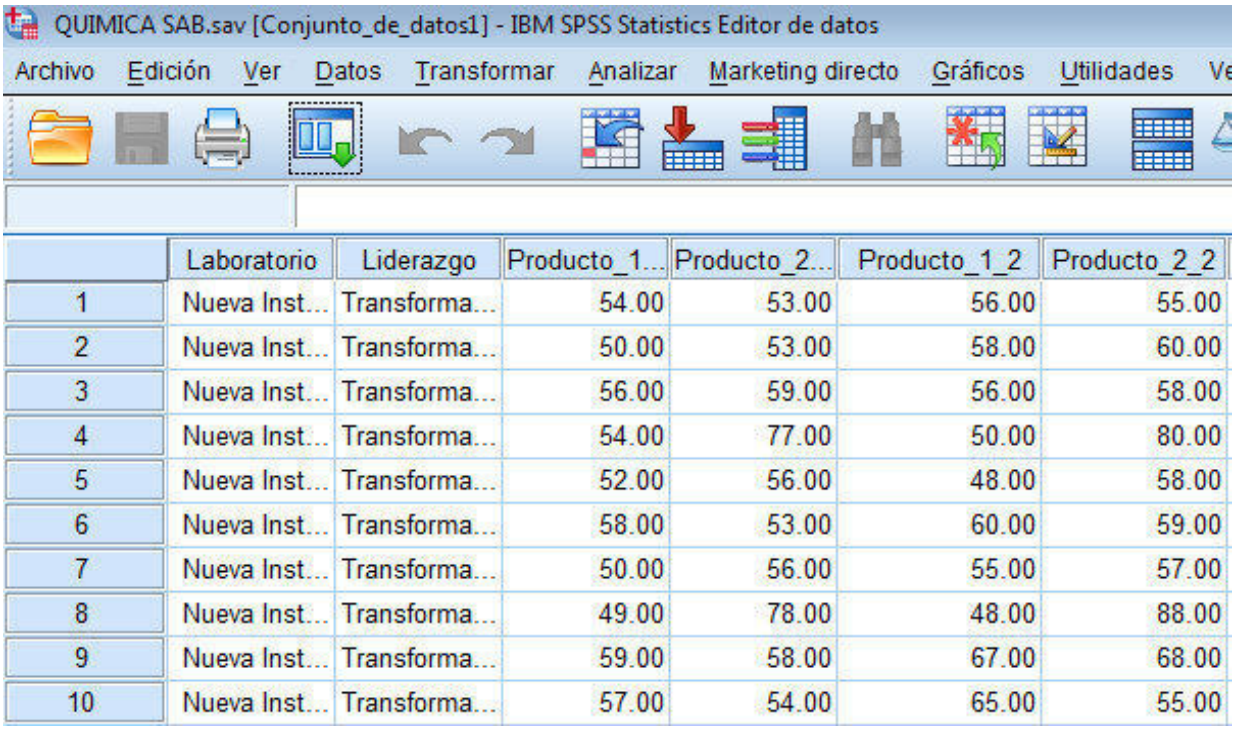
Figura 8.104. Visor de Variables de QUIMICA SAB.sav



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	Laboratorio	Numérico	4	0	Laboratorio	{0, Instalaci...	Ninguna	8	Derecha	Nominal	Entrada
2	Liderazgo	Numérico	4	0	Liderazgo	{0, Transac...	Ninguna	8	Derecha	Nominal	Entrada
3	Producto_1_1	Numérico	8	2	Cantidad Produ...	Ninguna	Ninguna	8	Derecha	Escala	Entrada
4	Producto_2_1	Numérico	8	2	Cantidad Produ...	Ninguna	Ninguna	8	Derecha	Escala	Entrada
5	Producto_1_2	Numérico	8	2	Cantidad Produ...	Ninguna	Ninguna	10	Derecha	Escala	Entrada
6	Producto_2_2	Numérico	8	2	Cantidad Produ...	Ninguna	Ninguna	9	Derecha	Escala	Entrada

Fuente: SPSS 20 IBM

Figura 8.105. Visor de Datos de QUIMICA SAB.sav



	Laboratorio	Liderazgo	Producto_1...	Producto_2...	Producto_1_2	Producto_2_2
1	Nueva Inst...	Transforma...	54.00	53.00	56.00	55.00
2	Nueva Inst...	Transforma...	50.00	53.00	58.00	60.00
3	Nueva Inst...	Transforma...	56.00	59.00	56.00	58.00
4	Nueva Inst...	Transforma...	54.00	77.00	50.00	80.00
5	Nueva Inst...	Transforma...	52.00	56.00	48.00	58.00
6	Nueva Inst...	Transforma...	58.00	53.00	60.00	59.00
7	Nueva Inst...	Transforma...	50.00	56.00	55.00	57.00
8	Nueva Inst...	Transforma...	49.00	78.00	48.00	88.00
9	Nueva Inst...	Transforma...	59.00	58.00	67.00	68.00
10	Nueva Inst...	Transforma...	57.00	54.00	65.00	55.00

Fuente: SPSS 20 IBM

Paso 2: Diseño

- En adelante para efectos de aprendizaje, sólo se tomarán de forma directa los datos con el fin de aprender a utilizar la técnica.
- **SPSS** reporta información adicional, como los cálculos de **ANOVA univariada** para el efecto de las variables independientes sobre cada una de las dependientes de forma separada. As , es posible usar estas tablas cuando se obtenga un hallazgo significativo en la **MANOVA** y ver donde se encuentran las variables independientes que produzcan los efectos ma s grandes. Sin embargo, se debe tomar en cuenta que el cálculo es de **ANOVA univariada** (idéntico a tener que hacerlos sin **MANOVA**) que es previsible **corregir para el incremento de riesgo de errores de Tipo I** dado que la **MANOVA** se calculó primero (por ejemplo, realice una corrección de ***Bonferroni*** en el nivel de significancia). La **MANOVA**, al ser una versio n ma s compleja de la **ANOVA**, requiere de los mismos supuestos de una **ANOVA de mediciones repetidas.**
- **SPSS** reporta **4 estadísticos** de resultados típicos de MANOVA:
  1. ***Pillai's Trace***
  2. ***Hotelling's***
  3. T^2
  4. ***Lambda de Wilks***
  5. ***Raíz mayor de Roy*.**

  La razón de esto es que cada una es una diferente fórmula que **intentan calcular la proporción de la variabilidad en las variables dependientes explicadas por las variables independientes .** Al seleccionar la estadística apropiada que se requiere,

claramente se avanza en el conocimiento del propósito de la **MANOVA**, del que se recomienda básicamente**:**

1. Buscar en el análisis si todas las 4 estadísticas convergen en la significancia del efecto y si lo hacen, entonces podemos confiar que es correcta.

2. Las 4 estadísticas difieren en su potencia dependiendo del tipo de datos y en los tamaños de la muestra que al escoger la apropiada, serpa dependiente de las diferencias que estén siendo examinadas. Sin embargo la ***Lambda de Wilks*** , se considera como una de las mejores por los valores medios que toma. Con un **ANOVA,** a menudo se necesita realizar comparaciones para descubrir específicamente el lugar de las diferencias causantes de manera global del **valor *F* significante**. El **MANOVA** es similar en que una ***Lambda de Wilks*** **nos reporta que hay un efecto de la variable independiente sobre las variables dependientes, sin saber exactamente donde se ubica el efecto.**

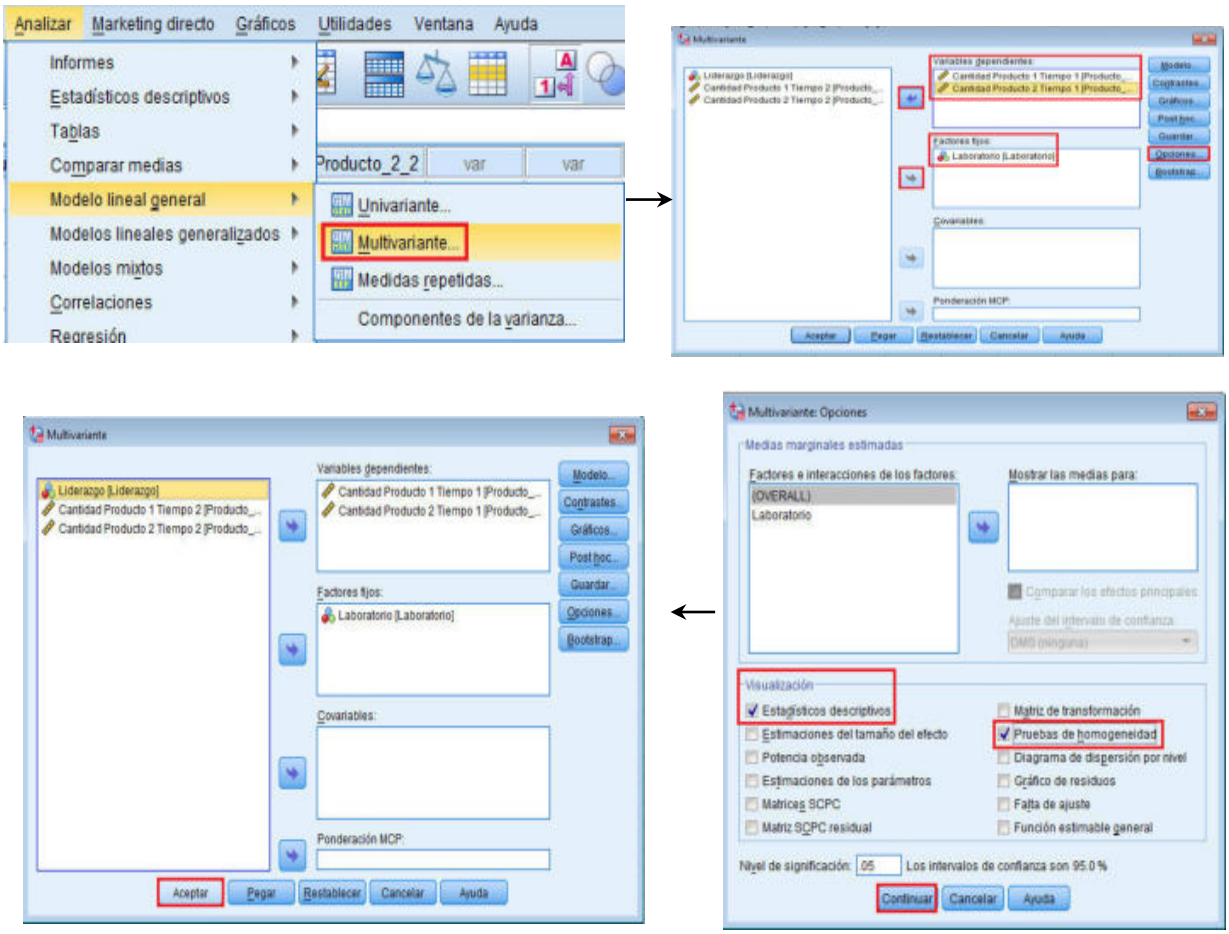
Paso 3: Condiciones de Aplicabilidad

- No olvidar realizar los análisis previos de inspección visual de la base de datos para detectar ausentes y/o atípicos (corregir en su defecto), confiabilidad, normalidad, homocedasticidad y linealidad

Paso 4: Estimación y ajuste

-Teclear: Analizar-> Modelo lineal general->Multivariante->Variables dependientes: ingresar métricas (más de una); en nuestro caso : Cantidad Producto 1 Tiempo 1 y Cantidad Producto 2 Tiempo 1->Factores fijos: ingresar métricas/nominales (en nuestro caso: Laboratorio)->Opciones;Visualización; Estadísticos descriptivos; Pruebas de homogeneidad->Continuar-> Ver Figura 8.106.

Figura 8.106. Proceso MANOVA de mediciones independientes



Fuente: SPSS 20 IBM

Nota: observe que en este ejemplo no se usan pruebas ***post hoc*** dado que la variable independiente solamente tiene 2 niveles. Para 2 factores de medición independientes de **MANOVA**, donde 1 o ambos factores tienen 3 o más niveles, de ***click*** en el comando ***post hoc*** y seleccione la prueba apropiada.

Paso 5: Interpretación

La primer tabla que **SPSS** genera es la **Factores inter-sujetos** que reporta cuantos participantes estuvieron en cada grupo y cuantos niveles de nuestro factor independiente tiene. Ver **Figura 8.107**

Figura 8.107 Tabla Factores inter-sujetos

Factores inter-sujetos

		Etiqueta del valor	N
Laboratorio	0	Instalación Actual	10
	1	Nueva Instalación	10

Factores Independientes

Niveles de los factores

Fuente: SPSS 20 IBM

La siguiente tabla que el **SPSS** genera es la de Estadísticos descriptivos, la cual reporta la Media, Desviación típica (desviación estándar) y el número de participantes de cada grupo. Ver Figura **8.108.**

Figura 8.108. Tabla Estadísticos descriptivos

Estadísticos descriptivos

	Laboratorio	Media	Desviación típica	N
Cantidad Producto 1 Tiempo 1	Instalación Actual	47.7000	5.55878	10
	Nueva Instalación	53.9000	3.57305	10
	Total	50.8000	5.54977	20
Cantidad Producto 2 Tiempo 1	Instalación Actual	62.5000	8.60555	10
	Nueva Instalación	59.7000	9.61538	10
	Total	61.1000	8.99649	20

Fuente: SPSS 20 IBM

- La Figura **8.108** despliega la Media de las puntuaciones para cada **Laboratorio** en sus resultados de **Cantidad de Producto 1 Tiempo 1** y **Cantidad de Producto 2 Tiempo 1**. Como no nos interesa comparar ambas puntuaciones (por ejemplo, las 2 variables dependientes), ambas aparecen en sus renglones correspondientes. Sin embargo, sí estamos interesados en comparar las puntuaciones de los 2 tipos de laboratorios en cada una de las variables dependientes. Por ejemplo, al observar los resultados de **Cantidad de Producto 1 Tiempo 1 tenemos que los Nuevas Instalaciones tienen una media más alta** (**53.9000**) **que las Instalaciones actuales** (**47.7000**).

- La **Media Total** nos reporta la puntuación de la media global en las pruebas a través de ambos grupos.
- La Desviación típica (desviación estándar) muestra que las puntuaciones dispersas se realizaron en ambas pruebas
- La siguiente tabla es **la Prueba de *Box* sobre la igualdad de las matrices de covarianzas,** la cual reporta si nuestros datos violan el supuesto de la **igualdad de la covarianza**
- Uno de los supuestos de la **MANOVA** es que NO existe una diferencia en la covarianza de las variables dependientes a través de los grupos independientes.
- La **Prueba de *Box* sobre la igualdad de las matrices de covarianzas** indica se tenemos el problema con la covarianza. Ver **Figura 8.109.**

Figura 8.109. Prueba de *Box* sobre la igualdad de las matrices de covarianzas

Prueba de Box sobre la igualdad de las matrices de covarianzas[a]

M de Box	4.114
F	1.206
gl1	3
gl2	58320.000
Sig.	.306

Contrasta la hipótesis nula de que las matrices de covarianza observadas de las variables dependientes son iguales en todos los grupos.

a. Diseño: Intersección + Laboratorio

Fuente: SPSS 20 IBM

- Recuerde, como con la **ANOVA de mediciones repetidas**, es siempre importante revisar el supuesto de homocedasticidad así como el de **homogeneidad de la covariancia**. Dentro de **MANOVA** estos supuestos se realizan a través de checar la **prueba Contraste de *Levene* sobre la igualdad de las varianzas error y la prueba de *Box*.** Si se encuentra un resultado significativo, este hallazgo sugeriría que el supuesto ha sido violado. Como se observa, hemos encontrado este supuesto en la **prueba de *Box* que NO es significativo** (***p* > 0.05**).

- Con nuestras **MANOVA** de mediciones independientes , analizaremos si nuestras 2 variables dependientes: **Cantidad de Producto 1 Tiempo 1 y Cantidad de Producto 2 Tiempo 1**, juntas están siendo influenciadas por la estrategia de implementación de la

- innovación organizacional. Esta información está contenida en, la tabla de **Contrastes multivariados.**Ver **Figura 8.110.**

Figura 8.110. Contrastes multivariados

Contrastes multivariados[a]

Efecto		Valor	F	Gl de la hipótesis	Gl del error	Sig.
Intersección	Traza de Pillai	.994	1418.673[b]	2.000	17.000	.000
	Lambda de Wilks	.006	1418.673[b]	2.000	17.000	.000
	Traza de Hotelling	166.903	1418.673[b]	2.000	17.000	.000
	Raíz mayor de Roy	166.903	1418.673[b]	2.000	17.000	.000
Laboratorio	Traza de Pillai	.352	4.612[b]	2.000	17.000	.025
	Lambda de Wilks	.648	4.612[b]	2.000	17.000	.025
	Traza de Hotelling	.543	4.612[b]	2.000	17.000	.025
	Raíz mayor de Roy	.543	4.612[b]	2.000	17.000	.025

a. Diseño: Intersección + Laboratorio

b. Estadístico exacto

Fuente: SPSS 20 IBM

- Dentro de la tabla se observa que la prueba estadística ***Lambda de Wilks*** **se encuentra como significativa**, lo cual podría indicar que sobretodo, existe un efecto global significativo de la implementación estratégica de la innovación organizacional en ambas variables dependientes.
- Así también, se observa que sobretodo, existe un efecto global significativo de la implementación estratégica de la innovación organizacional en la combinación de ambas variables dependientes:

$$F(2,17) = 4.612,\ p < 0.05;\ \textit{Lambda de Wilks} = 0.648$$

- **La prueba de Contraste de *Levene* sobre la igualdad de las varianzas error** para la homogeneidad de las varianzas para cada variable dependiente. **Figura 8.111.**

Figura 8.111. Contraste de *Levene* sobre la igualdad de las varianzas error

Contraste de Levene sobre la igualdad de las varianzas error[a]

	F	gl1	gl2	Sig.
Cantidad Producto 1 Tiempo 1	1.028	1	18	.324
Cantidad Producto 2 Tiempo 1	.080	1	18	.780

Contrasta la hipótesis nula de que la varianza error de la variable dependiente es igual a lo largo de todos los grupos.

a. Diseño: Intersección + Laboratorio

Fuente: SPSS 20 IBM

- Esta prueba nos permite observar el supuesto de la homocedasticidad de cada variable dependiente. **Un resultado significativo indica que el supuesto ha sido violado.**
- Como se observa en el ejemplo, se encontraron resultados no-significativos para ambas variables dependientes variables, o sea se encuentran con $p > 0.05$.
- La tabla siguiente es de las más representativas y conocidas el de las **Pruebas de los efectos inter-sujetos,** el cual nos permite analizar cada una de las variables independientes individualmente. Ver **Figura 8.112.**

Figura 8.112. Pruebas de los efectos inter-sujetos

Pruebas de los efectos inter-sujetos

Origen	Variable dependiente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
Modelo corregido	Cantidad Producto 1 Tiempo 1	192.200[a]	1	192.200	8.803	.008
	Cantidad Producto 2 Tiempo 1	39.200[b]	1	39.200	.471	.501
Intersección	Cantidad Producto 1 Tiempo 1	51612.800	1	51612.800	2363.945	.000
	Cantidad Producto 2 Tiempo 1	74664.200	1	74664.200	896.807	.000
Laboratorio	Cantidad Producto 1 Tiempo 1	192.200	1	192.200	8.803	.008
	Cantidad Producto 2 Tiempo 1	39.200	1	39.200	.471	.501
Error	Cantidad Producto 1 Tiempo 1	393.000	18	21.833		
	Cantidad Producto 2 Tiempo 1	1498.600	18	83.256		
Total	Cantidad Producto 1 Tiempo 1	52198.000	20			
	Cantidad Producto 2 Tiempo 1	76202.000	20			
Total corregida	Cantidad Producto 1 Tiempo 1	585.200	19			
	Cantidad Producto 2 Tiempo 1	1537.800	19			

a. R cuadrado = .328 (R cuadrado corregida = .291)

b. R cuadrado = .025 (R cuadrado corregida = -.029)

Fuente: SPSS 20 IBM

Paso 5: Interpretación

- Como se observa, en las puntuaciones de nuestra variable dependiente: **Cantidad de Producto 1 Tiempo 1 SÍ** se encuentra un efecto significativo de la implementación estratégica de la innovación organizacional, por lo que:

$$F(1,18) = 8.803; p < 0.01.$$

- Por otro lado se observa, en las puntuaciones de nuestra variable dependiente: **Cantidad de Producto 2 Tiempo 1 NO** se encuentra un efecto significativo de la implementación estratégica de la innovación organizacional, por lo que:

F(1,18) = .471; p > 0.01.

- La **Suma de Cuadrados** reporta una medida de la variabilidad en las puntuaciones debida a una fuente particular de variabilidad. La **Media cuadrática** es la cantidad de varianza producida como resultado de esa fuente.
- Los valores de **R cuadrado** abajo de la tabla, indican la cantidad de variación en cada variable dependiente que pueda contarse por medio del factor independiente. Por ejemplo, la **R Cuadrado** para **Cantidad de Producto 1 Tiempo 1** es **0.328**, lo cual muestra que la estrategia de la implementación innovación organizacional cuenta en un **32.8%** de la variación de las puntuaciones de **Cantidad de Producto 1 Tiempo 1.**
- Debido a posibles afectaciones de **errores Tipo 1**, valdría la pena considerar llevar a cabo correcciones de ***Bonferroni*** en los resultados de **ANOVA.**
- **Conclusión: se acepta H_1= La introducción de la innovación organizacional en las nuevas instalaciones NO tienen mayores niveles de productividad en los dos productos estrella de la empresa que las que tienen los procesos actuales ya que sólo mejora al producto 1 y no al producto 2.**

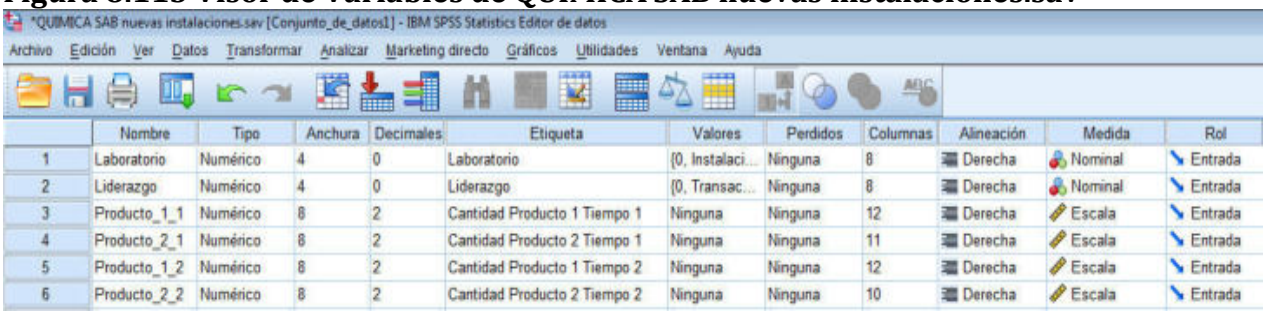
8.23. MANOVA de mediciones repetidas. Resumen

En este tipo de **MANOVA** se repiten mediciones sobre las variables independientes (no se tiene una agrupación de variables).Por ejemplo, una compañía minera está interesada en investigar el efecto de atención en la calibración en zona de seguridad, en 3 tipos de máquinas que hacen la misma tarea de extracción. Un grupo de trabajadores son ubicados en un simulador de control que permite ver en 3 diferentes monitores el despliegue de información para realizar ajustes cuando las máquinas detectan que la tarea de extracción, por descalibración, está a punto de salir a zona de baja seguridad. Así, el mismo conjunto de participantes es medido en sus respuestas en las 3 máquinas en 2 diferentes periodos de tiempo: al comienzo de una jornada de trabajo de 12 horas como primer turno y al final de un cambio de un segundo turno de trabajo.
La **MANOVA** mostrará si existe un efecto de atención en la calibración sobre la variable dependiente compuesta , que consiste en las respuestas diferentes a las 3 tareas. Si se encuentra un efecto significante en la **MANOVA**, podemos analizar las respuestas de **3 ANOVAS de mediciones repetidas de mediciones de un factor**, las cuales mostrarán el efecto de la variable independiente en cada una de las variables dependientes separadamente (no llevaríamos a cabo un análisis de función discriminante ya que no hay agrupación de variables en este análisis de mediciones repetidas.

Paso 1: Objetivos
Problema 8: El consejo de dirección de la empresa de Química SAB quiere evaluar si la estrategia de implementación de una innovación organizacional en sus 10 nuevas instalaciones, tienen mayores niveles de productividad en los dos productos estrella de la empresa, y se sostienen en un periodo 2 de 6 meses. Los datos se **califican de 10 a 100** y encuentran en **QUIMICA SAB nuevas instalaciones.sav**
H_0= La introducción de la innovación organizacional en las instalaciones nuevas genera un factor subyacente con efecto significativo sobre los dos productos estrella de la empresa

$_1$= La introducción de la innovación organizacional en las instalaciones nuevas NO genera un factor subyacente con efecto significativo sobre los dos productos estrella de la empresa.Ver Figuras 8.113 y 8.114.

Figura 8.113 Visor de Variables de QUIMICA SAB nuevas instalaciones.sav



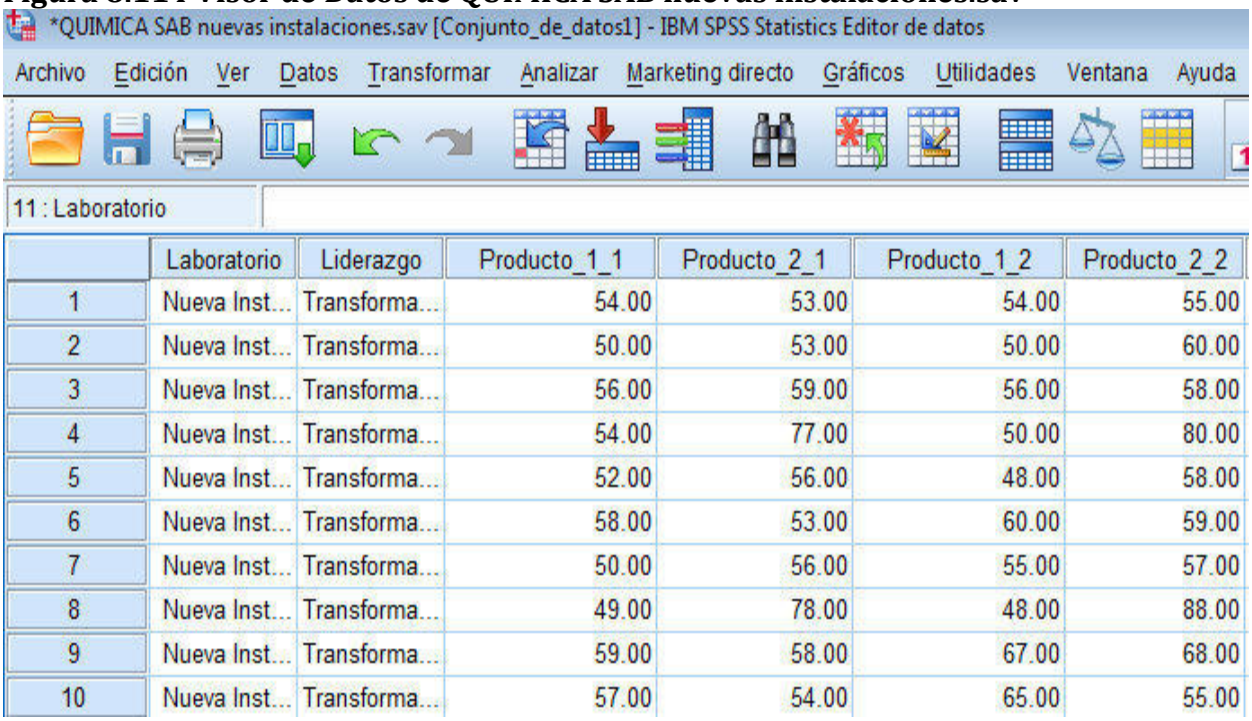
*QUIMICA SAB nuevas instalaciones.sav [Conjunto_de_datos1] - IBM SPSS Statistics Editor de datos

Archivo Edición Ver Datos Transformar Analizar Marketing directo Gráficos Utilidades Ventana Ayuda

	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	Laboratorio	Numérico	4	0	Laboratorio	{0, Instalaci...	Ninguna	8	Derecha	Nominal	Entrada
2	Liderazgo	Numérico	4	0	Liderazgo	{0, Transac...	Ninguna	8	Derecha	Nominal	Entrada
3	Producto_1_1	Numérico	8	2	Cantidad Producto 1 Tiempo 1	Ninguna	Ninguna	12	Derecha	Escala	Entrada
4	Producto_2_1	Numérico	8	2	Cantidad Producto 2 Tiempo 1	Ninguna	Ninguna	11	Derecha	Escala	Entrada
5	Producto_1_2	Numérico	8	2	Cantidad Producto 1 Tiempo 2	Ninguna	Ninguna	12	Derecha	Escala	Entrada
6	Producto_2_2	Numérico	8	2	Cantidad Producto 2 Tiempo 2	Ninguna	Ninguna	10	Derecha	Escala	Entrada

Fuente: SPSS 20 IBM

Figura 8.114 Visor de Datos de QUIMICA SAB nuevas instalaciones.sav



*QUIMICA SAB nuevas instalaciones.sav [Conjunto_de_datos1] - IBM SPSS Statistics Editor de datos

Archivo Edición Ver Datos Transformar Analizar Marketing directo Gráficos Utilidades Ventana Ayuda

11 : Laboratorio

	Laboratorio	Liderazgo	Producto_1_1	Producto_2_1	Producto_1_2	Producto_2_2
1	Nueva Inst...	Transforma...	54.00	53.00	54.00	55.00
2	Nueva Inst...	Transforma...	50.00	53.00	50.00	60.00
3	Nueva Inst...	Transforma...	56.00	59.00	56.00	58.00
4	Nueva Inst...	Transforma...	54.00	77.00	50.00	80.00
5	Nueva Inst...	Transforma...	52.00	56.00	48.00	58.00
6	Nueva Inst...	Transforma...	58.00	53.00	60.00	59.00
7	Nueva Inst...	Transforma...	50.00	56.00	55.00	57.00
8	Nueva Inst...	Transforma...	49.00	78.00	48.00	88.00
9	Nueva Inst...	Transforma...	59.00	58.00	67.00	68.00
10	Nueva Inst...	Transforma...	57.00	54.00	65.00	55.00

Fuente: SPSS 20 IBM

Paso 2: Diseño

- En adelante para efectos de aprendizaje, sólo se tomarán de forma directa los datos con el fin de aprender a utilizar la técnica.

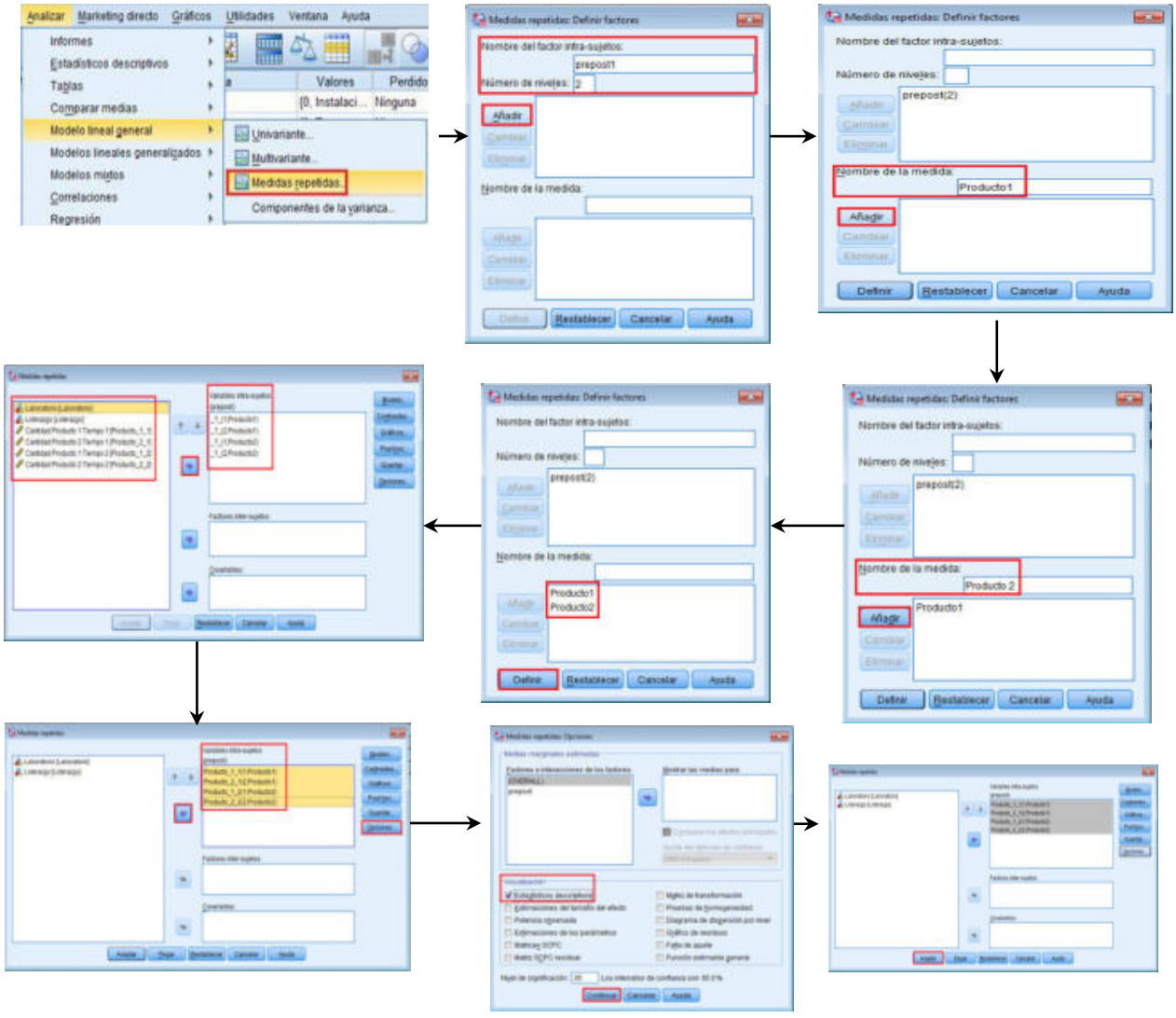
Paso 3: Condiciones de Aplicabilidad

- No olvidar realizar los análisis previos de inspección visual de la base de datos para detectar ausentes y/o atípicos (corregir en su defecto), confiabilidad, normalidad, homocedasticidad y linealidad

Paso 4: Estimación y ajuste

- **Teclear: Analizar->Modelo lineal general->Medidas repetidas>Nombre del factor intra-sujetos (nombre de la variable subyacente sugerida): prepost; Número de niveles: 2-> Añadir->Nombre de la medida (ingresar nombre de la variable dependiente sugerida): Producto1->Añadir->Nombre de la medida (ingresar nombre de la variable dependiente sugerida): Producto2->Definir->Variables intra-sujetos (ingresar por orden las variables dependientes en momento 1 y 2 respectivamente)->Opciones->Visualización: Estadísticos descriptivos->Continuar->Aceptar. Ver Figura 8.115.**

Figura 8.115. Proceso de MANOVA de mediciones repetidas



Fuente: SPSS 20 IBM

Paso 5: Interpretación

La primera tabla generada por **SPSS** es la de **Factores intra-sujetos** la cual reporta una descripción de las 2 variables dependientes analizadas por el cálculo de **MANOVA.** La primer columna muestra ambas variables dependientes: **Producto 1 y Producto 2.** La segunda columna despliega los niveles de la medición de factor repetidas. Ver **Figura 8.116**

Figura 8.116. Tabla Factores intra-sujetos

Factores intra-sujetos

Medida	prepost	Variable dependiente
Producto1	1	Producto_1_1
	2	Producto_2_1
Producto2	1	Producto_1_2
	2	Producto_2_2

Fuente: SPSS 20 IBM

- Se observa de dicha tabla que **Producto_1_1** es la etiqueta para **Producto1** al comienzo del estudio
- Se ha etiquetado como **Producto1** las puntuaciones al final del estudio de **Producto 2_1**.
- Las otras combinaciones de niveles de los factoresestpan asignados de manera similar
- La siguiente tabla a considerar es la de **Estadísticos descriptivos** en la que se pueden analizar el conjunto de nuestros datos para potenciales diferencias en las pruebas de puntuación. Ver **Figura 8.117**

Figura 8.117. Tabla Estadísticos descriptivos

Estadísticos descriptivos

	Media	Desviación típica	N
Cantidad Producto 1 Tiempo 1	53.9000	3.57305	10
Cantidad Producto 2 Tiempo 1	59.7000	9.61538	10
Cantidad Producto 1 Tiempo 2	55.0000	4.52155	10
Cantidad Producto 2 Tiempo 2	73.0000	9.51023	10

Fuente: SPSS 20 IBM

- Se tiene que la media de las puntuaciones para la **Cantidad Producto 1 Tiempo 2** al final del estudio después de haber sido implementada la innovación organizacional es más grande (**55.0000**) que al comienzo del estudio cuando **Cantidad Producto 1 Tiempo 2** (**53.9000**).
- Similarmente, la media inicial de **Cantidad Producto 2 Tiempo 1** con puntuaciones (**59.7000**) se ha incrementado (**73.0000**) como **Cantidad Producto 2 Tiempo 2** al final del estudio
- Como los 2 conjuntos de puntuaciones fueron medidas en diferentes escalas no se considera apropiado comparar las puntuaciones de Producto 1 con Producto 2
- La **Deviación típica** (**desviación estándar**) indica que hubo una gran dispersión de las puntuaciones dentro de **Cantidad Producto 2**

Fuente: SPSS 20 IBM La siguiente tabla muestra la prueba de **Contrastes multivariados.** Esta tabla es generada por **SPSS** durante todo el proceso del modelo general lineal de medidas repetidas mostrando sus principales hallazgos.Ver **Figura 8.118.**

Figura 8.118. Tabla de Contrastes multivariados

Contrastes multivariados[a]

Efecto			Valor	F	Gl de la hipótesis	Gl del error	Sig.
Entre sujetos	Intersección	Traza de Pillai	.995	797.865[b]	2.000	8.000	.000
		Lambda de Wilks	.005	797.865[b]	2.000	8.000	.000
		Traza de Hotelling	199.466	797.865[b]	2.000	8.000	.000
		Raíz mayor de Roy	199.466	797.865[b]	2.000	8.000	.000
Intra-sujetos	prepost	Traza de Pillai	.891	32.743[b]	2.000	8.000	.000
		Lambda de Wilks	.109	32.743[b]	2.000	8.000	.000
		Traza de Hotelling	8.186	32.743[b]	2.000	8.000	.000
		Raíz mayor de Roy	8.186	32.743[b]	2.000	8.000	.000

a. Diseño: Intersección
Diseño intra-sujetos: prepost

b. Estadístico exacto

Fuente: SPSS 20 IBM

También se observa la **prueba estadística de *Lambda de Wilks* que es significativa**, el cual, indica que globalmente hay un efecto producido por la innovación organizacional en ambas **variables dependientes Producto 1 y Producto 2:**

F(2,8) = 32.743, p < 0.001; Wilks' lambda = 0.109

La prueba de *esfericidad* de *Mauchly*, muestra las pruebas de *esfericidad*. **Figura 8.119**

Figura 8.119. Prueba de *esfericidad* de *Mauchly*

Prueba de esfericidad de Mauchly[a]

Efecto intra-sujetos	Medida	W de Mauchly	Chi-cuadrado aprox.	gl	Sig.	Epsilon[b]		
						Greenhouse-Geisser	Huynh-Feldt	Límite-inferior
prepost	Producto1	1.000	.000	0	.	1.000	1.000	1.000
	Producto2	1.000	.000	0	.	1.000	1.000	1.000

Contrasta la hipótesis nula de que la matriz de covarianza error de las variables dependientes transformadas es proporcional a una matriz identidad.

a. Diseño: Intersección
Diseño intra-sujetos: prepost

b. Puede usarse para corregir los grados de libertad en las pruebas de significación promediadas. Las pruebas corregidas se muestran en la tabla Pruebas de los efectos inter-sujetos.

Fuente: SPSS 20 IBM

Usted advertira que la columna **Sig.** aparece en blanco, sin reporte de grados de libertad. Esto es porque **la *esfericidad* es solamente problema si Usted tiene 2 o más condiciones en sus factores de medidas repetida**s. Nuestro factor de medidas repetidas "***prepost***" tiene solamente 2 niveles y por lo tanto *esfericidad* no sera un problema con nuestros datos.

- Una segunda tabla multivariada también es producida, la cual es la **Prueba de efectos intra-sujetos.** En nuestro ejemplo estos valores son los mismo como en las tablas de prueba multivariadas previas.. Ver la **Figura 8.120.**

Figura 8.120. Tabla Multivariante

Multivariante[a,b]

Efecto intra-sujetos		Valor	F	Gl de la hipótesis	Gl del error	Sig.
prepost	Traza de Pillai	.891	32.743[c]	2.000	8.000	.000
	Lambda de Wilks	.109	32.743[c]	2.000	8.000	.000
	Traza de Hotelling	8.186	32.743[c]	2.000	8.000	.000
	Raíz mayor de Roy	8.186	32.743[c]	2.000	8.000	.000

a. Diseño: Intersección
Diseño intra-sujetos: prepost

b. Las pruebas se basan en las variables promediadas.

c. Estadístico exacto

Fuente: SPSS 20 IBM

La siguiente tabla es la de **Contrastes univariadas**, el cual nos permite analizar cada una de las variables dependientes individualmente. Ver la **Figura 8.121.**

Figura 8.121. Contrastes univariados

Contrastes univariados

Origen	Medida		Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
prepost	Producto1	Esfericidad asumida	168.200	1	168.200	2.657	.138
		Greenhouse-Geisser	168.200	1.000	168.200	2.657	.138
		Huynh-Feldt	168.200	1.000	168.200	2.657	.138
		Límite-inferior	168.200	1.000	168.200	2.657	.138
	Producto2	Esfericidad asumida	1620.000	1	1620.000	27.458	.001
		Greenhouse-Geisser	1620.000	1.000	1620.000	27.458	.001
		Huynh-Feldt	1620.000	1.000	1620.000	27.458	.001
		Límite-inferior	1620.000	1.000	1620.000	27.458	.001
Error(prepost)	Producto1	Esfericidad asumida	569.800	9	63.311		
		Greenhouse-Geisser	569.800	9.000	63.311		
		Huynh-Feldt	569.800	9.000	63.311		
		Límite-inferior	569.800	9.000	63.311		
	Producto2	Esfericidad asumida	531.000	9	59.000		
		Greenhouse-Geisser	531.000	9.000	59.000		
		Huynh-Feldt	531.000	9.000	59.000		
		Límite-inferior	531.000	9.000	59.000		

Fuente: SPSS 20 IBM

- Los renglones importantes son el de ***Esfericidad*** **asumida.**
- **De la tabla anterior se aprecia un efecto significante de la variable subyacente "*prepost*" sobre las dos variables dependientes Producto1 y Producto2**
- Los renglones importantes son el de ***Esfericidad*** **asumida.**
- **De la tabla anterior se aprecia un efecto significante de la variable subyacente *"prepost"* sobre las dos variables dependientes Producto1 y Producto2**.
- Así, se encuentra que **NO existe un efecto principalmente significante para nuestra variable subyacente "*prepost*" sobre la variable dependiente Producto1 en:**

  ***F(1,9) = 2.657; p =0.138> 0.05*.**
- Así, se encuentra que **SÍ existe un efecto principalmente significante para nuestra variable subyacente "*prepost*" sobre la variable dependiente Producto2 en:**

  F(1,9) = 27.458; p=0.001 < 0.01

La tabla de **Pruebas de contrastes intra-sujetos** es generada por el **SPSS** durante el procedimiento de cálculo de mediciones repetidas del modelo lineal general (**GLM**), y es un tópico de análisis. Ver **Figura 8.122**

Figura 8.122. Pruebas de contrastes intra-sujetos

Pruebas de contrastes intra-sujetos

Origen	Medida	prepost	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
prepost	Producto1	Lineal	168.200	1	168.200	2.657	.138
	Producto2	Lineal	1620.000	1	1620.000	27.458	.001
Error(prepost)	Producto1	Lineal	569.800	9	63.311		
	Producto2	Lineal	531.000	9	59.000		

Fuente: SPSS 20 IBM

La tabla analiza y reporta información en cuanto al mejor ajuste de los datos por el modelo subyacente

•Dado que solamente se tienen 2 niveles para nuestro factor de mediciones repetidas, las únicas posibles tendencias son aquellas que siguen un modelo lineal

• En nuestro ejemplo podemos observar que la variable **Producto2 sigue una significativa** tendencia lineal. Esto es esperado, debido a que el factor solamente tiene 2 niveles.

La tabla de **Pruebas de los efectos inter-sujetos** es generada por **SPSS** en el proceso de cálculo de **MANOVA de mediciones repetidas**. Como nuestro factor es de mediciones repetidas, no tenemos sujetos entre ellos como factores y la información que la tabla reporta se refiere a la intersección **Figura 8.123**

Figura 8.123. Pruebas de los efectos inter-sujetos

Pruebas de los efectos inter-sujetos

Variable transformada: Promedio

Origen	Medida	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
Intersección	Producto1	64524.800	1	64524.800	1539.563	.000
	Producto2	81920.000	1	81920.000	1578.758	.000
Error	Producto1	377.200	9	41.911		
	Producto2	467.000	9	51.889		

Fuente: SPSS 20 IBM

- En esta ocasión, como se tienen mediciones de variables no independientes solamente se encuentra la intersección, la cual nos reporta que nuestra media global para ambas variables dependientes es significativamente diferente de cero.

Conclusión:

- **H_0= La introducción de la innovación organizacional en las instalaciones nuevas genera un factor subyacente con efecto significativo sobre los dos productos estrella de la empresa….Se rechaza**
- **H_1= La introducción de la innovación organizacional en las instalaciones nuevas NO genera un factor subyacente con efecto significativo sobre los dos productos estrella de la empresa….Se acepta**

8.24. ANOVA de 1 factor para datos no paramétricos. ¿Qué es?

Los investigadores a menudo prefieren utilizar una prueba paramétrica en lugar de una prueba no paramétrica, debido a:

1. U**na prueba paramétrica es más potente**.
2. Podemos **calcular las medias y desviaciones estándar**, que proporcionan un buen resumen claro de los datos. De hecho, incluso cuando tenemos calificaciones (como el juicio de la gente de lo que es felicidad) algunos investigadores realizan una **prueba paramétrica**, argumentando que tratarán los datos como una **escala de intervalo.**
3. Lo anterior es recomendable para los investigadores que confían y tienen experiencia en el manejo de sus datos, a fin de sopesarlos riesgos de las decisiones a tomar.
4. Hay una serie de razones para decidir que no es apropiado usar un parámetro, **especialmente cuando se viola uno o más de los supuestos de una prueba paramétrica.** Por ejemplo: si Usted sabe que la población de la cual usted retira los datos no se distribuye normalmente o Usted sabe que los datos no son de una variable continua entonces usted emprendería una **prueba no paramétrica.** Por ejemplo, si una empresa como **MKT Digital** califica a sus jugadores beta en el esfuerzo para una prueba de su videojuego especial, con **100 puntos de escala** y sólo clasificamos a los jóvenes entre **0 y 40 o 60 y 10.** Podemos ver que **la variable no es continua**, ya que el rango medio no está siendo usado y ciertamente **No parecen estar distribuidos normalmente.**
5. Hay **dos pruebas NO paramétricas** que podemos usar **en lugar del ANOVA de un factor:**

   - La **prueba de *Friedman*** es un equivalente **NO paramétrico del ANOVA de un factor de medidas repetidas**.

   -La prueba de ***Kruskal-Wallis*,** que es la **prueba no paramétrica** utilizada en lugar del **ANOVA de un factor de medidas independientes.**

8.24.1 Prueba de *Kruskal-Wallis* para muestras independientes

Cuando queremos realizar un análisis **NO paramétrico** y tenemos sólo factor de medida independiente (variable independiente) con más de dos muestras, es que elegimos la prueba de ***Kruskal-Wallis***. Por ejemplo, la empresa **MKT Digital** solicita a un investigador que seleccione cuatro ciudades de diferentes tamaños y habitantes les pide que califiquen el nivel de felicidad que producen sus videojuegos en una escala de 100 puntos. El investigador está interesado en verificar si existe un efecto del tamaño de la ciudad en las

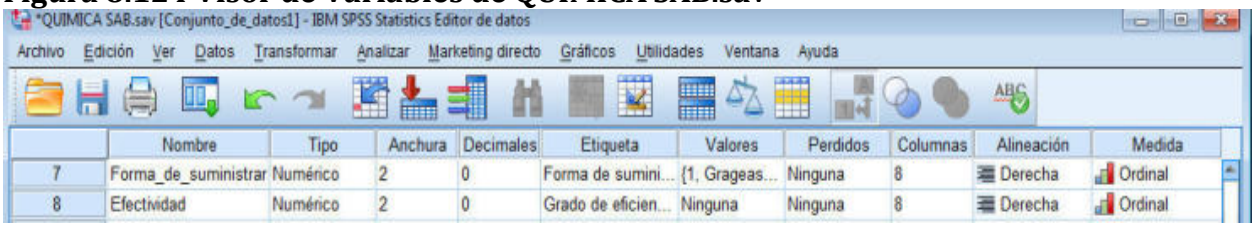
calificaciones. La característica clave de muchas pruebas No paramétricas es que los datos se tratan como ordinales y la primera parte del análisis consiste en clasificar los datos. La prueba de ***Kruskal-Wallis*** no es diferente. Todas las puntuaciones (de todas las condiciones) se clasifican de menor a mayor. Después que un análisis similar al ANOVA se lleva a cabo en las filas. **La estadística *H* (en lugar de *F* en el ANOVA)** da una medida de la fuerza relativa de la variabilidad en las filas entre las condiciones comparadas con un valor estándar para este número de participantes. Dentro de la fórmula de ***Kruskal-Wallis*** hay cálculos donde los elemento que interviene, pueden causar un problema. Por ejemplo, con sólo tres puntuaciones tenemos las filas 1, 2 y 3. Si las elevamos al cuadrado, obtenemos 1, 4 y 9, lo que da un total de 14. Si los dos primeros valores estuvieran empatados, Tendría que dar a las filas 1,5, 1,5 y 3. Cuando estos son cuadrados obtenemos 2.25, 2.25 y 9, dando un total de 13.5. Si solo hay unos pocos vínculos, normalmente no nos preocupamos. Sin embargo, con muchos vínculos**, la prueba de *Kruskal-Wallis* puede ser inapropiada**.

Paso 1: Objetivos

Problema 9: La empresa **QUIMICA SAB** está próxima a lanzar un medicamento antipirético, del cual, está interesado en conocer cuál es la percepción que un grupo de 18 personas le da a sus presentaciones al público planeadas en forma de: grageas con capa entérica, jarabe y pastillas. A los participantes se les pidió, evaluaran la efectividad de la medicina presentada una escala de 0-50, donde: 0.- mucho peor, hasta 25.-sin cambio y hasta 50, mucho mejor que antes. Seis personas probaron el método de las grageas, 5 utilizaron el jarabe y siete las pastillas. La pregunta es: ¿Existe alguna relación de la forma de presentación en sus calificaciones?

Ver Figuras 8.124 y 8.125

Figura 8.124 Visor de Variables de QUIMICA SAB.sav



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida
7	Forma_de_suministrar	Numérico	2	0	Forma de sumini...	{1, Grageas...	Ninguna	8	Derecha	Ordinal
8	Efectividad	Numérico	2	0	Grado de eficien...	Ninguna	Ninguna	8	Derecha	Ordinal

Fuente: SPSS 20 IBM

Figura 8.125 Visor de Datos de QUIMICA SAB.sav

	Forma_de_suministrar	Efectividad
1	Grageas	14
2	Grageas	10
3	Grageas	18
4	Grageas	22
5	Grageas	14
6	Grageas	20
7	Jarabe	29
8	Jarabe	38
9	Jarabe	27
10	Jarabe	25
11	Jarabe	26
12	Pastillas	44
13	Pastillas	30
14	Pastillas	40
15	Pastillas	28
16	Pastillas	33
17	Pastillas	36
18	Pastillas	42

Fuente: SPSS 20 IBM

Paso 2: Diseño

- En adelante para efectos de aprendizaje, sólo se tomarán de forma directa los datos con el fin de aprender a utilizar la técnica.
- La prueba estadística de ***Kruskal-Wallis H*** es la más adecuada para la mayoría de las pruebas de esta naturaleza. Para realizarlo, **SPSS** nos proporciona dos alternativas de ejecución:
  -La menos poderosa **prueba de la mediana** y
  -La altamente poderosa prueba ***Jonckheere-Terpstra***, la cual se utiliza si Usted estuviera buscando diferencias ordenadas entre sus grupos (una **tendencia ascendente o descendente**).
- El cálculo por defecto realizado para el ***p* valor** es el ***p* asintótico**, que es una estimación del verdadero ***p* valor**. Este es generalmente un método adecuado para calcular la ***p* valor** pero es posible calcularlo exactamente. Esto se puede lograr a través de oprimir el **botón Exacta.**
- Se prefiere el método de ***Monte Carlo*** para el cálculo de ***p* valor,** si no es posible calcularlo, ya que **el tamaño de la muestra es demasiado grande** (simplemente tomaría demasiado tiempo hacer ejercicio). El método de ***Monte Carlo*** da una **estimación imparcial del valor exacto de *p*.**
- En el botón **Opciones** es la selección para calcular las medias y desviaciones estándar. Mientras que estos Estadística descriptiva se puede realizar en conjuntos de datos ordinales, se recomienda Grado de cautela al calcularlas e interpretarlas

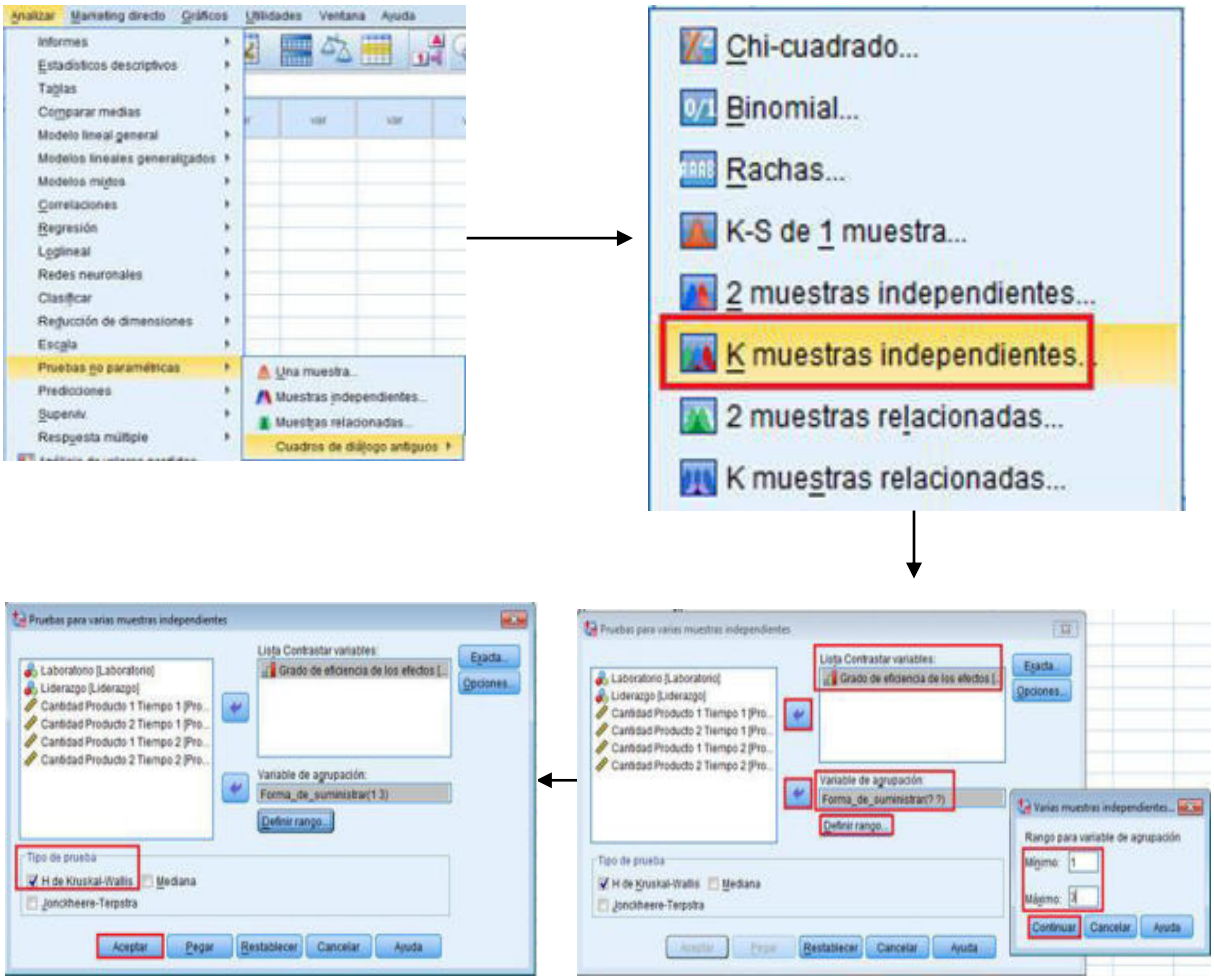
Paso 3: Condiciones de Aplicabilidad

- No olvidar realizar los análisis previos de inspección visual de la base de datos para detectar ausentes y/o atípicos (corregir en su defecto), confiabilidad, normalidad, homocedasticidad y linealidad

Paso 4: Estimación y ajuste

- **Teclear: Analizar->Pruebas no paramétricas->Cuadros de diálogo antiguos->K muestras independientes-> Lista Contrastar variables: Grado de eficiencia de los efectos->Variable de agrupación: Forma de suministrar->Definir rango* ->Mínimo: 1; Máximo: 3->Continuar->Tipos de prueba: H de Kruskal-Wallis->Aceptar. Ver Figura 8.126.**
- *** Rango:** 1, 2, 3 de los grupos declarados en la variable Forma de suministrar: grageas, jarabe, pastillas

Figura 8.126. Proceso de Prueba de *Kruskal-Wallis* para muestras independientes



Fuente: SPSS 20 IBM

Paso 5: Interpretación

La primera tabla generada por **SPSS** es la **tabla Rangos**, que es una descripción de los datos reportando el número de participantes así como el **Rango promedio** de cada grupo. Ver **Figura 8.127**

Figura 8.127. Tabla Rangos

Rangos

	Forma de suministra medicina	N	Rango promedio
Grado de eficiencia de los efectos	Grageas	6	3.50
	Jarabe	5	10.00
	Pastillas	7	14.29
	Total	18	

Fuente: SPSS 20 IBM

- **N** es el número de participantes en cada grupo y el número total de participantes.
- El **Rango promedio** indica lo propio de los puntajes dentro de cada grupo.
- **Si no hubiera diferencias entre las calificaciones de los grupos, es decir, que la hipótesis nula fuera cierta, entonces cabría esperar que los rangos medios fueran casi iguales en los tres grupos.**
- Podemos ver a partir de nuestro ejemplo anterior que los **tres grupos no parecen ser iguales en sus calificaciones del Grado de la eficiencia de los efectos.**

Para determinar si la diferencia en la clasificación es significativa, debemos buscar en la tabla Estadísticos de contraste. **Figura 8.128.**

Figura 8.128. Tabla Estadísticos de contraste

Estadísticos de contraste[a,b]

	Grado de eficiencia de los efectos	
Chi-cuadrado	13.262	**Prueba estadística**
gl	2	
Sig. asintót.	.001	***p* Valor**

a. Prueba de Kruskal-Wallis

b. Variable de agrupación: Forma de suministra medicina

Fuente: SPSS 20 IBM

- El estadístico de prueba a reportar es el ***Chi-cuadrado*** **de** ***Kruskal-Wallis***, que en el ejemplo tiene un valor de **13.262.**

- El **Sig. Asintót.** Nos reporta el valor de probabilidad.
- En el ejemplo, la sintaxis a reportar es ***χ 2 = 13.262, gl = 2, p=0.001 <0.01***.
- **Conclusión: la diferencia entre las calificaciones de los tres grupos es significativo.**
- **SPSS** siempre presenta los resultados como un ***Chi-cuadrado* en lugar de la *Kruskal-Wallis H.***
- Esto se debe a que la **distribución de *H* se aproxima mucho a la del *Chi-cuadrado.***
- El procedimiento ***Kruskal-Wallis*** en **SPSS NO** ofrece la oportunidad de realizar pruebas de comparación múltiple ***post hoc***. Hay una serie de pruebas que podrían hacerse pero estos tendrían que ser calculados sin la ayuda de **SPSS**.

8.24.2 Prueba de *Friedman* para muestras relacionadas

Paso 4: Estimación y ajuste

Utilizamos la **prueba de *Friedman*** como un equivalente de **prueba no paramétrico de ANOVA de un factor de medidas repetidas** en los casos en que los supuestos para el **ANOVA** no se cumplan. Por ejemplo, si tenemos calificaciones y **consideramos que los datos no son intervalo o de razón, entonces, se recomienda utilizar la prueba de *Friedman*.** Por ejemplo, un investigador está interesado en las preferencias de los televidentes para diferentes tipos de programas y pide a **20** personas que califiquen las siguientes **4** categorías de programa en términos de su interés en una escala de **1 a 10**: telenovela, documental, acción / aventura, actualidad. Una persona podría calificar los programas como sigue: telenovela: **8**, documental: **3**, acción / aventura: **9**, asuntos de actualidad: **1**. Una segunda persona podría producir las siguientes calificaciones: telenovela: **6**, documental: **5**, acción / aventura: **7**, asuntos de actualidad: **4**. A pesar de que las calificaciones son muy diferentes ambos participantes han proporcionado el mismo orden de calificación para los **4** tipos de programa. **Es el orden de los resultados de los participantes, lo que analiza la prueba de *Friedman*.**

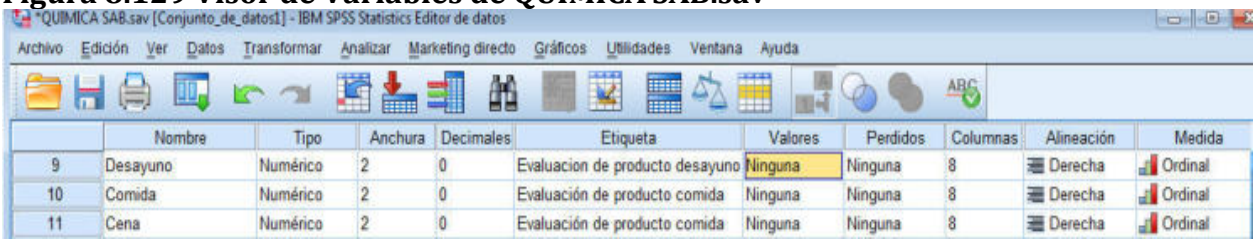
La primera parte de una **prueba de *Friedman*** es clasificar el orden de los resultados de cada participante separadamente. Por lo tanto, ambos participantes descritos anteriormente tendrían las siguientes filas: **3, 2, 4,1**. Estos rangos se analizan de manera similar a un **ANOVA**. SE destaca que no es exactamente el mismo porque, como estamos tratando con rangos, ciertos valores son fijos y hace que nuestro análisis algo más fácil. **La prueba de *Friedman*** produce una estadística de ***Chi-cuadrado***, con una gran valor que indica que hay una diferencia entre los puntajes de uno o más de las condiciones. Si bien no hay los supuestos que necesitemos **ANOVA de un factor de medidas repetidas**, debemos examinar cuántos rangos atados tenemos en la prueba de *Friedman*. Los menos rangos vinculados más apropiado es el análisis. Como Usted puede imaginar, en el anterior ejemplo las calificaciones de una persona de 5, 7, 7, 7 darían filas de 1, 3, 3, 3, y esto no son datos muy informativos para este tipo de análisis. Afortunadamente, ya que estamos dentro de cada participante en vez de a través de todos los participantes, normalmente no tenemos varios puntajes empatados en **una prueba de *Friedman***.

Paso 1: Objetivos

Problema 10: La empresa **QUIMICA SAB**, desea realizar una prueba de producto de su línea de alimentos a 10 personas a quienes se les solicita valorar la calidad de los alimentos para las tres comidas: desayuno, almuerzo y cena, con calificación de basada en una escala

de 0 a 100 (de malo a bueno). Así, la empresa se pregunta si ¿existe alguna diferencia significativa entre las tres comidas en su evaluación de calidad? **Ver Figuras 8.129 y 8.30**

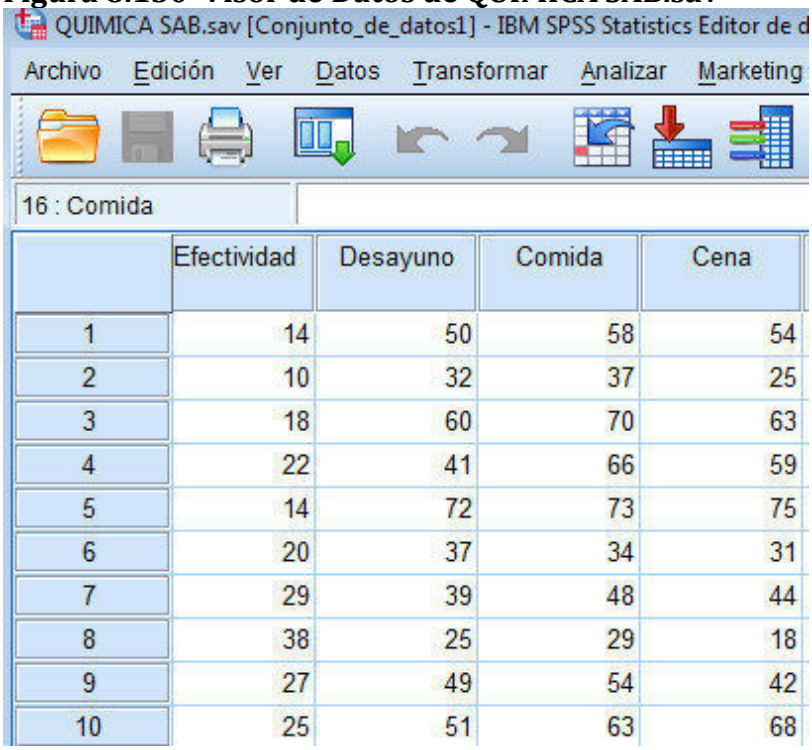
Figura 8.129 Visor de Variables de QUIMICA SAB.sav



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida
9	Desayuno	Numérico	2	0	Evaluacion de producto desayuno	Ninguna	Ninguna	8	Derecha	Ordinal
10	Comida	Numérico	2	0	Evaluación de producto comida	Ninguna	Ninguna	8	Derecha	Ordinal
11	Cena	Numérico	2	0	Evaluación de producto comida	Ninguna	Ninguna	8	Derecha	Ordinal

Fuente: SPSS 20 IBM

Figura 8.130 Visor de Datos de QUIMICA SAB.sav



	Efectividad	Desayuno	Comida	Cena
1	14	50	58	54
2	10	32	37	25
3	18	60	70	63
4	22	41	66	59
5	14	72	73	75
6	20	37	34	31
7	29	39	48	44
8	38	25	29	18
9	27	49	54	42
10	25	51	63	68

Fuente: SPSS 20 IBM

Paso 2: Diseño

- En adelante para efectos de aprendizaje, sólo se tomarán de forma directa los datos con el fin de aprender a utilizar la técnica.
- La prueba de *Friedman* será adecuada para la mayoría de los análisis estadísticos, aunque existen las siguientes oportunidades están disponibles en **SPSS**:
- ***La prueba estadística W de Kendall*** produce un coeficiente de concordancia entre los evaluadores. Los
- coeficientes oscilan entre **0 y 1**, siendo **1** un acuerdo completo y **0** sin acuerdo.

- **La prueba estadística Q de *Cochran*,** la cual es una extensión de las pruebas de ***McNemar***. Esta opción se recomienda usar cando hay más de **dos variables que son categóricas o dicotómicas por naturaleza.**
- El cálculo por defecto realizado para el ***p* valor** es el ***p* asintótico**, que es una estimación del verdadero ***p* valor**. Este es generalmente un método adecuado para calcular la ***p* valor** pero es posible calcularlo exactamente. Esto se puede lograr a través de oprimir el **botón Exacta.**
- Se prefiere el método de ***Monte Carlo*** para el cálculo de ***p* valor,** si no es posible calcularlo, ya que **el tamaño de la muestra es demasiado grande** (simplemente tomaría demasiado tiempo hacer ejercicio). El método de ***Monte Carlo*** da una **estimación imparcial del valor exacto de *p*.**
- En el botón **Opciones** es la selección para calcular las medias y desviaciones estándar. Mientras que estos Estadística descriptiva se puede realizar en conjuntos de datos ordinales, se recomienda Grado de cautela al calcularlas e interpretarlas.

Paso 3: Condiciones de Aplicabilidad

- No olvidar realizar los análisis previos de inspección visual de la base de datos para detectar ausentes y/o atípicos (corregir en su defecto), confiabilidad, normalidad, homocedasticidad y linealidad.

Paso 4: Estimación y ajuste

Teclear: Analizar->Pruebas no paramétricas->Cuadros de diálogo antiguos->K muestras relacionadas-> Variables de contraste: Evaluación de producto desayuno; Evaluación de producto comida; Evaluación de producto cena->Tipo de prueba: Friedman->Aceptar. Ver Figura 8.131

Figura 8.131. Proceso Prueba de Friedman para muestras relacionadas

Fuente: SPSS 20 IBM

Paso 5: Interpretación

La primera tabla generada por **SPSS** es la **tabla Rangos**, que es una descripción de los datos reportando el número de participantes así como el **Rango promedio** de cada grupo. **Ver Figura 8.132**

Figura 8.132. Tabla Rangos

Rangos

	Rango promedio
Evaluacion de producto desayuno	1.50
Evaluación de producto comida	2.70
Evaluación de producto comida	1.80

Fuente: SPSS 20 IBM

- El **Rango promedio** indica lo propio de los puntajes dentro de cada grupo.
- **Si no hubiera diferencias entre los puntajes de la calidad de los productos alimenticios en las tres comidas, la hipótesis nula sería verdadera y esperaríamos a que el rango promedio fuera aproximadamente igual entre las 3 comidas.**
- Podemos ver a partir de nuestro ejemplo anterior que **las tres comidas no parecen ser iguales, ya que la Evaluación de producto comida, es la que recibe una calificación más alta que las otras dos.**

Con el fin de determinar si la diferencia en estos puntajes es **significativa**, debemos analizar la **tabla Estadísticos de contraste**.**Ver Figura 8.133.**

Figura 8.133. Tabla Estadísticos de contraste

Estadísticos de contraste[a]

N	10
Chi-cuadrado	7.800
gl	2
Sig. asintót.	.020

a. Prueba de Friedman

Prueba estadística

***p* Valor**

Fuente: SPSS 20 IBM

- El estadístico de prueba a reportar es el ***Chi-cuadrado* de *Friedman***, que en el ejemplo anterior tiene un valor de **7.800.**
- La importancia del resultado se verifica examinando el **p Valor o Sig. asintót.**
- En el ejemplo, la sintaxis a reportar es: ***χ 2 = 7.800, gl = 2; p=0.02 < 0.05.***
- **Conclusión:** El disfrute de la comida se ve afectada por la hora del día, con la Evaluación producto comida con la calificación más alta.
- La **prueba de *Friedman*** en **SPSS NO** ofrece la oportunidad de realizar pruebas de comparación múltiple ***post hoc***. Hay una serie de pruebas que podrían hacerse, como la **prueba de *Nemenyi*, pero este se tendría que realizar fuera de SPSS.**

Referencias

Cattell, R. B., ed. (1966), *Handbook of Multivariate Experimental Psychology*. Chicago: Rand Mc- Nally.

Cohen, J. (1977), *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

Cole, D. A, Maxwell, S. E., Avery, R., y Salas, E. (1994), How the Power of MANOVA Can 8oth lncrease and Decrease as a Function of the Intercorrelations among Dependent Variables. *Pychological Bulletin* 115: 465-74.

Cooley, W. W., y Lohnes, P. R. (1971), *Multivariate Data Analysis*. New York: Wiley.

Oreen, P. E., y Tull, D. S. (1979), *Research for Marketing Decisions*, 3d ed. Upper Saddle River, N.J.: Prentice Hall.

Hair , J.F.; Anderson, R.E.; Tatham, R.L.; Black W.C. (1999).*Análisis Multivariante.*5a. Ed. España. Prentice Hall.

Harris, R. J. (1975), *A Primer of Multivariate Statistics*. New York: Academic Press.

Hubert, C. J., y MmTis, J. D. (1989). Multivariate Analysis versus Multiple Univariatc Analyses. *Psvchological Bulletin* 105: 302-8.

Huitema. B. (1980), The Analisis of Covriance all(/A!ternatives. New York: Wiley.

IBM (2011a). *IBM SPSS Statistics Base 20.* EUA. .Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Base.pdf

IBM (2011b).*Guía breve de IBM SPSS Statistics 20.* EUA.Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Brief_Guide.pdf

IBM (2011c). IBM SPSS Missing Values 20. EUA. .Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Missing_Values.pdf

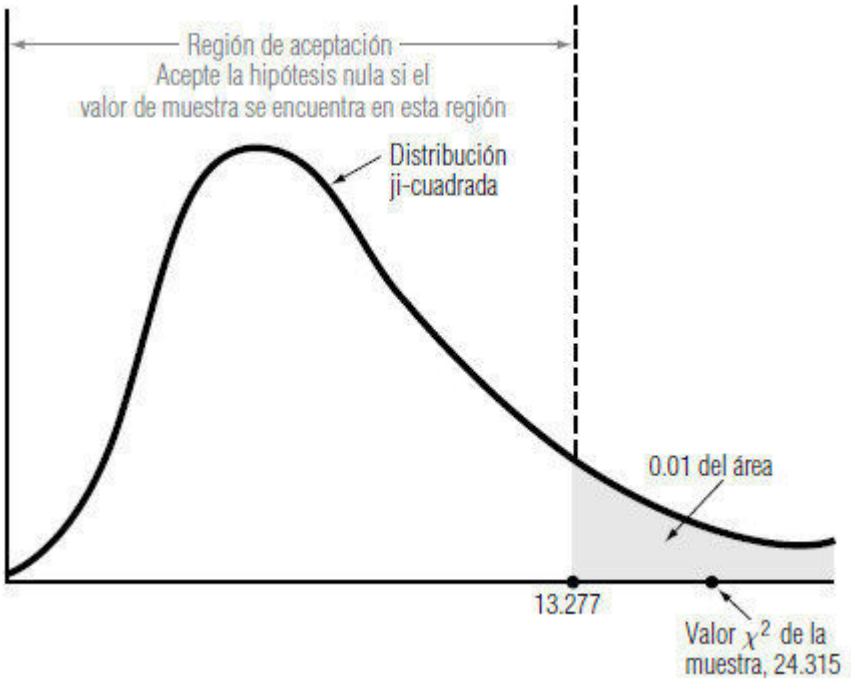
Koslowsky, M., y Caspy, T. (1991). Analysis of Variance. *Joumal of Organizational Behavior* 12: 555-59.

Lauter, J. (1978), Sample Size Requirements for the F Test of MANOVA (Tables for One-Way Classification). *Biometrical Journal* 20: 389-406.

Meyers, J. L. (1975), *Fundamentals of Experimental Design*. Boston: Allyn y Bacon.

Morrison, D. F. (1967), *Multivariate Statistical Methods*. New York: McOraw-Hill.
18. Rao, C. R. (1978*), Linear Statistical lnference and its Application*, 2d ed. New York: Wilcy.
Stevens, J. P. (1972), Four Methods of Analyzing bct- ween Variations for the k-Group MANOVA Problcm. *Multivariate Behavinral Rescarch* 7 (October): 442-54.
Stevens, J. P. (1980), Power of the Multivariate Analysis of Variance Tests. *Psychological Bulletin* 88: 728-37.
Tatsuoka, M. M. (1971), *Multivariate Analysis: Techniques for Education and Psychological Research*. New York: Wiley.
Wilks, S. S. (1932), *Certain Generalizations in the Analysis of Variance*. Biometrika 24: 471-94.
Winer, B. J. (1962), *Statistical Principies in Experimental Design*. New York: McGraw-Hill.

Capítulo 9. Cruce-tabular y Chi-Cuadrada



9.1. Cruce-tabular y *Chi-cuadrada*: ¿Qué es?

Un investigador orientado al fenómeno de la innovación en las Pymes del sector electrónico, diseña y aplica un cuestionario para descubrir los principales factores que la originan e impulsan desde varios puntos de vista como lo es la estrategia, la cultura, la organización e incluso, la incertidumbre del entorno (Hinton et al. 2004). Habrá una gran cantidad de datos producidos, sobre todo, si hay una gran cantidad de indicadores para los gerentes que atienden la operación interna y externa de las Pymes. Una vez que los datos son capturados y procesados vía computador, se tendrán los primeros resultados a nivel de estadística descriptiva, involucrando edad, género, experiencia en el puesto, resultados financieros producto de la innovación, mercado potencial y cautivo satisfecho, etc. Sin embargo, el investigador querrá ir más allá de simplemente resumir los hallazgos de cada pregunta, por lo que es aquí donde entra en juego el **cruce-tabular** (para saber más, vea: IBM, 2011a; IBM, 2011b; IBM, 2011c) Podemos combinar los resultados de diferentes preguntas en una tabla, con los resultados de una pregunta a nivel de las filas y los resultados de otra pregunta a nivel de las columnas. El cruce-tabular más común es cuando los datos son recuentos de frecuencias, por lo que en nuestro ejemplo, el investigador puede hacer el **cruce-tabular** de la pregunta de "***género***"¿es hombre o mujer? , con la pregunta de **edad *"¿cuál es su edad en años?***" Esto nos proporcionará una tabla que muestra cuántos varones y mujeres hay por edad. Podemos **crear una tabla con más de dos variables** (preguntas o indicadores). Podríamos añadir los resultados de otra pregunta **posgrado**: "***¿tiene posgrado?***" a la tabulación. Ahora esto puede parecer difícil de imaginar, ya que donde se escribe esto, de inicio sólo tiene dos dimensiones y, después de colocar las dos primeras variables como filas y columnas, la tercera variable necesita una tercera dimensión. Hay maneras de dibujar

el gra fico (como podemos dibujar un objeto tridimensional como un cubo en un pedazo de papel). Una forma es dibujar una tabla de **género vs edad** para aquellos que tienen **posgrado** y una segunda tabla de **edad vs género** para aquellos que no lo tienen. Podemos referirnos a estas dos tablas como "***capas***" producidas por la **tercera variable**. Estamos produciendo un cruce-tabular con **3 variables** an adiendo e sta como una capa despue s de usar las filas y columnas. Podemos agregar ma s variables, si se desean como capas adicionales. As que, aunque estamos creando una tabla con muchas variables, **NO** es un problema para **SPSS** porque agrega las variables como **capas** a la tabla. Ahora, **¿por qué debemos de producir un cruce-tabular?** La respuesta es **por que nos permitirá examinar la asociación entre las variables. Podríamos haber pre-planeado Predicciones o podríamos simplemente examinar si existe evidencia de una asociación.** Tomemos algunas variables diferentes del cuestionario de investigacio n. Una pregunta es "***¿cómo acostumbra más informarse del estado de innovación de su compañía: por medios físicos, electrónicos u otros***?" Queremos ver si esto difiere segu n la edad, y para simplificar las cosas podr amos decidir que agregar amos los resultados de los 25 a los 30 an os de edad como **gerentes jóvenes** y de los de 31-35 años más de edad como **gerentes maduros**. Imagine que la mayor a de las respuestas se dieron por **medios electrónicos.** Así, se tendrá en el **cruce-tabular**, en las **columnas: medios físicos, electrónicos u otros y en filas: gerentes jóvenes y gerentes maduros.** Los datos de la tabla serán el número de personas en cada grupo de elección de medios. Podríamos tener una predicción de que los **medios electrónicos** serán elegidos por una mayor cantidad de **gerentes jóvenes y el resto por los gerentes maduros , pero No tiene que hacerlo.** Cuando probamos la asociacio n entre la **edad** y **el medio de cómo informarse del estado de innovación de la empresa** estamos examinando tambie n la **hipótesis nula de que no hay diferencia entre los grupos de edad vs la elección de del medio.** Así que, **cuando la hipótesis nula es verdadera, deberíamos ver aproximadamente la misma proporción de cada edad eligiendo cada una de los medios de información, pero ¿qué prueba estadística lo realiza mejor?**

La prueba estadística del *Chi-cuadrado* examina estas proporciones y presenta la probabilidad de obtener este patro n cuando no hay diferencia en de opciones. **Un gran valor del *Chi-cuadrado* indica una gran diferencia entre los grupos. Si La probabilidad es muy pequeña (p <0.05), entonces podemos concluir que hay una diferencia en la elección del medio de información para los diferentes grupos de edad.** Si hubie ramos hecho una prediccio n espec fica que ma s del grupo de **gerentes jóvenes elegirían los medios electrónicos**, entonces estar amos haciendo un **esfuerzo unilateral (Una cola o direccional)** y tendr amos que **comprobar los resultados para asegurarnos que el patrón predicho es el que surgió en lugar de alguna otra diferencia**. **La suposición clave de la prueba del *Chi-cuadrado* es que hay independencia de las observaciones,** es decir, los resultados en cada seccio n **(o "*célula*")** de la tabla son independientes entre s . As que si 45 **gerentes jóvenes** eligen medios de **información electrónicos** y **28 eligen a los medios de información físicos**, los **45 y 28 no tienen relación**. No podr amos permitir que un encuestado dijera que le gustaban los dos medios, igualmente y anotarlos en ambas células.

Hay una segunda suposicio n de que el ***Chi-cuadrado* calculado** es **"*continuo*"** as como el cuadrado la distribucio n **es continua.**A fin de satisfacer este supuesto es mejor tener

grandes números en cada celda en lugar de números pequeños. Cuando la frecuencia "***esperada***" de una célula es inferior a 5 nos preocupa que esta suposición sea violada. Además, cuando tener una tabla con sólo dos columnas y dos filas (**2 × 2**) existe un riesgo real de que violan esta suposición. Para compensar esto se genera un valor de "***corrección de continuidad***" en **SPSS, por lo que debe tener cuidado al analizar una tabla 2 × 2**. Viéndolo positivamente, con una tabla tan pequeña, **la prueba exacta de Fisher** (también producido por **SPSS**) se puede utilizar para un cálculo exacto.
Por lo general, con tablas de ma s de**2 × 2**y **frecuencias celulares mayores de 5 (y preferiblemente más de 10)** usamos el valor ***Chi-cuadrado*** estándar. Si las celdas tienen números pequeños (esperado Frecuencia < 5), pero la **tabla es grande,** entonces podría valer la pena recopilar más datos o, si No es posible, combinando celdas. Si estuviéramos buscando en las **medios de información favoritos**, en vez de analizar cada **medio individualmente**, podríamos agruparlas en **tablet, smartphone, laptop, etc.** para hacer que los números en cada celda sean más grandes. Si no queremos hacer esto, se debe elegir el valor de **"*corrección de continuidad*"**. Si tenemos una **tabla de 2 × 2** tomamos la el **valor exacto de *Fisher***, ya que es so lo eso, un valor exacto para nuestra tabla. Finalmente, hemos visto que el ***Chi-cuadrado*** compara patrones de resultados para ver si s**similares o diferentes**. Tambie n podemos usar la prueba del ***Chi-cuadrado*** como una**prueba de bondad de ajuste.** Esto significa *"¿el patrón de resultados coincide (o encaja) con un patrón predicho?"* Por ejemplo, si creen que los gerentes más jóvenes preferirán a los medios de información electrónicos sobre los **medios de información físicos** por 2 a 1. Entonces tenemos un modelo donde esperamos que haya **dos** personas eligiendo a **medios electrónicos de información vs, cada uno eligiendo medios físicos** . Ponemos las predicciones del modelo en nuestro tabla de datos de **SPSS**. Podemos entonces comparar los hallazgos reales (**las frecuencias**) con el modelo, para ver si se ajusta a los datos. En este caso. (Hinton et al., 2005; Levin yRubin, 2004).

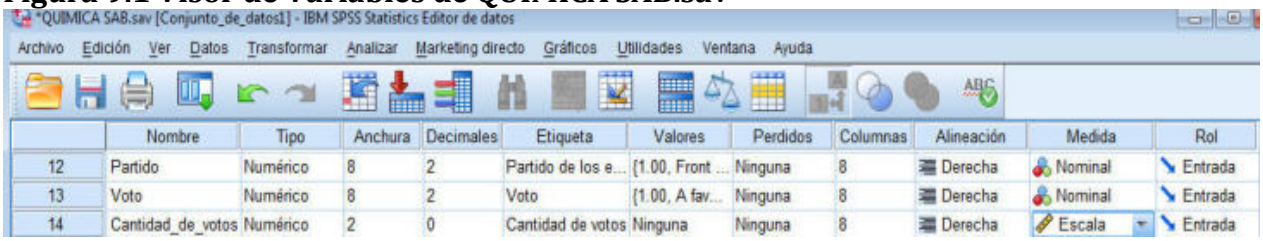
1. **Un valor significativo del *Chi-cuadrado*** nos diría que los datos **son significativamente diferentes al patrón predicho por el modelo.**
2. **Un valor NO significativo *Chi-cuadrado*** indicaría que **los datos NO difieren** significativamente de los modelos patrón predicho.

9.2. Cruce-tabular y *Chi-Cuadrada*: Ejemplo 1

Paso 1: Objetivos

Problema 1: La empresa **QUIMICA SAB**, desea realizar un estudio de opinión interno acerca de la fusión se realizaría con la empresa número 2 del ramo de biotecnología , entre los empleados considerados liberales ("***front office***") y los conservadores ("***back office***"). Se identifican a 120 empleados como conservadores y 80 como liberales. La pregunta a contestar es: ¿estaría de acuerdo d la fusión con la empresa X? el diseño de respuestas en 3: de acuerdo, en desacuerdo y no sé. El investigador cree que la afiliación por de los encuestados (conservador/liberal) influye en su visión de fusión. **Ver Figuras 9.1 y 9.2**

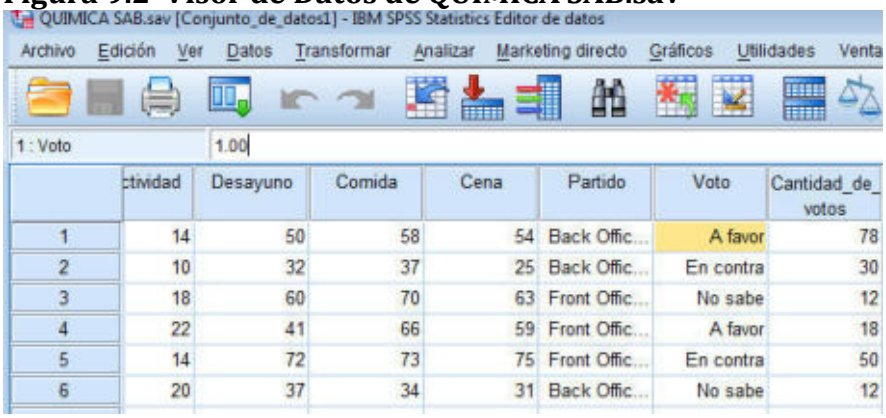
Figura 9.1 Visor de Variables de QUIMICA SAB.sav



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
12	Partido	Numérico	8	2	Partido de los e...	{1.00, Front ...	Ninguna	8	Derecha	Nominal	Entrada
13	Voto	Numérico	8	2	Voto	{1.00, A fav...	Ninguna	8	Derecha	Nominal	Entrada
14	Cantidad_de_votos	Numérico	2	0	Cantidad de votos	Ninguna	Ninguna	8	Derecha	Escala	Entrada

Fuente: SPSS 20 IBM

Figura 9.2 Visor de Datos de QUIMICA SAB.sav



	ctividad	Desayuno	Comida	Cena	Partido	Voto	Cantidad_de_votos
1	14	50	58	54	Back Offic...	A favor	78
2	10	32	37	25	Back Offic...	En contra	30
3	18	60	70	63	Front Offic...	No sabe	12
4	22	41	66	59	Front Offic...	A favor	18
5	14	72	73	75	Front Offic...	En contra	50
6	20	37	34	31	Back Offic...	No sabe	12

Fuente: SPSS 20 IBM

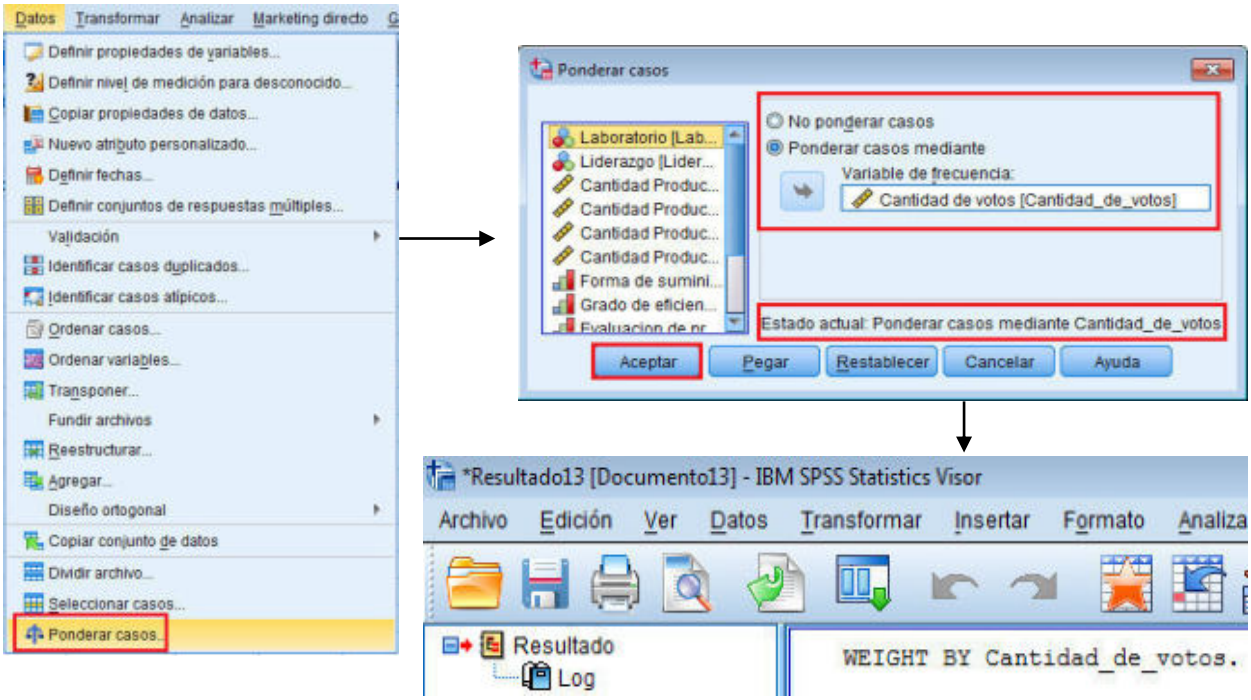
Paso 2: Diseño

- En adelante para efectos de aprendizaje, sólo se tomarán de forma directa los datos con el fin de aprender a utilizar la técnica.
- Sólo para ingresar los datos y analizarlos, se puede realizar por 2 métodos:

Método 1: Se ingresan los datos por cada persona por separado, recordando codificar adecuadamente las categorías, ya que usaremos datos nominales. Recuerde que la selección de las etiquetas **Ver** y **Valor** puede mostrar las etiquetas de valor dadas en el procedimiento de introducción de datos.

Método 2, que implica tomar del estadístico datos para manipular la secuencia de comandos, **Teclear: Datos-> Ponderar casos->Variable de frecuencia: Cantidad de votos ->Aceptar. Ver de la Figura 9.3**

Figura 9.3. Proceso de entrada de datos para el cruce-tabular y la *Chi-cuadrada* con ponderación de casos



Fuente: SPSS 20 IBM

Nota: Usted puede estar seguro de que el procedimiento se ha llevado a cabo con éxito al verificar que muestre **Estado actual: Ponderar casos mediante Cantidad_de_votos**

Paso 3: Condiciones de Aplicabilidad

- No olvidar realizar los análisis previos de inspección visual de la base de datos para detectar ausentes y/o atípicos (corregir en su defecto), confiabilidad, normalidad, homocedasticidad y linealidad.

Paso 4: Estimación y ajuste

Para producir la tabla de **cruce-tabular** y el ***Chi-cuadrado***, siga el procedimiento:

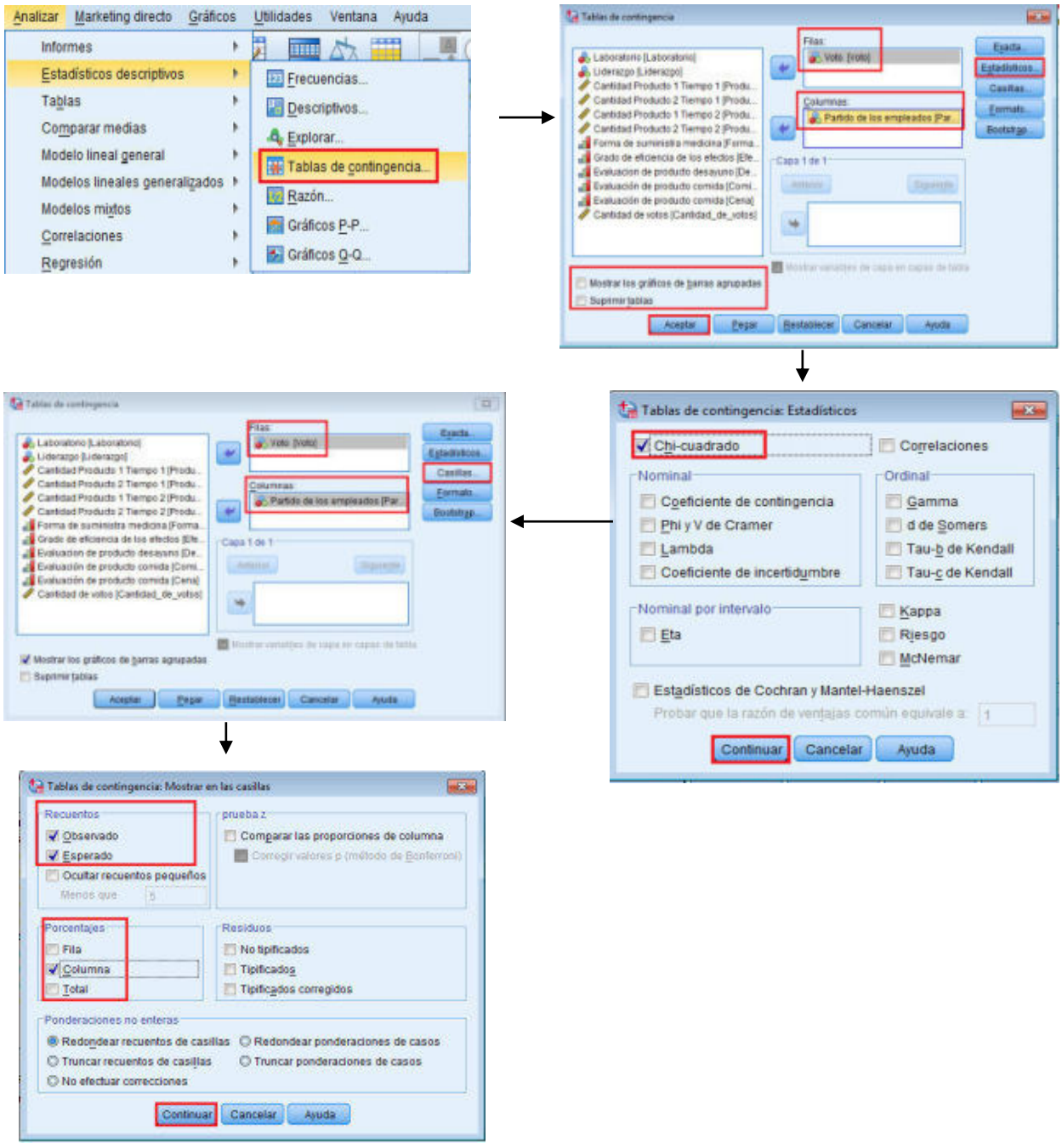
-Teclear: Analizar->Estadísticos descriptivos->Tablas de contingencia.>* Filas: Voto->Columnas: Partido de los empleados-> Estadísticos->*Chi-cuadrada*-Continuar->Casillas->Recuentos; Observado; Esperado->Porcentaje: Columna->Continuar->Aceptar**

Nota: * Variable influenciada; ** Variable influenciadora (normalmente en la columna ya que, por convención, la lectura se hace a través de la tabla y no se hace en otro sentido).

Un gráfico de barras agrupado es el modo más apropiado de mostrar los resultados de su **cruce-tabular**, y esto puede ser generado como una parte integral del procedimiento. Si tu desea crear este gráfico, coloque una marca contra Mostrar los gráficos de barras agrupados.

• Recuerde que cuando realizamos una tabla cruzada y ***Chi-cuadrado*** **estamos comparando cuántas personas tienen en una categoría con cuántas personas esperamos tengan esa categoría si la hipótesis nula es verdadera. Ver Figura 9.4**

Figura 9.4. Proceso de cálculo de datos para el cruce-tabular y la *Chi-cuadrada*



Fuente: SPSS 20 IBM

Paso 5: Interpretación

- La primera tabla que genera **SPSS**, es **Resumen del procesamiento de los casos**, el cual reporta una descripción del conjunto de datos. **Ver Figura 9.5.**

Figura 9.5. Tabla Resumen del Procesamiento de los casos

Resumen del procesamiento de los casos

	Casos					
	Válidos		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Voto * Partido de los empleados	200	100.0%	0	0.0%	200	100.0%

Fuente: SPSS 20 IBM

- Esta tabla nos reporta del número y el porcentaje de las puntuaciones faltantes en nuestro conjunto de datos. Como podemos ver, no hay puntuaciones en nuestro conjunto de datos que falten y por lo tanto estamos considerando todos los **200 participantes** al hacer juicios sobre la asociación entre las dos variables.
- Otra tabla generada por **SPSS**, es la de **Contingencia filas*columnas. Ver Figura 9.6**

Figura 9.6. Tabla de contingencia filas*columnas Cruce tabular 2x3 y *Chi-cuadrado*

Tabla de contingencia Voto * Partido de los empleados

Variable influenciadora

			Partido de los empleados		
			Front Office= Liberal	Back Office=Conservador	Total
Voto	A favor	Recuento	18	78	96
		Frecuencia esperada	38.4	57.6	96.0
		% dentro de Partido de los empleados	22.5%	65.0%	48.0%
	En contra	Recuento	50	30	80
		Frecuencia esperada	32.0	48.0	80.0
		% dentro de Partido de los empleados	62.5%	25.0%	40.0%
	No sabe	Recuento	12	12	24
		Frecuencia esperada	9.6	14.4	24.0
		% dentro de Partido de los empleados	15.0%	10.0%	12.0%
Total		Recuento	80	120	200
		Frecuencia esperada	80.0	120.0	200.0
		% dentro de Partido de los empleados	100.0%	100.0%	100.0%

Variable dependiente

El conteo que esperaríamos encontrar si la hipótesis nula fuera verdadera

Número total de participantes

Fuente: SPSS 20 IBM

- El **Recuento** representa los conteos reales de la frecuencia obtenida en la encuesta.

- La **Frecuencia esperada** representa cuántos esperaríamos encontrar en esta combinación si la **hipótesis nula es verdadera**, es decir, si la distribución de frecuencias fuera **completamente aleatorio.**
- Los porcentajes son una mejor representación de los datos debido a los números desiguales en las categorías. Todas las columnas suman un total del **100%** como pedimos porcentajes para las columnas, que es donde enviamos la variable influenciadora **Partido de los empleados**
- Los porcentajes representan lo que el recuento es como una proporción del número total de la columna, por ejemplo **18** votantes liberales quienes están a favor de la fusión, lo cual representa el **22.5%** del total de electores liberales (Front Office), es decir, **80.**
- Los porcentajes totales, que aparecen en la columna **Total**, representan el porcentaje de la población que caía en cada categoría de preferencia, por ejemplo el **48.0 %** de la población total eran para la fusión de la compañía con independencia de la preferencia de voto.
- Como hemos totalizado nuestros porcentajes en las columnas debido a la colocación aquí de nuestra **variable de influencia**, leemos la tabla a continuación para determinar si existe una diferencia entre los porcentajes. Si parece haber una diferencia, entonces no puede ser una asociación entre las variables. La importancia de esto se evalúa a través de la prueba estadística ***Chi-cuadrado*.**
- En el ejemplo anterior podemos ver que dentro del grupo de personas que estaban a favor de **fusión**, el **65.0%** de ellos eran empleados conservadores (Back Office), frente al **22.5%** que eran liberales (Front Office). Esta tendencia se invierte dentro del grupo que está en contra de la fusión, con el **25.0%** siendo conservador (Back Office) y el **62.5%** liberal (Front Office). El grupo que está dudas respecto de sus opiniones sobre la fusión están razonablemente similares en sus votaciones (**10.0 % conservador y 15.0 % liberal**).
- Un examen de la tabla de los Pruebas de ***Chi-cuadrada*** nos permitirá determinar si la patrones identificados anteriormente son significativos**. Ver Figura 9.7.**

Figura 9.7. Tabla pruebas de *Chi-cuadrado*

Prueba estadistica **Pruebas de chi-cuadrado** ***p* Valor**

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	35.938[a]	2	.000
Razón de verosimilitudes	37.429	2	.000
Asociación lineal por lineal	22.908	1	.000
N de casos válidos	200		

a. 0 casillas (0.0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 9.60.

Fuente: SPSS 20 IBM

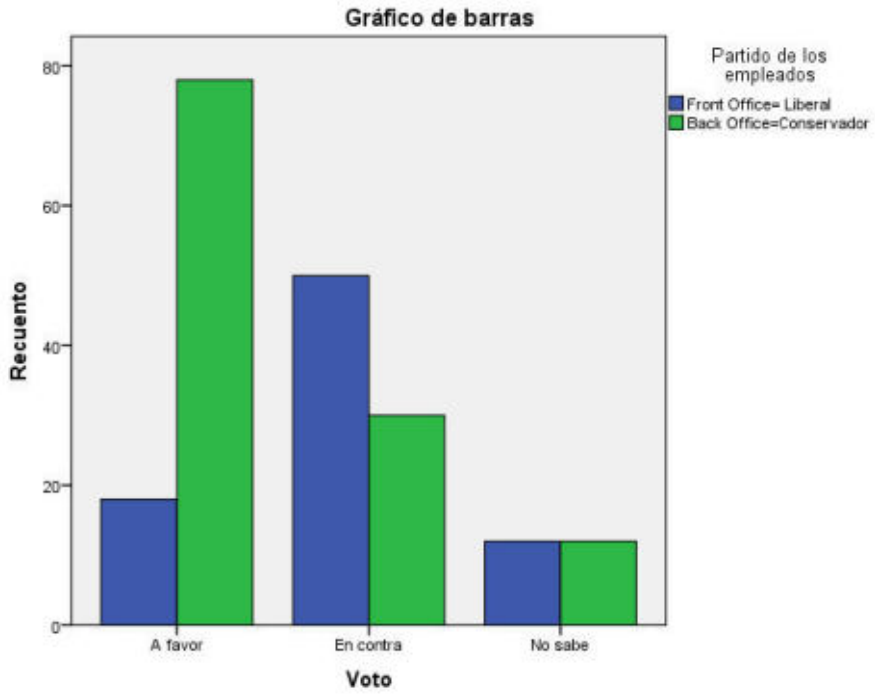
- La prueba estadística usualmente elegida para determinar si la asociación descrita por

el **cruce-tabular** es significativa, es el ***Chi-cuadrado* de *Pearson*= 35.938**, con **2** grados de libertad, que es significativa en el nivel de ***p* <0.001.**

- Los resultados del ***Chi-cuadrado*** se presentan típicamente como: ***χ 2 = 35.938, gl = 2, p < 0.001***
- De la tabla anterior se puede observar que el estadístico ***Chi-cuadrado*** **asume un comportamiento no direccional de la hipótesis,** que es indicada por el **Sig. asintótica (bilateral)** y se refiere a las aproximaciones realizadas en nuestros cálculos de ***p* valor**.
- La **Razón de verosimilitud** es una prueba alternativa a la ***Chi-cuadrada*** empleado un método diferente. Normalmente usamos el resultado del ***Chi-cuadrada***, pero ésta estadística a veces se prefiere **cuando el tamaño de la muestra es pequeño**.
- **La Asociación lineal por lineal** se usa cuando las categorías **son ordinales** y existe el interés de **encontrar una tendencia.**
- La nota en la parte inferior de la tabla de pruebas de ***Chi-cuadrado*** indica si alguna celda individual en el cruce-tabular tiene un conteo de **menos de cinco**. Para una **tabla 2 × 2**, **SPSS** realizará automáticamente **una corrección de continuidad**. Para una tabla más grande, si Usted tiene un número de valores **menor de cinco,** puede desear combinar las celdas para aumentar el tamaño.

- **SPSS** genera una última tabla es un resumen del conteo de frecuencias en la forma de es un resumen de los recuentos de frecuencia en forma de un gráfico de barras agrupado, lo cual proporciona una representación útil de los datos. Podemos ver cómo Las frecuencias de las diferentes opciones están vinculadas a los grupos de empleados. **Ver Figura 9.8**

Figura 9.8. Gráfico



Fuente: SPSS 20 IBM

9.3. Cruce-tabular y *Chi-Cuadrada*: Ejemplo 2

Paso 1: Objetivos

Problema 2: La empresa **MKT Digital**, tiene dos TV blogs: A y B que tratan asuntos de actualidad del software para aplicaciones, ambos muy populares en el país.. La empresa está interesada en saber si dicha popularidad está basada en la asociación que tiene el participante de vivir en las 2 regiones en las que está dividido el país: norte-sur.

Paso 2: Diseño

- En adelante para efectos de aprendizaje, sólo se tomarán de forma directa los datos con el fin de aprender a utilizar la técnica.
- Las dos primeras tablas tienen el mismo diseño que la **tabla de 2 × 3**, indicando el número de respuestas y cualquier posible falta de datos, junto con la tabla de **cruce-tabular** de frecuencias de las dos variables.

Paso 3: Condiciones de Aplicabilidad

- No olvidar realizar los análisis previos de inspección visual de la base de datos para detectar ausentes y/o atípicos (corregir en su defecto), confiabilidad, normalidad, homocedasticidad y linealidad.

Paso 4: Estimación y ajuste

- Se realizar de forma similar al ejemplo 1.

Paso 5: Interpretación

- Las tablas que produce **SPSS**, son **Resumen del procesamiento de los datos**, con una descripción del conjunto de datos. Ver **Figura 9.9**

Figura 9.9. Tabla Resumen del Procesamiento de los casos

Resumen del procesamiento de los casos

	Casos					
	Validos		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
REGION * BLOG	26	100.0%	0	.0%	26	100.0%

Fuente: SPSS 20 IBM

- Esta tabla nos reporta del número y el porcentaje de las puntuaciones faltantes en nuestro conjunto de datos. Como podemos ver, no hay puntuaciones en nuestro conjunto de datos que falten y por lo tanto estamos considerando todos los **26 participantes** al hacer juicios sobre la asociación entre las dos variables

- Otra tabla generada por **SPSS**, es la de **Contingencia filas*columnas. Ver Figura 9.10.**

Figura 9.10. Tabla de contingencia filas*columnas Cruce tabular 2x2 y *Chi-cuadrado*

Tabla de contingencia REGION * BLOG

			TV		
			BLOG A	BLOG B	Total
REGION	NORTE	Recuento	10	3	13
		Frecuencia esperada	7.5	5.5	13.0
		% dentro de BLOG	66.7%	27.3%	50.0%
	SUR	Recuento	5	8	13
		Frecuencia esperada	7.5	5.5	13.0
		% dentro de BLOG	33.3%	72.7%	50.0%
Total		Recuento	15	11	26
		Frecuencia esperada	15.0	11.0	26.0
		% dentro de BLOG	100.0%	100.0%	100.0%

Fuente: SPSS 20 IBM

- Se observa de la tabla que las personas que viven en el **NORTE** son más propensas a participar en el **BLOG A** (**66.7 %**) que el **BLOG B** (**27.3 %**), siendo lo inverso en el **SUR.**
- Conlcusión: del cruce-tabular se puede ver que al parecer **existe una asociación entre la región en la que vive una persona y su preferencia por el blog**

Otra tabla generada por **SPSS**, es Pruebas de ***Chi-Cuadrada*. Ver Figura 9.11**

Figura 9.11. Pruebas de *Chi-cuadrada*

Prueba estadistica

***p* Valor**

Pruebas de Chi-cuadrada

	Valor	gl	Sig. asintótica (bilateral)	Sig. Exacta (bilateral)	Sig. Exacta (unilateral)
Chi cuadrada de Pearson	3.939[b]	1	.047		
Corrección de continuidad[a]	4.521	1	.112		
Razón de verosimilitudes	4.057	1	.044		
Prueba exacta de Fisher				.111	.055
Asociación lineal por lineal	3.788	1	.052		
N de casos válidos	26				

[a]. Computado sólo para tabla 2x2

[b]. 0 celdas (0%) se han esperado contar en menos que 5 . El mínimo conteo esperado es 5.50

Fuente: SPSS 20 IBM

- La información en esta tabla nos **permitirá juzgar si el patrón de visualización es significativamente diferente del patrón que esperaríamos por el azar.**

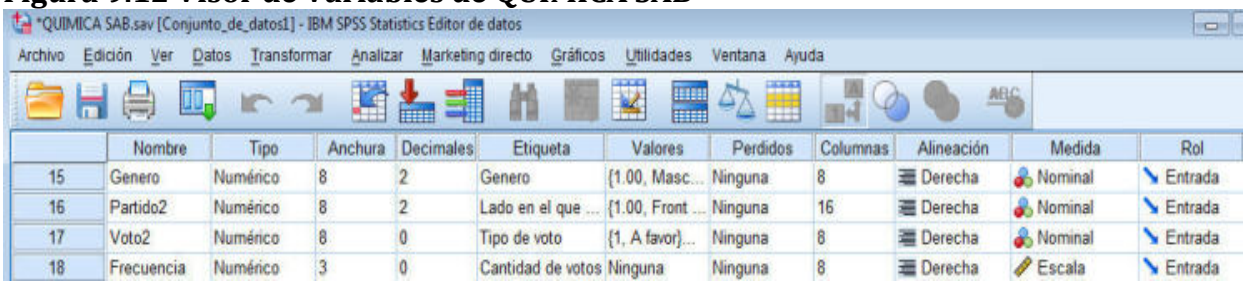
- El estadístico de prueba usualmente elegido para determinar si la asociación descrita por el **cruce-tabular** es significativa es el ***Chi-cuadrado* de *Pearson*,** el cual es **3.939**, con **1** grado de libertad, y como el ***p* valor** es menor que **0.05, podemos concluir que hay un patrón significativamente diferente de visualización en las dos regiones** (***p* <0.05**).
- Los resultados del ***Chi-cuadrado*** se presentan típicamente como sigue: $\chi 2 = 3.939, gl = 1, p=0.047 < 0.05$
- Tenemos una **cruce-tabular de 2 × 2** con valores bastante pequeños, más la probabilidad de que nuestro valor se encuentre cerca de un nivel de significatividad, por lo que se recomienda interpretar el resultado con precaución.
- La **Corrección de continuidad** es producida automáticamente por **SPSS** para una **tabla 2 × 2**, la cual realiza una **corrección de *Yates*.** Este es un ajuste conservador y su uso es controvertido, por lo que a menudo, **NO se recomienda**. Sin embargo, **SÍ se recomienda el uso de la Prueba exacta de *Fisher*** en el análisis de una **tabla 2 × 2 con una frecuencia de celda inferior a cinco** (vea abajo).
- La **Razón de verosimilitudes** es una prueba alternativa a la ***Chi-cuadrada*** que emplea un método diferente. Normalmente usamos el resultado del ***Chi-cuadrado*** pero la **Razón de verosimilitud** es preferido cuando el **tamaño de la muestra es pequeño.**
- La **Prueba exacta de Fisher** calcula la probabilidad exacta basada en un cálculo de todas las posibles combinaciones de la **distribución de frecuencia** en el **cruce-tabular**. **Esto sólo se calcula para una tabla 2 × 2 debido al aumento en las posibles combinaciones de las distribuciones de frecuencia proporcionadas por un cruce-tabular mayor**. Como se calcula la probabilidad real, a menudo es más conservadora. Como puede verse en el ejemplo, el valor ***Chi-cuadrado* de *Pearson* = 3.939**, con una probabilidad ***p* = 0.047**. Sin embargo, cuando los cálculos se basan en la **prueba exacta de *Fisher***, encontramos ***p*** = **0.111**.
- La **Asociación lineal por lineal** se utiliza cuando las categorías son **ordinales**. Esto es también interpretado de la misma manera que la estadística de la prueba del ***chi-cuadrado.***
- El **Sig. asintótica (bilateral)**, se calcula como la probabilidad por defecto por **SPSS**. Esta es una aproximación que se calcula para un conjunto de datos que es grande y se distribuye adecuadamente.

9.4. Cruce-tabular y *Chi-Cuadrada*: Ejemplo 3

Paso 1: Objetivos

Problema 3: Volviendo al problema de la empresa **Quimica SAB**, a veces es posible que desee **añadir una tercera variable** al cálculo de **cruce-tabular** para obtener una imagen más precisa de una asociación e investigar cualquier posible influencia de una tercera variable. El procedimiento para configurar el conjunto de datos y los pasos iniciales del análisis son idénticos a los descritos anteriormente. Una columna adicional debe ser añadido al conjunto de datos para la variable extra y las posibles combinaciones de los grupos creados. En este ejemplo, **N= 800.Ver Figura 9.12 y 9.13**

Figura 9.12 Visor de Variables de QUIMICA SAB

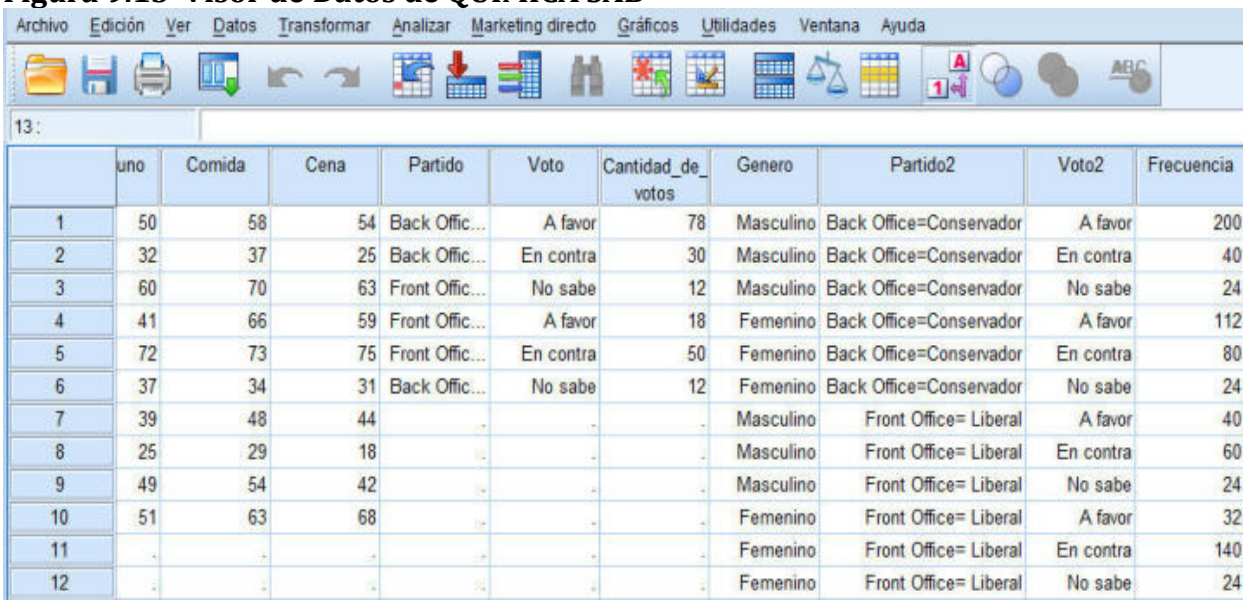

*QUIMICA SAB.sav [Conjunto_de_datos1] - IBM SPSS Statistics Editor de datos

Archivo Edición Ver Datos Transformar Analizar Marketing directo Gráficos Utilidades Ventana Ayuda

	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
15	Genero	Numérico	8	2	Genero	{1.00, Masc...	Ninguna	8	Derecha	Nominal	Entrada
16	Partido2	Numérico	8	2	Lado en el que ...	{1.00, Front ...	Ninguna	16	Derecha	Nominal	Entrada
17	Voto2	Numérico	8	0	Tipo de voto	{1, A favor}...	Ninguna	8	Derecha	Nominal	Entrada
18	Frecuencia	Numérico	3	0	Cantidad de votos	Ninguna	Ninguna	8	Derecha	Escala	Entrada

Fuente: SPSS 20 IBM

Figura 9.13 Visor de Datos de QUIMICA SAB


Archivo Edición Ver Datos Transformar Analizar Marketing directo Gráficos Utilidades Ventana Ayuda

13 :

	uno	Comida	Cena	Partido	Voto	Cantidad_de_votos	Genero	Partido2	Voto2	Frecuencia
1	50	58	54	Back Offic...	A favor	78	Masculino	Back Office=Conservador	A favor	200
2	32	37	25	Back Offic...	En contra	30	Masculino	Back Office=Conservador	En contra	40
3	60	70	63	Front Offic...	No sabe	12	Masculino	Back Office=Conservador	No sabe	24
4	41	66	59	Front Offic...	A favor	18	Femenino	Back Office=Conservador	A favor	112
5	72	73	75	Front Offic...	En contra	50	Femenino	Back Office=Conservador	En contra	80
6	37	34	31	Back Offic...	No sabe	12	Femenino	Back Office=Conservador	No sabe	24
7	39	48	44	.	.	.	Masculino	Front Office= Liberal	A favor	40
8	25	29	18	.	.	.	Masculino	Front Office= Liberal	En contra	60
9	49	54	42	.	.	.	Masculino	Front Office= Liberal	No sabe	24
10	51	63	68	.	.	.	Femenino	Front Office= Liberal	A favor	32
11	.	.	.	.	.	.	Femenino	Front Office= Liberal	En contra	140
12	.	.	.	.	.	.	Femenino	Front Office= Liberal	No sabe	24

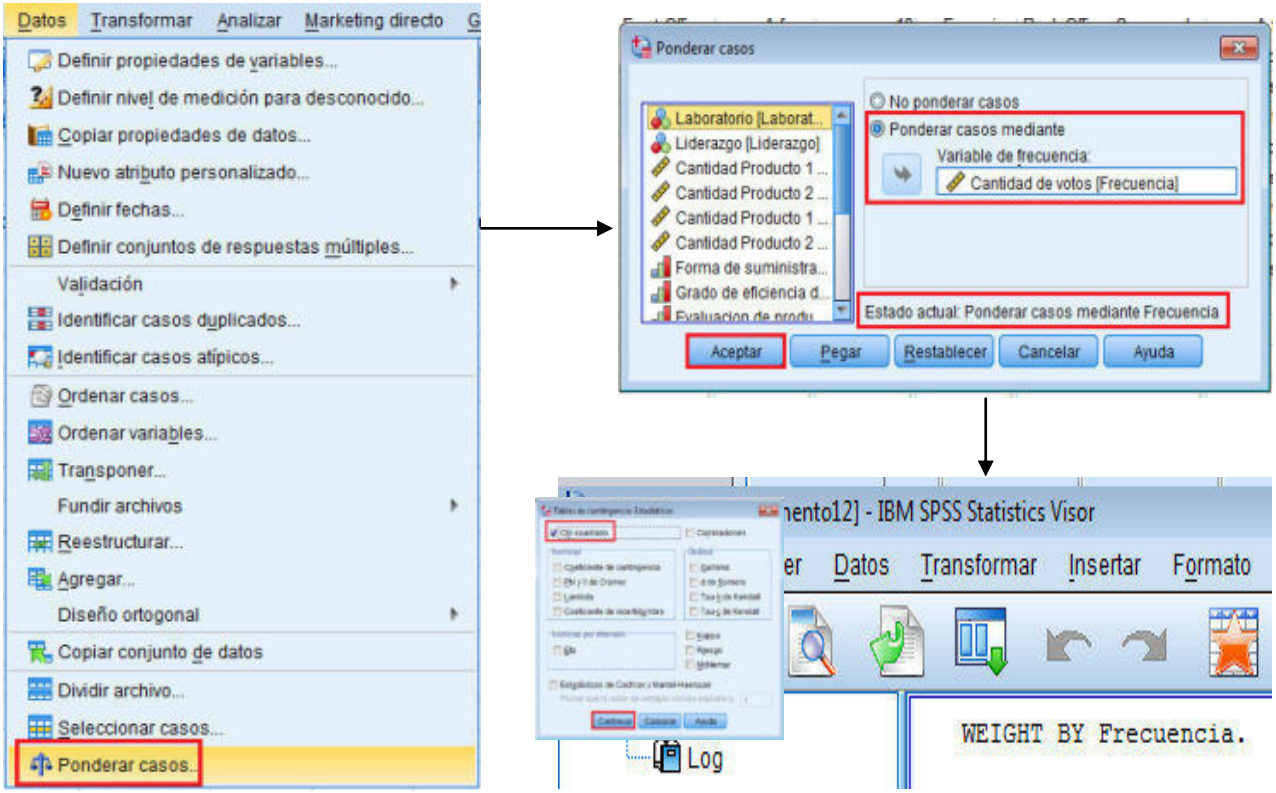
Fuente: SPSS 20 IBM

Paso 2: Diseño

- En adelante para efectos de aprendizaje , sólo se tomarán de forma directa los datos con el fin de aprender a utilizar la técnica.
- Este ejemplo se refiere al diseño previo de una **tabla de 2 × 3**, indicando el número de respuestas y cualquier posible falta de datos, junto con la tabla de **cruce-tabular** de frecuencias de las **dos variables.**
- Las diferencias en el procedimiento surgen cuando se crean **capas** para nuestro **cruce-tabular**, formadas por la adición de una **tercera variable**. En nuestro ejemplo seguimos analizando las preferencias de voto para fusionar la empresa o no y su relación con el **partido** de empleados que hay (liberales –***front-office***- o conservadores -***back-office***-) . Sin embargo, deseamos investigar esta relación analizando a influencia que el **género** puede tener. Así, Partido y Voto son relacionados y previamente descritos en el cruce-tabular anterior y en nuestro caso, se ingresa la variable **género**, como **primera capa.**

Se deberá actualizar el Método 2, que implica tomar del estadístico datos para manipular la secuencia de comandos, **Teclear: Datos-> Ponderar casos->Variable de frecuencia: Cantidad de votos ->Aceptar. Ver Figura 9.14**

Figura 9.14. Proceso de entrada **de datos para el cruce-tabular y la *Chi-cuadrada* con ponderación de casos**



Fuente: SPSS 20 IBM

Nota: Usted puede estar seguro de que el procedimiento se ha llevado a cabo con éxito al verificar que el Estado actual muestre, **Estado actual: Ponderar casos mediante Frecuencia.**

Paso 3: Condiciones de Aplicabilidad

- No olvidar realizar los análisis previos de inspección visual de la base de datos para detectar ausentes y/o atípicos (corregir en su defecto), confiabilidad, normalidad, homocedasticidad y linealidad

Paso 4: Estimación y ajuste

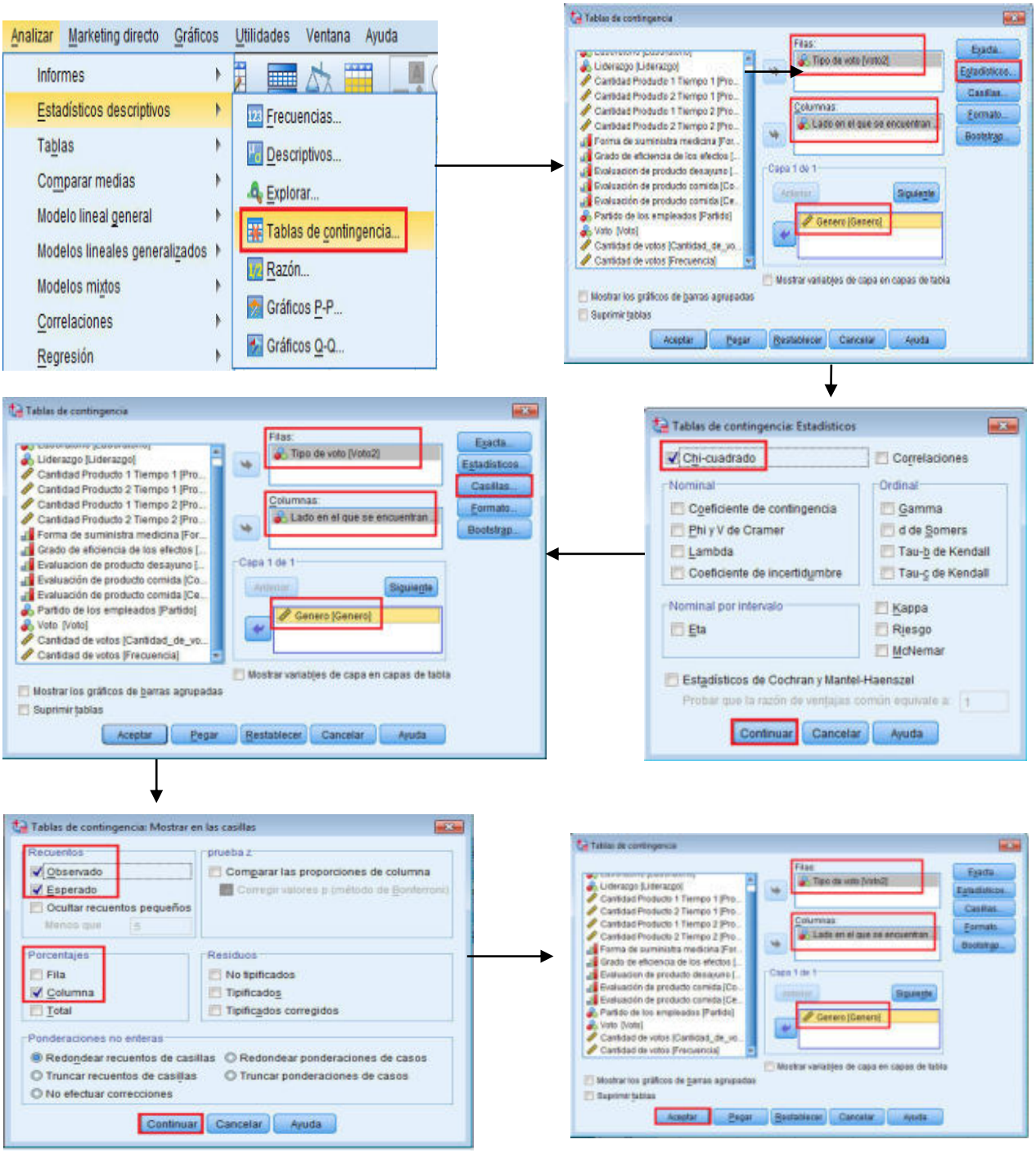
Para producir la tabla de **cruce-tabular** y el ***Chi-cuadrado***, siga el procedimiento:

-Teclear: Analizar->Estadísticos descriptivos->Tablas de contingencia.>* Filas: Voto->Columnas: Partido de los empleados-> Capa 1 de 1|: Género->Estadísticos->Chi-**

cuadrada->Continuar->Casillas->Recuentos; Observado; Esperado->Porcentaje: Columna->Continuar->Aceptar. Ver Figura 9.15.

Nota: * Variable influenciada; ** Variable influenciadora (normalmente en la columna ya que, por convención, la lectura se hace a través de la tabla y no se hace en otro sentido).

Figura 9.15. Proceso de cálculo de datos para el cruce-tabular y la *Chi-cuadrada*



Fuente: SPSS 20 IBM

Paso 5: Interpretación

La tabla **Resumen de procesamiento de los casos** indica de nuevo el número de casos en nuestro conjunto de datos. Como antes, no tenemos casos que falten y por lo tanto estamos considerando los **800** casos al hacer nuestros juicios sobre la relación entre los tres variables**. Ver Figura 9.16**

Figura 9.16. Tabla Resumen del Procesamiento de los casos

Resumen del procesamiento de los casos

	Casos					
	Válidos		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Tipo de voto * Lado en el que se encuentran los empleados * Genero	800	100.0%	0	0.0%	800	100.0%

Fuente: SPSS 20 IBM

- La **tabla de contingencia** indica cualquier patrón en nuestros datos. Como podemos ver esto es En la tercera variable **"*Género*". Ver Figura 9.17**

Figura 9.17 Tabla de contingencia filas*columnas Cruce tabular 2x3 y *Chi-cuadrado*

Tabla de contingencia Tipo de voto * Lado en el que se encuentran los empleados * Genero

Genero				Lado en el que se encuentran los empleados: Front Office= Liberal	Lado en el que se encuentran los empleados: Back Office=Conservador	Total
Masculino	Tipo de voto	A favor	Recuento	40	200	240
			Frecuencia esperada	76.7	163.3	240.0
			% dentro de Lado en el que se encuentran los empleados	32.3%	75.8%	61.9%
		En contra	Recuento	60	40	100
			Frecuencia esperada	32.0	68.0	100.0
			% dentro de Lado en el que se encuentran los empleados	48.4%	15.2%	25.8%
		No sabe	Recuento	24	24	48
			Frecuencia esperada	15.3	32.7	48.0
			% dentro de Lado en el que se encuentran los empleados	19.4%	9.1%	12.4%
	Total		Recuento	124	264	388
			Frecuencia esperada	124.0	264.0	388.0
			% dentro de Lado en el que se encuentran los empleados	100.0%	100.0%	100.0%
Femenino	Tipo de voto	A favor	Recuento	32	112	144
			Frecuencia esperada	68.5	75.5	144.0
			% dentro de Lado en el que se encuentran los empleados	16.3%	51.9%	35.0%
		En contra	Recuento	140	80	220
			Frecuencia esperada	104.7	115.3	220.0
			% dentro de Lado en el que se encuentran los empleados	71.4%	37.0%	53.4%
		No sabe	Recuento	24	24	48
			Frecuencia esperada	22.8	25.2	48.0
			% dentro de Lado en el que se encuentran los empleados	12.2%	11.1%	11.7%
	Total		Recuento	196	216	412
			Frecuencia esperada	196.0	216.0	412.0
			% dentro de Lado en el que se encuentran los empleados	100.0%	100.0%	100.0%
Total	Tipo de voto	A favor	Recuento	72	312	384
			Frecuencia esperada	153.6	230.4	384.0
			% dentro de Lado en el que se encuentran los empleados	22.5%	65.0%	48.0%
		En contra	Recuento	200	120	320
			Frecuencia esperada	128.0	192.0	320.0
			% dentro de Lado en el que se encuentran los empleados	62.5%	25.0%	40.0%
		No sabe	Recuento	48	48	96
			Frecuencia esperada	38.4	57.6	96.0
			% dentro de Lado en el que se encuentran los empleados	15.0%	10.0%	12.0%
	Total		Recuento	320	480	800
			Frecuencia esperada	320.0	480.0	800.0
			% dentro de Lado en el que se encuentran los empleados	100.0%	100.0%	100.0%

Fuente: SPSS 20 IBM

- La tabla contingencia se divide entre el **Género** para obtener una comprensión más completa de las posibles relaciones entre las variables.
- Al examinar la tabla podemos ver que las relaciones entre la votación de preferencias y opiniones sobre el asunto principal de la fusión de la empresa no son las mismas cuando se toma en consideración el **Género**.
- **La relación entre las preferencias de voto y la fusión de la empresa identificaba anteriormente que la mayoría de las personas que estaban a favor de la ley tributaria eran conservadoras, mientras que la mayoría de los que estaban en contra de la ley eran liberales. Del grupo que no estaba seguro de Sus opiniones, había una división razonablemente uniforme en las preferencias de voto.**
- Al examinar a los votantes varones podemos ver que la mayoría de los encuestados a favor de la fusión, con independencia de sus preferencias de voto **(61.9%** a favor Comparado con el **25.8 %** en contra y el **12.4 %** sin opinión).
- Al examinar a las mujeres votantes podemos ver que esta tendencia se invierte, con la mayoría **Las mujeres están en contra de la fusión**, independientemente de sus preferencias de voto (**35,0%** a favor, **53.4%** en contra y **11.7%** sin opinión).
- Dentro del grupo a favor de la fusión, la mayoría eran conservadores en sus preferencias de voto; esto era verdad para los hombres y mujeres.
- Dentro del grupo contra la fusión, la mayoría era liberal (empleados ***Front-office***) en sus preferencias de voto; esto era verdad para los hombres y mujeres.
- De la **tabla de contingencias** podemos resumir que la relación entre la votación de preferencias y opiniones sobre el derecho a fusionarse identificadas en las tabulaciones cruzadas 2 × 3 el Género es una variable determinante. Sin embargo, el género en esta muestra difiere en su opinión general sobre la fusión de la empresa; una vez que las preferencias de voto están separadas, y la mayoría de los hombres están a favor de la política de fusión, y la mayoría de las mujeres están en contra de la misma. Esto nos lleva a una relación más complicada entre la afiliación de lo que piensan los trabajadores y la decisión de fusionarse.
- Un examen de la tabla de **Pruebas de *Chi-cuadrados*** nos permitirá comprobar si los patrones identificados anteriormente son significativos. V**er Figura 9.18.**

Figura 9.18. Pruebas de *Chi-cuadrada*

Pruebas de chi-cuadrado

Genero		Valor	gl	Sig. asintótica (bilateral)
Masculino	Chi-cuadrado de Pearson	69.155[b]	2	.000
	Razón de verosimilitudes	68.795	2	.000
	Asociación lineal por lineal	48.904	1	.000
	N de casos válidos	388		
Femenino	Chi-cuadrado de Pearson	59.979[c]	2	.000
	Razón de verosimilitudes	62.673	2	.000
	Asociación lineal por lineal	33.459	1	.000
	N de casos válidos	412		
Total	Chi-cuadrado de Pearson	143.750[a]	2	.000
	Razón de verosimilitudes	149.714	2	.000
	Asociación lineal por lineal	91.977	1	.000
	N de casos válidos	800		

p Valor

Prueba estadística

a. 0 casillas (0.0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 38.40.

b. 0 casillas (0.0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 15.34.

c. 0 casillas (0.0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 22.83.

Fuente: SPSS 20 IBM

- El estadístico de prueba generalmente elegido para determinar si la asociación descrita por tabla de contingencia, es ***Chi-cuadrado* de *Pearson*.**
- En el ejemplo anterior tenemos dos valores para esto como estamos examinando la relación a través de dos grupos de personas, hombres y mujeres, lo que crea una **capa cruce-tabular**. Para los participantes masculinos, puede verse que el valor estadístico de la prueba para ***Chi-cuadrado*** de ***Pearson*** es **69.155**, con **2** grados de libertad y como, el valor de ***p*** es más pequeño de **0.001**, podemos concluir que este patrón de puntuaciones es significativo.
- Los resultados del ***Chi-cuadrado*** se presentan típicamente como sigue:

$$\chi 2 = 69.155,\ gl = 2,\ p < 0.001$$

- Para las participantes femeninas, la relación es también significativa, con un ***Chi-cuadrada* de *Pearson*=59.979**, 2 grados de libertad y, como el valor ***p*** es menor **0.001**. que
- **Podemos concluir que este patrón de puntuaciones es significativo.**
- Los resultados del ***Chi-cuadrado*** se presentan típicamente como sigue:

$$\chi 2 = 59.979,\ gl = 2,\ p < 0.001$$

- De la tabla anterior se puede observar que el estadístico ***Chi-cuadrado*** asume un comportamiento no direccional Hipótesis, que es indicada por el **Sig. Asintótica (bilateral)**
- La razón de verosimilitud es una prueba alternativa a la ***Chi-cuadrada*** que emplea un método. Normalmente usamos el resultado del ***Chi-cuadrado*** pero la proporción de verosimilitud es a veces preferido cuando el tamaño de la muestra es pequeño.
- La **Asociación Lineal por Lineal** se usa cuando las categorías son ordinales y nosotros requerimos identificar tendencias.
- La nota en la parte inferior de la tabla de pruebas ***Chi-cuadrado*** indica si alguna célula individual en el **cruce-tabular** han esperado contar menos de cinco. Para ambos sexos dél Género, ninguno de los las cuentas esperadas son menos de cinco (**0 %**).

9.5. *Chi-Cuadrada* como bondad de ajuste: Ejemplo 4

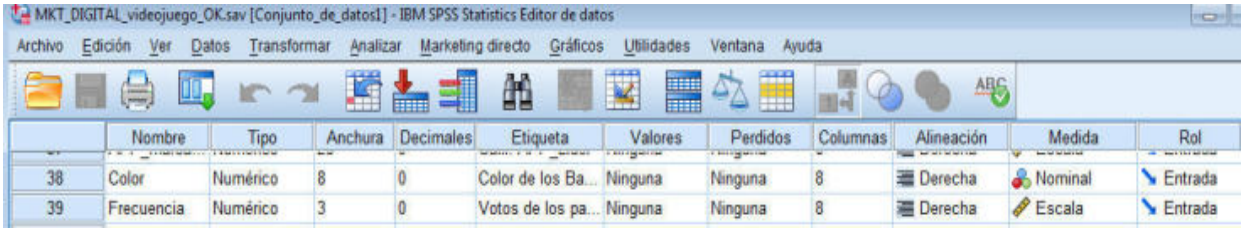
Paso 1: Objetivos

El uso de C***hi-cuadrada*****,** puede también orientarse a las **pruebas de bondad de ajuste.** Esto evalúa si el modelo de resultados se ajusta a un modelo de predicción. **Problema 4:** La empresa **MKT Digital,** se propuso probar si hay una diferencia en la preferencia de color de la gente para el diseño de Banners. Cien participantes recibieron cuatro modelos de Banners , idénticas en forma pero no en color., y se les pidió que declararan su preferencia. Los colores presentados eran **Rojo, Azul, Negro y Blanco**. Si no hubiera ninguna preferencia, entonces se esperaría que cada color se eligiera igualmente, por lo que esperamos que la probabilidad de que cada categoría sea elegida para un cuarto o p = 0.25 cuando la **hipótesis nula es verdadera.** Con un total (**N**) de **100 participantes,** esperaríamos que cada categoría fuera elegida por N * p participantes, es decir, **100 * 0.25,** sea **25** participantes. Al realizar el experimento, el investigador encuentra que **48** participantes eligen el ***Banner*** rojo, **15** el azul, **10** el negro y **27** el blanco. Así, ¿**éstas frecuencias observadas difieren significativamente, de las frecuencias esperadas**?. **Ver Figura 9.19 y 9.20**

H_0 =Los cuatro colores de Banner son igualmente preferidos

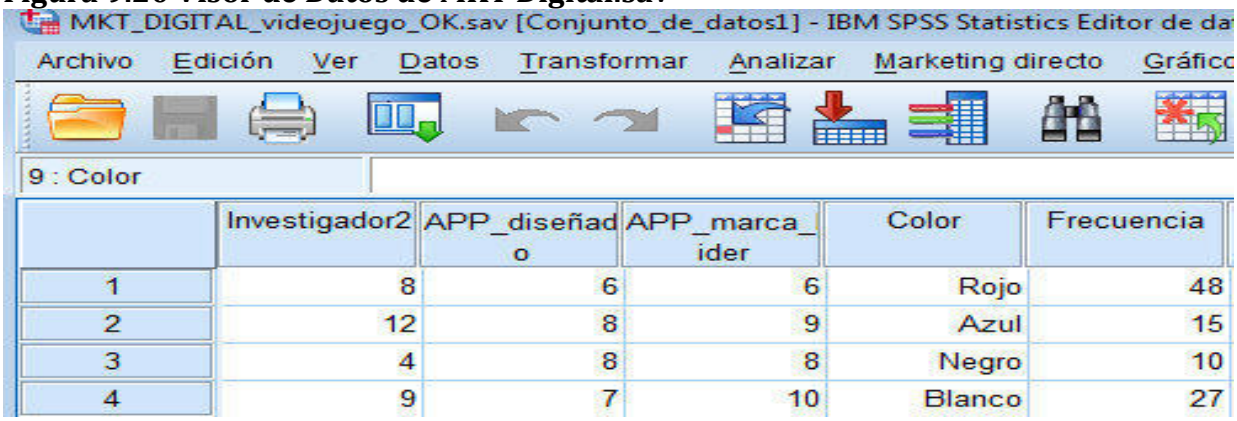
$H_{1=}$ Los cuatro colores de Banner No son igualmente preferidos

Figura 9.19 Visor de Variables de MKT Digital.sav



Fuente: SPSS 20 IBM

Figura 9.20 Visor de Datos de MKT Digital.sav

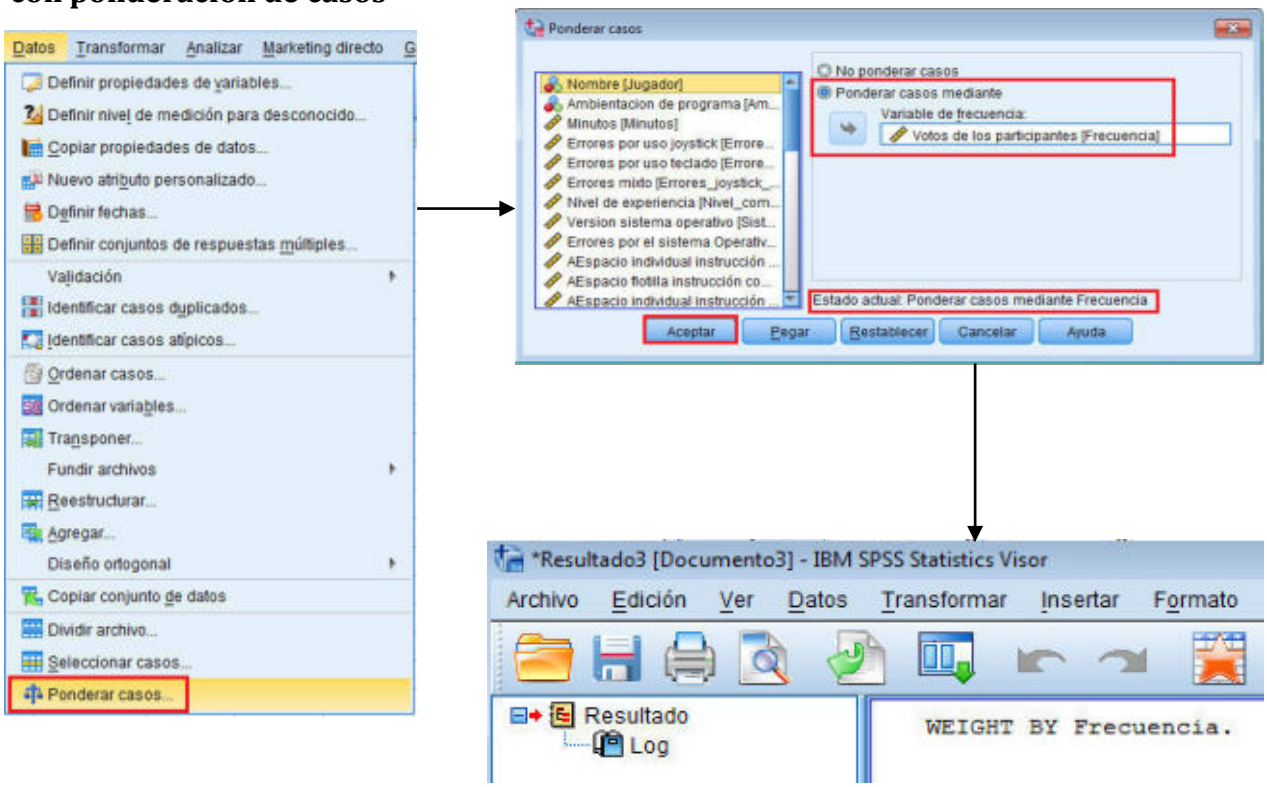


Fuente: SPSS 20 IBM

Paso 2: Diseño

- En adelante para efectos de aprendizaje, sólo se tomarán de forma directa los datos con el fin de aprender a utilizar la técnica.
- Este procedimiento debe ser tratado como conteo de frecuencia con la opción de ponderación de casos con la secuencia de comandos, **Teclear: Datos-> Ponderar casos->Variable de frecuencia: Frecuencia ->Aceptar. Ver Figura 9.21**

Figura 9.21. Proceso de entrada de datos para el cruce-tabular y la *Chi-cuadrada* con ponderación de casos



Fuente: SPSS 20 IBM

Nota: Usted puede estar seguro de que el procedimiento se ha llevado a cabo con éxito al verificar que el Estado actual muestre, **Estado actual: Ponderar casos mediante Frecuencia.**

Paso 3: Condiciones de Aplicabilidad

- No olvidar realizar los análisis previos de inspección visual de la base de datos para detectar ausentes y/o atípicos (corregir en su defecto), confiabilidad, normalidad, homocedasticidad y linealidad

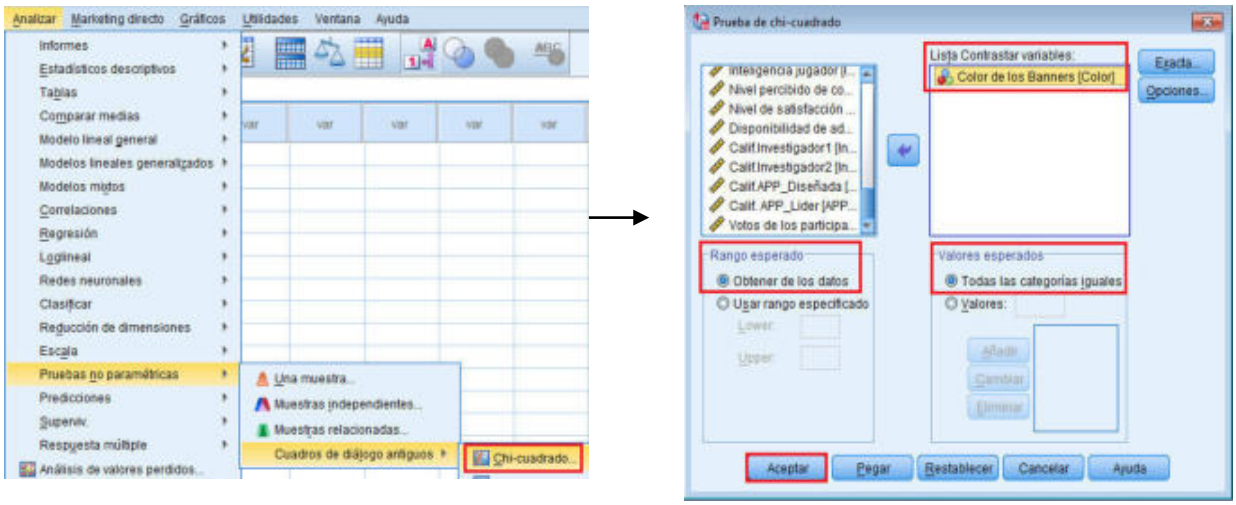
Paso 4: Estimación y ajuste

Teclear: Analizar->Pruebas no paramétricas ->Cuadro de diálogo antiguos->Chi-cuadrada-> Lista contrastar variables: Color->Aceptar. Ver Figura 9.22

Notas:

- La variable de prueba "***Color***" es la que se selecciona para la **Lista contrastar variables**.
- La variable "***Frecuencia***" ya no se manipula más ya que se ingresó vía **Ponderación de casos.**
- Como estamos probando nuestra **Hipótesis nula** de que el número de participantes de cada color son iguales, podemos dejar en **Valores esperados = Todas las categorías iguales**

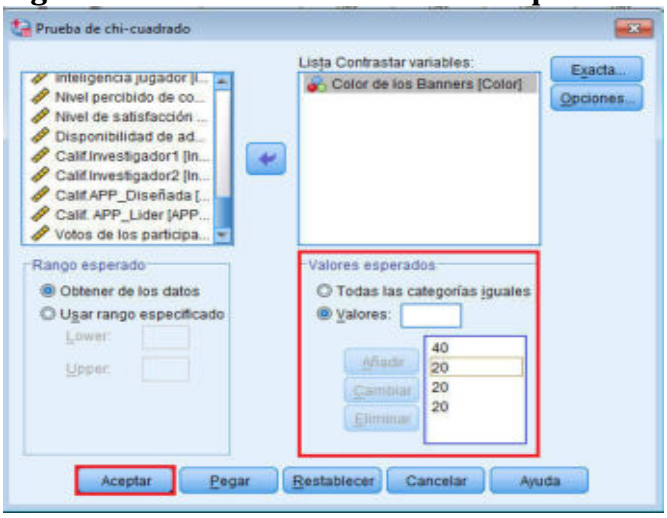
Figura 9.22. Proceso de entrada de datos para el cruce-tabular y la *Chi-cuadrada* con ponderación de casos para bondad de ajuste



Fuente: SPSS 20 IBM

- Si estuviéramos probando el supuesto de que las **proporciones no sean iguales, a partir de un patrón específico,** por ejemplo, que el **40 %** de los participantes prefieran **Banners de negros** y el resto de las **preferencias iguales,** ingresaríamos los respectivos valores de frecuencias a través de la opción **Valores esperados**.Haga ***click*** en **Agregar. Ver Figura 9.23**

Figura 9.23. Selección de valores de ponderación diferentes por variable



La primera tabla producida por **SPSS** muestra los recuentos observados y esperados para los cuatro colores de Banners. Ver **Figura 9.24**

Fuente: SPSS 20 IBM

Figura 9.24. Tabla de datos observados vs esperados de la variable no paramétrica color

Color de los Banners

	N observado	N esperado	Residual
Rojo	48	25.0	23.0
Azul	15	25.0	-10.0
Negro	10	25.0	-15.0
Blanco	27	25.0	2.0
Total	100		

Fuente: SPSS 20 IBM

- Los conteos **N observados** nos muestran cuántas personas preferirían cada color de Banner.
- Los conteos **N esperados** confirman la cantidad de personas que esperaríamos eligieran el color del Banner si la **hipótesis nula fuera verdadera**. Así, se observa de la tabla que las **N observadas** y los **N esperados** **son muy diferentes entre sí**. Esto se confirma mediante la columna **Residual**. Los valores residuales indican que sólo el recuento de **N observado vs N esperado** de las preferencias de los **Banner Blancos se aproximan al recuento esperado. Los Banner de color Azul y Negro** se quedan cortos de la cuenta esperada como se indica por la cifra **con signo negativo**, mientras que los **Banner Rojos** exceder el conteo esperado en **23.0.**

A fin de evaluar si la distribución encontrada a través de la recolección de datos difiere del patrón esperado, la tabla **de Estadísticos de contraste** debe ser analizada.Ver **Figura 9.25**

Figura 9.25. Tabla Estadísticos de contraste

Estadísticos de contraste

	Color de los Banners
Chi-cuadrado	34.320[a]
gl	3
Sig. asintót.	.000

Prueba estadística

p Valor

a. 0 casillas (0.0%) tienen frecuencias esperadas menores que 5. La frecuencia de casilla esperada mínima es 25.0.

Fuente: SPSS 20 IBM

- Como puede verse, de la tabla se tiene: ***χ 2 = 34.320, df = 3, p=0.000 < 0.001***. Recuerde que el último cero cambia a 1si el ***p* Valor** es **0.000.**
- **Conclusión:** existe una diferencia significativa entre el conteo de los N observados y los N esperados, por lo que **rechazamos la hipótesis nula**; **Los cuatro colores No son igualmente preferidos.**

Referencias

Hinton, P.R.; Brownlow, Ch.; McMurray,I y Cozens, B. (2004). *SPSS Explained*. USA: Routledge,.Taylor y Francis Group

IBM (2011a). *IBM SPSS Statistics Base 20.* EUA. .Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Base.pdf

IBM (2011b).*Guía breve de IBM SPSS Statistics 20.* EUA.Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Brief_Guide.pdf

IBM (2011c). IBM SPSS Missing Values 20. EUA. .Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Missing_Values.pdf

Levin, R.,I.; Rubin, D.S. (2004). *Estadística para Administración y Economía*.7ª. Edición. México: Prentice-Hall

Capítulo 10 Análisis de Conjunto

10.1. Análisis de Conjunto. ¿Qué es?

Es una técnica que a partir de la década de los 70de siglo XX, recibió una gran atención por parte de las ciencias de la administración, dado que permitía descubrir en forma clara, las decisiones de los consumidores respecto de productos y servicios, con gran cantidad de atributos para la toma de decisiones de venta y compra [Huber, 1987]. Ya para los años 80, se tiene registro de una amplia aceptación y uso en diversas industrias [Wittink, y Cattin, 1989] y para fines del siglo XX, se convirtió en una de las herramientas de administración que apoya a la mercadotecnia y a la toma de decisiones estratégica para el diseño de productos y servicios, haciendo una derivación en particular a la **mercadotecnia industrial** [Mahajan y Wind, 1991].

El análisis conjunto, tiene la cualidad de acercarse lo más posible a la realidad a través de escenarios experimentales, con software especializado [Bretton-Clark 1988, Intelligent

Marketing Systems, 1993, SAS Institute, 1992, Sawtooth Software 1993, Smith, Scott M. 1989, SPSS, Inc. 1990]. Aún más, la conversión de desarrollos de investigación teórica en programas accesibles para un PC sigue produciéndose [Carroll y Green 1995], y el interés en esos programas está aumentando [Carmone, y Schaffer, 1995, Oppewal, H. 1995]. Por ejemplo, en el diseño de un producto alimentario, un investigador trabaja en un nuevo proceso donde requiere conocer de los efectos humedad y densidad de componentes en el proceso de fabricación, ejercerán en los atributos nutricionales del producto resultante. Una vez que se han llevado a cabo los experimentos, se podrían analizar con los procedimientos **ANOVA** y responder preguntas como: ¿qué condiciones de temperatura y densidad de componentes cumplen con la normatividad nutricional?, ¿es posible retirar calorías sin comprometer los costos? ¿se alteran características de textura y de sabor si se retiran aditivos?. En el caso de las ciencias de la administración, se requiere implementar también experimentos, como en las situaciones de toma de decisiones, por ejemplo, a través de factores ambientales de control. Por ejemplo, ¿qué condiciones ambientales deben prevalecer para que el personal proponga ideas de innovación?, ¿qué características del personal deben asociarse para desarrollar capacidades de innovación de procesos, mercadotecnia, organización, producto/servicio?. (Para saber más, ver : IBM, 2011a; IBM, 2011b; IBM, 2011c).
El análisis de conjunto se desarrolla basándonos en la necesidad de analizar los efectos de los factores bajo control (**variables independientes**), caracterizados a menudo de forma cualitativa **con baja y/o nula fiabilidad** [Green y Srinivasan, 1978: Green y Wind, 1975]. El análisis conjunto es, en realidad, una familia de técnicas y métodos, en teoría basados en los modelos de información, integración y medida funcional [Louvierc, 1988]. En términos de los modelos de dependencia básica:

$$\boldsymbol{Y_1 = X_1 + X_2 + X_3 + \ldots + X_N}$$

(No Métrica, Métrica) (No Métrica)

Las técnicas alrededor del **análisis de conjunto** actualmente se orientan más a estudios de administración de la mercadotecnia, en los que se ajusta para hacer una mejor comprensión de las reacciones de los consumidores asociadas a evaluaciones de una serie de combinaciones de atributos prediseñados, los cuales constituyen a los productos y /o servicios potenciales. Esto trae como ventaja un alto nivel de flexibilización y unidad de respuesta por el nivel de la realidad que se aborda. Esto le reporta a los investigadores una amplia visión de cómo están compuestas las preferencias de los consumidores, motivados por :

1. Su capacidad de usar variables dependientes métricas y no métricas,
2. Usar variables predictoras categóricas y
3. Supuestos bastante generales acerca de las relaciones entre las variables dependientes y las independientes.

Como técnica multivariante, en el campo de la **administración de la mercadotecnia** tiene el uso específico para entender cómo los consumidores desarrollan sus preferencias respecto a productos o servicios, basándose en evaluaciones que éstos hacen sobre el **valor** (**producto/servicio)** con la **idea** que tienen de él (**real o hipotética**) con una combinación de cantidades separadas de valor que cada atributo es capaz de proporcionar. La **utilidad**,

base conceptual para medir el valor en el análisis conjunto, **es un juicio subjetivo de preferencia única para cada individuo. Cubre todos los atributos de un producto o servicio (tangible e intangible)**, y como tal es la medida de la preferencia global. De esta forma:

1. Los atributos tienen diferentes niveles que se expresan en una relación en la que se refleja cómo se formula la utilidad, para cualquier combinación de los atributos y sumar por tanto, cada uno de los valores de la utilidad asociados a cada atributo del producto o servicio para **llegar a la utilidad conjunta**.
2. Se asume que productos/servicios con los mayores valores de utilidad, son más preferidos y tienen mayor posibilidad de ser elegidos.
3. Como técnica multivariante, el análisis de conjunto es el único que le ofrece al investigador **planear, diseñar y/o construir primero un conjunto real o hipotético de productos/servicios combinando diversos niveles escogidos de cada atributo.**
4. Ese diseño de asociaciones por combinación se presenta a los encuestados (consumidores reales y/o potenciales),para capturar sus evaluaciones globales, por lo que es importante realizar una operación lo más de elección dentro de un conjunto de productos/servicios.
5. Dado que **el investigador diseña y construye** los productos/servicios hipotéticos en una forma específica, se forma un indicador **influencia de atributo-valor del atributo- juicio de utilidad de un encuestado**, el cual es determinante para la clasificación final del encuestado.
6. El éxito, depende de que el investigador sea capaz de describir y asociar el producto/ servicio bajo los términos de sus atributos como de todos los valores relevantes a cada atributo.
7. El uso del término **factor** se refiere a la descripción de un atributo específico u otra característica del producto/servicio. Los valores posibles para cada factor se denominan **niveles**. En términos conjuntos, describimos un producto o servicio respecto a su nivel en el conjunto de factores que lo caracteriza. Por ejemplo, el nombre de la **marca** dos niveles (**marca X y marca Y**) y el **precio** puede tener cuatro niveles (70 pesos, 85 pesos, 90 pesos y 100 pesos) pueden ser **dos factores del análisis conjunto**.
8. **Cuando el investigador selecciona los factores y los niveles** para describir un producto/servicio de acuerdo con un plan específico, la combinación se conoce como **tratamiento o estímulo**.

Con lo anterior, suponga que una empresa de telecomunicaciones está diseñando un nuevo servicio de acceso por internet. Ver **Figura 10.1**

Figura 10.1 Tabla de Estímulos para el diseño de un nuevo servicio de telecomunicaciones

Factor	**Nivel**	
Nombre de la marca	Original	Submarca
Velocidad de acceso	120 Mbps	160 Mbps
Medio de acceso	Fibra Óptica	Inalámbrico

Fuente: propia

Se puede construir un servicio de telecomunicaciones hipotético seleccionando un nivel de cada atributo. Para los tres atributos (**factores**) con dos valores (**niveles**), se pueden formar ocho combinaciones (**2 X 2 X 2**), cuyos tres ejemplos de las **8** combinaciones (**estímulos**):

• Original-160Mbps-Fibra Óptica
• Submarca- 120 Mbps-Inalámbrico
• Original-120 Mbps-Inalámbrico

Los investigadores solicitan a los encuestados de la empresa de telecomunicaciones que hagan ordenamiento de **8 estímulos** de su preferencia o que hagan clasificación en una **escala de preferencia** (por ejemplo, de **1 a 10**). As , se puede entender como el análisis de conjunto se le llama también ***análisis "trade-off"***, dado que cuando los encuestados emiten un juicio sobre producto/servicio hipote tico, deben considerar a 360 grados los atributos, es decir, los "***buenos***" y los "***malos***" al formar una preferencia y ponderar todos los atributos simultáneamente al hacer sus juicios. En el diseño y/o construccio n de las combinaciones específicas (**estímulos**), los investigadores esta n realizando un intento por entender la estructura de preferencias del encuestado. La **estructura de preferencias** "***explica***" no so lo lo importante que representa cada factor en la decisio n global, sino tambie n co mo los **diferentes niveles de un factor** influyen en la formacio n de una preferencia conjunta (**utilidad**). En el ejemplo, la te cnica evaluar a el impacto relativo de cada nombre de marca (Original o Submarca), velocidad de acceso (120 Mbps o 160 Mbps), y los diferentes medios de acceso (Fibra O ptica o Inala mbrico) en la determinacio n de la utilidad de un usuario consumidor. Es posible deducir que esta **utilidad**, al representar un "***valor***" total o conjunto de una preferencia por un objeto, esta basada en los **componentes parciales de la utilidad total para cada nivel**. De esta forma, un modelo conjunto puede verse as :

$(\boldsymbol{Valor\ total\ del\ producto/servicio})_{ij\,n}$ = Componente parcial de utilidad total de nivel ***i*** para **factor l**+Componente parcial de utilidad total de nivel ***j*** para **factor 2**+ ...
+ Componente parcial de utilidad total de nivel ***n*** para factor **m**
Donde:

1. El producto/servicio tiene **m** atributos, y cada atributo **n** niveles.
2. El producto/servicio consiste en el **nivel i** del **factor 2**, el **nivel j** del **factor 2** y así sucesivamente, **hasta el nivel n del factor m**.

En el ejemplo, un **modelo aditivo simple** representaría la **estructura de preferencia** del nuevo servicio de telecomunicaciones basándose en **3 factores:**

Utilidad= efecto de marca+ velocidad de acceso+ medio de acceso.

La preferencia por un servicio de telecomunicaciones específico puede calcularse directamente a partir de los valores de los componentes parciales de la utilidad total. Por ejemplo, la preferencia de la **Submarca-160 Mbps-Fibra Óptica**, es:

Utilidad= Componente parcial de utilidad total de **efecto de marca: Submarca**+ Componente parcial de utilidad total de **velocidad de acceso a: 160 MBps**
+ Componente parcial de utilidad total de **medio de acceso: Fibra Óptica**

Al determinar los componentes parciales de la utilidad total, la preferencia de un consumidor puede estimarse en cualquier combinación de factores. Todavía más, revelaría qué **factores** son más importantes en la determinación de la **utilidad conjunta** y la elección del producto/servicio. Al combinarse las elecciones de varios encuestados, también puede representar todo el ambiente competitivo al que se enfrentaría el producto/serviciode manera real.

10.2. Análisis de conjunto: acción por las ciencias de la administración.

Ésta técnica es fundamental para la **toma de decisiones y la planeación estratégica**, ya supone que cualquier **conjunto de conceptos** (por ejemplo: marca, presentaciones, modalidades, etc.) se evalúen como un conjunto de **atributos**. Una vez determinadas la contribuciones de cada factor a la **valoración global del consumidor**, el investigador de la administración de la mercadotecnia, es capaz de:

1. Hacer una definición del concepto con la combinación óptima de factores.
2. Demostrar de los atributos, sus contribuciones relativas y cada nivel de valoración conjunta del concepto.
3. De los juicios del consumidor, usar estimaciones para hacer predicciones de las preferencias entre conceptos con diferentes conjuntos de características (manteniendo constante el resto).
4. De los grupos aislados, formar grupos de clientes potenciales que aporten diferentes niveles de importancia a los atributos, para definición de segmentos potenciales en alto y bajo nivel.
5. Hacer una identificación de las oportunidades de mercadotecnia, al explorar el mercado potencial para combinaciones de atributos no disponibles en ese momento.

La determinacio n y conocimiento de la **estructura de preferencias** para cada consumidor en ana lisis, permite a los investigadores una ilimitada flexibilidad en el examen tanto de las reacciones agregadas como individuales en un amplio rango de aspectos de los productos/ servicios. Examinaremos algunas de las aplicaciones más habituales al final de este capítulo.

10.3. Análisis de conjunto vs. otras técnicas multivariantes.

Hair (2014), refiere que ésta técnica difiere de otras técnicas multivariantes en:

1. Su naturaleza de **técnica de descomposición,**
2. El hecho de que las estimaciones pueden hacerse a nivel individual, y
3. Su flexibilidad en términos de relaciones entre variables dependientes e independientes.

Ver Figura 10. 2

Figura 10.2. Tabla de diferenciación del análisis de conjunto respecto a otras técnicas multivariantes

Tipo de diferenciación	Descripción	Diferencia
Técnicas de descomposición	El análisis conjunto se denomina modelo de descomposición ya que el investigador inicia con **una preferencia global** del consumidor para un	Así, la técnica difiere de los modelos de composición, como: el análisis discriminante y varias aplicaciones de la regresión, en los que el investigador obtiene calificaciones

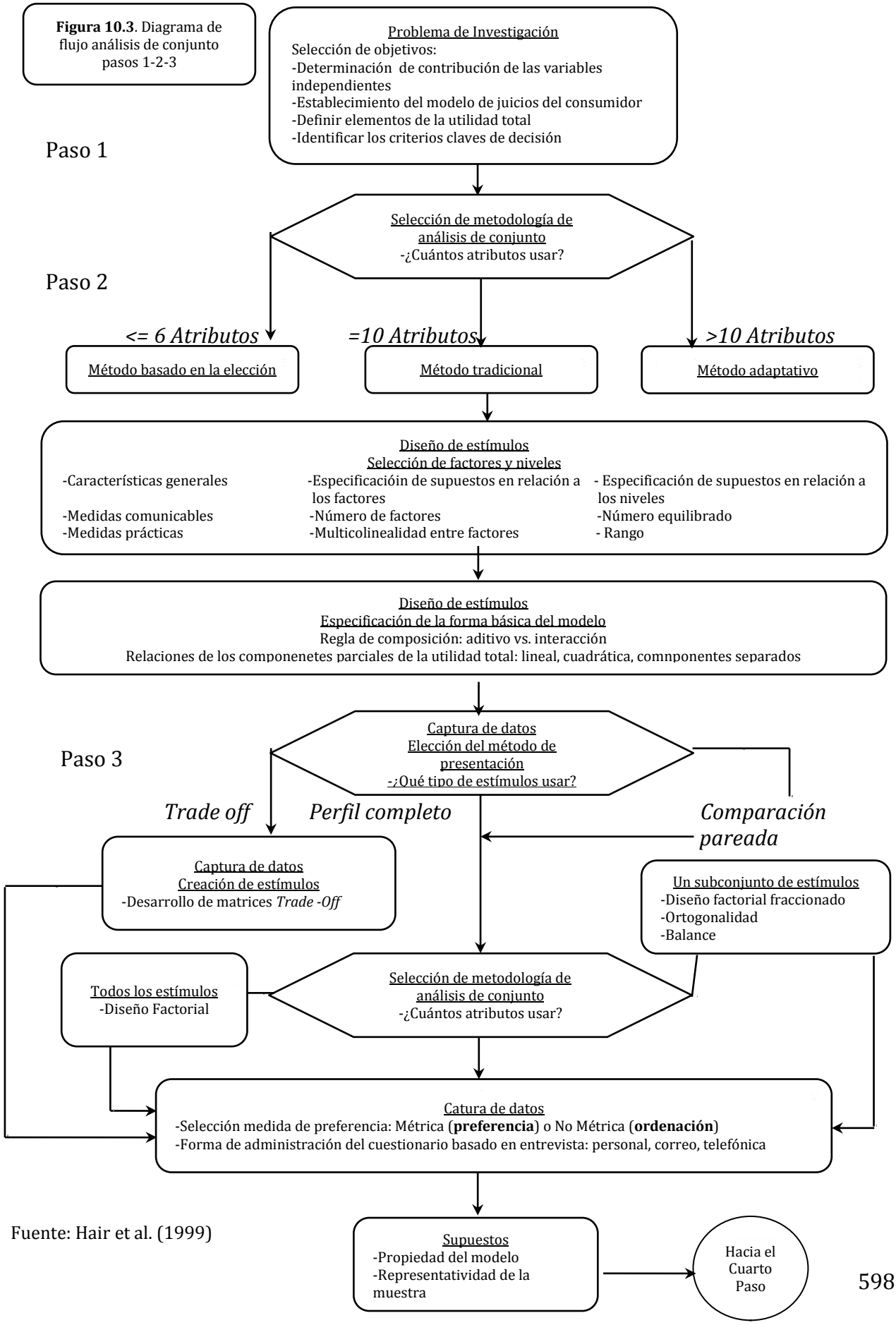
	concepto creado a través especifica los **valores (niveles)** para cada **atributo (factor)**. Con lo anterior, **se descompone la preferencia para determinar el valor de cada atributo.**	del consumidor sobre numerosos atributos del producto/ servicio (por ejemplo, preferencias por, velocidades de acceso, medio de acceso, texturas, olores, sabores, colores, estilos, atributos específicos, etc.) para, a continuación relacionarlas con alguna calificación de preferencia global y desarrollar un modelo predictivo. El investigador no conoce previamente las calificaciones sobre los atributos del producto/servicio, pero las obtiene del encuestado. Con la regresión y el análisis discriminante los resultados del encuestado y las preferencias conjuntas se analizan para componer la preferencia conjunta década uno de los atributos del producto/servicio.
Valor teórico	**Como toda técnica multivariante, se distingue por determinar el valor teórico** que en conjunto es una combinación lineal de efectos de las **variables independientes (factores)** sobre una **variable dependiente.**	La diferencia importante está en que en el valor teórico conjunto, **el investigador especifica tanto las variables independientes (factores) como sus valores (niveles).** La información proporcionada por el encuestado es la **medida dependiente**. El investigador propone los niveles especificados que se utilizan para descomponer las respuestas del encuestado en efectos para cada nivel. Esto también se realiza en el **análisis de la regresión** para cada **variable independiente.** Es aquí, que se presentan características comunes del análisis conjunto y la experimentación, donde el diseño del proyecto es un paso crítico para la realización, de tal forma que si una variable o efecto no se tomó en consideración de forma anticipada en el diseño de la investigación, entonces, no se encontrará disponible para el análisis. Por esta razón, un investigador puede estar tentado de incluir un número de variables que sean irrelevantes, pero que la técnica modera ya que está limitado al número de variables que se pueden incluir. Con esto se evita que el investigador incluya cuestiones adicionales para compensar la falta de claridad en la concepción del problema.
Separación de modelos por individuo	Las estimaciones pueden ser hechas de forma individual (**desagregadas**) o de grupos de individuos representando un segmento de mercado o del mercado entero (**agregado**). A **nivel desagregado**, cada encuestado califica los suficientes estímulos para que el análisis sea realizado para cada persona. **La precisión predictiva se calcula para cada persona, en lugar de hacerlo sólo para la muestra total.** Los resultados individuales pueden agregarse para componer un modelo conjunto. Muchas veces, el investigador es capaza de seleccionar un método de análisis agregado que realice la estimación de los componentes parciales de la utilidad total para el grupo de encuestados en conjunto, de tal forma que la técnica proporciona: (1) Un medio para reducir la tarea de recogida de datos mediante diseños más complejos (2) Métodos para estimar interacciones (es decir, análisis conjunto basado en elecciones), y (3) Mayor eficacia estadística al utilizar más observaciones en la estimación. Al seleccionar entre el **análisis conjunto agregado y el desagregado**, el investigador debe sopesar los beneficios obtenidos por los métodos desagregados frente a las perspectivas proporcionadas por los modelos aislados obtenidos por los métodos desagregados.	El análisis conjunto **difiere de casi cualquier otro método multivariante** en que: (1) Tiene la posibilidad de realizarse a nivel individual, o sea, el investigador genera un "***modelo***" separado para predecir las preferencias de cada encuestado. (2) La mayoría de los métodos multivariantes toman una única medida de preferencia (**observación**) de cada encuestado y a continuación desarrollan el análisis considerando a todos los encuestados simultáneamente. De hecho, muchos métodos requieren que un encuestado ofrezca una única observación (**el supuesto de independencia)** y a continuación desarrollan un modelo común para todos los encuestados, ajustando a cada encuestado con varios grados de precisión (representada por los errores de precisión para cada observación, tal como los residuos en una regresión).
Tipos de relaciones	La técnica no está limitada a los tipos de relaciones exigidas entre las variables dependientes e independientes. Las técnicas de los métodos de dependencia suponen que existe una relación lineal cuando la variable diente aumenta (o disminuye) en cantidades iguales para cada unidad en la variable dependiente.	**El análisis conjunto, sin embargo, puede hacer predicciones separadas para los efectos de cada nivel sobre la variable independiente y no supone que estén relacionadas en modo alguno.** El análisis conjunto puede tratar fácilmente con relaciones no lineales incluso la compleja relación **curvilinea**l, en la cual un valor es positivo, el siguiente negativo y el tercero positivo otra vez .

Fuente: Hair, 2014 con adpatación propia.

10.4. Análisis de conjunto: el experimento

Usted como investigador deberá decidir en ciertos puntos clave tanto del diseño del experimento como en el análisis de resultados ya que el cuestionario no le exige al encuestado más que el responder al número de preguntas con cierto tipo de la **Figura 11.3** (pasos 1-3) y la **Figura 11.4** (pasos 4-7) describen de manera general el flujo de proceso que se siguen en el diseño y ejecución de un experimento de conjunto. Se parte del objetivo de análisis y dado que técnica es muy parecida a un experimento, ésta etapa es crítica para su éxito. Una vez definidos los objetivos, se abordan los más relacionados con el diseño de la investigación en sí, para evaluar los supuestos. Con lo anterior, el proceso de decisión considera: la estimación efectiva del análisis conjunto, la interpretación los resultados y los métodos utilizados para validar los resultados. La discusión termina con un análisis del uso de los resultados del análisis conjunto en posteriores análisis tales como la segmentación de mercado y la elección de simuladores. Cada una de estas decisiones proviene de utilización de la cuestión a investigar y el análisis conjunto como herramienta para entender las preferencias de los consumidores y su proceso de valoración. Se destaca la importancia de definir la metodología a utilizar de análisis de conjunto, que son a saber: **basados en elección y los basados en conjunto adaptativo.** A continuación, se desglosarán los pasos de cómo abordar un experimento de análisis de conjunto a través de las **Figuras 10.3 y 10.13**

Figura 10.3. Diagrama de flujo análisis de conjunto pasos 1-2-3



Fuente: Hair et al. (1999)

10.5. Análisis de conjunto: Paso 1. Objetivos

¿Qué es lo que se va a investigar?. Como se observa, en el análisis conjunto, el diseño que se exige es experimental y en las decisiones del consumidor se buscan 2 objetivos (Hair et al. 1999):

1. De las variables predictor determinar las contribuciones así como sus niveles en el nivel de las preferencias del consumidor. Por ejemplo, ¿en qué medida contribuye la textura del producto a la disposición a comprarlo? ¿Qué nivel de valor es el mejor? ¿En qué medida se puede tener en cuenta el cambio en la disposición a contratar el servicio nuevo de telecomunicaciones por la diferencia en las velocidades de acceso?

2. Fijar un modelo válido de los juicios del consumidor. Estos modelos nos permiten **predecir** la aceptación por parte del consumidor de cualquier combinación de atributos, incluso de aquellos no originariamente evaluados por los consumidores. Para lograrlo, se debe cuestionar: ¿existe una relación lineal entre las variables predictor y las elecciones? ¿Es suficiente un **modelo simple** que "***sume***" los valores de cada atributo o se requiere añadir evaluaciones más complejas de preferencia para reflejar el proceso de valoración más adecuadamente? Los encuestados reaccionan sólo a lo que el investigador les proporciona en términos de **estímulos (combinaciones de atributos)**. ¿Son estos los atributos que se utilizan en realidad en el proceso de toma de decisiones? ¿Son otros atributos, particularmente atributos de una naturaleza más cualitativa tales como reacciones emotivas, también importantes? Así, se requiere formular el cuestionario alrededor de dos temas principales:

1. ¿Es posible describir todos los atributos que dan utilidad o valor al producto o servicio que se está estudiando?
2. ¿Cuáles son los criterios clave implicados en el proceso de elección del consumidor, para este tipo de producto/servicio? Estas cuestiones necesitan ser resueltas antes de proceder a la fase de diseño del análisis conjunto, dado que ofrecen una guía fundamental en cada paso.

10.5.1. Utilidad total del objeto y su definición

Usted como investigador debe, en primer lugar asegurarse de **definir la utilidad total del objeto.** Para la valoración de las respuestas del encuestado con precisión, debe **incluirse todos los atributos** que crean o sustraen utilidad al producto/servicio y a 360 grados, es decir, considerando lo negativo y/o bueno de atributos, ya que: requiere:

1. centrarse sólo en los valores positivos distorsionaría seriamente los juicios de los encuestados y
2. Es posible que ingresen factores negativos de los encuestados de manera inconsciente incluso aunque no los proporcionen en la encuesta, situación que anularía el experimento. Por ejemplo, en ciertas dinámicas de grupo exploratorias para valorar los tipos de características consideradas cuando se evalúa el objeto, **el investigador debe asegurarse que estudia lo que hace el objeto poco atractivo o muy atractivo.** Afortunadamente, la omisión de un factor aislado sólo tiene un impacto pequeño sobre la estimación de otros factores cuando se utiliza en un **modelo aditivo** , pero puede tener algún impacto si es importante.

10.5.2. Factores determinantes y su especificación

Por último, Usted como investigador debe asegurarse de incluir todos los factores determinantes (deducidos del concepto de ***atributos determinantes*** [Alpert, 1971]). El objetivo es incluir los **factores que mejor diferencian entre los objetos**. Pueden ser considerados varios atributos importantes pero también **pueden no diferenciarse en la realización de elecciones porque no varían sustancialmente entre objetos**. Por ejemplo, la calidad de servicio digital de las empresas de telecomunicaciones, es un atributo muy importante, **pero no sería determinante en la mayoría de los casos porque todas las empresas del ramo cumplen estrictamente los reglamentos administrativos de gobierno y por tanto se consideran seguros, al menos en un nivel aceptable**. Sin embargo, otras características, como consumo de energía, rendimiento, ruido o precio **son más importantes** y probablemente serán mucho más utilizados a la hora de elegir entre diferentes modelos de coches. Por tanto, el investigador deberá siempre intentar identificar las variables determinantes claves porque son las piedras angulares en la decisión

10.6. Análisis de conjunto: Paso 2: Diseño

Con base a los objetivos, se debe enfocar el esfuerzo de Usted como investigador a definir las condiciones del experimento, cuestionándose:

1. Métodos altenativos : ¿de los métodos alternativos disponibles cuál debe elegirse?
2. Resolver condiciones específica: ¿qué combinaciones específicas de niveles y atributos deben presentarse a los consumidores para su evaluación?
3. Combinaciones (**estímulos**) del experimento: ¿qué atributos incluir?, ¿cuántos niveles de cada uno?, ¿cómo medir la preferencia?
4. La captura de datos y qué procedimiento de estimación utilizar.
   La fase más importante y critica del diseño de estos aspectos, es el factor clave del éxito de la técnica de análisis de conjunto, por lo que es vital prever todos los detalles potenciales que generen un estudio mal diseñado.

10.6.1. Análisis de conjunto y las metodologías alternativas

Una vez que los atributos del producto/servicio se han determinado, se procede a resolver el cuestionamiento: ¿De las metodologías básicas del análisis conjunto (**tradicional, adaptativa o aditiva**), cuál usar? En este punto, se deberá tomar en cuenta **3 características básicas de la investigación** basada en la técnica: **el número de atributos propuestos, el nivel del análisis y forma del modelo permitida.** Ver Figura **10.4**

Figura 10.4. Tabla comparativa de las 3 metodologías alternativas del análisis de conjunto

Características	Tradicional	Adaptativo de conjunto	Basado en elección
Número máximo de atributos	9	30	6
Nivel del análisis	Individual	Individual	Agregado
Forma del modelo	Aditivo	Aditivo	Aditivo + efectos de iteración

Fuente: Hair, 1999

El **análisis conjunto tradicional ha sido el pilar de la técnica del análisis de conjunto**, y en referencia al ejemplo anterior, se caracteriza por un **modelo aditivo simple** con **9 factores** estimados para **cada individuo**.
El **método adaptativo de conjunto**, es desarrollado para considerar un gran número de factores (**generalmente > 30**) que **NO es factible en un análisis conjunto tradicional.**
El **método basado en la elección** se caracteriza por no sólo emplear una forma única de presentar los estímulos en conjunto (en lugar de uno a uno), sino que también difiere al incluir directamente las interacciones y debe ser estimado a nivel agregado.
Estas alternativas ofrecen mejores resultados que la **técnica de análisis de conjunto tradicional**, que permiten en muchas ocasiones, resolver con mayor alcance los objetivos del investigador a situaciones más diversas.

10.6.2. Los estímulos y su diseño

Es de gran relevancia el disen o de est mulos en la técnica de ana lisis de conjunto dada su practicidad experimental, la cual comprende: el valor teo rico especificado en conjunto basado en la selección de factores y sus niveles, para la construcción de los estímulos. Los supuestos se deben tratar cuando se tienen definidos los factores y los niveles, a fin de relacionar relacionar el carácter general de cada medida, a la par de que el resto de las consideraciones son específicas a cada uno de los factores y niveles. La efectividad de los estímulos depende de las especificaciones del diseño ya que afectan a la precisión de los resultados y por lo tanto al impacto práctico del estudio.

10.6.3. Factores y niveles. Características que los definen

Es importante asegurar que cuando los **factores y niveles** se hacen operativos, éstos tengan:
-Medidas comunicables, de manera fácil y clara para una evaluación realista. Así, una escala de una encuesta puede tener aún sesgos y/o deficiencias en el abanico de posibilidades para detectar el pleno de los atributos bajo estudio, al grado que las sensaciones producidas por texturas, olores, sabores, colores, etc. se han capturado en formas específicas de **realidad virtual, aumentada y/o de neuromarketing** [Research Triangle Institute , 1996; Loosschilder et al. 1995, Sawtooth Technologies, 1997] para una mejor apreciación de las mismas.
-Medidas prácticas, lo que significa que los atributos deben ser distintos para representar un concepto e implementarlo de forma precisa. No debe existir vaguedad en atributos como la satisfacción, calidad o el valor. Además, se requiere que los niveles **NO** se especifiquen en términos amplios, difusos o imprecisos como: **bajo, alto o moderado**, ya que lo que es bajo para una persona no significa lo mismo para otra. Por otro lado, **los conceptos NO son fáciles de poner en práctica**; lo que trae en consecuencia que el investigador pierda seguridad sobre si el diseño del producto/servicio realmente representa lo evaluado por el encuestado. Por lo anterior y para reducir dicha incertidumbre, se recomienda hacer un proceso de **2 etapas**:
1. Un estudio preliminar conjunto que determina los juicios de esos factores (calidad o satisfacción).
2. Los factores identificados como importantes en el estudio preliminar se incluyen en un estudio de mayor alcance con términos más precisos.

10.6.4. Los factores como base de la especificación de los supuestos

Ya obtenidos **los atributos** a incluir como **factores** y asegurado que las medidas sean **comunicables y prácticas,** Usted debe aclarar **3 conceptos** que apoyan la definición de **factores:**

1. **Número de factores.** Este concepto afecta directamente a la **eficiencia estadística y a la fiabilidad de los resultados.** Así, **añadir más factores y niveles,** trae como consecuencia un creciente número de parámetros a estimar que exige: **o bien un número mayor de estímulos o bien una reducción de la fiabilidad de los parámetros.** Se sugiere que el **número mínimo de estímulos** a evaluar por un encuestado a nivel individual es:

   Número mínimo de estímulos = número total de niveles para todos los factores - número de factores + 1

   Suponga, un análisis conjunto con **5 factores** y con **3 niveles cada uno** (un total de **15 niveles**) necesitará por lo tanto un mínimo de **11 estímulos** (**15 - 5 + 1**). Ver **Figura 10.5.**

   Figura 10.5. Tabla ejemplo de estímulos a partir de factores y niveles propuestos

Número mínimo de estímulos	Número de Factores	Número de Niveles
2	1	2
3	2	2
7	3	3
9	4	3
11	5	3
19	6	4
22	7	4
25	8	4
28	9	4
41	10	5

   Fuente: propia

   Nota: Este problema es muy **similar al que se presenta en la regresión cuando el número de observaciones era insuficiente para estimar coeficientes válidos**. Como cada encuestado genera el número exigido de observaciones, el problema no puede "***resolverse***" añadiendo más encuestados, de ahí la importancia de este criterio en la técnica.

2. **Multicolinealidad entre factores.** Esto representa un problema a ser solucionado. La correlación del entorno o inter atributos, mejor conocida como **correlación entre los factores,** refleja la **falta de independencia conceptual entre los factores,** que afectan e impactan a los parámetros estimados como en el caso de las **regresiones.** La colinealidad llega a complicarse para terminar muy frecuentemente en **combinaciones no creíbles de dos o más factores.** Por ejemplo, la potencia de cómputo y el consumo de energía eléctrica se suponen que en una correlación negativa, por lo que una relación positiva ya no es creíble, esto el problema no se encuentra en los **niveles** *per se* sino en el hecho de que existen combinaciones que no se asocian de forma realista, aspecto que se requiere para estimar los parámetros. La multicolinealidad al crear **estímulos irreales**, motiva a que Usted elija de **2 opciones:**

-La más directa es crear "*super atributos*" que combinen los aspectos de los atributos correlacionados. En el ejemplo de la potencia de cómputo y el gasto eléctrico, se deba sustituir por el **factor de"*rendimiento*".** Otro ejemplo de atributos correlacionados positivamente, los **factores de diseño un equipo modem de telecomunicaciones, diseño ergonómico y facilidad de instalación** se pueden tratar mejor con "***portabilidad***". En todos los casos de consideración de "***super atributos***", deben ser tan **específicos y prácticos** como sea posible. De no definirlos en términos más amplios, entonces se vera **obligado a eliminar uno de ellos.**

-La segunda opción implica dos modificaciones posibles a la metodología subyacente del análisis conjunto. La primera modificación comprende **diseños experimentales refinados y técnicas de estimación**, que crean **estímulos cuasi-ortogonales,** que pueden ser utilizados para eliminar cualquier estímulo resultante de una **correlación ínter-atributos** [Steckel et al. 1991]. **La segunda modificación es restringir la estimación de los componentes parciales de la utilidad total**. Estas restricciones pueden ser entre factores así como relacionadas con los niveles dentro de cualquier factor aislado [Srinivasan et al. 1983, Van der Lans, y Heiser 1992]. Se debe intentar primero con las soluciones directas ya que cualquiera de estas dos modificaciones a la metodología **añade una considerable complejidad al diseño y estimación del análisis conjunto.**

3. **El impacto del precio como factor**. El precio es un factor generalmente inclu do en la mayor a de los ana lisis de conjunto, ya que representa un componente distinto de valor para muchos productos/servicios a estudiar. Por otro lado, no es un factor ma s en su relacio n con otros factores [Johnson y Olberts. 1991]. **Por lo general, tiene un alto grado de correlación ínter-atributos con otros factores,** donde al aumentar la cantidad del atributo se asocia un aumento en el precio, por lo que **un nivel de precios descendente es considerado irreal.** Así también, la relación **calidad-precio** al operar entre ciertos factores, puede provocar ciertas combinaciones **irreales**. Otros muchos factores **"*positivos*"** (satisfaccio n, calidad, fiabilidad) pueden incluirse en la definicio n de la **utilidad** del producto/servicio. Sin embargo, al definir lo que se **"*renuncia*"** por esa utilidad (**el precio**), **sólo se incluye un factor, lo que inherentemente esto puede disminuir la importancia del precio**. Finalmente, el **precio puede entrar en interacción con otros factores más intangibles** como el nombre de la **marca.** As , el **impacto de una interacción** en esta situacio n es que un **cierto nivel del precio tiene diferentes significados** para diferentes marcas -en un caso puede ser un **"*premio*" de "*marca*"** y en otro un **"*descuento*" de "*marca*".** A pesar de todas las acepciones en el atributo singular del precio como factor, **no deberían provocar un rechazo del uso de precio por parte del investigador**, sino que debera Usted intentar **anticipar los impactos y ajustar el diseño y la interpretación exigida**, mediante:
-Modelizaciones específicas, tales como **el análisis de valor conjunto (CVA)** donde el precio es el factor clave de estudio [Sawtooth Software 1993].
-De considerarse importantes las **interacciones del precio con otros factores, las relaciones cuantitativas de dichas interacciones se aprecian mejor en los métodos basados en la elección o los multietapas** [Pinnell, 1994] Estos supuestos en la definición de niveles de precios y en la interpretación de los resultados, los puede considerar Incluso si no se hace un ajuste específico.

10.6.5. Los niveles y su relación con la especificación de supuestos

Un aspecto altamente crítico es la definición de los niveles debido a que son las medidas reales utilizadas para formar los estímulos. Así, además de ser prácticos y comunicables, se debe considerar:

1. **Niveles equilibrados en número.** Usted deberá intentar en lo posible, **equilibrar o igualar el número de niveles para todos los factores.** Se ha encontrado que **la importancia relativa estimada de una variable aumenta a medida que el número de niveles lo hace, incluso si los extremos siguen siendo los mismos.** Se ha conjeturado que la categorización muy desagregada llama la atención sobre el atributo y **provoca que los consumidores se centren en unos factores más que en otros.** Con la importancia relativa conocida a priori de los factores, entonces Usted puede **desear aumentar los niveles de los factores más importantes para evitar una di- solución de la importancia y obtener información adicional sobre los factores más importantes** [Wittink et al. 1992].
2. **Rango de los niveles de un factor**. Los niveles al tener un rango (**amplio o reducido**) se sugieren fijarse de alguna forma, quizás fuera de los valores existentes **pero nunca en niveles improbables.** Realizarlo así, **reduciremos la correlación ínter-atributos, aunque también se puede reducir la credibilidad,** por lo que no deben ser muy extremos los niveles. Se pueden provocar problemas sustanciales con niveles inaceptables y deberían por lo tanto, eliminarse. Por lo anterior, Usted debe asegurarse de que sean eso, verdaderamente inaceptables, antes de excluir un nivel, dado que muchas veces los encuestados seleccionan productos/servicios que tienen lo que ellos denominan **niveles inaceptables**. De encontrarse un nivel inaceptable, al haberse realizado el experimento, se recomienda como solución: **o bien eliminar todos los estímulos que hacen los niveles inaceptables o reducir las estimaciones de los componentes parciales de la utilidad total del nivel excesivo a un nivel tan bajo que cualquier objeto que contenga ese nivel no sea elegido.** En la definición de los niveles Usted debe aplicar **los criterios de relevancia práctica y factibilidad.** Cabe destacar que pueden aceptar artificialmente a los resultados, niveles que sean impracticables o que nunca se utilizarían en situaciones reales. Por ejemplo, en un caso hipotético suponga que en el curso normal de la actividad de negocios de una región, el rango de precios varía un 15% alrededor de un precio medio de mercado. Si se incluyera un nivel de precio por debajo del 30 % pero que no se fuera a ofrecer en realidad, su inclusión **distorsionaría** marcadamente los resultados. Los encuestados lógicamente serían más favorables hacia ese nivel de precios. De la utilidad total, cuando se realizan estimaciones de los componentes parciales y se calcula la importancia de los precios, éstos aparecerán artificialmente como más importantes de lo que serían en la realidad de las decisiones del día a día. Usted deberá aplicar los criterios de **factibilidad y relevancia práctica para todos los niveles de atributos para asegurar que los estímulos no sean creados**, de tal forma que se viesen de forma favorable por el encuestado pero que nunca tendrían la posibilidad real de ocurrir.

10.6.6.El Modelo, forma y especificación básica.

El investigador debe tomar **2 decisiones claves** en relación con el modelo conjunto subyacente, **para que el análisis conjunto explique la estructura de preferencias del encuestado sólo a partir de las evaluaciones conjuntas de un conjunto de estímulos**. Estas decisiones afectan tanto al diseño de los estímulos como al análisis de las evaluaciones del encuestado.

1. **Selección de modelo aditivo vs uno interactivo**
   La regla de composición describe cómo combina el encuestado **los componentes parciales de la utilidad total de los factores para obtener el valor con junto.** La decisión de más alcance a que tiene que enfrentarse Usted, le implica la especificación de la regla de composición del encuestado, el cual se divide en los siguientes modelos:
   -1.a. El modelo aditivo, el cual es la regla de composicio n ma s comu n, simple y ba sica, con el cual **el encuestado simplemente "*suma*" los valores de cada atributo (los componentes parciales de la utilidad total) para conseguir el valor total de una combinación de atributos (productos/servicios).** Por ejemplo, supongamos que un producto tiene **2 factores con componentes parciales de 4 y 5.** Por tanto, **la utilidad total sería simplemente 9.** Dado que este modelo tiene en cuenta la mayor a (**80% al 90 %) de la variación de la preferencia** en casi todos los casos, es suficiente para la mayor a de las aplicaciones. **Es también el modelo básico subyacente** tanto en el **análisis conjunto tradicional** como en el **adaptativo** (Ver **Figura 10.5**).
   -1.b. Incorporación de los efectos interacción. Es muy similar a la forma aditiva, ya que supone que el consumidor suma los componentes parciales de la utilidad total de todo el conjunto de atributos. Su mayor diferencia esta en que **permite que ciertas combinaciones de niveles sean superiores o inferiores a la suma.** As , del ejemplo previo, **un modelo interactivo** permite que la suma de los dos niveles, (superior o inferior a **9**) sea **el resultado del modelo aditivo**. En nuestro ejemplo del servicio de telecomunicaciones, un encuestado puede realmente preferir a la **submarca**, pero so lo a cierta velocidad de acceso (**120 Mbps**). En este caso, los atributos por separado se presentan con un nivel de baja utilidad hacie ndose so lo interesantes cuando se combinan**.** Decimos que **submarca** y **velocidad de acceso** esta n **interactuando** y | utilizamos los **efectos aditivos para cada factor**. Un enunciado que relaciona la forma interactiva es: "***el conjunto es mayor (o menor) que la suma de sus partes***." Muchas veces, la incorporacio n de te rminos de interaccio n al modelo disminuye el poder predictivo porque **la reducción de la eficacia estadística (más estimaciones de componentes parciales) no se ve compensada por aumentos en el poder predictivo ganados con las interacciones** (o incapacidad para representar las diferencias efectivas entre ciertos atributos, al estar las partes "***inexplicadas***" asociadas so lo a ciertos niveles de un atributo.)**.** E stas predicen una varianza sustancialmente menor que los **efectos aditivos**, que a menudo no exceden de un **5% a 10%** de aumento de la varianza explicada. Los te rminos de interaccio n es ma s probable sean relevantes en casos en los que los atributos son menos tangibles, particularmente cuando se encuentran reacciones este ticas y emocionales que tienen un papel importante. Por ejemplo, los efectos de interaccio n que tienen lugar muchas veces entre **precio-marca**, **que es menos tangible, pero que genera percepciones concretas.** La importancia del te rmino interaccio n

Un ejemplo de los efectos interacción sobre estimaciones de componentes parciales de la utilidad total. Volviendo a nuestro ejemplo anterior del **servicio de telecomunicaciones**, es posible plantear una situación donde las interacciones parecen influir en las elecciones. Suponga que un **tercer encuestado hace el siguiente orden de preferencias. Ver Figura 10.6.**

Figura 10. 6.Tabla de datos servicio de telecomunicaciones

Marca	Velocidad de acceso	Medio de acceso	
		Alámbrico	Inalámbrico
Original	120 Mbps	1	2
	140 Mbps	3	4
Submarca	120 Mbps	7	8
	140 Mbps	5	6

Fuente: propia

Suponiendo que este **3er** encuestado prefiere la marca **Original** y normalmente prefiere velocidades de acceso de **140 Mbps** que de **120 Mbps** por medio de acceso **inalámbrico** . Sin embargo, una mala experiencia con un la marca **Submarca** hace que el encuestado seleccione **120 Mbps** por medio de acceso **inalámbrico** A este proceso de elección se denomina **efecto interacción entre los factores de maca- velocidad de acceso.** Si consideramos sólo un modelo aditivo, obtenemos los siguientes coeficientes. Ver **Figura 10.7.**

Figura 10.7 . Tabla coeficientes

Medio de acceso		Velocidad de acceso		Marca	
AlámbrIco	Inalámbrico	120 Mbps	140 Mbps	Original	Submarca
0.42	-0.42	0.0	0.0	1.68	1.68

Fuente: propia

Sara calcular los órdenes de preferencia de las combinaciones, se usan los coeficientes por lo que obtenemos lo siguiente. Ver Figura **10.8.**

Figura 10. 8. Tabla de órdenes de preferencia de las combinaciones

Calif.real	1	2	3	4	5	6	7	8
Calif. prevista	1.5	3.5	1.5	3.5	5.5	7.5	5.5	7.5

Fuente: propia

Como se observa **las predicciones son menos precisas**, dado que sabemos que **existe la interacción.** También, los **coeficientes están equivocados** por que los efectos principales de **marca y velocidad de acceso** están confundidos por las **interacciones.** Si procediéramos sólo con un **modelo aditivo**, **estaríamos violando uno de los supuestos principales y realizando predicciones potencialmente imprecisas.**

Analizar las interacciones es una tarea razonablemente simple. Para los datos de preferencia previos, se realiza lo siguiente:

1. Se forman **3 matrices de orden de preferencia de doble entrada.** Por ejemplo, en la **primera matriz,** se enumeran los **dos órdenes de preferencias (uno para cada marca)** para cada combinación de **velocidad de acceso** y después se suman.

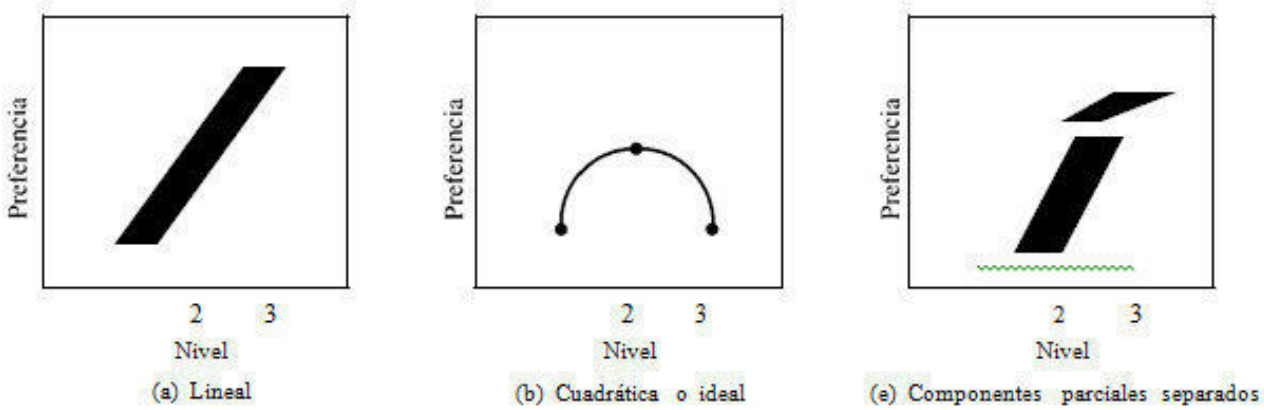
2. Para **comprobar las interacciones**, **se suman a continuación los valores de la diagonal y se calcula la diferencia**. Si el **total es cero**, entonces **no existe interacción.** Como se ve en el ejemplo, la única interacción encontrada es entre **marca y velocidad de acceso**.
3. A medida que la diferencia se hace mayor, **el impacto de la interacción aumenta, y depende del investigador decidir cuándo plantean las interacciones** suficientes problemas en la predicción para justificar el **aumento de la complejidad de los coeficientes a estimar para los efectos interacción.**

1.c. Seleccionar el tipo de modelo. Usted no conocerá con certeza la mejor forma del modelo, pero sí deberá entender las implicaciones de cada elección tanto en el diseño del estudio como de los resultados obtenidos. Elegir una **regla de composición** determina para el encuestado que evalúa , el **tipo y número de tratamientos** o estímulos, junto con la **forma del método de estimación** utilizado:
-Una **forma aditiva** es más fácil de obtener las estimaciones para los componentes parciales de la utilidad total y exige menor número de evaluaciones por parte del encuestado. Si se selecciona una forma de **modelo aditiva**, **no es posible estimar los efectos interacción.** Esto no quiere decir que el investigador deba incluir siempre los efectos interacción, en la medida en que añaden una mayor complejidad al proceso de estimación y, en la mayoría de los casos, provocan que el análisis se desarrolle a un **nivel agregado en lugar de a un nivel individua**l.
-Una **forma interactiva** es una representación **más precisa** de cómo los encuestados valoran realmente un producto/servicio.

2. Relaciones de los componentes parciales de la utilidad total: Lineal, cuadrática o componentes parciales separados y su selección.
El investigador decide qué factores están relacionados con otros en el proceso de decisión del encuestado. Al definir el tipo de relación de los componentes parciales de la utilidad total, el investigador se centra en cómo están relacionados los niveles de un factor.

2a. Utilidad total y tipos de relaciones de los componentes parciales. El análisis conjunto le ofrece **3 alternativas,** que van desde la **más restrictiva** (**relación lineal**) a la **menos restrictiva** (**componentes parciales de la utilidad total separados**), y el **punto ideal, o modelo cuadrático,** entre ambas alternativas. La **Figura 10.9** ilustra las diferencias entre los tres tipos de relaciones:

Figura 10.9. Los tres tipos de relaciones básicas entre los niveles de los factores en el análisis conjunto.



-El modelo lineal es el más simple aunque el más restrictivo, debido a que estima un único componente (de la misma forma que un **coeficiente de regresión**), que se multiplica por el valor del nivel para llevar a valores de los componentes parciales separados para cada nivel.

-En la forma cuadrática, también conocida como **modelo ideal**, se relaja el supuesto de estricta linealidad, de tal forma que **se plantea relación curvilínea** (hacia arriba o hacia abajo).

-Finalmente, la alternativa de los **componentes aislados** (o de **los componentes parciales**) es la más general, al permitir **estimaciones aisladas para cada nivel**. Cuando se usa este último método, el número de valores estimado es el más alto y **se incrementa rápidamente a medida que añadimos más factores y niveles, porque cada nivel tiene una estimación propia.**

La forma de relación de los componentes parciales de la utilidad total puede especificarse para cada factor separadamente, e incluso es posible una mezcla de formas si es necesario. Esta elección no afecta a cómo se crean los estímulos o los tratamientos, ya que se siguen calculando los componentes parciales para cada nivel. Afecta, sin embargo, a cómo y a qué tipos de componentes parciales de la utilidad total son estimados por el análisis conjunto.

Si se especifican formas **lineales o cuadráticas**, entonces los valores de los componentes parciales de la utilidad total para cada nivel se estiman a partir de la relación, con estimaciones de los componentes hechas por separado. Si podemos reducir el número de componentes parciales de utilidad total estimados por cualquier conjunto de estímulos, utilizando un número más restringido de relaciones de los componentes parciales de la utilidad total (es decir, **forma lineal o cuadrática**), los cálculos serán más efectivos y factibles desde el punto de vista de la estimación estadística. Pero debemos considerar el equilibrio entre estas ganancias y la posibilidad de una representación más precisa de cómo el consumidor forma realmente las preferencias conjuntas si empleamos relaciones de componentes parciales de la utilidad total menos restrictivas.

2b. Relación de componentes parciales de la utilidad total y cómo seleccionar. Para cada factor, Usted, como investigador tiene varias formas de decidir sobre el tipo de relación. Así, para establecer el tipo de relación puede basarse en modelos conceptuales o

de investigación previos. Empíricamente, de manera previa el modelo conjunto puede ser estimado como un modelo de componentes parciales de la utilidad total, y examinar visualmente las diferentes estimaciones de los componentes de la utilidad total para **detectar si es más apropiada una forma lineal o una forma cuadrática**. La forma general, en muchos casos, es aparente, **y el modelo puede ser reestimado con relaciones especificadas para cada variable de forma justificada.** Puede evaluar los cambios en la capacidad predictiva, alternativamente bajo diferentes combinaciones de relaciones para una o más variables. Sin embargo, **esto no se recomienda sin al menos cierta evidencia empírica o teórica** de los posibles tipos de relaciones consideradas (es decir, estimaciones previas de los componentes parciales de la utilidad total). **Se debe evitar que los resultados tengan una alta capacidad predictiva pero escaso interés en la toma de decisiones.** En cualquier caso, Usted debe sopesar la capacidad predictiva, el trasfondo conceptual, el grado de relevancia práctica e interpretación necesaria con el uso que se pretende dar al estudio.

10.6.7. Representación de estímulos, tipo de variable respuesta y captura de datos.

Especificados **factores, niveles**, **forma básica del modelo**, Usted deberá decidir: **el tipo de representación de los estímulos (*trade-off; perfil completo o comparación pareada*), el tipo de variable respuesta y el método de captura de datos.** El objetivo: transmitir, lo más apegado a la realidad al encuestado, las combinaciones de atributos (estímulos) de la forma más eficiente posible, generalmente, los estímulos se presentan de forma escrita por descripción, pero sirven modelos físicos y psíquicos adicionales.

Se destaca, que los **3** grandes métodos de ***trade-off, perfil completo y comparación pareada*** representan mejores técnicas para el **diseño de los estímulos** más frecuentemente utilizados en el análisis conjunto, aunque difieran notablemente en la cantidad y forma de la información presentada al encuestado. Ver **Figuras 10.10.**

Figura 10.10. Métodos de presentación de estímulos

Método Trade-Off			**Factor 1: Valor**		
			Nivel 1	**Nivel 2**	**Nivel 3**
Factor 2: Composición			**$120,000**	**$200,000**	**$250,000**
	Nivel 1	**Automático**			
	Nivel 2	**Manual**			
	Nivel 3	**Semiautomático**			

Fuente: propia

Método Perfil Completo
Automático: Transmisión Variable Continua (CVT) Precio: $120,000 Vehículo: Sedán, SUV, Crossover

Fuente: propia

Método Perfil Completo		
Automático: Transmisión Variable Continua (CVT) Precio: $120,000 Vehículo: Sedán, SUV, Crossover	**VS**	Automático: Transmisión Variable Continua (CVT) Precio: $250,000 Vehículo: Sedán, SUV, Crossover

Fuente: propia

La elección entre los métodos depende del tipo de estimación que se está empleando y de los supuestos sobre la cantidad de datos que se están procesando durante el desarrollo del análisis conjunto. Ver **Figura 10.11.**

Figura 10.11. Descripción de los métodos de presentación de estímulos

1. Método *Trade-Off*
Este método **compara dos atributos al mismo tiempo a partir de clasificar todas las combinaciones de niveles** (Ver **Figura 10.10**). Sus ventajas: es sencillo y fácil para el encuestado ; evita sobrecarga de información al presentar sólo **2 atributos al mismo tiempo.** Sin embargo, su uso ha disminuido drásticamente en los últimos años, debido a ciertas limitaciones: **(a)** es de **bajo nivel de realismo** debido al uso de sólo **2 factores al mismo tiempo,** (**b**) **gran número de juicios** necesarios incluso para un número reducido de niveles, (**c**) **tiende a confundir a los encuestados** o seguir un **tipo de respuesta rutinaria** a causa de la **fatiga**, (**d**) **baja capacidad de estímulos gráficos o no literarios** (**e**) uso exclusivo de **respuestas no métricas** y (**f**) **incapacidad para utilizar los diseños de estímulos de factorial fraccionado** para reducir el número de comparaciones realizadas. Recientes estudios han mostrado que la tercera aproximación, **las comparaciones pareadas, han desplazado al método de trade-off** a un segundo lugar en aplicaciones comerciales [Wittink, et al. 1990].
2. Método de presentación de perfil completo
Es el método de presentación más habitual, principalmente por su **realismo en la percepción y su capacidad para reducir el número de comparaciones a través del uso de diseños factoriales fraccionales.** Cada estímulo se describe por separado, a menudo en una ***tarjeta de perfiles*** (ver **Figura 11.10**). Esta aproximación, **obtiene pocos juicios**, pero cada uno es más complejo y los **juicios pueden ser clasificados o calificados**. Entre sus ventajas están: (**a**) **descripción más realista** conseguida por la definición de un estímulo en términos de **un nivel para cada factor**, (**b**) es un retrato más explícito de los ***trade-off*** entre todos los factores y **las correlaciones ambientales** existentes entre los atributos, y (**c**) uso posible de **más tipos de juicios de preferencia,** tales como las **intenciones de compra, probabilidad del juicio y oportunidades de dar marcha atrás -todas difíciles de realizar con el método del *trade-off*** Tiene **2 limitaciones principales:** (**a**) A medida que el número de factores aumenta, también lo hace la posibilidad de **sobrecarga de información**, esto es, el encuestado puede verse tentado de simplificar el proceso centrándose en sólo unos pocos factores, cuando en una situación real se considerarían todos los factores. (**b**) Existe alta tendencia al que **el orden** en que se presenta la relación de los factores en la tarjeta de estímulos **puede tener su impacto en la evaluación**. Por tanto, deberá **alternar los factores** entre los encuestados cuando sea posible **para minimizar los efectos del orden.** Se recomienda el **método de perfil completo cuando el número de factores no sea superior a seis.** Cuando el número de factores var a **de 7 a 10 a diez**, entonces el **método trade-off se convierte en una posible solución** respecto al me todo de perfil completo. **Si el número de factores es > 10, entonces considere un método alterno** (**conjunto adaptativo**) [Green, 1990].
3. El método de presentación de combinaciones pareadas
Este método combina los dos anteriores ya que es una comparación de dos perfiles (ver **Figura 10.10**), presentando a menudo al encuestado, una escala de calificación para **indicar la fuerza de la preferencia por un perfil sobre otro [**Johnson,1975**]**. Una característica distintiva, es que el perfil normalmente no contiene todos los atributos, como ocurre en el método de perfil completo, sino que sólo se seleccionan en un momento unos **pocos atributos** en la construcción de los perfiles. Es **similar al método de trade-off** en que los pares se evalúan, pero en el caso del **método de trade-off** los pares evaluados son atributos, mientras que

en el **método de comparación pareada** los pares son perfiles con múltiples atributos. El **método de comparación pareada** es también instrumental en muchos diseños conjuntos especializados, tales como el **análisis conjunto adaptativo (ACA) [**Sawtooth Software 1993**], que se utiliza en conjunción con un mayor número de atributos**

Fuente: Hair, 1999 con adapatacion propia

10.6.8. Estímulos y su creación

Seleccionados los **factores y los niveles**, **elegido el método de presentación**, deberá volver a **crear los estímulos** a evaluar por los encuestados. Para cualquier método de presentación, **siempre deberá enfrenta a un aumento de la carga de información al encuestado** según el número de **factores y niveles aumenta**. El investigador debe **ponderar los beneficios de aumentar el esfuerzo frente a la información adicional** generada.

1. **Método *trade-off*** . Se utilizan todas las posibles combinaciones de atributos que producen matrices basadas estrictamente en el número de factores calculándose como:

   Número de matrices *trade-off*= (N(N-1)/2)

   N es el número de factores. Por ejemplo, **6 factores** resultarían en **15 matrices *trade-off*** (**5*4/2 = 15**). Se debe recordar, sin embargo, que **cada matriz de *trade- off*** comprende un número de respuestas igual al producto de los niveles de los factores. Por ejemplo, una **matriz *trade-off*** con factores de **3 niveles cada uno exige nueve evaluaciones (3*3) en cada matriz aislada.** Si los **6 factores** de nuestro ejemplo tienen cada uno **3 niveles**, entonces el encuestado evaluaría **15 matrices de *trade-off***, cada una con **9 evaluaciones**, para una total de **135 evaluaciones conjuntas**. Como se observa, **este método de presentación puede llevar rápidamente a sobrecargar excesivamente de información al encuestado a medida que el número de atributos o niveles aumenta**. Su contraparte de compensación, es que este método sigue haciendo las **cosas sencillas al preguntar al encuestado que evalúe sólo dos factores al mismo tiempo**, mientras que los otros métodos de presentación pueden llegar a verse bastante comprometidos por la **complejidad de los estímulos**
2. **Los métodos de presentación de combinación pareada y perfil completo.** Estos 2 métodos finales, comprenden la **evaluación de un estímulo cada vez (perfil completo) o pares de estímulos (comparación pareada).** En un análisis conjunto simple **con un reducido número de factores y niveles** (tales como los discutidos previamente para **3 factores con 2 niveles cada uno que resultaban en ocho combinaciones**), **el encuestado puede evaluar todos los posibles estímulos.** A esto se le conoce como un **diseño factorial** cuando se utilizan todas las combinaciones. Pero a **medida que aumenta el número de factores y niveles, el diseño se hace impracticable** de forma similar a la mostrada por el **método de *trade- off***. Si está interesado en evaluar el impacto de **4 variables** con **4 niveles** para cada variable, se crearían **256 estímulos** (4 niveles* 4niveles*4 niveles *4 niveles) en un diseño **factorial completo** para el **método de perfil completo.** Esto es demasiado para que lo evalúe un solo encuestado con respuestas consistentes y significativas. Se crearía un número incluso **superior de pares de estímulos para las combinaciones pareadas** de perfiles con diferentes números de atributos. Lo que se necesita es un método de desarrollo de un conjunto de los estímulos totales que pueda ser evaluado y siga

ofreciendo la información necesaria para hacer predicciones precisas y fiables de los componentes parciales de la utilidad total.

10.6.9. Definición de conjuntos de estímulos.

El método más común de definición de un conjunto de estímulos a evaluar, es a través del **diseño factorial fraccionado,** ya que selecciona una **muestra de posibles estímulos**, donde **el número de estímulos depende del tipo de regla de composición que se supone que usan los encuestados.** Utilizando el **modelo aditivo**, que supone sólo los efectos principales para cada factor **sin interacciones**, un estudio que utilice **el método de perfil completo** con **4 factores** a **4 niveles** requiere sólo de **16 estímulos** para estimar los efectos principales. Ver **Figura 10.12.**

Figura 10.12.Diseños factoriales fraccionados de un modelo aditivo (efectos principales) 4 factores a 4 niveles cada uno

	Diseño 1: niveles para (a)				Diseño 1: niveles para (a)			
Estímulo	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**
1	3	2	3	1	2	3	1	4
2	3	1	2	4	4	1	2	4
3	2	2	1	2	3	3	2	1
4	4	2	2	3	2	2	4	1
5	1	1	1	1	1	1	1	1
6	4	3	4	1	1	4	4	4
7	1	3	2	2	4	2	1	3
8	2	1	4	3	2	4	2	3
9	2	4	2	1	3	2	3	4
10	3	3	1	3	3	4	1	2
11	1	4	3	3	4	3	4	2
12	3	4	4	2	1	3	3	3
13	1	2	4	4	2	1	3	2
14	2	3	3	4	3	1	4	3
15	4	4	1	4	1	2	2	2
16	4	1	3	2	4	4	3	1

(**a**) Los números que aparecen en las columnas bajo el factor 1 hasta el 4 son los **niveles para cada factor**. Por ejemplo, el primer estímulo del diseño **1 consiste del nivel 3 para el factor 1, nivel 2 para el factor 2, nivel 3 para el factor 3 y nivel 1 para el factor 4.**

Fuente: Hair, 1999

Los **16 estímulos** pueden ser cuidadosamente construidos para asegurar la correcta estimación de los **efectos principales**. Los dos diseños de la **Figura 11.13** son **diseños óptimos, en la medida en que son ortogonales (no existe correlación entre los niveles y atributos) y equilibrados (cada nivel aparece en el factor el mismo número de veces).** La creación de un **diseño óptimo, con ortogonalidad y equilibrado** no significa, sin embargo, que todos los estímulos de ese diseño **sean aceptables para ser evaluados.** Existen varias razones para que produzcan **estímulos inaceptables**:

1. La creación de **estímulos "*obvios*",** es decir estímulos cuya evaluación es obvia debido a la combinación de niveles. Los ejemplos más comunes son los estímulos que tienen en

todos los niveles los **valores más altos o más bajos.** En estos casos, **los estímulos en realidad proporcionan poca información acerca de la elección y pueden crear una percepción de incredulidad por parte del encuestado.**

2. La creación de **estímulos increíbles** a causa de la **correlación ínter-atributos**, que pueden crear estímulos con combinaciones de niveles de atributos (**alto consumo de gasolina, alta aceleración) que no son realistas**.
3. Finalmente, se pueden **encontrar restricciones en las combinaciones de atributos**. En cualquiera de estos casos, **los estímulos inaceptables presentan elecciones increíbles para el encuestado y deberían ser eliminadas** para asegurar una estimación válida del proceso así como una percepción de credibilidad de la elección entre los encuestados.

10.6.10. Eliminando los estímulos inaceptables

Existen varias alternativas:

Usted puede **generar otro diseño factorial fraccional** y evaluar la aceptabilidad de sus estímulos. Si todos los diseños contienen estímulos inaceptables y no se puede encontrar una mejor alternativa de diseño, entonces deberían **destruirse los estímulos inaceptables**. Aunque el diseño no sea totalmente o**rtogonal (es decir, que esté de alguna forma correlacionado y se denomina cuasi-ortogonal),** no se viola ningún supuesto del análisis conjunto. Esto creará problemas similares a los de la **multicolinealidad** en la regresión (es decir, inestabilidad de las estimaciones cuando hay cambio ligero de los niveles y una menor capacidad para evaluar el impacto aislado de cada atributo). Todos los diseños **cuasi ortogonales** deberían ser evaluados por su **eficiencia del diseño**, que es una medida de la correspondencia del diseño en términos de ortogonalidad y equilibrio para un diseño óptimo [Kuhfeld et al.1994]. Con escalas de **100 puntos** (**diseño óptimo= 100**), se pueden evaluar los **diseños no ortogonales alternativos y seleccionar el diseño más eficiente con todos los estímulos aceptables.** La mayoría de los programas de análisis conjunto **evalúan la eficacia** de los diseños para desarrollar diseños cuasi ortogonales.Las correlaciones ínter-atributos se deben tratar en términos conceptuales. Los **estímulos inaceptables** debidos a las **correlaciones ínter-atributos** pueden darse en diseños óptimos y ortogonales, y Usted debe acomodarlos dentro del diseño. En términos prácticos, las correlaciones ínter-atributos deberían minimizarse pero no hacerse necesariamente cero si correlaciones reducidas (0,20 o menos) añaden realismo. La mayoría de los problemas se encuentran en el caso de **correlaciones negativas**, como entre el **consumo de gasolina y caballos de potencia.** La adición de factores no correlacionados puede reducir la correlación media ínter-atributos, de tal forma que con un **número realista de factores (es decir, 6 factores),** la intercorrelación media estaría cercana a **0.20,** lo que tendría efectos sin consecuencias. Pero el investigador debería evaluar siempre la credibilidad de un estímulo como medida de relevancia práctica.

Se necesitan los **240 posibles** estímulos de nuestro ejemplo que **no están incluidos en el diseño factorial fraccional s**eleccionado si se van a estimar los **11 términos de interacción.** Usted puede decidir que las interacciones seleccionadas son importantes y deberían incluirse en la estimación del modelo. En este caso, el **diseño factorial fraccionado debería incluir estímulos adicionales para acomodar las interacciones.** Las guías publicadas para los diseños factoriales fraccionados o componentes de los

programas de análisis conjunto **diseñan los subconjuntos de estímulos para mantener la ortogonalidad o eficiencia máxima en diseños "*cuasi*" ortogonales, haciendo la generación de estímulos de perfil completo bastante fácil** [Addelman, 1962, Conner, y Zelen, 1959, Hahn y Shapiro 1966, MeLean y Anderson, 1984]. Si el **número de factores** se hace muy grande y **la metodología adaptativa conjunta no es aceptable**, se puede emplear un **diseño de puente** [Baalbaki y Malhotra, 1995]. En este diseño, los factores se dividen en subconjuntos de tamaño apropiado, donde **algunos atributos se solapan** entre los conjuntos de tal forma que cada conjunto tiene un factor(es) en común con otros conjuntos de factores. Los estímulos se construyen para cada subconjunto de tal forma que **el encuestado nunca ve el número original de factores en un perfil único.** Cuando se estiman **los componentes parciales de la utilidad total**, se combinan los conjuntos separados de perfiles y se proporciona un único conjunto de estimaciones. Los programas informáticos manejan la **división de los atributos, creación de estímulos y su recombinación para la estimación [**Bretton-Clark 1988**]**. Cuando se utilizan **comparaciones pareadas, el número puede ser muy grande y complejo**, de modo que muy a menudo se utilizan los programas informáticos interactivos que seleccionan los conjuntos óptimos de pares a medida que se realiza el cuestionario.

10.6.11. Preferencia del consumidor y selección de su medida.

El investigador debe también seleccionar la medida de preferencia: clasificación de orden frente a calificación (**una escala de 1 a 10**). Aunque el método ***trade-off*** emplea sólo datos de clasificación, el método de comparación pareada puede evaluar la preferencia, bien mediante la obtención de una clasificación de preferencia de un estímulo frente a otro o bien como una medida binaria, lo que es preferible. El método de perfil completo también se acomoda tanto al método de clasificación como al método de calificación. Cada medida de preferencia tiene ciertas ventajas y limitaciones. Obtener una medida de preferencia de clasificación de orden (es decir, ordenación de los estímulos desde el más preferido al menos preferido) tiene dos ventajas principales: (1) es probable que sea más fiable porque la ordenación es más fácil que la calificación para un número razonablemente reducido (menos de 20) de estímulos, y (2) proporciona más flexibilidad en la estimación de los diferentes tipos de reglas de composición. Tiene, sin embargo, una desventaja principal: es difícil de administrar, dado que el proceso de ordenación se realiza normalmente clasificando tarjetas que recogen los estímulos y este procedimiento sólo puede realizarse mediante una entrevista personal.

La alternativa es obtener una calificación de preferencia en una escala métrica. Las medidas métricas se analizan y se administran más fácilmente, incluso por correo, y permiten realizar estimaciones conjuntas mediante regresiones multivariantes. Sin embargo, los encuestados pueden ser menos discriminantes en sus juicios de lo que lo serían con una ordenación. También, dado el elevado número de estímulos evaluados, es útil aumentar el número de categorías de respuesta por encima del considerado en la mayoría de los estudios del consumidor. Como norma general se toman 11 categorías (es decir, calificaciones de 0 a 10 o 0 a 100 en incrementos de 10) para 16 o menos estímulos y se expande a 21 categorías para más de 16 estímulos [Louvierc,1988].

10.6.12. El estudio y su realización

Antes de la era del procesamiento de cómputo personalizado, la complejidad de la técnica se basaba a menudo en entrevistas de profundidad personal para obtener las respuestas, los cuales permiten explicar a los encuestados los a veces difíciles métodos asociados con el análisis conjunto. Sin embargo, los desarrollos de los últimos años en los métodos de entrevista, permiten el uso de la tecnología para aplicar los cuestionarios tanto por correo electrónico como por teléfono. Si el estudio se diseña para asegurar que el encuestado asimile y procese apropiadamente el estímulo, entonces todos los métodos de encuesta arrojan relativamente la misma precisión predictiva [Akaah, 1991]. El uso de entrevistas por email, por ejemplo, ha simplificado las demandas que el análisis conjunto exigía al encuestado y ha hecho muy asequible desarrollos tales como los **análisis conjuntos adaptativos** [Sawtooth Software 1993].
Se destaca que una cuestión altamente delicada de cualquier estudio conjunto, **es la carga que se pone en el encuestado** debido al número de estímulos conjuntos evaluados. Obviamente, el encuestado no podría evaluar los **256 estímulos** de nuestro anterior diseño factorial, pero **¿cuál es el número apropiado de tareas en un análisis conjunto?** Estudui9s de mercadotecnia afirman que es factible que los **encuestados podían completar fácilmente hasta 20 evaluaciones conjuntas [**Johnson y Orme. 1996**]** y que, rebasando esta cantidad de evaluaciones, las respuestas se vuelven **menos creíbles y menos representativas de la estructura de referencia subyacente**. Al momento, **NO existe un número mínimo o máximo absoluto de evaluaciones de estímulos**, por lo que se sugiere que el investigador se esfuerce en **utilizar los menos posibles a la vez que se mantiene la eficiencia en el proceso de estimación.** Así también, **NO existe método sustituto de contraste previo de un estudio conjunto que evalúe la carga que soporta el encuestado, ni el método de realización ni la aceptabilidad de los estímulos.**

10.7. Análisis de conjunto: Paso 3. Condiciones de aplicabilidad

La técnica **tiene el grupo de supuestos menos restrictivo en la estimación del modelo conjunto** ya que la naturaleza generalizada del modelo así como el diseño experimental estructurad, hace innecesarios la mayoría de los test realizados en otros métodos de dependencia. Por lo anterior:

1. Las pruebas estadísticas de **normalidad, homocedasticidad e independencia que se desarrollaron en otras técnicas de dependencia no son necesarios.**
2. La aplicación de **diseños de estímulos con base estadística** asegura también que la estimación no está confundida y que los resultados sean **interpretables bajo la regla de composición asumida.**
3. Incluso aunque hay pocos **supuestos estadísticos**, los **supuestos conceptuales** son, probablemente mayores que con cualquier otra técnica multivariante.
4. El investigador **debe especificar la forma general del modelo** (**efectos principales vs. a modelo con interacciones**) antes de diseñar la investigación. Con ello, se "***construye***" esta decisión y **se hace imposible contrastar modelos alternativos** una vez que la investigación está diseñada y los datos se han recogido. El análisis conjunto **no es como la regresión,** por ejemplo, **donde los efectos adicionales (términos de interacción o no lineales) pueden ser evaluados fácilmente.** El investigador debe tomar esta decisión teniendo en cuenta la forma del modelo y debe diseñar la

investigación consecuentemente. Por tanto, la técnica, aunque tiene pocos supuestos estadísticos, **está muy determinado por la teoría en su diseño, estimación e interpretación.**

10.8. Análisis de conjunto: Paso 4. Estimación y ajuste

Dados los avances de los últimos años, las técnicas de estimación han aumentado, de forma que el desarrollo de técnicas en conjunción con métodos especializados de presentación de estímulos (por ejemplo, **el análisis conjunto adaptativo o basado en elección**) representa un tipo de mejora de esta clase. Usted, al obtener los resultados de un estudio de análisis conjunto, sin embargo, debe tratar los supuestos de selección del método de estimación y evaluación de los resultados (véase **Figura 11.13**) y siguiendo:

1. **Eligiendo una técnica de estimación. En primera instancia,** las evaluaciones de ordenación exigen una forma modificada de análisis de **varianza específicamente diseñado para datos ordinales,** residentes en desarrollos de software tales como: **el MONANOVA (análisis monotómico de la varianza)** [Johnson, 1975, Kruskal, 1965] y **LINMAP** [Schocker y Srinivasan 1977]. Estos programas dan **estimaciones de los atributos de los componentes parciales de la utilidad total**, de tal forma que **ordenar su suma (valor total)** para cada tratamiento, se **correlacione lo más cercanamente posible a la clasificación observada.** Si se obtiene una **medida métrica de referencia** (**calificaciones vs. clasificaciones**), entonces muchos métodos, **incluido el análisis de regresión,** pueden servir para estimar los componentes parciales de la utilidad total para cada nivel. En conclusión, la mayoría de los programas informáticos existentes hoy en día pueden utilizar tanto el tipo de evaluación (**calificaciones o clasificaciones**), así como estimar cualquiera de los **3 tipos de relaciones** (**lineal, punto ideal y componentes parciales de la utilidad total**).
2. **Evaluación de la bondad del ajuste del modelo.**
   La **precisión** de los resultados de la técnica se evalúa tanto a nivel individual como agregado. Así, se debe tener como objetivo el **determinar la consistencia de predicción del modelo del conjunto de evaluaciones de preferencias dadas por cada persona**:
   1. Para los **datos de clasificación,** se utilizan las **correlaciones** basadas en las clasificaciones previstas y efectivas (es decir, la ***ro de Spearman*** o la ***tau de Kendall***).
   2. De obtener una **calificación métrica, entonces es apropiada. Una simple correlación de *Pearson*, la misma que se utiliza en la regresión, junto con una comparación de las calificaciones previstas y actuales.**
   3. En casos de **predicción a nivel individual,** las preferencias previstas y efectivas están **correlacionadas para cada persona y contrastadas para la significación estadística.**
   4. Por otro lado, en la mayor a de los experimentos conjuntos, **el número de estímulos no excede sustancialmente del número de parámetros,** y siempre existe la posibilidad de **"*sobreajustan*" los datos**.
   5. Se recomienda a los investigadores **medir la precisión del modelo no sólo sobre los estímulos originales sino también con un conjunto de estímulos *holdout* o de validación.** En un procedimiento similar a la muestra ***holdout*** en el análisis

discriminante, el investigador **prepara más tarjetas de estímulo que las necesarias para la estimación de los componentes parciales de la utilidad total, y los encuestados las califican todas al mismo tiempo.**

6. Se utilizan entonces parámetros del modelo conjunto estimado para **predecir la preferencia sobre el nuevo conjunto de estímulos, que se compara con las respuestas efectivas para evaluar la fiabilidad del modelo**. Los individuos que tienen un ajuste predictivo muy bajo para la **muestra *holdout* se pueden eliminar del análisis**.
7. La **muestra *holdout*** también ofrece al investigador una oportunidad de evaluación directa de estímulos de interés para el estudio.
8. Si se utiliza una ***técnica de estimación agregada,*** el investigador puede usar una muestra ***holdout*** de encuestados en cada grupo para evaluar la precisio n predictiva. Este método **no es factible** para **resultados desagregados** porque **NO hay un modelo "*generalizado*" que se aplique a la muestra *holdout***, en la medida en que cada encuestado de la muestra de estimación tiene estima- ciones de los componentes parciales de la utilidad total individualizadas.

10.9. Análisis de conjunto: Paso 5. Interpretación

Esta se da mediante:

1. **Análisis agregado vs. análisis desagregado.** Para interpretar el análisis de conjunto, se deberá verificar:

   (a) La aproximación normal para interpretar el análisis conjunto es la **desagregada;** esto es, cada encuestado se **modeliza separadamente y los resultados del modelo se examinan para cada encuestado.**

   (b) El método más común de interpretación es el **examen de las estimaciones de los componentes parciales para cada factor, evaluando su magnitud y su pauta tanto a efectos de relevancia práctica como a efectos de correspondencia con relaciones teóricas entre niveles. Cuanto mayor sea el componente parcial (positiva o negativa), mayor será el impacto que tenga sobre la utilidad total.**

   (**c**) Los valores de los **componentes parciales de la utilidad total** pueden ser representados gráficamente para identificar las pautas.

   (**d**) Convertir con software especializado, **las estimaciones de los componentes parciales de la utilidad total a una escala común** (por ejemplo, de un máximo de **100 puntos**) para que se puedan **realizar comparaciones entre los factores individuales, así como entre individuos.**

   (**e**) Esta conversión ofrece un medio de utilizar los componentes parciales de la utilidad total en otras técnicas multivariantes tales como el **análisis cluster**.

   (**f**) **La interpretación también se puede realizar con resultados agregados.** Si la estimación del modelo se hace a **nivel individual y después agregado**, o se hacen estimaciones agregadas para un conjunto de encuestados, el análisis ajusta un modelo al agregado de respuestas. **Este proceso generalmente ofrece unos resultados pobres cuando se intenta predecir lo que haría cualquier encuestado aislado o cuando se quieren interpretar** los componentes parciales de la utilidad total para cualquier encuestado aislado. A menos que el investigador esté tratando con una población que muestre concluyentemente un comportamiento homogéneo con

referencia a los factores, no debería utilizarse el análisis agregado como método de análisis. Sin embargo, muchas veces el análisis agregado predice mejor un comportamiento agregado como la cuota de mercado. Por tanto, el investigador debe identificar el objetivo primordial del estudio y emplear el nivel apropiado de análisis o una combinación de los niveles de análisis.

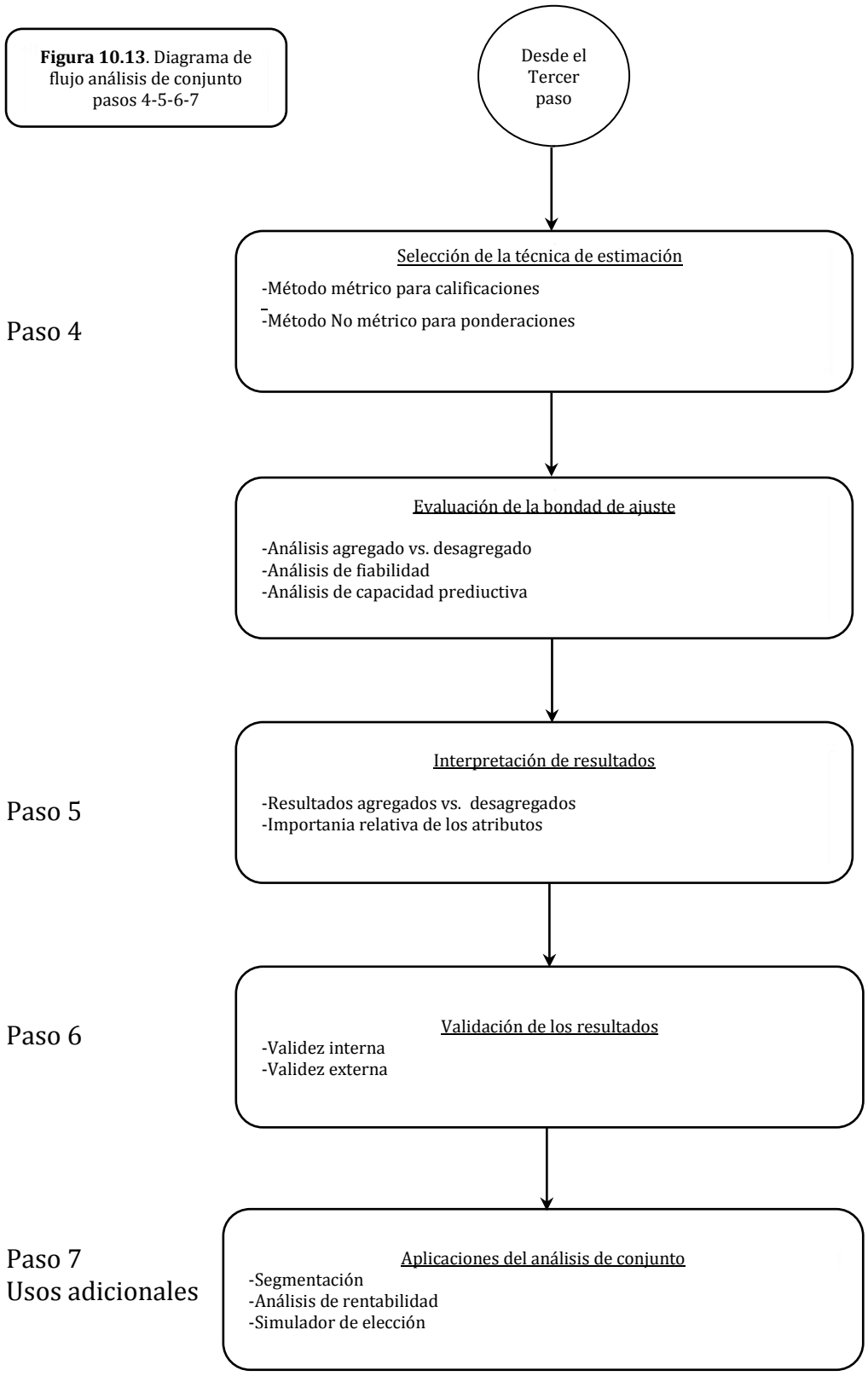
2. **Importancia relativa de los atributos y s evaluación. La técnica puede evaluar la importancia relativa de cada factor , además de representar el impacto de cada nivel con las estimaciones de los componentes parciales de la utilidad total.** Dado que normalmente se convierten a una escala común las estimaciones de los **componentes parciales de la utilidad total,** la mayor contribución a la utilidad conjunta y (por tanto el factor más importante) es el **factor con el mayor rango** (alto o bajo) de los componentes parciales de la utilidad total. Esto permite la comparación entre los encuestados en una escala común así como dar significado a la magnitud de la puntuación de importancia. **El investigador debe considerar el impacto sobre los valores de importancia de un nivel extremo o prácticamente improbable.** Si se encuentra un nivel así, debería ser eliminado del análisis o los valores de importancia se verían reducidos a reflejar sólo el rango de los niveles factibles.

10.10. Análisis de conjunto: Paso 6: Validación

Los resultados de la técnica**, se validan interna y externamente:**

1. **La validación interna** implica la confirmación de que la regla de composición seleccionada (**aditiva vs. interactiva**) es la apropiada. Sin embargo, Usted está normalmente **limitado nada más que a la evaluación empírica de la validez de la forma del modelo seleccionado en un estudio completo**, debido a las necesidades de datos para contrastar ambos modelos. **Esta evaluación se realiza más efectivamente mediante la comparación de modelos alternativos** (**aditivos vs. interactivos**) en un examen previo para confirmar qué modelo es el apropiado.
2. **La validación externa** implica en general la capacidad del análisis conjunto de **predecir elecciones efectivas**, y de forma más específica la representatividad de la muestra.
3. Aunque ésta técnica se ha empleado muy frecuentemente en estudios de los últimos 20 años, existe escasa investigación científica centrada en su **validez externa.** Un estudio confirmó que el análisis conjunto se corresponde estrechamente con los resultados de la tradicional contrastación de concepto, una metodología aceptada para la predicción de las preferencias del cliente [Tumbush, J. J. 1991].No hay aún **evaluación del error de muestreo en los modelos de nivel individual**, por lo que debe asegurarse siempre de que **la muestra sea representativa de la población a estudiar.** Esto es especialmente importante cuando los resultados conjuntos se utilizan para la **segmentación o simulación de elecciones**.

Figura 10.13. Diagrama de flujo análisis de conjunto pasos 4-5-6-7



Fuente: Hair et al. (1999)

10.11. Análisis de conjunto y sus aplicaciones.

Los modelos conjuntos estimados a **nivel individual** (**un modelo por individuo**), generalmente se utilizan en una o más de las siguientes áreas de apoyo a la toma de decisiones con **resultados a nivel individual [**Veiens et al. 1996**].** Los resultados agregados conjuntos pueden representar grupos de individuos y ofrecer también un medio de predecir sus decisiones en cualquier tipo de situación. Las aplicaciones más comunes, con su representación de la estructura de preferencias del consumidor incluyen :

1. **La segmentación**. A nivel individual, uno de los usos más comunes de los resultados del análisis conjunto es agrupar a los encuestados con **componentes parciales de la utilidad total o valores de importancia similares para identificar los segmentos**. Los componentes parciales estimados de la utilidad total pueden usar **por separado o en combinación** con otras variables (demográficas) para obtener **agrupaciones de encuestados que son más parecidos en sus preferencias** [Green y Kreiger 1991]. En el ejemplo del **servicio de telecomunicaciones** es posible encontrar un grupo para el que la **marca** es lo más importante, mientras que otro grupo puede valorar más la **velocidad de acceso**. De estar interesado en conocer la presencia de tales grupos y de sus magnitudes relativas, e**xisten varias aproximaciones a la segmentación, todas con diferentes fortalezas y debilidades.**
2. **El análisis de rentabilidad.** Un complemento a l**a toma de decisiones del diseño del producto** es un análisis de **rentabilidad marginal** del diseño del producto propuesto. **Si se conoce el coste de cada característica**, el coste de cada "***producto***" puede combinarse con la **cuota de mercado esperada y el volumen de ventas para predecir su viabilidad**. Este proceso puede señalar a una **combinación de atributos** con una r**educida cuota de mercado como la más rentable** debido a un aumento del margen de beneficio resultante del bajo coste de determinados componentes. También debe incluirse **la evaluación de la sensibilidad del precio,** que puede ser tratada a través de diseños de investigación específicos [Pinnell, 1994] o programas especializados [Sawtooth Software 1993]. **Pueden utilizarse, tanto los resultados individuales como los agregados.**
3. **Simuladores conjuntos.** ¿cómo consigue el análisis conjunto sus otros objetivos primarios utilizando el ana lisis ***"qué-pasa-si"*** para prediccio n de **la cuota de preferencias** que es probable que capture un estímulo (real o hipotético) en varios escenarios competitivos de interés para la gestión'? Este es el papel que juegan los simuladores de elección, que siguen un proceso de tres pasos:

   l. **Estimar y validar los modelos conjuntos para cada encuestado (o grupo).**

   2. **Seleccionar los conjuntos de estímulos a contrastar de acuerdo a los posibles escenarios competitivos.**

   3. **Simular las elecciones de todos los encuestados (o grupos) para los conjuntos específicos de estímulos y predecir las participaciones en la preferencia de cada estímulo agregando sus elecciones**.Una vez que se ha estimado el modelo conjunto, el investigador puede especificar cualquier número de conjuntos de estímulos para la simulación de las elecciones del consumidor. Entre los usos posibles están: (**a**) la evaluación del impacto de añadir un producto a un mercado existente; (**b**) el aumento potencial de una estrategia multimarca o multiproducto, incluyendo estimaciones de canibalismo comercial o (**c**) el impacto de eliminar una marca o un producto del

mercado. En cada caso, el investigador proporciona el conjunto de estímulos que representan al mercado y se simulan las elecciones de cada encuestado.Los simuladores de elección utilizan normalmente **2 tipos de reglas en la predicción de la elección de un estímulo** [Green y Kreiger 1988]:

a. **El modelo de máxima utilidad**, que asume que el encuestado elige el estímulo con la mayor utilidad prevista. Este modelo es el más apropiado en casos de mercados con individuos de preferencias muy diferentes y en situaciones que comprendan compras esporádicas y no rutinarias.
b. **La regla de elección alternativa** es una medida de probabilidad de compra, en la que la predicción de la suma de las probabilidades de elección suma el **100 por ciento** sobre el conjunto de estímulos contrastado. Esta aproximación se ajusta a **situaciones de compra repetitivas**, para las que las compras pueden estar más atadas a situaciones habituales en el tiempo. **Los dos métodos más comunes de hacer estas predicciones son los modelos *logit* y los BTL (Bradford-Terry-Luce**), que realizan predicciones bastante parecidas en casi todas las situaciones [Huber, J. y Moore 1979].

4. Se advierte a Usted, sobre la suposición de **que la participación de la preferencia en una simulación conjunta se traslade directamente a la participación en el mercado**. La simulación conjunta representa sólo aspectos del producto y quizá del precio interesantes para la gestión de la mercadotecnia, omitiendo todos aquellos factores de mercadotecnia (por ejemplo, **publicidad y promoción, distribución y respuestas competitivas**) **que afectan a la cuota de mercado**. La simulación conjunta presenta una visión del mercado de producto y la dinámica de preferencias que puede verse en la muestra bajo estudio.

10.12. Otras metodologías comparables con el análisis de conjunto.

Hasta este momento, se ha tratado con las aplicaciones del análisis conjunto incluidas en la metodología tradicional del análisis conjunto**. Sin embargo, las aplicaciones del mundo real, en muchas ocasiones implican 20 o 30 atributos o exigen un marco de decisión más realista de lo que hemos utilizado en discusiones anteriores**. La investigación reciente se ha dirigido a solucionar estos problemas encontrados en muchos estudios conjuntos, con **2 nuevas metodologías** que se están desarrollando:

(**a**) **un modelo adaptativo conjunto** para tratar con un gran número de atributos, y
(**b**) **un análisis conjunto basado en la elección** que proporciona elecciones más realistas. Estas áreas representan el eje fundamental de la investigación actual del análisis conjunto [Carrolly Green. 1995; Green y Srinivasan, 1990].

1.- **Análisis conjunto adaptativo:** Análisis conjunto con un gran número de factores los métodos de perfil completo y trade-offempiezan a ser inmanejables cuando consideran **entre 6 y 9 atributos**, aunque muchos estudios conjuntos necesitan **ingresar entre 20 y 30 atributos.** En estos casos, se utiliza una **forma adaptada o reducida** del análisis conjunto para simplificar el esfuerzo de la recogida de datos y representar una decisión realista. Las **2** opciones son los **modelos auto-explicados y los modelos adaptativos e híbridos.**

(a) Modelos conjuntos auto-explicados. El encuestado ofrece una calificación de la

atraccio n de cada nivel de un atributo y a continuacio n califica la importancia relativa del atributo conjunto. **Los componentes parciales de la utilidad total** se calculan mediante una combinación de los dos valores [Srinivasan,1988]. Se trata de un me todo composicional donde las calificaciones **se realizan sobre los componentes de la utilidad** en lugar de sobre una preferencia global. Como variante importante del ana lisis con- junto y ma s cercana a los modelos multiatributos tradicionales, esta forma de modelizacio n presenta varios temas de interés. Así , ¿pueden los encuestados evaluar la importancia relativa de los atributos con precisio n cuando la investigación muestra que estos pueden sub-estimarse en modelos multiatributos dado que los encuestados quieren dar respuestas socialmente deseables? En segundo lugar, las correlaciones inter-atributos juegan un papel importante y provocan sesgos sustanciales en los resultados debido a la "***doble contabilidad***" de los factores correlacionados. Finalmente, los encuestados nunca realizan una elección (califican el conjunto de combinaciones hipotéticas de atributos), y esta falta de realismo es una limitación crítica de las aplicaciones sobre nuevos productos. La investigación reciente, sin embargo, ha demostrado que este método puede tener una capacidad predictiva idónea cuando se compara con los métodos de análisis conjunto tradicionales [Green et al. 1991]. Por tanto, si el número de factores no puede reducirse a un nivel práctico aceptable en los métodos de análisis conjunto tradicional, entonces un modelo auto explicado puede ser un método viable alternativo al análisis conjunto tradicional.

(**b**) **Modelos de análisis conjunto híbridos o adaptativos.** Denominado as porque combina los modelos conjuntos de componentes parciales de la utilidad total y los auto-;explicados [Green, 1984_ Green et al.1981]. Utiliza valores auto-explicados al crear un conjunto reducido de est mulos (de tres a nueve) seleccionados a partir de un disen o factorial fraccional. Los estímulos se evalúan a continuacón de forma similar al análisis conjunto tradicional. Los conjuntos de est mulos difieren entre los encuestados, y aunque cada encuestado evalu a so lo un nu mero reducido, se evalu an colectivamente todos los est mulos por una parte de los encuestados. La aproximacio n de integrar informacio n del encuestado para simplificar o aumentar los trabajos de elección ha llevado a que recientemente se hayan realizado diversas investigaciones sobre diferentes aspectos del disen o de investigacio n [Allenby et al 1995, Jedidi, et al. 1996, Srinivasan,1997, Van der Lans, 1992]. Una de las variantes más comunes de esta aproximaciones **ACA**, un software de ana lisis conjunto desarrollado por **Sawtooth Software** [Sawtooth Software 1993]. **ACA** emplea calificaciones auto-explicadas para reducir el taman o del disen o factorial y hacer el proceso más manejable. Su relativa capacidad predictiva se ha mostrado comparable a la del análisis conjunto tradicional, y es una alternativa adecuada cuando el nu mero de atributos es elevado [Green et al.1991, Johnson, 1991, Tumbush, 1991, Wittink et al. 1994].Cuando nos enfrentamos con un nu mero de factores que no pueden ser admitidos en los me todos conjuntos discutidos hasta este momento, los modelos h bridos o adaptativos y auto explicados preservan al menos una parte de los principios que subyacen al ana lisis conjunto. Al comparar estas dos extensiones, los me todos auto-explicados tienen relativamente una baja fiabilidad, aunque existen investigaciones recientes encaminadas a mejorarla. Cuando los modelos h bridos y los auto-explicados se comparan con los me todos de perfil completo, los resultados son mixtos, aunque funciona ligeramente mejor el **método híbrido o adaptativo, particularmente el ACA** [Huber et al. 1992]. Aunque se

necesita ma s investigacio n para confirmar las comparaciones entre los mtodos, los estudios emp ricos indican que tanto los me todos h bridos o adaptativos como las nuevas formas de modelos auto-explicados ofrecen alternativas viables al análisis tradicional conjunto cuando se trata con un número elevado de factores.

2. Análisis conjunto basado en elección: Introducir otro toque de realismo

En los últimos años, muchos investigadores del área del análisis conjunto han dirigido sus esfuerzos hacia una nueva metodología conjunta que proporcione una mayor dosis de realismo en la tarea de elegir. Con el imperioso objetivo de entender el proceso de toma de decisiones del encuestado y predecir el comportamiento en el mercado, el análisis conjunto tradicional asume que el juicio, basado en clasificaciones o calificaciones, recoge la elección del encuestado. Algunos investigadores han argumentado que no es el modo más realista de representar el proceso de decisión efectivo, y otros investigadores han señalado la falta de una teoría formal que vincule estos juicios de medida de la elección [Louviere et al. 1988]. Lo que ha surgido es una metodología conjunta alternativa, conocida como análisis conjunto basado en elecciones, con la validez en principio inherente de preguntar al encuestado que elija un estímulo de perfil completo de un conjunto de estímulos conocidos como un conjunto de elección. Esta situación es mucho más representativa del proceso real de selección de un producto de entre un conjunto de productos competidores. Además, el análisis conjunto basado en la elección ofrece una opción de no elección de ninguno de los estímulos presentados al incluir una opción de no elección en el conjunto de elección. Mientras que el análisis conjunto tradicional supone que las preferencias de los encuestados se asignan siempre entre el conjunto de estímulos, la aproximación basada en la elección permite una contracción del mercado si todas las alternativas del conjunto de elección no son atractivas.

10.13. Análisis conjunto de perfil completo vs. análisis basado en la elección

Un simple ejemplo sólo para efectos ilustrativos. Una compañía de teléfonos celulares desea estimar el mercado potencial de tres opciones de servicio que pueden añadirse a la tarifa del servicio básico de 14.95$ al mes y 0.50$ al minuto por tiempo de llamada:

TMX. Factura detallada con un cargo de 2.75$ al mes.
Axtl. Llamada en espera con cargo de 3.50$ al mes.

Iusacl. Llamada a tres con un cargo mensual de 3.50$. El análisis conjunto tradicional se realiza con estímulos de perfil completo que representan varias combinaciones de servicio, que van desde el servicio base al servicio base con las tres opciones. En la Figura **Figura 10.14** se muestra el conjunto completo de perfiles (diseño factorial). El estímulo 1 representa el servicio base sin opciones, el estímulo 2 es el servicio base más la factura detallada, y así sucesivamente hasta el estímulo 8, que es el servicio base con las tres opciones (factura detallada, llamada en espera y llamada a tres). Se pide al encuestado que clasifique o califique estos ocho perfiles.

En una aproximación basada en la elección, se muestra al encuestado una serie de conjuntos de eleccio n, cada uno teniendo varios est mulos de perfil completo. Un disen o basado en la elección se muestra también en la **Figura 10.14**. El primer conjunto de eleccio n consiste en cinco de los est mulos de perfil completo (est mulos l, 2, 4, 5 y 6) y una opcio n de "***ninguno de éstos***". A continuacón el encuestado escoge sólo uno de los perfiles .

del conjunto de elección ("***más preferido***" o "***más deseado***") o la opción "***ninguno de éstos***". En la Tabla **10.15** se muestra un ejemplo de la preparación de un conjunto de elección para la elección por parte del encuestado, en particular para el conjunto 6. La preparación de los estímulos y los conjuntos de elección se basa en los principios de diseño experimental [Jedidi et al.1996 y Louviere, 1983] y es el objeto de un considerable esfuerzo investigador para refinar y mejorar la tarea de elegir [Allenby et al. 1995, Carroll, 1995, Huber, 1996; Jedidi et al.1996, Pinnell, J. 1994].

Figura 10.14. Tabla de comparación de diseño de estímulos: cálculo tradicional vs. basado en la elección

Análisis conjunto tradicional				**Análisis basado en elección**	
Estímulos	**Niveles de los factores**				
	TMX	***Axtl***	***IUSACl***	**Conjunto de elección**	**Estímulos del conjunto de elección**
1	0	0	0	1	1, 2, 4, 5, 6 y no elección
2	1	0	0	2	2, 3, 5, 6, 7 y no elección
3	0	1	0	3	1, 3, 4, 6, 7, 8 y no elección
4	0	0	1	4	2, 4, 5, 7, 8 y no elección
5	1	1	0	5	3, 5, 6, 8 y no elección
6	1	0	0	6	4. 6, 7 y no elección
7	0	1		7	1, 5, 7, 8 y no elección
8	1	1		8	1, 2, 6, 8 y no elección
				9	1, 2, 3, 7 y no elección
				10	2, 3, 4, 8 y no elección
				11	1, 3, 4, 5 y no elección

Fuente: Hair, 1999

Figura 10.15. Ejemplo de un conjunto de elección en un análisis conjunto basado en elección

¿Qué sistema de teléfono elegiría?			
1	**2**	**3**	**4**
Sistema base a 14,95$/mes y 0.50/$minuto más: **+TMX--**IIamada a tres por sólo 3.50$/mes	Sistema base a 14,95$/mes y 0.50$/minuto más: **Axtl**-facturación detallada por sólo 2.75$/mes	Sistema base a 14.95$/mcs y 0.50$/minuto más: **IUSACL**-Ilamada en espera por sólo 3,50$/mes y +TWC-Ilamada a tres por sólo 3.50$/mes	Ninguno de éstos

Fuente: Hair, 1999

El nu mero de perfiles var a a lo largo de los conjuntos de eleccio n. Tambie n, el nu mero de elecciones hechas (una elección para cada uno de los 11 conjuntos de eleccio n) es en realidad superior en este caso a lo que se requerir a en el diseñoo factorial. Pero a medida que el nu mero de factores y niveles aumenta, el disen o basado en la eleccio n requiere considerablemente menos evaluaciones (recordemos que nuestro ejemplo anterior de cuatro factores con cuatro niveles generaba 256 estímulos). Las ventajas de la aproximacio n basada en la elecćon consisten en el realismo adicional que introduce y la capacidad para estimar los términos de interacción, lo que no es posible con el ana lisis conjunto tradicional. Despue s de que cada encuestado ha escogido un est mulo para cada conjunto de elección, se agregan los datos para todos los encuestados (segmentos u otra agrupación homogénea de encuestados) para estimar los componentes parciales de la utilidad total para cada nivel y los te rminos de interaccio n. Desde estos resultados, podemos evaluar las contribuciones de cada factor y la interacción nivel-factor y estimar las cuotas de mercado probables de mercado de perfiles competidores.

10.14. Características únicas del análisis conjunto basado en la elección

La naturaleza básica de los modelos conjuntos basados en la elección y sus precedentes en el campo teórico de b integración de la información [Louvierc, J 1988] ha llevado a una perspectiva en cierta forma más técnica de lo que se puede encontrar en otras metodologías conjuntas. Mientras que las otras metodologías están basadas en sólidos principios experimentales y estadísticos, la complejidad adicional tanto en el diseño de estímulos como en la estimación ha sido objeto de una gran cantidad de investigaciones con el fin de desarrollar estas áreas. Gracias a estos esfuerzos, los investigadores tienen ahora una comprensión más clara de los supuestos implícitos en cada etapa. Las siguientes secciones detallan algunas de las áreas y temas en los cuales el análisis conjunto basado en la elección es distinto del resto de las metodologías conjuntas.

10.14.1. Tipo de proceso de toma de decisiones representada.

El análisis tradicional conjunto ha estado asociado siempre con una aproximación de información intensiva al proceso de toma de decisiones, caracterizado por el escrutinio de los estímulos de perfil completo compuesto de los niveles de cada atributo. Cada atributo está representado por igual y considerado en un perfil único. Pero en el análisis conjunto basado en la elección, los investigadores llegan a la conclusión de que la tarea de elegir

puede implicar un tipo diferente de proceso de toma de decisiones. Al hacer elecciones entre estímulos, los consumidores parecen elegir entre un conjunto reducido de factores sobre los cuales se hacen las comparaciones y, en última instancia, eligen [Huber et al. 1992]. Esto es similar a los tipos de decisiones asociadas con la restricción del tiempo o estrategias simplificadoras, caracterizadas por un bajo nivel de reflexión. Por tanto, cada metodología conjunta ofrece diferentes visiones en los procesos de toma de decisiones. Dado que los investigadores pueden no estar dispuestos a elegir una sola metodología en este momento, una nueva estrategia es emplear ambas metodologías y deducir perspectivas únicas de cada una de ellas [Huber et al. 1992].

1. **Diseño de estímulos** Quizá la mayor ventaja del modelo conjunto basado en la elección esta en el realismo del proceso de eleccio n representado por el conjunto de eleccio n. Dos desarrollos han mejorado aún más la tarea de elegir. En primer lugar, existe una eleccio n ma s realista e informativa entre alternativas parecidas y comparables, en vez de la situacón en la que uno de los estímulos es claramente inferior o superior. Sin embargo, el proceso de diseño de estímulos se centra en conseguir la ortogonalidad y el equilibrio entre los atributos. Un reciente trabajo ha mostrado co mo puede crearse el conjunto de eleccio n para asegurar el equilibrio no entre los niveles de los factores sino entre las utilidades de los est mulos [Huber, y Zwerina 1996]. Esto proporciona un esquema ma s realista, que puede aumentar la implicacio n del consumidor y ofrecer mejores resultados. En segundo lugar, en un método que implica informacio n adicional acerca de los encuestados, se crea un conjunto de eleccio n que ajuste las preferencias espec ficas de cada individuo y conseguir una mejor precisio n predictiva en las situaciones basadas en el mercado [Bretton-Clark 1988]. Estos dos desarrollos son caracter sticos de los esfuerzos que intentan mejorar la tarea de elegir haciendo de ella un método de evaluación incluso más realista y eficiente de las preferencias del consumidor.
2. **Técnicas de estimación** El fundamento conceptual del análisis conjunto basado en la elección es psicológico [Luce, 1959], pero fue el desarrollo de una técnica de estimación ***logit*** multinomial [McFadden, 1974], que ofrece un método operativo de estimación de estos tipos de modelos de elección. Aunque se han hecho esfuerzos considerables para refinar y hacer más asequible la técnica, todavía representa una metodología más compleja que aquellas asociadas con las otras metodologías conjuntas. Un aspecto particular que todavía queda sin resolver es la propiedad de IIA (independencia de alternativas irrelevantes), un supuesto que hace problemática la predicción de alternativas muy similares. Aunque explorar todos los supuestos que implica IIA está más allá del objeto de esta discusión, se aconseja al investigador que cuando utilice el análisis conjunto basado en la elección entienda las ramificaciones de este supuesto

10.14.2. Limitaciones del análisis conjunto basado en la elección y ventajas.

La creciente aceptación del análisis conjunto entre los practicantes de la investigación de mercados se debe fundamentalmente a la creencia de que la obtención de preferencias, teniendo los encuestados que elegir un u nico est mulo preferido entre un conjunto de est mulos, es más realista y por tanto un me todo mejor-para aproximarse al proceso de decisión efectivo. Adema s de an adir realismo a la tarea de eleccio n el investigador se encuentra con un número de ***"trade-off"*** que el investigador debe considerar antes de

elegir el modelo conjunto basado en la elección. La tarea de elegir cada conjunto de elección contiene varios estímulos y cada estímulo contiene varios factores a diferentes niveles, igual que los estímulos de perfil completo. Por tanto, el encuestado debe procesar una cantidad considerablemente mayor de información que otras metodologías del análisis conjunto a la hora de tomar una decisión en cada conjunto de elección. Sawtooth Software, promotores del sistema basado en la elección (***CBC***), creen que aquellas elecciones que implican más de seis atributos es probable que confundan y abrumen al encuestado [Sawtooth Software 1993]. Aunque el método basado en la elección reproduce las decisiones reales, la inclusión de muchas alternativas implica una tarea formidable que normalmente termina con menos información de la que se habría ganado a través de la calificación de los estímulos individualmente.

10.14.3. Precisión predictiva.

En la pra ctica, las **3 metodologías conjuntas** permiten tipos parecidos de ana lisis, simulaciones e información, incluso aunque los procesos de estimacio n sean diferentes. Aunque los modelos basados en la elección todavía tienen que someterse a más contrastaciones empíricas, algunos investigadores creen que tienen ventaja en la predicción de comportamiento decisor. Sin embargo, las contrastaciones empíricas indican una escasa diferencia entre los modelos basados en la calificación de niveles ajustados con la posibilidad de no elección y los modelos basados en la elección de tipo ***logit*** multinominal generalizado [Oliphant et al. 1992]. La conclusión fue que tanto el modelo basado en calificaciones como el modelo basado en la eleccio n predicen igualmente bien. Otra investigación comparo las dos aproximaciones al ana lisis conjunto (basado en las calificaciones o basado en la elección) en términos de la capacidad de predecir participaciones en una muestra ***holdout*** [Elrod et al. 1992]. Ambas aproximaciones predicen las elecciones de la muestra ***holdout*** bien, sin que ninguna aproximación sea dominante y los resultados se mezclen en diferentes situaciones. En última instancia, la decisión de utilizar un método sobre el otro está dictada por los objetivos y el alcance del estudio, la familiaridad del investigador con cada método y el software disponible para analizar apropiadamente los datos.

10.14.4. Aplicaciones prácticas.

Los modelos basados en la eleccio n estimados a nivel agregado ofrecen los valores y significación estadística de todas las estimaciones, producen fa cilmente predicciones de la cuota de mercado realistas para estímulos nuevos [Jedidi et al. 1996], y ofrecen seguros adicionales de que se utilizan "***elecciones***" entre estímulos para calibrar el modelo. Sin embargo, los modelos de elecćon agregados dificultan la segmentación del mercado. El análisis conjunto basado en la elecćon no es capaz de estimar un modelo conjunto diferenciado para cada encuestado, haciendo imposible agrupar a los encuestados de acuerdo a los resultados de sus modelos conjuntos tal y como se ha discutido previamente. Como contraste, los modelos basados en calificación descritos previamente se ajustan bien a los estudios de segmentación pero se enfrentan al problema de tests incómodos de significación estadística de las estimaciones de los componentes parciales de la utilidad total. Por tanto, los resultados pueden ser difíciles de resumir y la simula- ción de las

participaciones de la elección puede ser problemática.

10.14.5. Disponibilidad de los programas informáticos.

Existen varios programas informáticos basados en la elección a disposición de los investigadores que asisten en todas las fases del diseño de investigación, estimación del modelo e interpretación [Intelligent Marketing Systems, Inc. 1993, Sawtooth Software 1993]. Sin embargo, la investigación académica y aplicada se integra lentamente en esos programas disponibles comercialmente. La mayoría de los avances todavía se encuentran limitados a un reducido dominio y no están disponibles para un uso más amplio. Estas mejoras y nuevas capacidades, después de una rigurosa validación por la comunidad académica, deberían convertirse en parte habitual de todos los programas basados en la elección.

10.15. Análisis de Conjunto. Resumen para aplicar

El análisis conjunto permite responder preguntas que apuntan a las ciencias de la administración aplicadas a la mercadotecnia como : ¿qué atributos de producto/ servicio son importantes para el consumidor?, ¿cuáles son los atributos de producto/ servicio más atractivos al consumidor?, ¿cuál es la cuota de mercado de predferencia de los productos de los competidores en comparación con nuestro producto/servicio propuesto o existente? Para realizarlo, se requiere considerar:

1. **Enfoque de perfil completo:** con tolos encuestados clasifican , ordenan o dan puntuación a un conjunto de perfiles (conjunto de atributos del producto/servicio), que está dispuesto a evaluar. Cada perfil describe un producto/servicio completo y consta de una combinación diferente de niveles de factores para todos los factores (atributos) de interés.
2. **Matriz ortogonal:** es muy probable que surjan problemas con el enfoque de perfil completo si hay varios factores en juego y cada uno este compuesto por más de un par de niveles.
3. **Diseño factorial fraccional**, el cual presenta una fracción adecuada de todas las posible combinaciones de niveles de los factores. El conjunto resultante, se le denomina **matriz ortogonal,** el cual está diseñado para recoger los efectos principales de cada nivel de factor.
4. **La generación del diseño ortogonal,** crea una matriz ortogonal que suele utilizarse como punto de partida de un análisis conjunto.
5. La creación de los perfiles de los productos/servicios se simplifica a través del procedimiento **mostrar el diseño.** Ese procedimiento utiliza el diseño generado por la instrucción : **Generar diseño ortogonal o uno introducido por el usuario , generando un conjunto de perfiles de producto/servicio en un formato de fácil uso.**
6. Para **generalizar los resultados**, se selecciona una muestra aleatoria de los sujetos de la población de destino de manera que se tenga accesos a examinar los resultados de grupo. Un tamaño de muestra sugerido, de acuerdo a varios de los estudios de conjunto comerciales suele oscilar entre 100-10000 siendo el intervalo de **300-550** de lo más típico.

Para realizarlo, se sugiere además :

1. **Primer método,** se pide a los sujetos que asignen una **puntuación de preferencia** a cada perfil. Es habitual cuidando se utiliza una escala de Likert o cuando se solicita a los sujetos asignen un número del **1 al 100** para indicar dicha preferencia.
2. **Segundo método,** se puide a los sujetos que asignen un rango a cada perfil de **1 al número total de perfiles**
3. **Tercer métodos**, se pide a los sujetos que ordenen los perfiles según preferencia.
4. **Identificación y selección de atributos relevantes.**Esto es básico para categorizar al producto/servicio
5. **Definición de niveles** u opciones para cada atributo
6. **Definición de atributos y niveles para el potencial producto/servicio.** Debe definirse la combinación de atributos a ser evaluada,
7. **Recolección de datos.** Generalmente la evaluación de los productos/servicios se realiza en las **formas alternativas:**
    a. **De a pares,**
    b. **Solicitar al encuestado elija entre 2 productos/servicios o**
    c. **Repartir puntos entre ellos**, ordenando los productos en un ranking y calificándolos en una escala de 0-10 con un número mayor para mayor nivel de preferencias.
8. **Selección del método para obtener los valores de utilidad, con el SPSS**
9. **Requisitos:** se requiere de 2 archivos, uno de ***archivo de datos*** **y otro,un** ***archivo del plan*****.** Este último consta del conjunto de perfiles de productos/servicios que van a evaluar los sujetos y se debe generar mediante el procedimiento general: **Generar-diseño-ortogonal.** El archivo de datos contiene las clasificaciones o puntuaciones de preferencia de estos perfiles recopilados de los sujetos.

10.16. Análisis de Conjunto. Ejemplos.

Paso 1. Objetivos

Problema 1. La empresa **QUIMICA SAB** ha diseñado 3 modelos de equipos inyectores de medicina para capsulas, con precios por inyección (usd), tiempos de inyección (mseg) y capacidad máxima (ml) por caja de 1000 cápsulas. Le interesa saber cuál es el inyector que tiene mayores posibilidades de venta en el mercado de fabricación de medicinas. Ver **Figura 11.16**

Paso 2. Diseño

Figura 10.16. Tabla de Atributos y Niveles

	ATRIBUTOS		
	Precio_usd	**Tiempo_de_inyección_ seg**	**Maxima_capacidad_ml**
NIVELES	400,000	4	200
	700,000	6	400
	1,000,000	8	600
		10	800

Fuente: propia

- Las combinaciones que se toman en cuenta, son : **3 (Precio)*4(Tiempo de inyección)**

***4 (Máxima capacidad)=48**

- El diseño del estudio, se divide en 2 etapas:
  Etapa 1.- Se aplica el **método de diseño ortogonal,** que permite generar las mejores combinaciones de los atributos y sus niveles para apreciación del consumidor.
  El resultado se guardará en subdirectorio **Escritorio** (**Desktop**) en archivo **Plan_QUIMICA SAB.sav.**
  Etapa 2.- Creación de tabla cuestionario, en la que se mostrará al consumidor potencial el producto/servicio, en todas las combinaciones de atributos-niveles posibles para su evaluación (escala de 1-10 donde 1 es la peor y 10 la mejor calificación)

Paso 3. Condiciones de aplicabilidad

- No olvidar realizar los análisis previos de inspección visual de la base de datos para detectar ausentes y/o atípicos (corregir en su defecto), confiabilidad, normalidad, homocedasticidad y linealidad

Paso 4. Estimación y ajuste

Etapa 1.- Diseño Ortogonal

-Teclear: Datos->Diseño ortogonal->Generar->* Nombre del factor: Precio_usd; Tiempo_de_inyeccion_mseg; Maxima_capacidad_ml-> Definir valores ->Para Precio_usd: Valor (1), Etiqueta(400,000 usd); Valor(2), Etiqueta(800,000 usd); Valor(3), Etiqueta (1,000,000 usd)->Continuar-> Definir valores-> Para Tiempo_de_inyeccion_mseg: Valor (1), Etiqueta (4 mseg); Valor (2), Etiqueta (6 mseg); Valor (3), Etiqueta (8 mseg); Valor (4), Etiqueta (10 mseg)->Continuar-> Definir valores-> Para Maxima_capacidad_ml: Valor (1), Etiqueta (200 ml); Valor (2), Etiqueta (400 ml); Valor (3), Etiqueta (600 ml); Valor (4), Etiqueta (800 ml)-> Crear nuevo archivo de datos->*** Buscar en: Desktop; Archivo: Plan QUIMICA SAB-> Guardar->**** Restablecer semilla de aleatorización a: 2000000->Aceptar. Ver Figura 10.17 y Figura 10.18**

Notas:

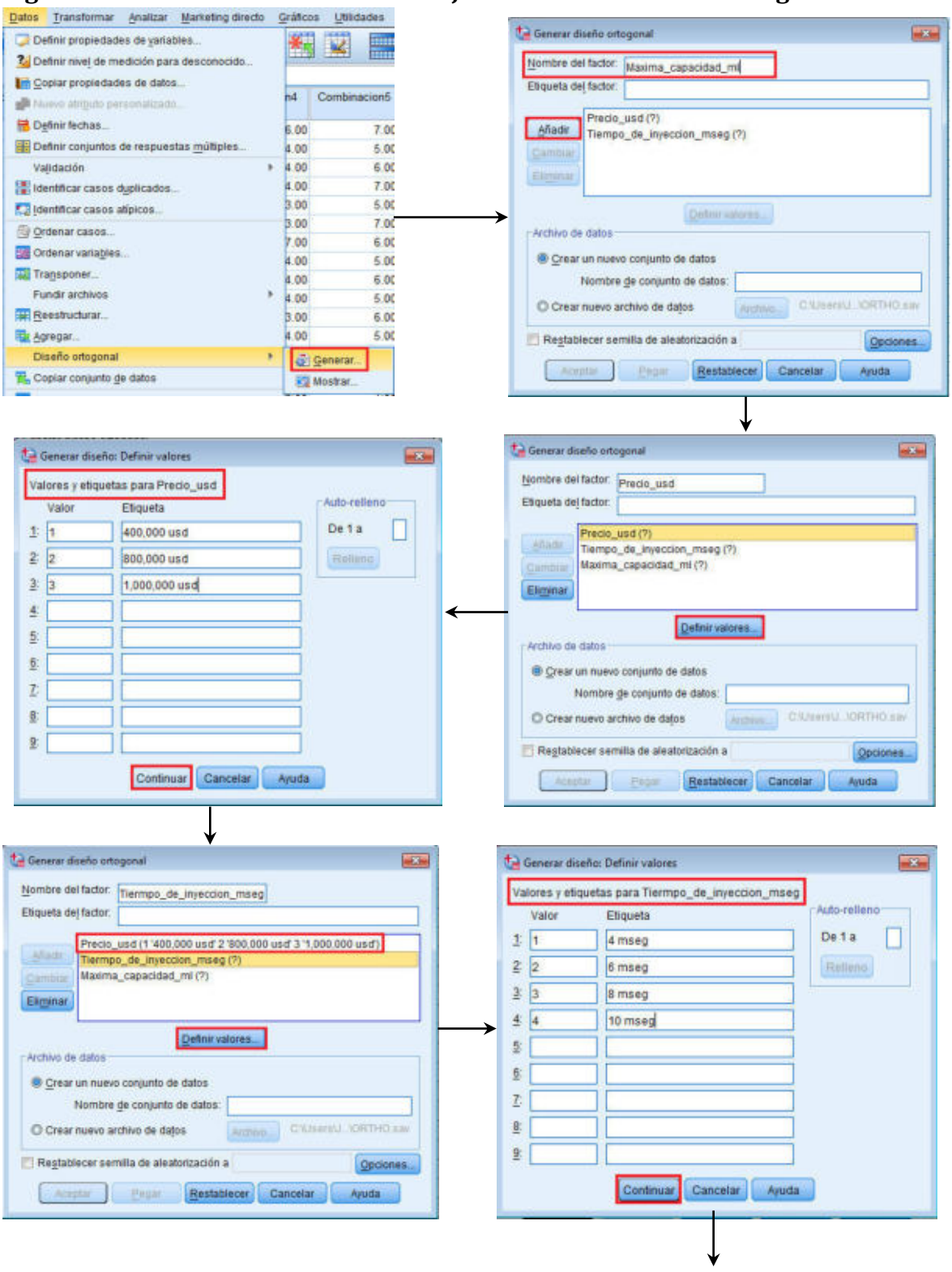
* .- Se dan los atributos y sus deferentes niveles

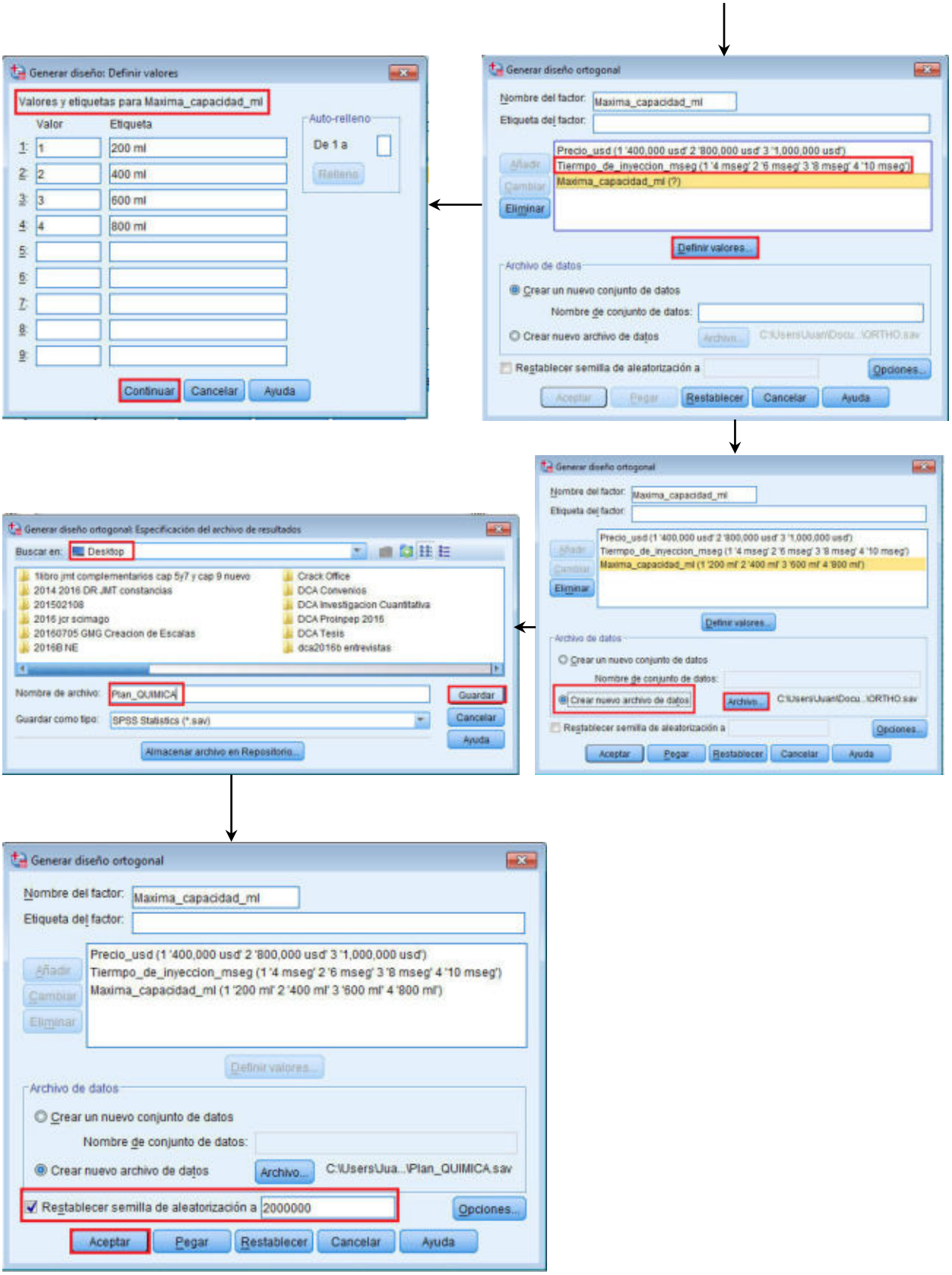
.- Se crea el archivo de datos en el que **SPSS sugerirá la cantidad de "***tarjetas***" , o sea, las combinaciones sugeridas a presentar a los consumidores

***.- Se asigna la ubicación y nombre de archivo a generar

******.-** Con 2,000,000 se recomienda generar resultados más precisos.

Figura 10.17.- Proceso Análisis de conjunto mediante diseño ortogonal





Fuente: SPSS 20 IBM

Figura 10.18. Resultado del plan ortogonal

```
*Generar diseño ortogonal.
SET SEED 2000000.
ORTHOPLAN
  /FACTORS=Precio_usd (1 '400,000 usd' 2 '800,000 usd' 3 '1,000,000 usd') Tiempo_de_inyeccion_mseg (1 '4 mseg' 2 '6 mseg' 3 '8 mseg'
4 '10 mseg') Maxima_capacidad_ml (1 '200 ml' 2 '400 ml' 3 '600 ml' 4 '800 ml')
  /OUTFILE='C:\Users\Juan\Desktop\Plan_QUIMICA.sav'.
```

Plan ortogonal

```
[Conjunto_de_datos1] C:\Users\Juan\Desktop\proy libro mc\QUIMICA SAB.sav
```

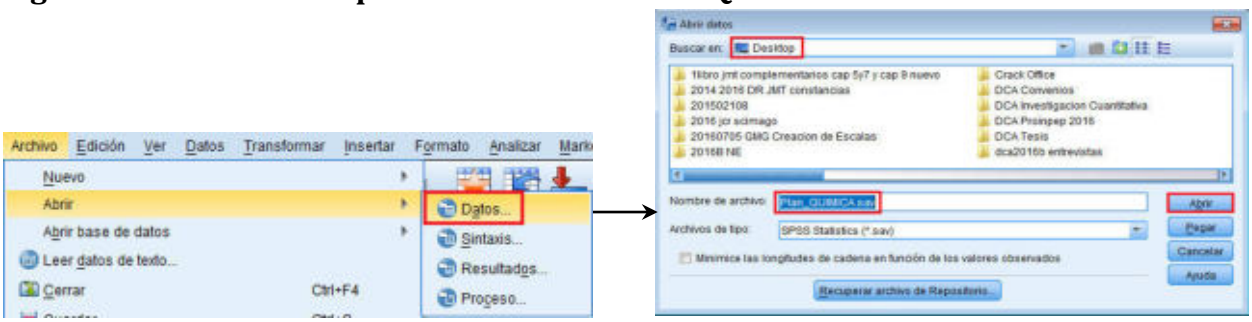
Advertencia
Se ha generado correctamente un plan con 18 tarjetas.

Fuente: SPSS 20 IBM

Etapa 2.

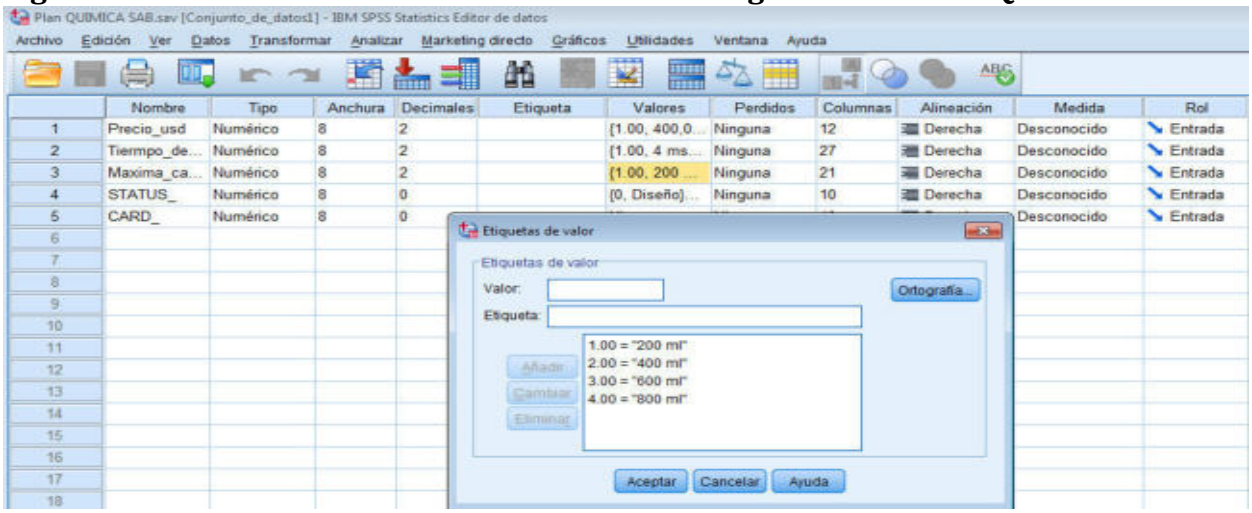
Teclear->Archivo->Abrir->Datos->Desktop->Nombre de Archivo: Plan QUIMICA SAB .sav->Abrir. Ver Figura 10.19 , Figura 10.20 y Figura 10.21.

Figura 10.19.- Proceso apertura de archivo Plan QUIMICA SAB.sav



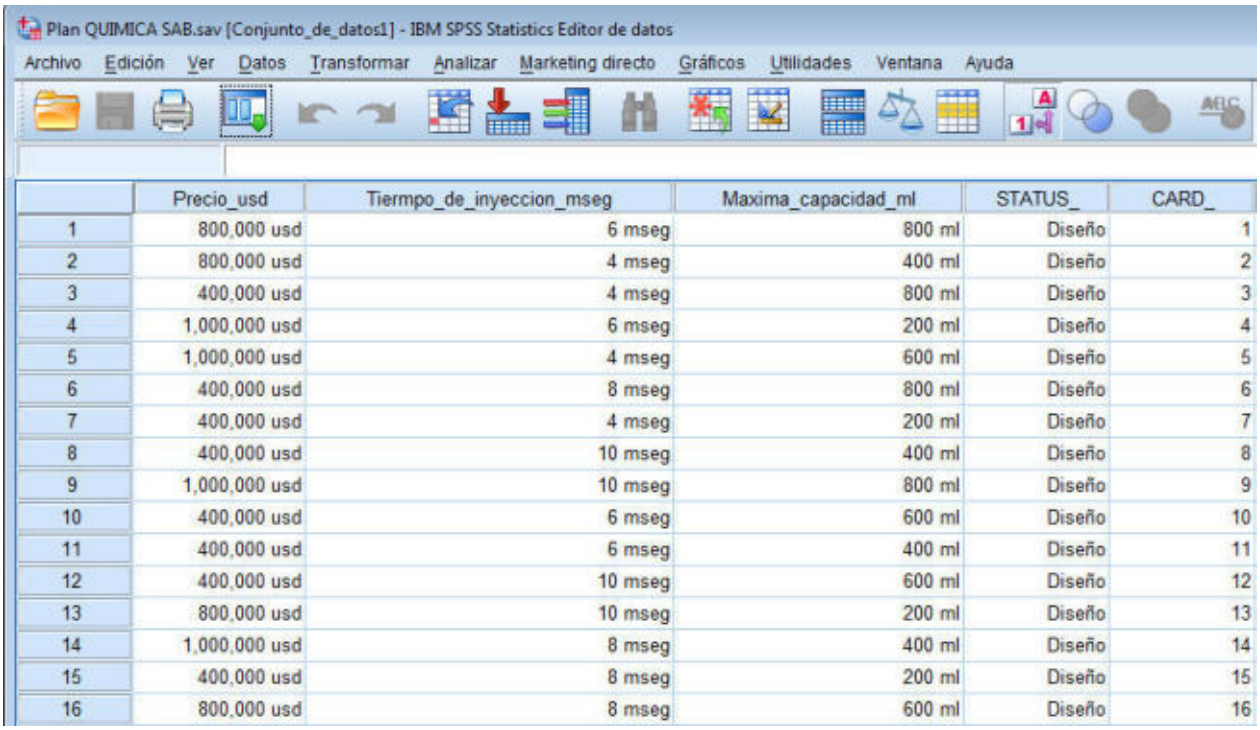
Fuente: SPSS 20 IBM

Figura 10.20.- Visor de variables del archivo generado Plan QUIMICA SAB.sav

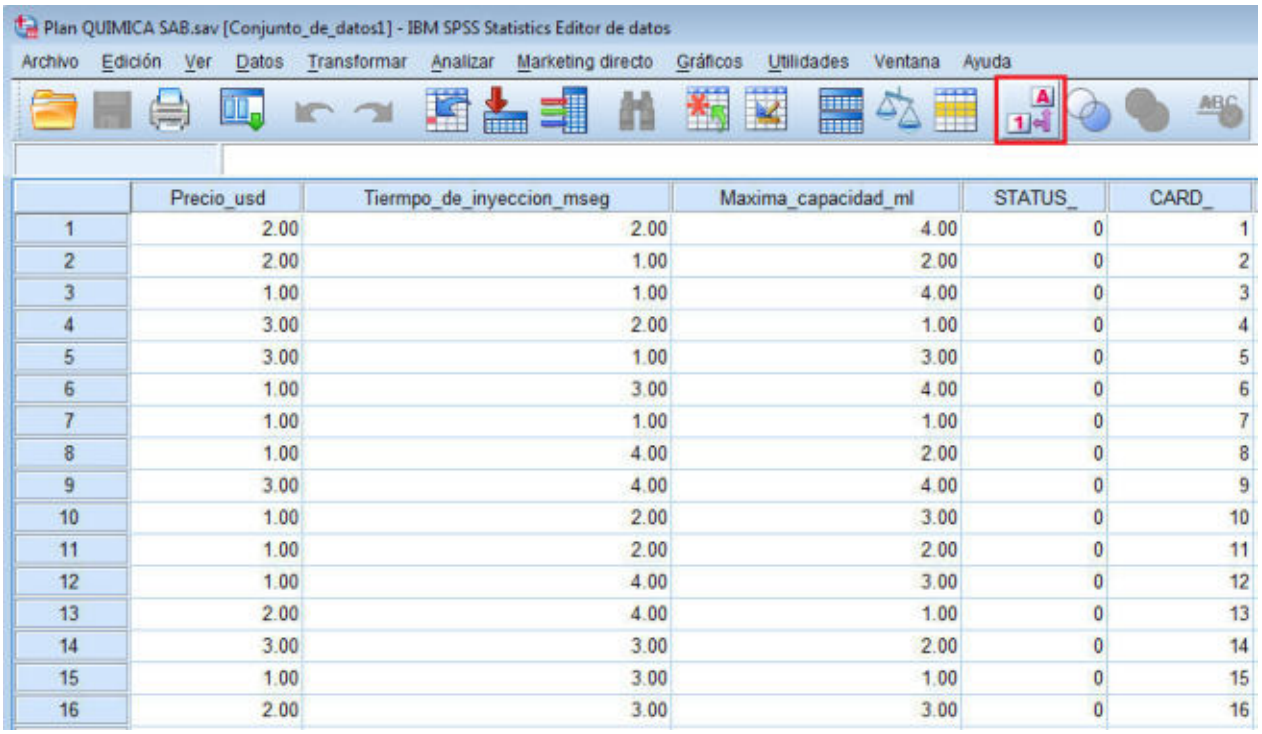


Fuente: SPSS 20 IBM

Figura 10.21.-Visor de datos del archivo generado Plan QUIMICA SAB.sav



	Precio_usd	Tiermpo_de_inyeccion_mseg	Maxima_capacidad_ml	STATUS_	CARD_
1	800,000 usd	6 mseg	800 ml	Diseño	1
2	800,000 usd	4 mseg	400 ml	Diseño	2
3	400,000 usd	4 mseg	800 ml	Diseño	3
4	1,000,000 usd	6 mseg	200 ml	Diseño	4
5	1,000,000 usd	4 mseg	600 ml	Diseño	5
6	400,000 usd	8 mseg	800 ml	Diseño	6
7	400,000 usd	4 mseg	200 ml	Diseño	7
8	400,000 usd	10 mseg	400 ml	Diseño	8
9	1,000,000 usd	10 mseg	800 ml	Diseño	9
10	400,000 usd	6 mseg	600 ml	Diseño	10
11	400,000 usd	6 mseg	400 ml	Diseño	11
12	400,000 usd	10 mseg	600 ml	Diseño	12
13	800,000 usd	10 mseg	200 ml	Diseño	13
14	1,000,000 usd	8 mseg	400 ml	Diseño	14
15	400,000 usd	8 mseg	200 ml	Diseño	15
16	800,000 usd	8 mseg	600 ml	Diseño	16



	Precio_usd	Tiermpo_de_inyeccion_mseg	Maxima_capacidad_ml	STATUS_	CARD_
1	2.00	2.00	4.00	0	1
2	2.00	1.00	2.00	0	2
3	1.00	1.00	4.00	0	3
4	3.00	2.00	1.00	0	4
5	3.00	1.00	3.00	0	5
6	1.00	3.00	4.00	0	6
7	1.00	1.00	1.00	0	7
8	1.00	4.00	2.00	0	8
9	3.00	4.00	4.00	0	9
10	1.00	2.00	3.00	0	10
11	1.00	2.00	2.00	0	11
12	1.00	4.00	3.00	0	12
13	2.00	4.00	1.00	0	13
14	3.00	3.00	2.00	0	14
15	1.00	3.00	1.00	0	15
16	2.00	3.00	3.00	0	16

Nota: La variable **CARD_** es considerada especial por lo que se sugiere mantener este nombre tanto en la realización de tabla cuestionario como en la programación de Sintaxis

Fuente: SPSS 20 IBM

Recopilando los datos, se crea tabla para que sea evaluada por los consumidores potenciales por ejemplo en **escala de 1a10**, donde 1 es la peor calificación y 1 la mejor. Ver la **Figura 10.22.**

Figura 10.22.-Tabla cuestionario que evaluará el consumidor potencial

Precio_usd	Tiempo_inyección_mseg	Maxima_capacidad_ml	Status	CARD_
800,000 usd	6 mseg	800 ml	Diseño	1
800,000 usd	4 mseg	400 ml	Diseño	2
400,000 usd	4 mseg	800 ml	Diseño	3
1,000,000 usd	6 mseg	200 ml	Diseño	4
1,000,000 usd	4 mseg	600 ml	Diseño	5
400,000 usd	8 mseg	800 ml	Diseño	6
400,000 usd	4 mseg	200 ml	Diseño	7
400,000 usd	10 mseg	400 ml	Diseño	8
1,000,000 usd	10 mseg	800 ml	Diseño	9
400,000 usd	6 mseg	600 ml	Diseño	10
400,000 usd	6 mseg	400 ml	Diseño	11
400,000 usd	10 mseg	600 ml	Diseño	12
800,000 usd	10 mseg	200 ml	Diseño	13
1,000,000 usd	8 mseg	400 ml	Diseño	14
400,000 usd	8 mseg	200 ml	Diseño	15
800,000 usd	8 mseg	600 ml	Diseño	16

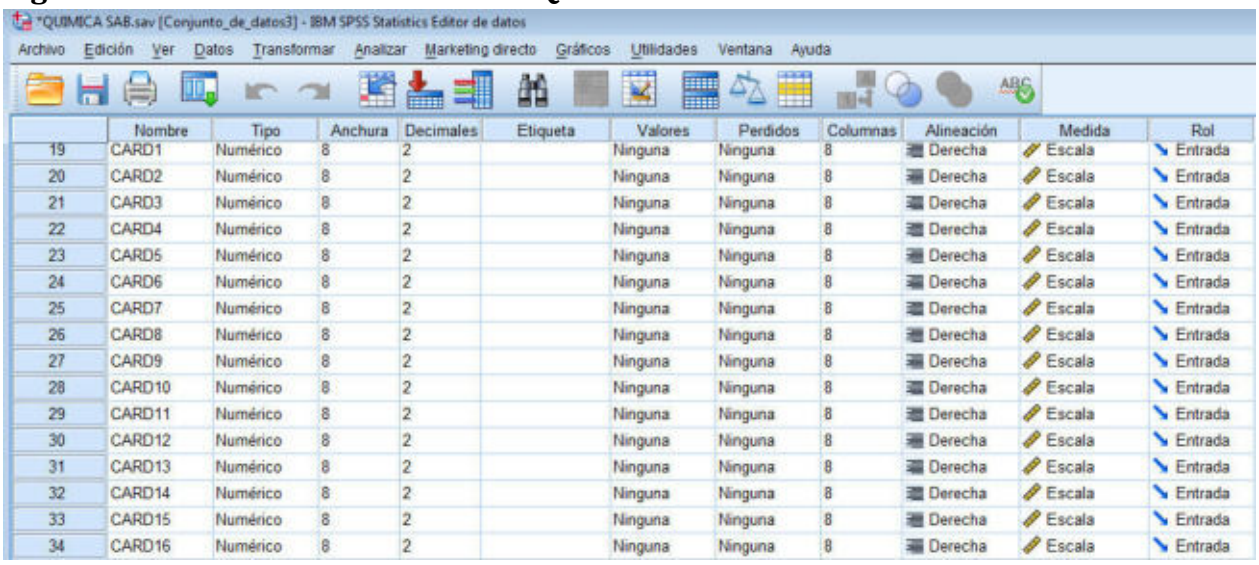
Nota: La variable **CARD_** es considerada especial por lo que se sugiere mantener este nombre tanto en la realización de tabla cuestionario como en la programación de Sintaxis

Fuente: propia

Etapa 3.

. Se crea archivo de captura del cuestionario. En nuestro ejemplo se tiene el registro de **24** casos, que califican las **16** combinaciones ubicándolas desde la combinación 1 a 16. (**Nota: NO DEBE REPETIRSE la asignación por parte del consumidor que evalúa, por registro**), contenidos en el archivo **QUIMICA SAB.sav**. Ver **Figura 10.23 y Figura 10.24**

Figura 10.23. Visor de datos archivo QUIMICA SAB.sav



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
19	CARD1	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala	Entrada
20	CARD2	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala	Entrada
21	CARD3	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala	Entrada
22	CARD4	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala	Entrada
23	CARD5	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala	Entrada
24	CARD6	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala	Entrada
25	CARD7	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala	Entrada
26	CARD8	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala	Entrada
27	CARD9	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala	Entrada
28	CARD10	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala	Entrada
29	CARD11	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala	Entrada
30	CARD12	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala	Entrada
31	CARD13	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala	Entrada
32	CARD14	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala	Entrada
33	CARD15	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala	Entrada
34	CARD16	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala	Entrada

Nota: La variable **CARD_** es considerada especial por lo que se sugiere mantener este nombre tanto en la realización de tabla cuestionario como en la programación de Sintaxis

Fuente: SPSS 20 IBM

Figura 10.24. Visor de variables archivo QUIMICA SAB.sav

	CARD1	CARD2	CARD3	CARD4	CARD5	CARD6	CARD7	CARD8	CARD9	CARD10	CARD11	CARD12	CARD13
1	1.00	2.00	3.00	6.00	7.00	8.00	9.00	10.00	4.00	5.00	13.00	11.00	12.00
2	2.00	1.00	3.00	4.00	5.00	6.00	7.00	9.00	8.00	10.00	11.00	12.00	13.00
3	3.00	1.00	2.00	4.00	6.00	5.00	8.00	9.00	7.00	11.00	10.00	12.00	13.00
4	1.00	2.00	4.00	3.00	7.00	6.00	5.00	9.00	8.00	10.00	12.00	11.00	13.00
5	1.00	2.00	4.00	3.00	5.00	6.00	7.00	8.00	10.00	9.00	11.00	12.00	13.00
6	1.00	2.00	4.00	3.00	7.00	5.00	9.00	6.00	8.00	10.00	15.00	11.00	16.00
7	2.00	1.00	3.00	7.00	6.00	5.00	11.00	12.00	4.00	8.00	16.00	14.00	15.00
8	2.00	1.00	3.00	4.00	5.00	6.00	8.00	7.00	10.00	9.00	13.00	11.00	12.00

Nota: La variable **CARD_** es considerada especial por lo que se sugiere mantener este nombre tanto en la realización de tabla cuestionario como en la programación de Sintaxis

Fuente: SPSS 20 IBM

. Por lo tanto, se requiere asociar tanto la base de datos de combinaciones (**Plan QUIMICA SAB.sav**) vs. el resultado del cuestionario levantado de los consumidores calificando los atributos-niveles del producto/servicio (**QUIMICA SAB.sav**). Lo anterior, se debe realizar con programación dentro de **SPSS** através del comando Sintaxis, desde el archivo que contenga los resultados a contrastar (en nuestro caso **QUIMICA SAB.sav**), así debe:

Teclear->Archivo->Nuevo->Sintaxis->

CONJOINT PLAN='C:\Users\Juan\Desktop\Plan QUIMICA SAB.sav'

Nota: Ubicación archivo previo de plan de combinaciones del diseño ortogonal a ofrecer a los consumidores para evaluación…es en esta sección que se genera la variable especial **CARD_**

/DATA='C:\Users\Juan\Desktop\QUIMICA SAB.sav '

Nota: Ubicación archivo de resultados de la evaluación que los consumidores hacen del producto/servicio que en este caso cada individuo posiciona desde la preferencia 1 hasta la 16…se sugiere hacer consecutivo a la variable **CARD_,** tantas veces como combinaciones se hayan encontrado en el archivo **Plan QUIMICA SAB.sav.** En nuestro caso 16; por lo tanto, se debe declarar dentro **del archivo QUIMICA SAB.sav, CARD1,CARD2,CARD3….CARD16**

/SEQUENCE=CARD1 TO CARD16

Nota: Secuenciación en la que el **SPSS** accesa a los datos y los analiza de acuerdo a la variable especial **CARD_**

/FACTORS=Precio_usd
Tiempo_de_inyeccion_mseg
Maxima_capacidad_ml

Nota: Declaración de los atributos que serán analizados por nivel provenientes del archivo **Plan QUIMICA SAB.sav**

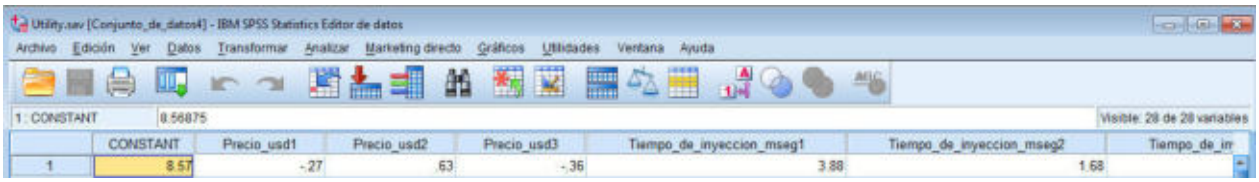
/PRINT=SUMMARYONLY
/UTILITY='C:\Users\Juan\Desktop\Utility.sav'
/PLOT=ALL

Nota:

PRINT.- Comando que habilita al **SPSS** mostrar todos los resultados producto de la ejecución de la Sintaxis

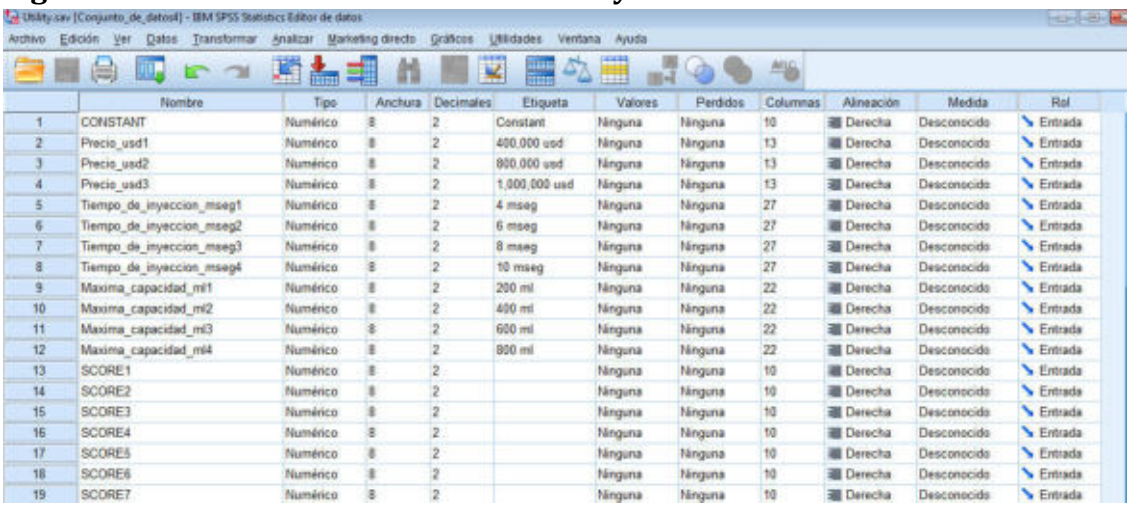
UTILITY.- Archivo .sav en la que se depositan de forma intermedia los resultados y **SCORES** (en nuestro caso **SCORE1…SCORE16**), producto de la ejecución de la Sintaxis. Ver **Figura 10.25 y Figura 10.26.**

Figura 10.25. Visor de variables archivo Utility.sav



Fuente: SPSS 20 IBM

Figura 10.26. Visor de datos archivo Utility.sav

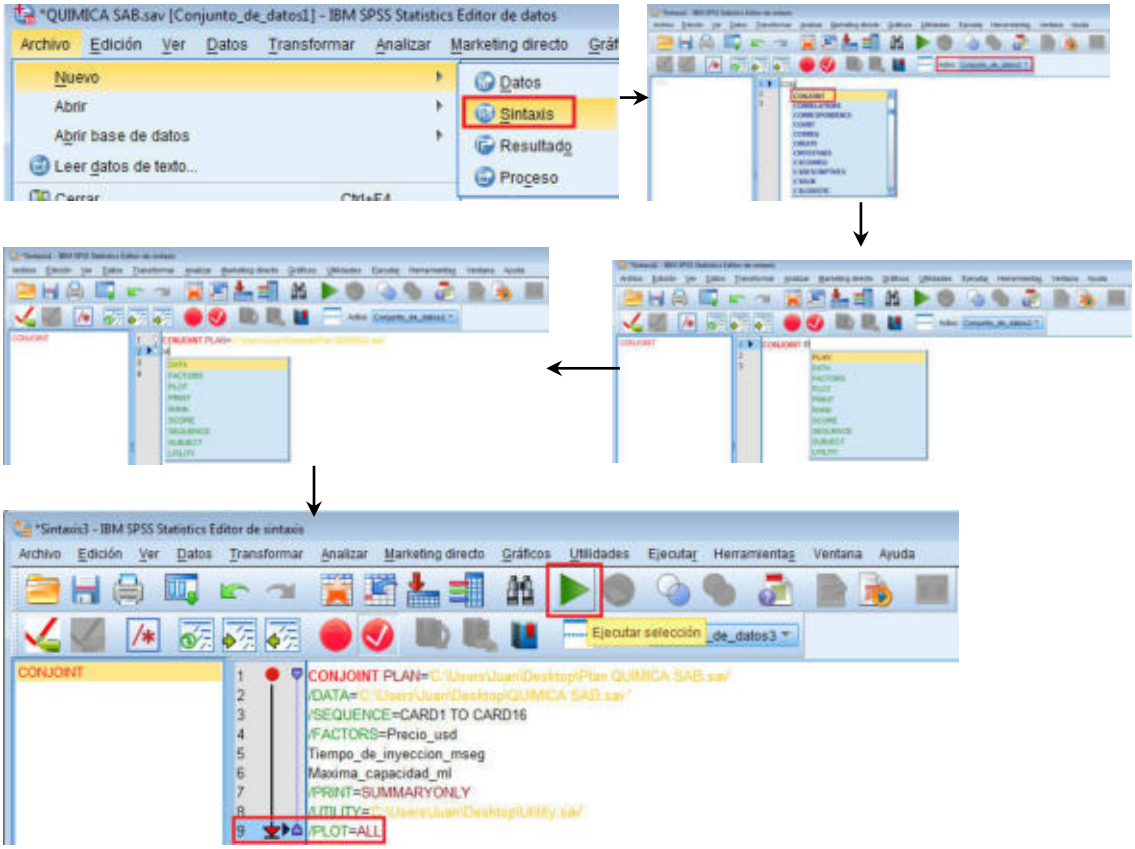


Fuente: SPSS 20 IBM

PLOT.-Comando que permite se generen los resultados por cada una de las variables bajo análisis producto de la ejecución de la Sintaxis.

->Ejecutar selección Ver Figura 10.27.

Figura 10.27. Proceso de ejecución de sintaxis de selección



Fuente: SPSS 20 IBM

Paso 5. Interpretación

. Con lo anterior **SPSS** genera un aviso de **Advertencia** para confirmar que el reporte NO generó ninguna inversión de variables respecto a los resultados obtenidos, es decir, confirma congruencia en los datos y resultados, con lo que proceso se realizó sin problemas. Así también, muestra la **tabla Descripción del modelo**, la cual despliega las variables o atributos del producto/servicio como los niveles de diferenciación que se trataron en el estudio. Ver **Figura 10.28.**

Figura 10.28. Advertencia proceso análisis de conjunto

Advertencia

No se ha producido ninguna inversión.

Descripción del modelo

	N° de niveles	Relación con rangos o puntuaciones
Precio_usd	3	Discreto
Tiempo_de_inyeccion_mseg	4	Discreto
Maxima_capacidad_ml	4	Discreto

Todos los factores son ortogonales.

Fuente: SPSS 20 IBM

. A continuación, **SPSS** muestra la tabla **Utilidades,** la cual es de los resultados más importantes ya que nos da los indicios sobre las preferencias de consumidores potenciales de la máquina de inyección propuesta por la empresa **QUIMICA SAB.**

Solución: En este caso, se observa como resultado que los consumidores potenciales se inclinan a comprar una máquina con **precio de 8000, 000 usd**, que tiene un **tiempo de inyección de medicamento de 4mseg**, con una capacidad de **800 ml** para una caja de **1000 capsulas,** por lo que es la sugerencia a lanzar a mercado. Ver **Figura 10.29**

Figura 10.29 Tabla Utilidades.

Utilidades

		Estimación de la utilidad	Error típico
Precio_usd	400,000 usd	-.275	1.046
	800,000 usd	.634	1.227
	1,000,000 usd	-.359	1.227
Tiempo_de_inyeccion_mseg	4 mseg	3.880	1.359
	6 mseg	1.683	1.359
	8 mseg	-3.980	1.359
	10 mseg	-1.582	1.359
Maxima_capacidad_ml	200 ml	-1.163	1.359
	400 ml	-.290	1.359
	600 ml	-2.592	1.359
	800 ml	4.045	1.359
(Constante)		8.569	.827

Fuente: SPSS 20 IBM

. Otra tabla muy apreciada que genera **SPSS**, es la de **Valores de importancia** la cual reporta cuál fue el atributo más importante que el consumidor determinó en el cuestionario de las combinaciones de presentación del producto/servicio. No es de extrañar que la alta competencia que existe en el sector de la industria química, muestre que en este caso, el

atributo más importante fuera el tiempo de inyección del medicamento (**50.742**), seguido de la máxima capacidad de inyección (**42.850**) y muy dejado atrás el precio (**6.407**). Ver **Figura 10.30.**

Figura 10.30. Tabla Valores de importancia

Valores de importancia

Precio_usd	6.407
Tiempo_de_inyeccion_mseg	50.742
Maxima_capacidad_ml	42.850

Puntuación promediada de la importancia

Fuente: SPSS 20 IBM

. La última tabla generada por **SPSS** es la de **Correlaciones**, que al tener altos valores de correlación, se asume que el modelo captura y predice de manera correcta las preferencias de los consumidores de la máquina de inyección de medicinas diseñado por la empresa **QUIMICA SAB.** Ver **Figura 10.31.**

Figura 10.31. Tabla de correlaciones

Correlaciones[a]

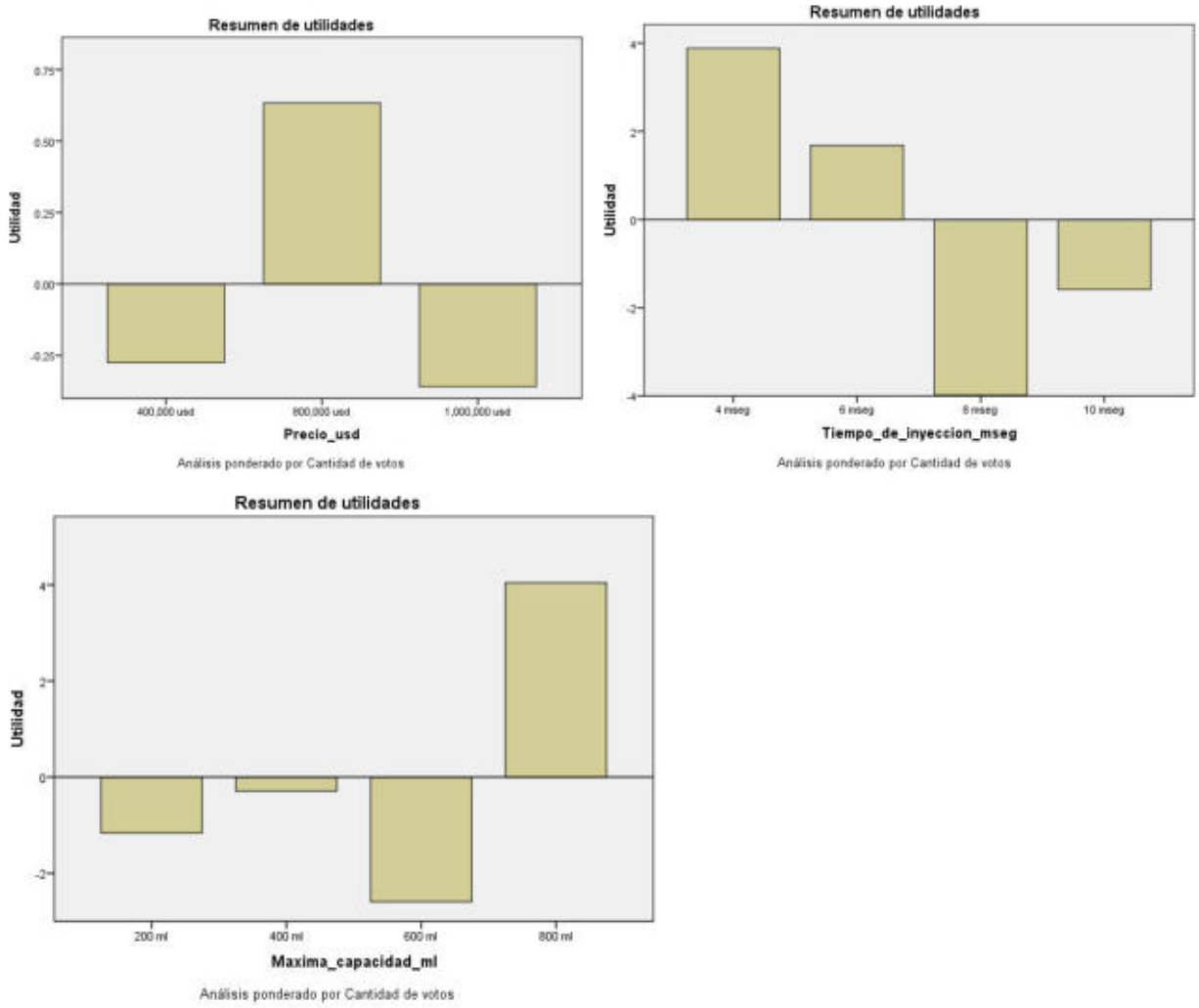
	Valor	Sig.
R de Pearson	.884	.000
Tau de Kendall	.700	.000

a. Correlaciones entre las preferencias observadas y las estimadas

Fuente: SPSS 20 IBM

. Finalmente, **SPSS** genera un grupo de gráficos que reforza de manera visual lo emitido en cada una de las tablas analizadas. Ver **Figura 10.32.**

Figura 10.32. Gráficos generados de cada uno de los atributos analizados



Fuente: SPSS 20 IBM

Referencias

Addelman, S. (1962), Orthogonal Main-Effects Plans for Asymmetrical Factorial Experiments. *Techno metrics* 4: 21-46.

Akaah, l. (1991), Predictive Performance of Self- Explicatcd, Traditional Conjoint, and Hybrid Conjoint Modcls undcr Alternative Data Collection Modcs. *Journal of the Academy of Marketing Science* 19:309-14.

Allenby, G. M., Arora, N., y Ginter, J. L. (1995), Incorporating Prior Knowledge into the Analysis of Conjoint Studies. *Journal of Marketing Research* 32 (May): 152-62.

Alpert, M. (1971), Definition of Determinan! Attributes: A Comparison of Methods. *Journal of Marketing Research 8*(2): 184-91.

Baalbaki, I. B., y Malhotra, N. K. (1995), Standardization versus Customization in Intemational Marketing: An Investigation Using Bridging Conjoint Analysis. *Journal of the Academy of Marketing Science* 23(3): 182-94.

Bretton-Clark (1988), Conjoint Analyzer. New York: Bretton-Clark.

Carmone, F. J., Jr., y Schaffer, C. M. (1995),Review of Conjoint Software. *Journal of Marketing Research* 32 (February): 113-20.

Carroll, J. D., y Green, P. E. (1995), Psychometric Methods in Marketing Research: Part 1, Conjoint Analysis. *Journal of Marketing Research* 32 (November): 385-91.

Conner, W. S., y Zelen, M. (1959*), Fractional Factorial Experimental Designs for Factors at Three Levels,* Applied Math Series S4. Washington, D.C.: National Bureau of Standards.

Elrod, T., Louviere, J. J. y Davey, K. S. (1992), An Empírical Comparison of Ratings-Based and Choice- Based Conjoint Models. *Journal of Marketing Research* 29: 368-77.

Green, P. E. (1984), Hybrid Models for Conjoint Analysis: An Exploratory Review. *Journal of Marketing Research* 21 (May): 155-69.

Green, P. E., Goldberg, S. M., y Montemayor, M. (1981), A Hybrid Utility Estimation Model for Conjoint Analysis. *Journal of Marketing 45* (Winter):33-41.

Green, P. E., Kreiger, A. M., y Agarwal M. K. (1991), Adaptive Conjoint Analysis: Sorne Caveats and Suggestions. *Journal of Marketing Research* 28 (May): 215-22.

Hahn, G. J., y Shapiro S. S. (1966*), A Catalog and Computer Program for the Design and Analysis of Orthogonal Symmetric and Asymmetric Fractional Factorial Experiments,* Report No. 66-C-165. Schenectady, N.Y.: General Electric Research and Development Center.

Hair , J.F.; Anderson, R.E.; Tatham, R.L.; Black W.C. (1999).*Análisis Multivariante.*5a. Ed. España. Prentice Hall.

Huber, J. (1987), Conjoint Analysis: How We Got Here and Where We Are, In *Proceedings of the Sawtooth Conference on Perceptual Mapping, Conjoint Analysis and Computer lnterviewing,* M. Metegrano, cd., Ketchum, Idaho: Sawtooth Software, pp. 2-6.

Huber, J., y Moore, W. (1979), A Comparison of Altemative Ways to Aggregate Individual Conjoint Analyses, In *Proceedings of the AMA Educator's Conference, L. Landon*, ed., pp. 64-68. Chicago: American Marketing Association.

Huber, J., Wittink, D. R., Fielder, J. A., y Miller, R. L. (1993), The Effectiveness of Altemative Preference Elicitation Procedures in Predicting Choice. *Journal of Marketing Research 30* (February): 105-14.

Huber, J., Wittink, D. R., Johnson, R. M., y Miller, R. (1992), *Leaming Effects in Preference Tasks: Choice-Based versus Standard Conjoint,* Sawtooth Software Conference Proceedings, M. Metegrano, ed. Ketchum, ID: Sawtooth Software, pp. 275-82.

Huber, J., y Zwerina, K. (1996), The Importance of Utility Balance in Efficient Choice Designs. *Journal of Marketing Research 33* (August): 307-17.

IBM (2011a). *IBM SPSS Statistics Base 20.* EUA. .Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Base.pdf

IBM (2011b).*Guía breve de IBM SPSS Statistics 20.* EUA.Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Brief_Guide.pdf

IBM (2011c). IBM SPSS Missing Values 20. EUA. .Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Missing_Values.pdf

Intelligent Marketing Systems, Inc. (1993), *CONSURV-Conjoint Analysis Software,* Version 3.0. Edmonton, Alberta: Intelligent Marketing Systems.

Jedidi, K., Kohli, R., y DeSarbo, W. S. (1996), Consideration Sets in Conjoint Analysis. *Journal of Marketing Research 33* (August): 364-72.

Johnson, R. M. (1991), Comment on Adaptive Conjoint Analysis: Sorne Caveats and Suggestions. *Journal of Marketing Research* 28 (May): 223-25.

Johnson, R. M., y Olberts, K. A. (1991), Using Conjoint Analysis in Pricing Studies: Is One Price Variable Enough? In *Advanced Research Techniques Forum Conference Proceedings*, pp. 164-73. Beaver Creek, Colo.: American Marketing Association, pp. 12-18.

Johnson, R. M., y Orme B. K. (1996), *How Many Questions Should You Ask in Choice-Based Conjoint Studies?* In Advanced Research Techniques Forum Confáence Proceedings, Beaver Creek, Colo.: American Marketing Association, pp. 42--49.

Kruskal, J. B. (1965), Analysis of Factorial Experiments by Estimating Monotone Transfonnations of the Data." *Journal of the Royal Statistical Societv* B27: 251-63.

Kuhfeld, W. F., Tobias, R. D., y Garrath, M. (1994), Efficient Experimental Designs with

Marketing Research Applications. *Journal of Marketing Research* 31 (November): 545-57.

Loosschilder, G. H., Rosbergen, E., Vricns, M., y Wittink, D. R. (1995), Pictorial Stimuli in Conjoint Analysis to Support Product Styling Decisions. *Journal of the Marketing Research Society* 37(1):17-34.

Louvierc, J. J. (1988), Ana(vzing Decision Making: Metric Conjoint Analysis. Sage University Paper Series on Quantitative Applications in the Social Sciences, vol. 67. Beverly Hills, Calif.: Sage.

Luce, R. D. (1959), *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley.

Mahajan, V., y Wind, J. (1991), *New Product Models: Practice, Shortcomings and Desired Improvements*-Report No. 91-125. Cambridge, Mass: Marketing Science Institute.

McFadden, D. L. (1974), *Conditional Logit Analysis of Qualitative Choice Bchavior. In Frontiers in Econometrics,* P. Zarcmbka, ed., pp. 105--42. New York: Academic Press.

MeLean, R., y Anderson, V. (1984), *Applied Factorial and Fractional Designs*. New York: Maree Dekker.

Oliphant, K., Eagle, T. C., Louviere, J. J., y Anderson, D. (1992), Cross-Task Comparison of Ratings-Based and Choice-Based Conjoint, In *Sawtooth Software Conjerence Proceedings*, M. Metegrano, ed., pp. 383--404. Ketchum, Idaho: Sawtooth Software.

Oppewal, H. (1995), A Review of Conjoint Software. *Journal of Retailing and Consumer Services* 2(1): 55-61.

Pinnell, J. (1994), Multistage Conjoint Methods to Measure Price Sensitivity, In *Advanced Research Techniques Forum,* pp. 65-69. Beaver Creek, Colo.: American Marketing Association.

Research Triangle Institute (1996), *Trade-Off VR.* Research Triangle Park, N.C.: Research Triangle Institute.

SAS Institute, Inc. (1992) *SAS Technical Report R/09: Conjoint Analysis Examples.* Cary, N.C.: SAS Institute, Inc.

Sawtooth Software.

Sawtooth Software (1993), *Conjoint Value Analysis*. Evanston, Ill.: Sawtooth Software.

Sawtooth Technologies (1997), *SENSUS*, Version 2.0. Evanston IL.: Sawtooth Technologies.

Schocker, A.D., y Srinivasan, V. (1977), *LINMAP* (Version II): *A Fortran IV Computer Program for Analyzing Ordinal Preference* (Dominance) Judgments Via Linear Programming Techniques for Conjoint Measurement. *Journal of Marketing Research 14* (February): 101-103.

SPSS, Inc. (1990), *SPSS Categories*. Chicago: SPSS, Inc.

Srinivasan, V. (1988), A Conjunctive-Compensatory Approach to the Self-Explication

of Multiattitudinal MPreference. *Decision Sciences* 19 (Spring): 295-305.

Srinivasan, V., Jain, A. K., y Malhotra, N. (1983), improving Predictive Power of Conjoint Analysis by Constrained Parameter Estimation. *Journal of Marketing Research* 20 (November): 433-38.

Srinivasan, V., y Park, C. S. (1997), Surprising Robustness of the Sclf-Explicated Approach to Customer Preference Structure Measurement. *Journal of Marketing Research 34* (May): 286-91.

Steckel, J., DeSarbo, W. S. y Mahajan V. (1991), On the Creation of Acceptable Conjoint Analysis Experimental Design. *Decision Sciences* 22(2): 435-442.

Tumbush, J. J. (1991), Validating Adaptive Conjoint Analysis (ACA) Versus Standard Concept Testing, In Sawtooth Software Conference Proceedings, M. Metegrano, ed. Ketchum, Idaho: Sawtooth Software, pp. 177-184.

Van der Lans, I. A., y Heiser, W. (1992), Constrained Part-Worth Estimation in Conjoint Analysis Using the Self-Explicated Utility Model. *lnternational Journal of Research in Marketing* 9: 325-44.

Veiens, M., Wedel, M. y Wilms T. (1996), Metric Conjoint Segmentation Methods: A Monte Cario Comparison. *Journal of Marketing Research 33* (February): 73-85.

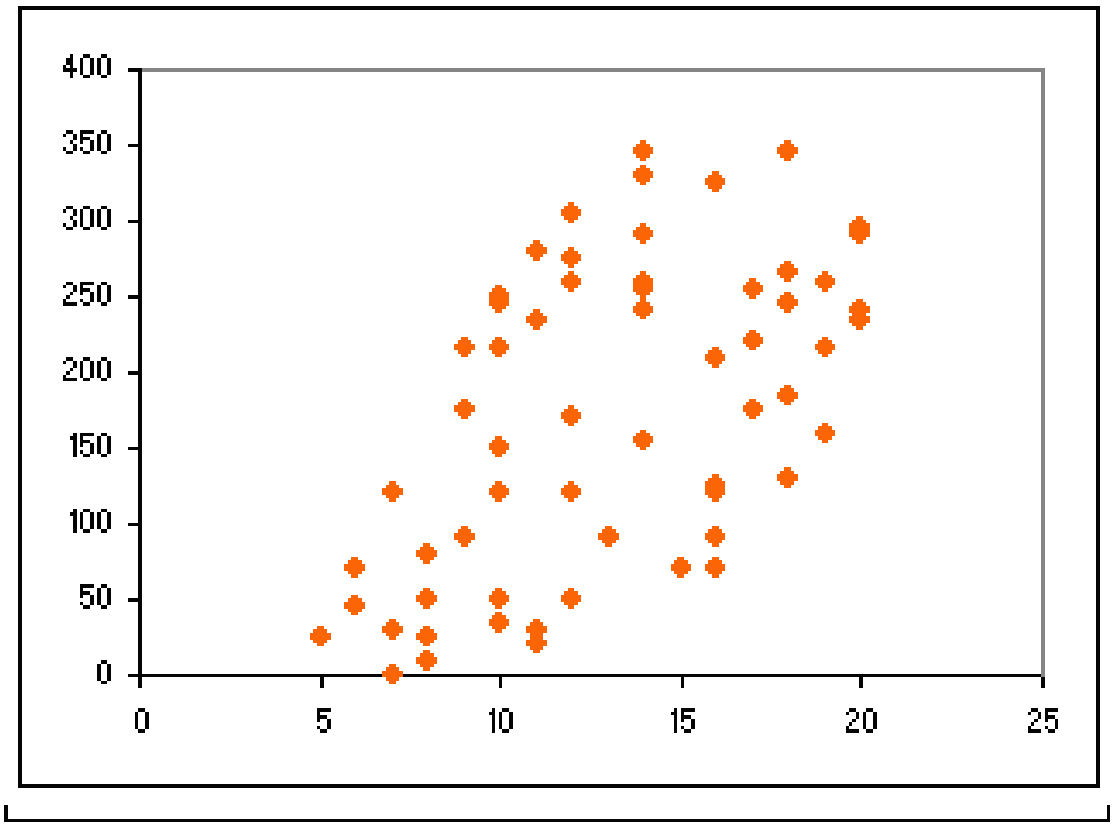
Wittink, D. R., y Cattin, P. (1989), Commercial Use of Conjoint Analysis: An Update. *Journal of Marketing 53* (July): 91-96.

Wittink, D. R., Huber, J. Zandan, P. y Johnson, R. M. (1992), *The Number of Levels Effect in Conjoint: Where Does It Come From, and Can It Be Eliminated?* In Sawtooth Software Conference Proceedings, M. Metegrano, ed. Ketchum, Idaho: Sawtooth Software, pp. 355-64.

Wittink, D. R., Krishnamurthi, L. y D. J. Reibstein (1990), The Effect of Differences in thc Number o Attribute Levels on Conjoint Results. *Marketin Letters* 1(2): 113-29.

Wittink, D. R., Vriens, M. y Burhenne, W. (1994) Commercial Use of Conjoint Analysis in Europe; Results and Critical Reflections. Internation . *Journal of Research in Marketing* 11: 41-52.

Capítulo 11. Análisis de Correlación Canónica



11.1. Correlación canónica. ¿Qué es?

Es una técnica estadística relativamente nueva y la disponibilidad de uso de software como el **SPSS** la hace ver con altas posibilidades para resolver problemas de investigación. **Es particularmente útil en situaciones donde se tienen múltiples variables dependientes** tales como: ventas, compras, índices, etc. Si las **variables predictoras** fueran exclusivamente **categóricas**, se podría emplear el análisis multivariante de la varianza (**MANOVA**) pero, ¿qué ocurre si las **variables predictoras son métricas**? **La correlación canónica es la respuesta,** ya que permite la valoración de la relación entre variables **predictoras métricas y múltiples medidas dependientes.** Como se discutió en los primeros capítulos, la **correlación canónica** es considerada como el modelo general en el que **se basan muchas otras técnicas mul tivariantes**, dado que se pueden emplear

tanto datos métricos como **no métricos** para **variables tanto dependientes como independientes.** Se expresa como::

$$\boldsymbol{Y_1 + Y_2 + Y_3 + \ldots + Y} \qquad \boldsymbol{X_1 + X_2 + X_3 + \ldots + X}$$
(Métrica, No Métrica) **(Métrica, No Métrica)**

En el **Capítulo 5** (sobre el **análisis de regresión múltiple**), se puede predecir el valor de una única variable criterio (**métrica**) a partir de una función lineal de un conjunto de variables predictoras **(independientes**). Para algunos problemas de investigación, el interés ya no es centrarse en una sola variable criterio (**dependiente**); sino en las **relaciones entre conjuntos de múltiples variables criterio y múltiples variables predictoras.** El **análisis de correlación canónica** es un modelo estadístico multivariante que facilita el estudio de las **interrelaciones entre múltiples variables criterio (dependientes) y múltiples variables predictoras (independientes)** ; en otras palabras, mientras que la **regresión múltiple predice una única variable dependiente a partir de un conjunto de múltiples variables independientes, la correlación canónica predice simultáneamente múltiples variables dependientes a partir de múltiples variables independientes.** La técnica establece **el menor número de restricciones sobre los tipos de datos con los que se trabaja.** Dado que las otras técnicas imponen restricciones más rígidas, se acepta generalmente que la información obtenida a partir de ellas es de una mayor calidad y que se puede interpretar más fácilmente. Por esta razón, se considera la **correlación canónica** como un **último intento,** para ser empleado cuando otras técnicas de mayor nivel no han funcionado correctamente. Pero en situaciones con múltiples variables dependientes e independientes, la correlación canónica es la técnica multivariante más apropiada y potente. Ha logrado la aceptación en muchos ámbitos y representa una herramienta útil para el análisis multivariante, especialmente dado el interés de considerar las variables múltiples dependientes. (Hair et al., 1999; IBM, 2011ª; IBM, 2011b; IBM, 2011c)

11.2.-Correlación canónica. Un caso supuesto.

Si tomamos en cuenta lo expuesto en el **Capítulo 5,** y a fin de aclarar mejor la naturaleza de la correlación canónica, consideremos el caso. Recuerde que los resultados de la encuesta de usaban el tamaño de la familia y la renta como predictores del número de tarjetas de crédito que tendría una familia. **El problema implicaba el examen de la relación entre dos variables independientes y una única variable dependiente.** Suponga que estuviera Usted interesado en un concepto más general del uso del crédito por los consumidores. Para medir este concepto, no solamente se requiere del número de tarjetas de crédito que tienen las familias, sino también los cargos medios mensuales en dinero que usen. Se llegó a la **conclusión de que estas dos medidas daban una mejor perspectiva sobre el uso de la tarjeta de crédito de la familia**. Por cierto, técnicas con igual tratamiento de datos:

1. Son el **análisis factorial**) y los **modelos de ecuaciones estructurales**. Ahora, **el problema incluye la predicción de dos medidas dependientes simultáneamente, el número de tarjetas de crédito y los cargos medios en dólares,** y la **regresión**

múltiple sólo es capaz de manejar una única variable dependiente.

2. Se podría utilizar el análisis multivariante de la varianza (**MANOVA**), pero únicamente si todas las variables independientes fuesen no métricas, no siendo éste el caso.
3. La **correlación canónica** es, por tanto, la única técnica disponible para examinar la relación con múltiples variables dependientes, que tienen una relación simple.

El problema de predecir el uso del crédito está reflejado en la **Figura 11.1**

Figura 11.1 . Tabla, correlación canónica del uso del crédito

Medidas de uso de crédito • Número de posesión de tarjetas de crédito • Gasto medio mensual de tarjetas de crédito en dólares		Caracteristicas de cliente • Tamaño familiar • Ingresos familiares
Compuesto de variables dependientes	Correlación canónica	Compuesto de variables independientes
Valor teórico canónico dependiente	***c***	Valor teórico canónico independiente

Nota: número de tarjetas de crédito y porcentaje de uso vs. perfil de clientes (tamaño familiar y renta familiar)

Fuente: Hair et al, 1999

Para lograrlo, las dos variables dependientes empleadas para medir el uso del crédito – (número de tarjetas de crédito que tienen las familias y gastos medios mensuales en realizados con las tarjetas) están representadas en el lado izquierdo. Las dos variables independientes seleccionadas para predecir el uso del crédito (el tamaño y la renta de las familias) se muestran en el lado derecho. Con el empleo del análisis de correlación canónica, se puede predecir una medida compuesta del uso del crédito que incluye ambas medidas dependientes, en lugar de tener que calcular una ecuación de regresión separada para cada una de las variables dependientes. Como resultado de aplicar la correlación canónica, se obtiene una medida de la validez de la relación entre los dos conjuntos de múltiples variables (valores teóricos). Esta medida se expresa como un **coeficiente de correlación canónica (Rc)** entre los dos valores teóricos. Ahora el investigador tiene dos resultados de interés; los va- lores teóricos canónicos que representan las combinaciones lineales óptimas de las variables dependientes e independientes y la correlación canónica que representa la relación entre ellas.

11.3. Correlación canónica. Análisis de las relaciones.

La técnica es el método más generalizado de la familia de las técnicas estadísticas multivariantes y se relaciona directamente con varios métodos de dependencia. Al igual que en **la regresión,** el objetivo de la correlación canónica es **cuantificar la validez de la relación, en este caso entre los dos conjuntos de variables (dependiente e independiente).** Se asemeja al **análisis factorial** en la creación de compuestos de variables. También se parece al **análisis discriminante** en su capacidad para determinar las dimensiones independientes (igual que las funciones discriminantes) para cada conjunto de

variables que produce la correlación máxima entre las dimensiones. De esta manera, la correlación canónica **identifica la estructura óptima o la dimensionalidad de cada conjunto de variables, que maximiza la relación entre los conjuntos de variables dependientes e independientes.**

La técnica trata con la asociación entre los **compuestos de conjuntos de variables múltiples dependientes e independientes.** Por ello, desarrolla varias funciones canónicas que **maximizan la correlación entre las combinaciones lineales, también conocidas como los valores teóricos canónicos, que son conjuntos de variables dependientes e independientes.** Cada función canónica se basa realmente en la correlación entre dos valores teóricos canónicos, **un valor teórico para las variables dependientes y otro para las variables independientes**. Otra característica única de la correlación canónica es que se obtienen los **valores teóricos de forma que se maximice su correlación**. Además, la correlación canónica no acaba con la obtención de una **relación simple** entre los conjuntos de variables. En su lugar, **se pueden conseguir varias funciones canónicas (pares de valores teóricos canónicos).**

11.4. Correlación canónica: Paso 1. Objetivos.

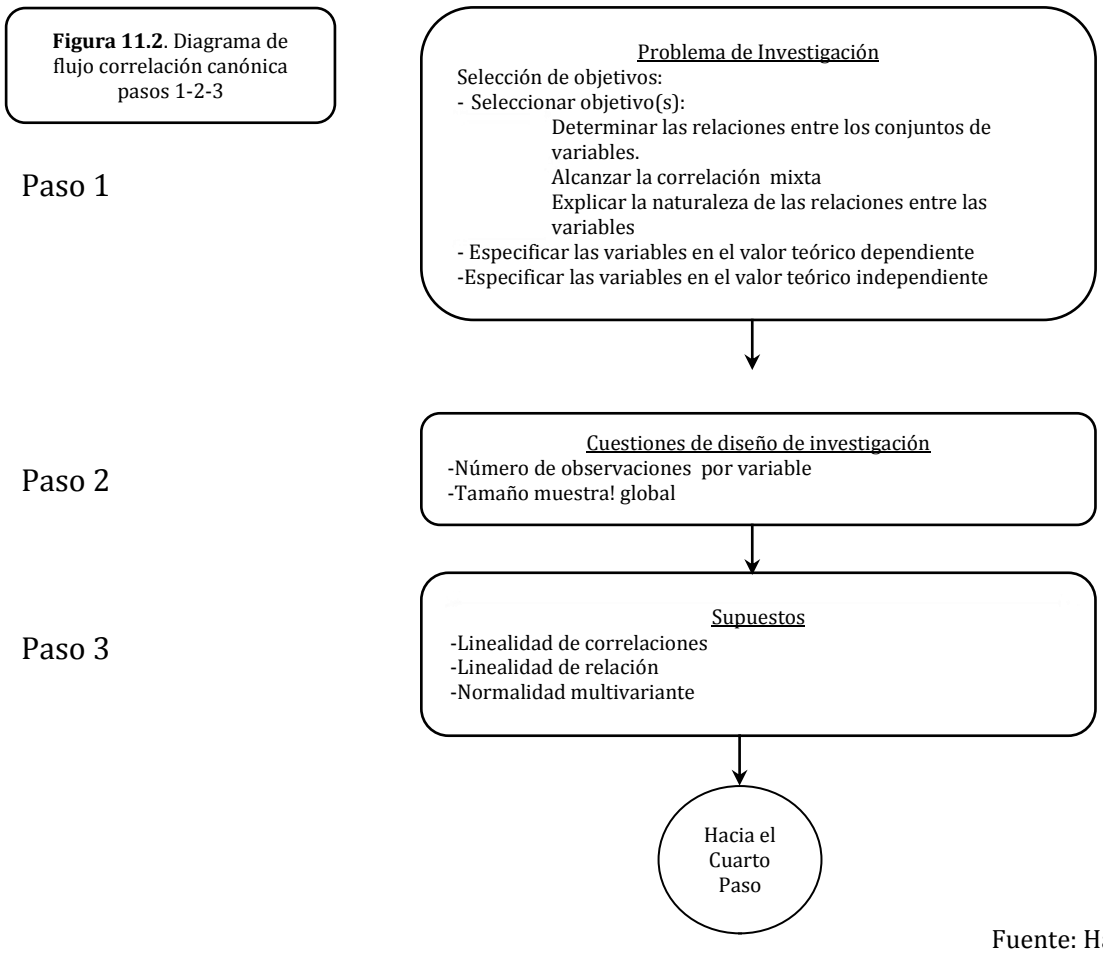
Los datos insumo para aplicar la técnica se basa en dos conjuntos de variableslos cuales tienen el suficiente sustento teórico, de talforma que un conjunto se ha definido como las variables independientes y el otro como las variables dependientes. Realizado lo anterior, la correlación canónica se lleva a a cabo con un amplio rango de objetivos, tales como, (Ver **Figura 11.2**):

l. Determinar si **2 conjuntos de variables** (medidas realizadas sobre los mismos objetivos) **son independientes uno de otro o**, **inversamente**, determinar la magnitud de las relaciones que pueden existir entre los **2 conjuntos.**

2. Obtener un conjunto de ponderaciones para cada conjunto de variables criterio y variables predictoras, para que las combinaciones **lineales** de cada conjunto estén correlacionadas de forma máxima. Las funciones lineales adicionales que maximizan la restante correlación son **independientes** del (los) conjunto(s) anteriores de combinaciones lineales.

3. Explicar la naturaleza de cualquiera de las relaciones existentes entre los conjuntos de variables criterio y variables predictoras. Se realiza midiendo la contribución relativa de cada variable a las **funciones canónicas (relaciones) que son extraídas.**

Figura 11.2. Diagrama de flujo correlación canónica pasos 1-2-3



Fuente: Hair et al. (1999)

11.5. Correlación canónica: Paso 2. Diseño

Esta técnica comparte las herramientas básicas comunes a todas las técnicas multivariantes, por lo que, todo lo relativo particularmente de la **regresión múltiple, el análisis discriminante y el análisis factorial.,** sobre: **el impacto del error de medición, los tipos de variables y sus transfommciones,** son también importantes en el ana lisis de correlacio n cano nica. As , se debe plantear y resolver cuestiones acerca del impacto del **tamaño muestral** (pequeño y/o grande), así como la **cantidad suficiente de observaciones por variable**. Puede existir la tentación de **incluir muchas variables, tanto en el conjunto de variables independientes como el de dependientes,** ignorando sus implicaciones en el tamaño muestral, lo que puede ocasionar:

1. Los tamaños muestrales que son muy pequeños, no representarán las correlaciones adecuadamente.
2. Como consecuencia, esconderán cualquier relación significante que pueda existir.
3. Los tamaños muestrales que son muy grandes, tendrán una tendencia a indicar una significación estadística en todas las instancias, incluso donde la significación práctica no está indicada.
4. También se alienta al investigador a mantener por lo menos **10 observaciones por variable para evitar el "*sobreajuste*" de los datos.**

Sobre las variables, en la correlación canónica deberá considerar, que:

1. La clasificación de las variables como **dependientes o independientes tiene poca importancia en la estimación estadística de las funciones canónicas,** ya que el análisis pondera ambos valores teóricos para **maximizar la correlación y no establece ningún énfasis particular en alguno de los valores teóricos**.
2. Aunque dado que la técnica produce valores teóricos que maximizan la correlación entre ellos, un valor teórico en cualquier conjunto relaciona a todas las otras variables en ambos conjuntos.
3. Con lo anterior, **se permite la incorporación o la supresión de una sola variable que afecte a la solución total, particularmente el otro valor teórico. La composición de cada valor teórico, ya sea dependiente o independiente, llega a ser muy importante,** por lo que, antes de aplicar el análisis de correlación canónica, **debe relacionar conceptualmente los dos conjuntos de variables.** De esta forma, la especificación de los valores teóricos dependientes frente a los independientes es esencial para establecer una base conceptual fuerte para las variables.

11.6. Correlación canónica: Paso 3. Condiciones de aplicabilidad

La generalidad del análisis de correlación canónica también se extiende a sus supuestos estadísticos básicos:

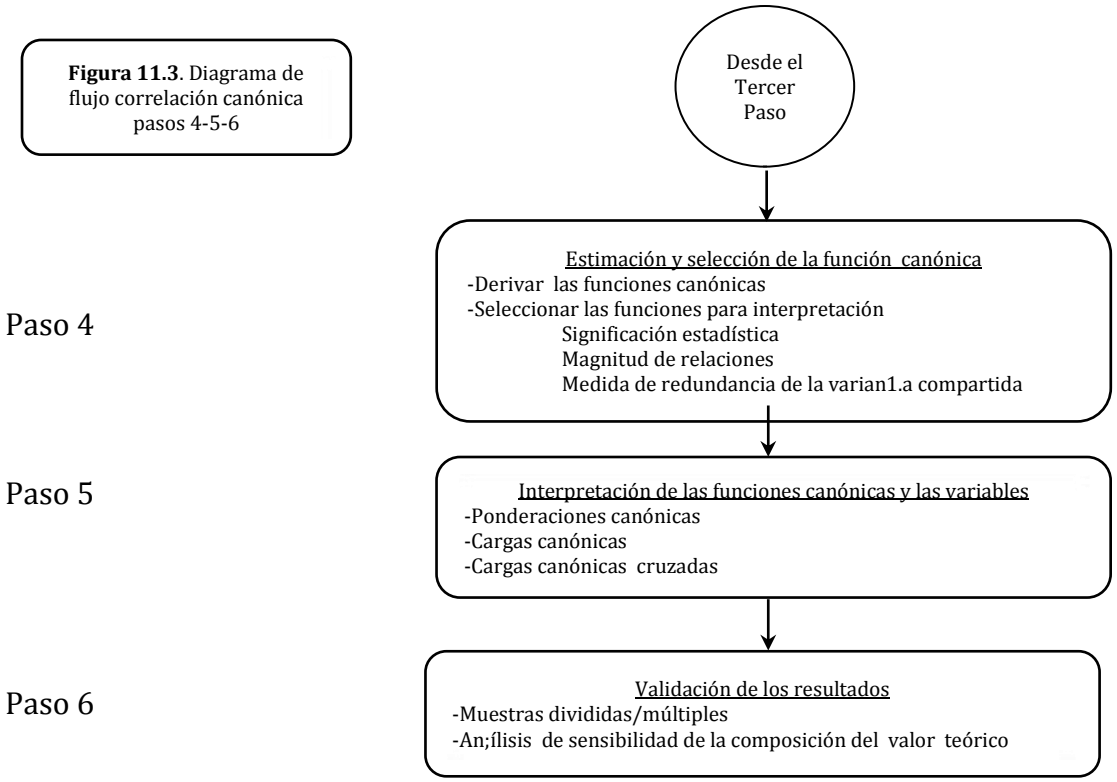
1. **El supuesto de linealidad** afecta a **2 aspectos** de los resultados de la correlación canónica.
    a. **Primero**, el coeficiente de correlación entre cualesquiera dos variables está basado en una relación lineal. **Si la relación no es lineal**, entonces se debe transformar una o ambas variables si fuera posible.
    b. **Segundo,** la correlación canónica es la relación lineal entre los valores teóricos. **Si los valores teóricos se relacionan de una manera no linea**l, la relación no será reflejada por la correlación canónica. De esta manera, aunque el análisis de correlación canónica es el método multivariante más extendido, está restringido a la identificación de relaciones lineales.
2. La técnica **puede emplear cualquier variable métrica sin que cumpla el estricto supuesto de normalidad.** La normalidad es deseable porque estandariza una distribución que nos permite una mayor correlación entre las variables. Pero en un estricto sentido, **el análisis de correlación canónica puede utilizar incluso variables no-normales si la forma de la distribución (por ejemplo., altamente asimétrica) no disminuye la correlación con otras variables. Con esto se permite la utilización de datos no métricos transformados** (en la forma de **variables ficticias**) también. Sin embargo, se requiere **normalidad multivariante para los contrastes de significación de inferencia estadística de cada función canónica**. Dado que los contrastes de **normalidad multivariante no se están disponibles fácilmente**, la línea a seguir que prevalece es asegurar **que cada variable presenta normalidad univariante.** De este modo, aunque estrictamente no se requiere normalidad, es **altamente recomendable que se compruebe la normalidad de todas las variables y que se transformen si fuese necesario**.
3. **La homoscedasticidad**, también debe ser estudiada, puesto que **disminuye la correlación entre las variables**.

4. Finalmente, **la multicolinealidad** entre algún conjunto de variables **distorsionará la capacidad de la técnica para aislar el impacto de cualquier variable única**, haciendo que la **interpretación sea menos fiable.** Los lectores no familiarizados con estos supuestos estadísticos, los contrastes para su diagnosis o las soluciones alternativas cuando los supuestos no se cumplen, deben repasar el **Capítulo 3.**

11.7. Correlación canónica: Paso 4. Estimación y ajuste. Caso Lineal

El primer paso de la técnica s la obtención de una o más funciones canónicas (ver **Figura 11.3**). **Cada función está formada por un par de valores teóricos,** uno que representa las **variables independientes** y el otro que representa las **variables dependientes.** El número máximo de valores teóricos canónicos (**funciones**) que se pueden obtener a partir de los conjuntos de variables es igual al **número de variables que hay en el conjunto de datos menor, ya sea dependiente o independiente**. Por ejemplo,si el problema de investigación incluye **5 variables independientes y 3 variables dependientes, el máximo número de funciones canónicas que se puede obtener es 3.**

Figura 11.3. Diagrama de flujo correlación canónica pasos 4-5-6



Fuente: Hair et al. (1999)

11.7.1. Funciones canónicas y su obtención

Similar al procedimiento empleado en el **análisis factorial sin rotación** (ver **Capítulo 12**), es la obtención de **sucesivos valores teóricos canónicos**, de forma que:

1. El primer factor extraído explica la máxima cantidad de varianza en el conjunto de variables.

2. Después se calcula el segundo factor para que explique lo más posible la varianza no explicada por el primer factor, y así sucesivamente, hasta que todos los factores han sido considerados.
3. Por tanto, los posteriores factores se calculan a partir de los residuos o de la varianza restante de los primeros factores.

   La técnica sigue un procedimiento similar, más que nada **centrándose en la explicación de la cantidad máxima de relación entre los dos conjuntos de variables, en lugar de en un solo conjunto**. El resultado es:

   1. Que el primer par de valores teóricos se calcula con el fin de obtener la mayor intercorrelación posible entre los dos conjuntos de variables.
   2. El segundo par de valores teóricos canónicos es obtenido después para que represente la **máxima relación entre los dos conjuntos de variables (valores teóricos) que no se ha explicado por el primer par de valores teóricos.**
   3. En resumen, los sucesivos pares de valores teóricos canónicos están basados en la **varianza residual**, y sus respectivas **correlaciones canónicas (que reflejan las interrelaciones entre los valores teóricos) disminuyen a medida que se calculan funciones adicionales**; es otras palabras, **el primer par de valores teóricos canónicos refleja la mayor intercorrelación, el siguiente par la segunda correlación mayor, y así sucesivamente.**
   4. Se debe tener en cuenta un aspecto adicional acerca de la obtención de valores teóricos canónicos. Ya se ha mencionado anteriormente que los sucesivos pares de valores teóricos canónicos están basados en la **varianza residual**. Por tanto, **cada uno de los pares de valores teóricos es ortogonal e independiente respecto a todos los otros valores teóricos obtenidos a partir del mismo conjunto de datos.**
   5. La **validez** de la relación entre los pares de valores teóricos se refleja en la correlación canónica. **Cuando se eleva al cuadrado, la correlación canónica representa la cantidad de varianza de un valor teórico explicada por el otro valor teórico.** A esto también se le puede definir como la **cantidad de varianza compartida entre los dos valores teóricos canónicos. Las correlaciones canónicas al cuadrado son denominadas raíces canónicas o autovalores.**

11.7.2. Funciones canónicas y su interpretación

Al igual que cualquier investigación que utiliza otras técnicas estadísticas, la práctica más común es analizar las funciones cuyos coeficientes de correlación canónica son estadísticamente significativos para un nivel, **normalmente >=005**. Si se consideran no significativas otras funcio nes independientes, éstas relaciones entre las variables no se interpretan. La interpretación de los valores teóricos canónicos en una función significativa está basada en la premisa de que las variables en cada conjunto, que contribuyen fuertemente a las varianzas compartidas por estas funciones, son consideradas como relacionadas unas con otras. Varios autores creen que el uso de un único criterio como el nivel de significación es **demasiado superficial.** En lugar de esto, recomiendan que sean empleados tres criterios conjuntamente para decidir qué funciones canónicas se deben interpretar. Los tres criterios son:

1. **Nivel de significación.** Generalmente se considera como el **mínimo aceptable** para la interpretación, el **nivel 0.05**, (y **el nivel 0.0l**) como el nivel más habitualmente aceptado para considerar que un **coeficiente de correlación es estadísticamente significativo,** debido a la disponibilidad de tablas para estos niveles. Sin embargo, estos niveles no son estrictamente requeridos en todas las situaciones, y los investigadores de diferentes disciplinas, frecuentemente deben basar sus resultados en niveles de significación menores. El contraste más habitualmente utilizado, y del que normalmente disponen todos los paquetes informáticos, **es el estadístico *F***, basado en la aproximación de **Rao** [Bartlett 1941]. Además de los **contrastes separados** para cada función canónica, también se puede emplear un **contraste multivariante de todas las raíces canónicas** para evaluar la significación de dichas raíces. Muchas de las medidas existentes para valorar la significación de las funciones discriminantes, incluyendo ***lambda de Wilks, la traza de Hotelling, la traza de Pillai y la mayor raíz de Roy***, también están disponibles. Véase **Capítulo 8** para un análisis de estas medidas.
2. **Magnitud de las relaciones canónicas.** También se tiene que tener en cuenta para decidir qué funciones interpretar, la significación práctica de las funciones canónicas, representada por el **tamaño de las correlaciones canónicas.** Aún no se han establecido líneas básicas que consideren cuáles son los tamaños aceptables para las correlaciones canónicas. En su lugar, la decisión se basa habitualmente en **la contribución de los resultados para una mejor comprensión del problema de investigación que se está estudiando.** Parece lógico que las líneas básicas sugeridas para las cargas factoriales significativas (véase **Capítulo 12**) podrían ser útiles en las correlaciones canónicas, particularmente cuando se considera que las correlaciones canónicas se refieren a la **varianza explicada** con los valores teóricos canónicos (**combinaciones lineales**), y no con las variables originales.
3. **Medida de la redundancia de la varianza compartida. Las correlaciones canónicas al cuadrado (raíces) proporcionan una estimación de la varianza compartida entre los valores teóricos canónicos.** Aunque está es una medida sencilla y atractiva de la varianza compartida, puede llevar a algunas **interpretaciones incorrectas**, dado que las correlaciones canónicas al cuadrado representan la varianza compartida por las **combinaciones lineales de los conjuntos de las variables criterio e independientes,** pero **NO** refleja la **varianza extraída de los conjuntos de variables** [Alpert et al. 1972]. Por ello, se puede obtener una correlación canónica relativamente **fuerte** entre las dos combinaciones lineales (valores teóricos canónicos), **incluso aunque estas combinaciones lineales no puedan extraer porciones significativas de varianza a partir de sus respectivos conjuntos de variables**. Dado que las correlaciones canónicas que se pueden obtener son considerablemente mayores que los coeficientes de correlación múltiple y bivariante anteriormente presentados, **puede existir la confusión de suponer que el análisis canónico ha encubierto relaciones importantes de significación teórica y práctica.** Antes de que estas conclusiones sean corroboradas, se deben llevar a cabo posteriores análisis que incluyan otras medidas distintas a la correlación canónica para determinar la cantidad de varianza de la variable ependiente explicada o compartida por las variables independientes [Lambert&Durand 1975] que ayuden **a superar el sesgo y la incertidumbre**

propios del empleo de raíces canónicas (o **correlaciones canónicas al cuadrado**), considerando:

a. Como una medida de la varianza compartida, se ha propuesto un **índice de redundancia** [Stewart y William 1968], que equivale por cierto, a calcular el **coeficiente de correlación múltiple al cuadrado entre el conjunto predictor total y cada una de las variables en el conjunto criterio, y después promediar estos coeficientes al cuadrado para obtener un R^2 medio.**
b. Lo anterior, proporciona una **medida resumen** de la capacidad del conjunto de las variables predictoras (consideradas como un **conjunto**) para explicar la variación de las variables criterio (consideradas **una a una**).
c. Como tal, la medida de redundancia es perfectamente **análoga al estadístico ^{2} de la regresión múltiple**, y su valor como ndice es similar. **El índice de redundancia de *Stewart-Love* calcula la cantidad de varianza de un conjunto de variables que puede ser explicada por la varianza de otro conjunto**. Este índice sirve como una medida de explicación de la varianza, **similar al cálculo del R^2 empleado en la regresión múltiple.**
d. El R^2 representa la cantidad de varianza de la variable dependiente explicada por la función de regresión de las variables independientes**. En la regresión, la varianza total de la variable dependiente es igual a 1, o al 100 %. Recuerde que la correlación canónica es diferente de la regresión múltiple por que no trabaja con sólo una variable criterio sino con un conjunto criterio que está compuesto de varias variables, y esta combinación tiene solo una porción de la varianza total de cada variable.** Por esta razón, no podemos suponer que el 100 por cien de la varianza en el conjunto criterio esté disponible para que sea explicada por el conjunto predictor.
e. El **conjunto de variables predictoras** se espera que explique solamente la varianza compartida del valor teórico canónico criterio. Por esta razón, el **cálculo del índice de redundancia** es un proceso en 3 **pasos**:
   -El primer paso comprende calcular la cantidad de **varianza compartida del conjunto de variables criterio incluída en el valor teórico canónico dependiente**. Para lograrlo, consideraremos primero como se calcula el estadísticoR^2 de la regresión. El R^2 es simplemente el cuadrado del coeficiente de **correlación R**, que representa la correlación entre la verdadera variable dependiente y el valor predicho. En el **caso canónico,** estamos interesados en la correlación entre el **valor teórico canónico criterio y cada una de las variables criterio**. Tal información puede ser obtenida a partir de las cargas criterio (L_1), que representan la correlación entre cada variable input y su propio valor teórico (analizado en más detalle en la siguiente sección). Elevando al cuadrado cada una de las cargas criterio (L_i^2), se puede obtener una medida de la variación en cada una de las variables criterio explicada por el valor teórico canónico criterio. Para calcular la cantidad de varianza compartida explicada por el valor teórico canónico, se emplea una simple media de las cargas al cuadrado.
   -El segundo paso consiste en calcular la cantidad de varianza en el valor teórico criterio que puede ser explicada por el valor teórico canónico independiente. Este paso del **proceso de redundancia** comprende el **porcentaje de la**

varianza en el valor teórico canónico criterio que puede ser explicado por el valor teórico canónico predictor, o sea, simplemente la correlación al cuadrado entre el valor teórico canónico predictor y el valor teórico canónico criterio, que se conoce de otra forma como la **correlación canónica,** que al cuadrado se la denomina habitualmente el R^2 cano nico.

-El tercer paso es calcular el índice de redundancia, que se calcula multiplicando estos dos componentes. El **índice de redundancia** se calcula multiplicando los **2 componentes** (**varianza compartida del valor teórico por la correlación canónica al cuadrado**) para obtener la cantidad de **varianza compartida**, que puede ser explicada por cada función canónica. Para obtener un **índice de redundancia alto**, se debe tener una **alta correlación canónica y un alto grado de varianza compartida explicada por el valor teórico dependiente**. Una alta correlación canónica por sí sola **NO** asegura una valiosa función canónica. Se calculan **índices de redundancia tanto para el valor teórico dependiente como para el independiente**, aunque en la mayoría de los casos el investigador está interesado solamente en la **varianza extraída del conjunto de variables dependientes**, la cual proporciona una **medida mucho más realista de la capacidad predictiva de las relaciones canónicas**. Usted deberá observar que mientras que la correlación canónica es la misma para los dos valores teóricos en la función canónica, **el índice de redundancia probablemente variará entre los dos valores teóricos**, dado que cada uno tendrá una cantidad de varianza compartida diferente.Con esto, se genera la cuestión: **¿cuál es el índice mínimo de redundancia aceptable necesario para justificar la interpretación de las funciones canónicas?** Al igual que con las correlaciones canónicas, **NO** existen unas líneas básicas a seguir. Deberá juzgar cada función canónica según su significación teórica y práctica para el problema de investigación que se está analizando, para determinar si el índice de redundancia es suficiente para justificar la interpretación. También se ha desarrollado un **contraste para la significación del índice de redundancia** [Alpert et al. 1975], aunque **NO** ha sido ampliamente utilizado.

11.8. Correlación canónica: Paso 5. Interpretación

Si la relación canónica resulta estadísticamente significativa, las magnitudes de la raíz canónica y del índice de redundancia son aceptables, el investigador aún necesita realizar interpretaciones de los resultados, que comprende el análisis de las funciones canónicas para determinar la importancia relativa de cada una de las variables originales en las relaciones canónicas. Se han propuesto **3 métodos:**

1. **Ponderaciones canónicas (coeficientes estandarizados).** Es el **enfoque tradicional para interpretar las funciones canónicas** que comprende el **examen** del **signo y la magnitud** de la ponderación canónica asociada a cada variable en su valor teórico canónico. Las variables con **ponderaciones relativamente mayores contribuyen más al valor teórico** y viceversa. Igualmente, las variables cuyas ponderaciones tienen **signos contrarios** presentan una **relación inversa** unas de otras, y las variables con ponderaciones del **mismo signo** presentan una **relación directa**. Sin embargo, la interpretación de la importancia o contribución relativa de una variable por su

ponderación canónica está sujeta a las mismas críticas asociadas con la interpretación de los **coeficientes beta en las técnicas de regresión,** con los siguientes puntos a considerar:

- **a.** Una **ponderación pequeña** puede significar o bien que su correspondiente variable es irrelevante para explicar la relación o bien que ha sido apartada de la relación debido un alto grado de multicolinealidad.
- **b.** Otro problema del uso de las ponderaciones canónicas es que están sujetas a una inestabilidad considerable (**variabilidad**) de una muestra a otra, la cual se da porque el procedimiento de cálculo del análisis canónico genera ponderaciones que **maximizan las correlaciones canónicas** para una muestra determinada de conjuntos de variables dependientes e independientes observadas [Lambert & Durand 1975] Este problema hace que el uso de las ponderaciones canónicas para interpretar los resultados de un análisis canónico,_ deba **realizarse con precaución**

2. **Cargas canónicas (correlaciones de estructura).** Su aplicación **ha sustituido al uso de ponderaciones canónicas como base de interpretación, debido a las deficiencias inherentes a estas últimas.** Las cargas canónicas, también denominadas **correlaciones de estructura canónica**, miden la **correlación lineal simple** entre una variable original observada del conjunto dependiente o independiente y el valor teórico canónico del conjunto. Las cargas canónicas reflejan la varianza que la variable observada comparte con el valor teórico canónico, y **puede ser interpretada como una carga factorial para valorar la contribución relativa de cada variable a cada función canónica.** Se considera cada función canónica independiente de forma separada, y se calcula la correlación dentro del conjunto entre variable y valor teórico. **Cuanto mayor sea el coeficiente, mayor es la importancia que tiene para calcular el valor teórico canónico.** Los criterios para determinar la significación de las correlaciones de estructura canónica también son los mismos que con las cargas factoriales (ver **Capítulo 12**). Las cargas canónicas, al igual que las ponderaciones, **pueden estar sujetas a una importante variabilidad de una muestra a otra**. Esta variabilidad sugiere que las cargas, y por tanto las relaciones asociadas a ellas, **pueden ser específicas de la muestra**, debido a la **aleatoriedad o a factores ajenos** [Lambert, y Durand 1975]. Aunque las cargas canónicas se consideran relativamente **más válidas que las ponderaciones como medios para interpretar la naturaleza de las ponderaciones canónicas**, Usted **deberá ser precavido** cuando emplea cargas para interpretar las correlaciones canónicas, particularmente al considerar la **validez externa** de los resultados.
3. **Cargas cruzadas canónicas.** Su aplicación se ha sugerido como una alternativa a las cargas convencionales [DillonyGoldstein 1984]. Este procedimiento **consiste en correlacionar cada una de las variables dependientes originales observadas directamente con el valor teórico canónico independiente, y viceversa.** Recuerde que las cargas convencionales correlacionan las variables originales observadas con sus respectivos valores teóricos después de que los dos valores teóricos canónicos se correlacionen de forma máxima uno con el otro. Esto puede resultar **algo parecido a la regresión múltiple pero difiere en que cada variable independiente,** por ejemplo, está correlacionada con el valor teórico dependiente en lugar de con una única variable

dependiente. De esta manera, **proporcionan una medida más directa de las relaciones entre las variables dependientes e independientes eliminando un paso intermedio incluido en las cargas convencionales**. Algunos análisis canónicos **NO calculan las correlaciones** entre las variables y los valores teóricos. En estos casos, se considera que **las ponderaciones canónicas son comparables pero NO equivalentes** para el propósito.

Tomando en cuenta los diferentes métodos para interpretar la naturaleza de las relaciones canónicas., subyace aún la pregunta: **¿qué método debe emplear el investigador?** En principio se sugiere utilizar el que se encuentre disponible en el software de aplicación. Aún así, se tiene como método usualmente más utilizado el de las **cargas cruzadas** y está presente en muchos programas de computador. Se sugiere incluso, **si las cargas cruzadas no están disponibles,** a que el investigador las calcule a mano, o bien a seleccionar otro método de interpretación. El enfoque de las cargas canónicas es de algún modo **más válido que el uso de las ponderaciones.** Por tanto, siempre que sea posible se recomienda el **enfoque de las cargas** como una **segunda alternativa al método de cargas cruzadas canónicas.**

11.9. Correlación canónica: Paso 6. Validación.

Como cualquier otra técnica multivariante, la técnica está sujeta a métodos de validación que aseguren que los resultados **no son solamente específicos de los datos de la muestra y que pueden ser generalizados a la población.** El procedimiento más directo:

1. Es crear **dos submuestras de los datos** (si el tamaño muestral lo permite) y llevar a cabo el análisis en cada submuestra de forma separada.
2. Después, los resultados **se pueden comparar para buscar la igualdad de las funciones canónicas, las cargas de los valores teóricos,** y demás aspectos.
3. Si se encuentran **importantes diferencias**, el investigador debe considerar el realizar una **investigación adicional** para asegurar que los resultados finales son representativos de los valores poblacionales, y **no solamente de una única muestra.**

Otro enfoque consiste en:

1. **Evaluar la sensibilidad de los resultados a la eliminación de una variable dependiente y/o independiente.**Dado que el procedimiento de correlación canónica **maximiza la correlación y no optimiza la interpretabilidad, las cargas y las ponderaciones canónicas pueden variar sustancialmente si una variable es eliminada de algún valor teórico.**
2. Para asegurar la **estabilidad de las cargas y de las ponderaciones canónicas.** Usted deberá estimar múltiples correlaciones canónicas, en donde en cada una se elimina una variable dependiente o independiente diferente.

Aunque existen escasos procedimientos de diagnosis desarrollados específicamente para el análisis de correlación canónica, el investigador debe observar los resultados teniendo en cuenta las limitaciones de la técnica. Dado a que existen con mayor impacto sobre los resultados y su interpretación los siguientes:

1. La **correlación canónica** refleja la varianza compartida por las combinaciones lineales de los conjuntos de variables, no la varianza extraída de las variables.
2. Las **ponderaciones canónicas** obtenidas para calcular las funciones canónicas están

sujetas a **gran inestabilidad.**

3. Las **ponderaciones canónicas** son obtenidas para maximizar la correlación entre las **combinaciones lineales**, no para la varianza extraída.
4. La interpretación de los valores teóricos canónicos puede ser difícil ya que éstos se calculan para maximizar la relación, **y no existen ayudas para la interpretación como pueden ser la rotación de los valores teóricos como se en el análisis factorial**.
5. **Es difícil identificar relaciones con significado entre los subconjuntos de variables dependientes e independientes** dado que aún **NO** se han desarrollado estadísticos precisos para interpretar el análisis canónico, y debemos utilizar medidas inadecuadas como las **cargas y las cargas cruzadas** [Lambert & Durand 1975].

Sin embargo, estas limitaciones **NO** deben desanimar al momento de utilizar la correlación canónica. Al contrario, se las menciona para aumentar la efectividad de sus acciones en la herramienta como recurso para la investigación.

11.10. Correlación canónica: Ejemplo 1.

Para ilustrar la aplicación de la **correlación canónica,** se tomará el ejemplo de la base de datos **WEB_MKT_Digital.sav** (esta base de datos se analiza con profundidad en el **Capítulo 12**). Recuerde que los datos consisten en una serie de medidas obtenidas a partir de una muestra de **100 clientes.** Las variables incluyen las clasificaciones de **WEB_MKT_Digital.sav** sobre **7 atributos** (X_1 **a** X_7) y dos medidas que reflejan sus efectos X_9 **y** X_{10}, **contrataciones y nivel de desempeño**; la discusión de esta aplicación de análisis de correlación canónica sigue un proceso en seis pasos , los que describimos a continuación

Paso 1. Objetivos

Problema 1. Para demostrar la aplicación de la correlación canónica, empleamos las **9 variables** como datos de entrada. Las clasificaciones de **WEB_MKT_Digital.sav** (X_1 **a** X_7) son designadas como el conjunto de múltiples **variables independientes o variables predictoras.** Las medidas de **contrataciones y nivel de desempeño** (variables X_9 **a** X_{10}) se especifican como el conjunto de múltiples **variables dependientes o variables criterio**.

Problema: a nivel estadístico**, la empresa MKT Digital requiere identificar cualquier relación latente entre las percepciones de sus clientes a través de analizar la base de datos WEB MKT Digital.sav relacionadas con el nivel de contratación y desempeño de sus servicios.**

Pasos2: Diseño y Paso 3: Condiciones de aplicabilidad

El diseño de las variables incluye **2 variables dependientes métricas y 7 variables independientes métricas.** La base conceptual de ambos conjuntos está bien establecida, por lo que no hay necesidad de realizar formulaciones de modelos alternativos para contrastar los diferentes conjuntos de variables. Las **7 variables** generan un ratio de observaciones frente a variables de **13 a 1**, por lo que excede correctamente el supuesto de **10** observaciones por variable. No se considera que el tamaño muestra de **100** afecte las estimaciones del error de muestreo notablemente y por tanto no debería tener impacto sobre la significación estadística de los resultados. Las variables dependientes e independientes se evalúan en el **Capítulo 12** para detectar los supuestos básicos sobre la

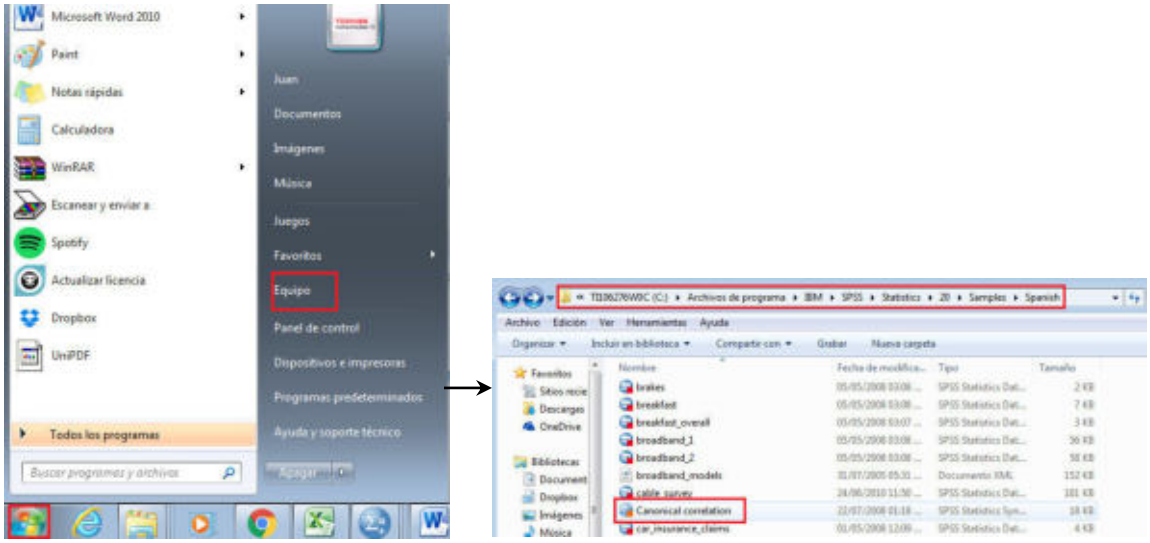
distribución que se deben dar en el análisis multivariante y pasen todos los tests estadísticos.

Paso 4: Estimación y ajuste

Para iniciar el análisis con el software SPSS, deberá realizar:

1. La activación del archivo Correlación canónica ubicado normalmente en el archivo de programas, de la siguiente forma:
   -Teclear: Bandera Windows->Equipo-> C->Program Files->IBM->SPS>Statistics ->20 ->Samples ->Spanish ->Canonical correlation.sps. Ver **Figura 11.4**

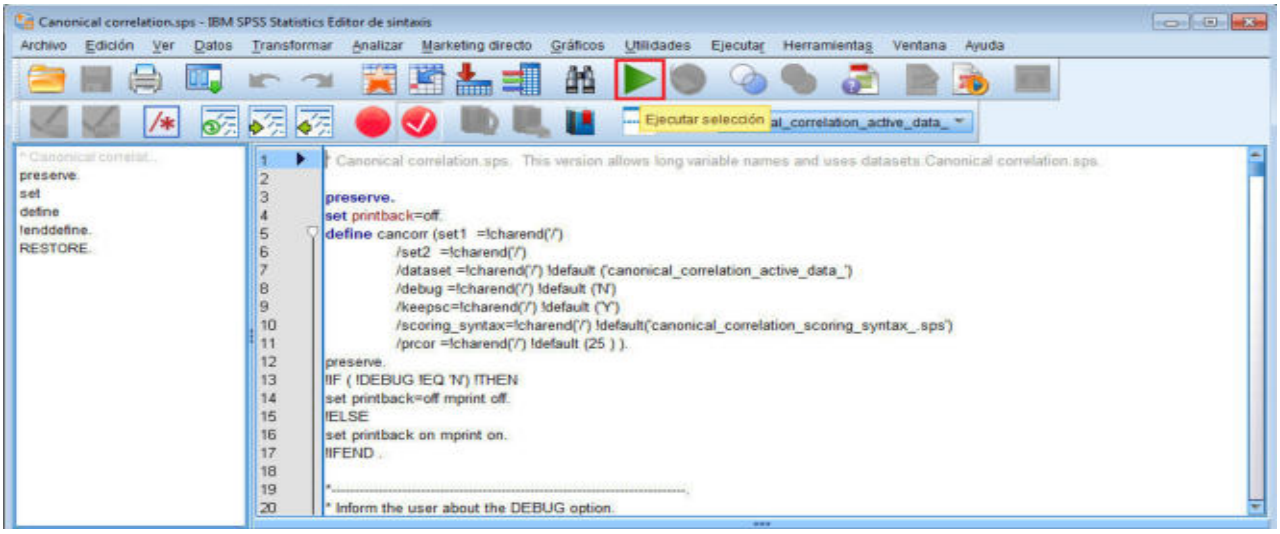
Figura 11.4.- Proceso para ubicar el archivo Canonical correlation.sps



Fuente: SPSS 20 IBM

2. Localizado el archivo, se requerirá activarlo oprimiendo 2 veces el ícono.
   Teclear: Canonical correlation.sps. (oprimir dos veces el ícono)-> Aparición del menú emergente Sintaxis de SPSS->Oprimir Ejecutar selección (Ver Figura 11.5)

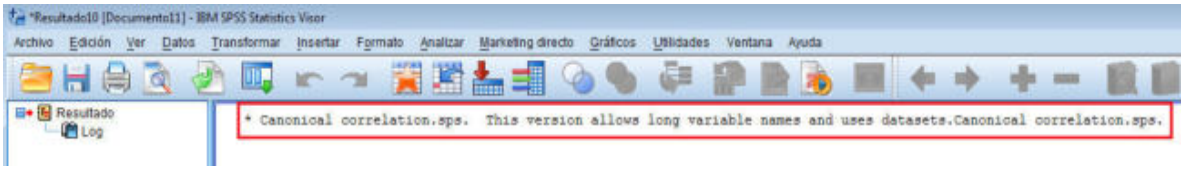
Figura 11. 5 Menú emergente Sintaxis SPSS



Fuente: SPSS 20 IBM

1. Aparición de aviso de activación del archivo **Canonical correlation.sps** a través del visor de resultados del SPSS. Ver **Figura 11.6**

   Figura 11.6. Aviso de activación del archivo Canonical correlation.sps a través del visor de resultados del SPSS.



   Fuente: SPSS 20 IBM

2. Con el fin de introducir los dos conjuntos de variables a analizar (dependientes e independientes), se deberá preparar el arreglo en forma de comandos de programación a través del menú Sintaxis de **SPSS**.

   -Teclear: Archivo->Nuevo->Sintaxis->

   INCLUDE'C:\Program Files\ IBM\ SPSS\ Statistics\ 20\ Samples\ Spanish\ Canonical correlation.sps'.

   Nota: INCLUDE linea de comando que ubica el archivo Canonical correlation.sps

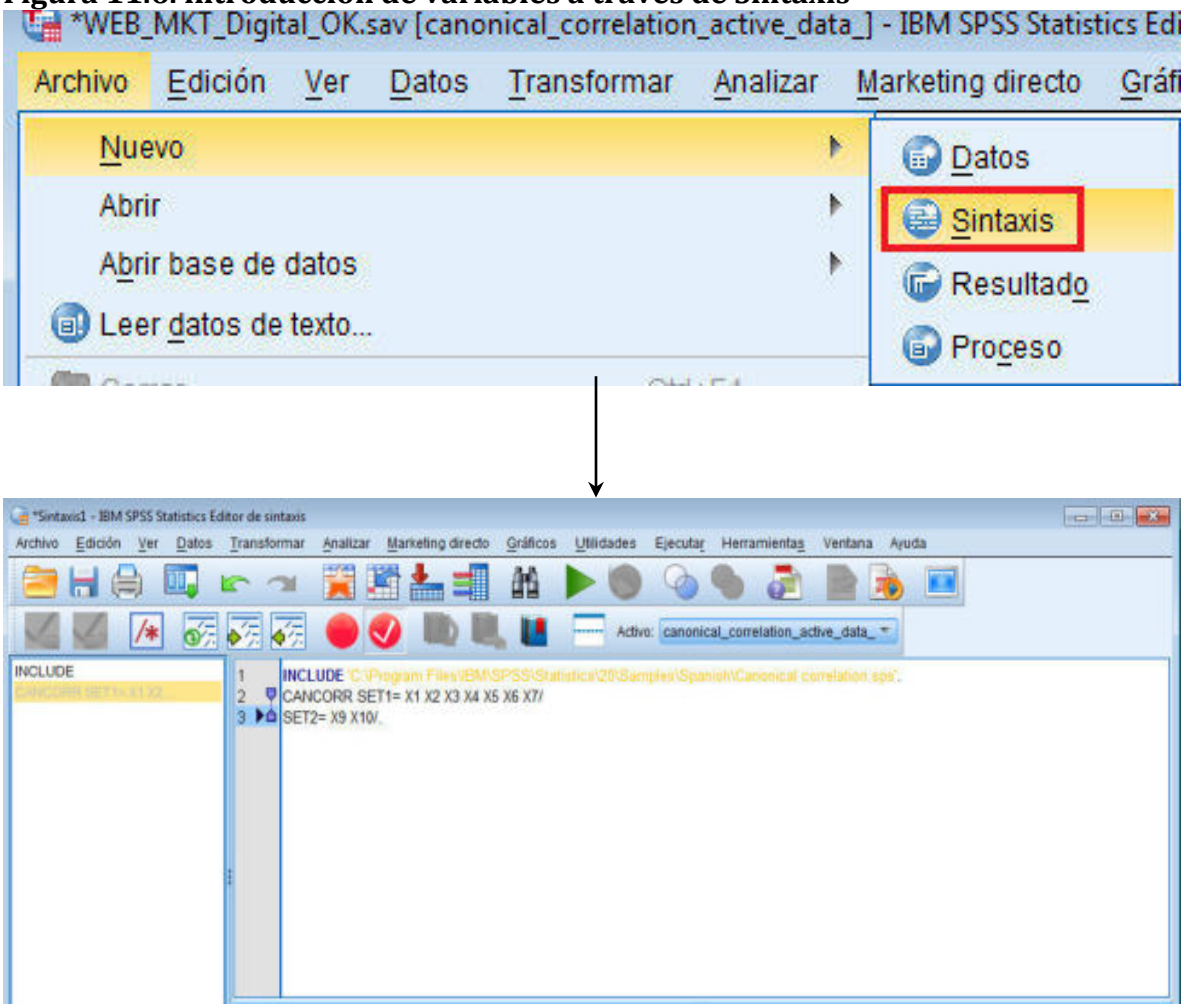
   CANCORR SET1= X_1 X_2 X_3 $X_4$$X_5$ X_6 X_7/

   Nota: CANCORR línea de comando que prepara el ingreso del **primer conjunto de variables,** se declaran tal y como se declaran en la base de datos **WEB MKT Digital.sav** separados por espacio; en este caso las independientes. Cierra con **/**

   SET2= X_9 X_{10} /

   Nota: declaración que prepara el ingreso del **segundo conjunto de variables**, se declaran tal y como se declaran en la base de datos **WEB MKT Digital.sav** separados por espacio; en este caso las independientes. Cierra con **/.**

Figura 11.6. introducción de variables a través de Sintaxis



Fuente: SPSS 20 IBM

El resultado se genera en forma de listado como se observa en la **Figura 11.7**

Figura 11.7. Listado de resultados de la correlación canónica de SPSS

```
Run MATRIX procedure:

Correlations for Set-1
        X1      X2      X3      X4      X5      X6      X7
X1  1.0000  -.3492   .5093   .0504   .6119   .0774  -.4826
X2  -.3492  1.0000  -.4872   .2722   .5130   .1853   .4697
X3   .5093  -.4872  1.0000  -.1161   .0666  -.0348  -.4481
X4   .0504   .2722  -.1161  1.0000   .2987   .7881   .2000
X5   .6119   .5130   .0666   .2987  1.0000   .2404  -.0552
X6   .0774   .1853  -.0348   .7881   .2404  1.0000   .1766
X7  -.4826   .4697  -.4481   .2000  -.0552   .1766  1.0000
```

Correlations for Set-2

	X9	X10
X9	1.0000	.7107
X10	.7107	1.0000

Correlations Between Set-1 and Set-2

	X9	X10
X1	.6765	.6506
X2	.0819	.0284
X3	.5590	.5248
X4	.2242	.4759
X5	.7007	.6312
X6	.2551	.3406
X7	-.1925	-.2833

(A) **Canonical Correlations**

1	**.937**
2	**.510**

(B) Test that remaining correlations are zero:

	Wilk's	Chi-SQ	DF	Sig.
1	.090	225.875	14.000	.000
2	.740	28.258	6.000	.000

(E) Standardized Canonical Coefficients for Set-1

	1	2
X1	-.225	.962
X2	-.103	.865
X3	-.569	-.161
X4	-.349	1.454
X5	-.445	-1.528
X6	.051	-.733
X7	-.001	-.479

Raw Canonical Coefficients for Set-1

	1	2
X1	-.170	.729
X2	-.086	.723
X3	-.410	-.116
X4	-.308	1.285
X5	-.593	-2.034
X6	.067	-.951
X7	.000	-.302

(F) Standardized Canonical Coefficients for Set-2

	1	2
X9	-.501	-1.330
X10	-.580	1.298

Raw Canonical Coefficients for Set-2

	1	2
X9	-.056	-.148
X10	-.678	1.517

(D)Canonical Loadings for Set-1

	1	2
X1	-.764	-.109
X2	-.061	-.142
X3	-.624	-.123
X4	-.415	.627
X5	-.765	-.222
X6	-.347	.201
X7	.278	-.219

(G) Cross Loadings for Set-1

	1	2
X1	-.716	-.056
X2	-.057	-.072
X3	-.584	-.063
X4	-.388	.319
X5	-.717	-.113
X6	-.325	.103
X7	.261	-.112

(C) Canonical Loadings for Set-2

	1	2
X9	-.913	-.408
X10	-.936	.352

(H) Cross Loadings for Set-2

	1	2
X9	-.855	-.208
X10	-.877	.179

Redundancy Analysis:

Proportion of Variance of Set-1 Explained by Its Own Can. Var.

	Prop Var
CV1-1	.276
CV1-2	.083

Proportion of Variance of Set-1 Explained by Opposite Can.Var.

	Prop Var
CV2-1	.242
CV2-2	.021

Proportion of Variance of Set-2 Explained by Its Own Can. Var.

	Prop Var

```
CV2-1               .855
CV2-2               .145


Proportion of Variance of Set-2 Explained by Opposite Can. Var.
               Prop Var
CV1-1               .750
CV1-2               .038

Fuente SPSS 20 IBM
```

Como el conjunto de variables dependientes **contiene solamente dos variables** la técnica se restringida a obtener **2 funciones canónicas** y para incluir en el paso de interpretación, el análisis se centra: en el **nivel de significación estadística**, la **significación práctica** de la correlación canónica, y los **índices de redundancia** para cada valor teórico. Con los datos de la **Figura 11.6**, es posible determinar (Nota: los valores negativos se generan por la forma de haber ingresado los datos, por lo que se sugiere invertirlos para sus consideraciones.):

1. **Significación estadística y práctica.** El primer contraste de significación estadística a través de **las correlaciones canónicas de cada una de las dos funciones canónicas.** En este ejemplo, ambas correlaciones canónicas son estadísticamente significativas (ver **Figura 11.7**).

Figura 11.7. Tabla de significatividad de las correlaciones canónicas (A)

Función Canónica	Correlación canónica (*R*)	^{2} canónica	Estadístico *F*	Probabilidad
1	**0.937**	**0.878**	30.235	0,000
2	**0.510**	**0.260**	5.391	0,000

Fuente SPSS 20 IBM con adaptación propia

También se llevan a cabo contrastes multivariantes para ambas funciones simultáneamente, junto a los tests para cada función canónica de forma separada. Los contrastes estadísticos empleados son mostrados en la **Figura 11.8** que presenta estos tests multivariantes, indicando que las funciones canónicas, consideradas conjuntamente, son estadísticamente significativas a un nivel de **0.0000 Figura 11.8.**

Figura 11.8. Tabla contraste de multivariantes y sus significatividad (B)

Función canónica	Wilk´s	Chi-Cuadrada	gl	Sig.
1	**0.090**	**225.875**	**14.000**	**0.000**
2	**0.740**	**28.258**	**6.000**	**0.000**

Fuente SPSS 20 IBM con adaptación propia

Además de la significación estadística, ambas correlaciones canónicas tienen el tamaño suficiente (**100 registros**) como para ser consideradas significativas de forma práctica. El último paso es la realización del **análisis de redundancia** en ambas funciones canónicas.

2. **Análisis de redundancia.**Se calcula un índice de redundancia para los valores teóricos dependiente e independiente de la primera función como se refleja en la **Figura 11.9 y Figura 11.10**

Figura 11.9. Cálculo de los índices de redundancia para la primera función canónica

Valor teórico/ variables	Carga canónica (CC) (C)		Promedio $\sum CC$	R^2 Canónica (de Figura 11.7)	Indice de redundancia Promedio $\sum CC^2 * R^2$ Canónica
Variables Dependientes					
X_9.- Web contrataciones de clientes	-0.913	0.834			
X_{10} .-Web satisfacción	-0.910	0.876			
Valor teórico dependiente		1.710	0.855 (I)	0.878	0.751 (J)
Variables independientes					
Valor teórico/ variables	Carga canónica (CC) (D)		Promedio $\sum CC$	R^2 Canónica (de Figura 11.7)	Indice de redundancia Promedio $\sum CC^2 * R^2$ Canónica
X_1.-Web tecnología	-0.764	0.584			
X_2.- Web precio del servicio	-0-.061	0.004			
X_3.- Web planeación estratégica	-0.624	0.389			
X_4.- Web imagen	-0.414	0.171			
X_5.- Web experiencia usuario	-0.765	0.585			
X_6.- Web calidad	-0.348	0.121			
X_7.- Web desempeño	-0.278	0.077			
Valor teórico independiente		1.931	0.276	0.878	0.242

Fuente SPSS 20 IBM con adaptación propia

Así, el **índice de redundancia del valor teórico** criterio es importante (**0.751**). Sin embargo, **el valor teórico predictor** tiene un índice de redundancia sustancialmente menor (**0.242**), aunque en este caso, dado que existe una clara explicación entre las variables dependientes e independientes, este menor valor **NO** es inesperado o problemático. La **baja redundancia del valor teórico predictor** se debe a la relativamente baja varianza compartida en el **valor teórico predictor** (**0.276**), y no al

***R^2*canónico.** De acuerdo al **análisis de redundancia** y a los **contrastes de significación estadística**, la **primera función debe ser aceptada**.
El análisis de redundancia para la **segunda función** genera unos resultados bastante diferentes (vea **Figura 11.10**).

Figura 11.10. Cálculo de los índices de redundancia para la segunda función canónica

Valor teórico/ variables	Carga canónica (CC) (C)	CC^2	Promedio $\sum CC^2$	R^2 Canónica (de Figura 11.7)	Indice de redundancia Promedio $\sum CC^2$* R^2 Canónica
Variables Dependientes					
X_9.- Web contrataciones de clientes	-0.408	0.166			
X_{10} .-Web satisfacción	0.352	0.124			
Valor teórico dependiente		0.29	0.145	0.260	0.038
Variables independientes					
Valor teórico/ variables	Carga canónica (CC) (D)	CC^2	Promedio $\sum CC^2$	R^2 Canónica (de Figura 11.7)	Indice de redundancia Promedio $\sum CC^2$* R^2 Canónica
X_1.-Web tecnología	-0.109	0.012			
X_2.- Web precio del servicio	-0.142	0.02			
X_3.- Web planeación estratégica	-0.123	0.015			
X_4.- Web imagen	0.627	0.393			
X_5.- Web experiencia usuario	-0.222	0.049			
X_6.- Web calidad	0.201	0.040			
X_7.- Web desempeño	-0.219	0.048			
Valor teórico independiente		0.577	0.0824	0.260	0.021

Fuente SPSS 20 IBM con adaptación propia

Por lo que, para realizar una mejor visualización de los resultados, tenemos la **Figura 11.11**

Figura 11.11. Análisis de la redundancia de los valores teóricos dependientes e independientes para ambas funciones canónicas

Varianza estandarizada de las variables independientes explicada por					
	Su propio valor teórico canónico (varianza compartida)			El valor teórico canónico opuesto (redundancia)	
Función canónica	Porcentaje	Porcentaje acumulado	R2 Canónica	Porcentaje	Porcentaje acumulado
1	0.855	0.855	0.878	0.751	0.751
2	0.145	1.000	0.260	0.038	0.789
Varianza estandarizada de las variables independientes explicada por					
	Su propio valor teórico canónico (varianza compartida)			El valor teórico canónico opuesto (redundancia)	
Función canónica	Porcentaje	Porcentaje acumulado	R2 Canónica	Porcentaje	Porcentaje acumulado
1	0.276	0.276	0.878	0.242	0.242
2	0.082	0.358	0.260	0.021	0.263

Fuente SPSS 20 IBM con adaptación propia

Primero, el R^2**canónico** es sustancialmente menor (0.260). Además, ambos con- juntos de variables presentan **baja varianza compartida** en la segunda función (**0.145** para el **valor teórico dependiente** y **0.0824** para el **valor teórico independiente**). Su combinación con la raíz canónica para obtener el índice de redundancia produce un valor de **0.038** para el **valor teórico dependiente** y **0.021** para el **valor teórico independiente**. De esta manera, mientras que la **segunda función es estadísticamente significativa**, **tiene escasa significatividad práctica.** Con ese pequeño porcentaje, se debe **cuestionar la aceptación de la función. Este es un excelente ejemplo de una función canónica estadísticamente significativa que no explica significativamente una gran parte de la varianza criterio.** Se sugiere revisar el **Capítulo 12** con atención al debate de desarrollo de escala ya que, de alguna manera la **correlación canónica es una forma de desarrollo de escala,** dado que los valores teóricos **dependientes e independientes** representan dimensiones de los conjuntos de la variable similares a las escalas desarrolladas con el **análisis factorial**. La **diferencia principal** es que se desarrolla estas dimensiones para **maximizar la relación entre ellos**, mientras que el **análisis factorial maximiza la explicación** (varianza compartida) del conjunto de variables.

Paso 5: Interpretación.

Hasta aquí, se han considerado estadísticamente **significativa la relación canónica** y aceptables la **magnitud de la raíz canónica** y el **índice de redundancia**, deberá a proceder a realizar **interpretaciones** de los resultados. Aunque la **segunda función podría ser considerada prácticamente no significativa,** debido al **bajo índice de redundancia, se incluye en la fase de interpretación por razones de explicación.** Estas interpretaciones comprenden el análisis de las funciones canónicas para determinar la importancia relativa de cada una de las variables originales para obtener las relaciones canónicas. Los **3 métodos de interpretación** son:

1. **Las ponderaciones canónicas (coeficientes estandarizados**). Ver **Figura 11.12**.

Figura 11.12. Ponderaciones canónicas para las dos funciones canónicas

Coeficientes estandarizadas para las variables independientes	**Función 1 (E)**	**Contribución 1ª. Función**	**Función 2 (E)**	**Contribución 2a. Función**
$_1$.-Web tecnología	-0.225	**4**	0.962	**3**
$_2$.- Web precio del servicio	-0.103	**5**	0.865	**4**
$_3$.- Web planeación estratégica	-0.569	**1**	-0.161	**7**
$_4$.- Web imagen	-0.349	**3**	1.454	**2**
$_5$.- Web experiencia usuario	-0.445	**2**	-1.528	**1**
$_6$.- Web calidad	0.051	**6**	-0.733	**5**
$_7$.- Web desempeño	-0.001	**7**	-0.479	**6**
Coeficientes estandarizadas para las variables dependientes	**Función 1 (E)**	**Contribución 1er grupo**	**Función 2 (E)**	**Contribución 2a. Función**
$_9$.- Web contrataciones de clientes	-0.501	**2**	-1.330	**1**
$_{10}$.-Web satisfacción	-0.580	**1**	1.298	**2**

Fuente SPSS 20 IBM con adaptación propia

Donde la tabla contiene las ponderaciones canónicas estandarizadas para cada valor teórico canónico, tanto para las **variables dependientes como independientes**. Como se expuso líneas arriba, **la magnitud de las ponderaciones representa su contribución relativa al valor teórico**. Basado en el tamaño de las ponderaciones, el orden de contribución de las variables independientes al primer valor teórico es $\mathbf{X_3, X_5, X_4}$, $\mathbf{X_1}$, $\mathbf{X_2}$, $\mathbf{X_6}$ y $\mathbf{X_7}$, mientras que el orden de las variables dependientes sobre el **primer valor** teórico es $\mathbf{X_{10}}$ y después $\mathbf{X_9}$. Clasificaciones similares se pueden encontrar para los valores teóricos de la **segunda función canónica.** Dado que las ponderaciones canónicas son **generalmente inestables**, particularmente en casos de **multicolinealidad**, debido a que se calculan exclusivamente para optimizar la correlación canónica, **se consideran más apropiadas las cargas canónicas y las cargas cruzadas canónicas.**

2. **Las cargas canónicas.** La **Figura 11.13** representa las **cargas canónicas** para los **valores teóricos dependiente e independiente para ambas funciones canónicas**. El objetivo de **maximizar** los valores teóricos para la correlación entre ellos tiene como resultado unos **valores teóricos "*optimizados*" no para la interpretación, sino para la predicción.** Con esto, la identificacio n de las relaciones **es más difícil**. En el primer valor teórico dependiente, ambas variables tienen cargas que exceden **0.90**, reflejándose en una alta varianza compartida (**0.855**). Con lo que se indica un alto grado de **intercorrelación** entre las dos variables y **sugiere que ambas medidas o bien una u otra son representativas de los efectos de los esfuerzos de la empresa MKT Digital**.

Figura 11.13. Estructura canónica de las dos funciones canónicas

	Ponderaciones canónicas	
Correlaciones entre las variables independientes y sus valores teóricos canónicos dependientes	**Función 1 (D)**	**Función 2 (D)**
X_1.-Web tecnología	-0.764	-0.109
X_2.- Web precio del servicio	-0.061	-0.142
X_3.- Web planeación estratégica	-0.624	-0.123
X_4.- Web imagen	-0.415	0.627
X_5.- Web experiencia usuario	-0.765	-0.222
X_6.- Web calidad	-0.347	0.201
X_7.- Web desempeño	0.278	-0.219
Correlaciones entre las variables dependientes y sus valores teóricos canónicos	**(C)**	**(C)**
X_9.- Web contrataciones de clientes	-0.913	-0.408
X_{10}.-Web satisfacción	-0.936	0.352
	Ponderaciones canónicas	
Correlaciones entre las variables independientes y sus valores teóricos canónicos dependientes	**Función 1 (G)**	**Función 1 (G)**
X_1.-Web tecnología	-0.716	-0.056
X_2.- Web precio del servicio	-0.057	-0.072
X_3.- Web planeación estratégica	-0.584	-0.063
X_4.- Web imagen	-0.388	0.319
X_5.- Web experiencia usuario	-0.717	-0.113
X_6.- Web calidad	-0.325	0.103
X_7.- Web desempeño	0.261	.0.112
Correlaciones entre las variables dependientes y sus valores teóricos canónicos	**(H)**	**(H)**
X_9.- Web contrataciones de clientes	-0.855	-0.208
X_{10}.-Web satisfacción	-0.877	0.179

Fuente SPSS 20 IBM con adaptación propia

El primer valor teórico independiente sigue un modelo bastante diferente, con cargas que varían desde **0.061** a **0.765**, con una variable independiente ($\mathbf{X_7}$) que incluso tiene una carga negativa, aunque es bastante pequeña y no de interés sustantivo. Las tres variables con las cargas más altas en cuanto al valor teórico independiente son $\mathbf{X_5}$ (web experiencia de usuario), $\mathbf{X_1}$ (web tecnología),y $\mathbf{X_3}$ (web planeación estratégica). Este

valor teórico **NO** corresponde a las dimensiones extraídas en el análisis factorial (ver **Capítulo 12**). **Aún así, no deberíamos esperar una gran coincidencia entre los resultados de ambas técnicas, porque los valores teóricos en la correlación canónica se extraen solamente para maximizar los objetivos de predicción**. Como tal, debería corresponder más a los resultados de otras técnicas de dependencia. Hay una correspondencia estrecha con la **regresión múltiple** (véase **Capítulo 5**). Dos de estas variables ($\mathbf{X_3}$ y $\mathbf{X_5}$) fueron incluidas en el análisis la regresión por etapas, en el que $\mathbf{X_9}$ (**una de las dos variables en el valor teórico dependiente**) era la **variable dependiente**. Por lo tanto, **la primera función canónica corresponde más estrechamente a los resultados de la regresión múltiple, con el valor teórico independiente representando el conjunto de variables que realiza la mejor predicción de las dos medidas dependientes.** Usted deberá también realizar un **análisis de sensibilidad** del valor teórico independiente en este caso para ver si las cargas cambian cuando es eliminada una variable independiente (**ver paso 6**). Los **bajos valores de redundancia del segundo valor teórico están reflejados por las cargas sustancialmente menores para ambos valores teóricos en la segunda función.** Por ello, la interpretabilidad más pobre reflejada en las menores cargas, unido a los bajos valores de redundancia, **refuerzan la baja significación práctica de la segunda función.**

3. **Cargas cruzadas.** La **Figura 11.13** también incluye las cargas cruzadas para las **2 funciones canónicas**. Al estudiar la primera función canónica, podemos observar que ambas variables dependientes ($\mathbf{X_9}$ y $\mathbf{X_{10}}$) presentan altas correlaciones con el valor teórico independiente (**función 1**): **0.855 y 0.877** respectivamente. Esto refleja la alta varianza compartida entre estas dos variables. **Elevando al cuadrado estos términos, hallamos el porcentaje de la varianza para cada una de las variables explicadas por la función l**. Los resultados muestran que el **73% de la varianza de $\mathbf{X_9}$ y el 77 % de la varianza de $\mathbf{X_{10}}$ está explicado por la función l.** Observando **las cargas cruzadas** de las variables independientes, vemos que las variables $\mathbf{X_1}$ y $\mathbf{X_5}$ tienen altas correlaciones de aproximadamente **0.72 con el valor teórico canónico criterio**. De esta información, observamos que el **52%** de la varianza de cada una de estas dos variables esta explicada por el **valor teórico criterio** (el **52%** se obtiene elevando al cuadrado el coeficiente de correlacio n; **0.72**). La correlación de X (0.584) puede parecer alta, pero despue s de elevar al cuadrado su correlación, solamente el 34 % de la variacio n esta incluida en el valor teo rico canónico. a u ltima cuestio n de la interpretacio n es examinar los signos de las cargas cruzadas. Todas las variables independientes excepto $\mathbf{X_3}$ (web planeación estratégica) tienen una relacio n directa positiva. Para la **segunda función,** dos variables predictoras ($\mathbf{X_4}$ **y** $\mathbf{X_6}$), más una variable criterio ($\mathbf{X_{10}}$) son negativas (se recomienda invertir dado la forma de ingreso de las variables al estudio). Por esto todas las relaciones son directas excepto para una relación inversa en la primera función. Por lo tanto todas las relaciones son directas excepto por una sola relación inversa para la primera función.

Paso 6: Validación

El último paso debe incluir una **validación de los análisis de correlación canónica por medio de uno de diferentes procedimientos.** Entre los enfoques disponibles estarían:

1. Dividir la muestra en muestras de estimación y de validación,
2. Análisis de sensibilidad del conjunto de variables independientes. La **Figura 11.14** contiene el resultado de este **análisis de sensibilidad donde se examina si las cargas canónicas son estables cuando se eliminan variables independientes individuales del análisis.** Como se observa, las cargas canónicas en nuestro ejemplo son **fuertemente estables y consistentes** en cada uno de los tres casos donde se **elimina** una variable independiente (**X_1, $_2$ o $_7$, se logra al retirar del SET1 las variables, del menú emergente Sintaxis**).

Las correlaciones canónicas totales también **permanecen estables.** Pero si el lector examina las ponderaciones canónicas (**NO** reflejadas en la tabla), existirán resultados muy diferentes dependiendo de qué variable sea eliminada. Esto refuerza el procedimiento de emplear la carga canónica y la carga cruzada con objetivos de interpretación.

Figura 11.14. Análisis de sensibilidad de los resultados de la correlación canónica al eliminar una variable pendiente

		Resultados después de la eliminación de:		
	Valor teórico completo	**1**	**2**	**7**
Correlación canónica R	**0.937 (A)**	0.936	0.937	0.937
Raíz canónica ^{2}	**0.878**	0.876	0.878	0.878
Valor teórico independiente cargas canónicas				
$_1$.-Web tecnología	-0.764 **(D)**	omitida	0.765	0.764
$_2$.- Web precio del servicio	-0.061 **(D)**	0.062	omitida	0.061
$_3$.- Web planeación estratégica	-0.624**(D)**	0.624	0.624	0.624
$_4$.- Web imagen	-0.415 **(D)**	0.413	0.414	0.415
$_5$.- Web experiencia usuario	-0.765 **(D)**	0.766	0.766	0.765
$_6$.- Web calidad	-0.347 **(D)**	0.348	0.348	0.348
$_7$.- Web desempeño	0.278 **(D)**	-0.278	-0.278	omitida
Varianza compartida	-0.276 **(D)**	0.225	0.322	0.309
Redundancia	-0.242 **(D)**	0.197	0.282	0.271
Valor teórico dependiente cargas canónicas				
$_9$.- Web contrataciones de clientes	0.913 **(C)**	0.915	0.914	0.913
$_{10}$.-Web satisfacción	0.936 **(C)**	0.934	0.935	0.936

Varianza compartida	0.855 **(I)**	0.855	0.855	0.855
Redundancia	0.751 **(J)**	0.749	0.750	0.750

Fuente SPSS 20 IBM con adaptación propia

El análisis de correlación canónica aborda **2 objetivos principales:**

1. La identificación de las dimensiones entre las variables dependientes e independientes que
2. Maximizan la relación entre las dimensiones. Desde una perspectiva gerencial, se proporciona al investigador una visión más detallada de la estructura de los diferentes conjuntos de variables relacionadas con una relación de dependencia. Primero, los resultados indican que solamente existe una relación simple, respalda- do por la baja significación práctica de la segunda función canónica.
   Con el estudio de esta relación, observamos:
   1. Que **las dos variables dependientes están estrechamente relacionadas y crean una dimensión claramente definida para representar los resultados de los esfuerzos de la empresa MKT Digital**.
   2. Cuando una serie de variables independientes actúan como un conjunto, se puede predecir esta dimensión de resultado adecuadamente. El valor de redundancia de **0.750** sería bastante aceptable para el R^2**comparable**. Cuando se interpreta el valor teórico independiente, observamos que **3 variables**, $\mathbf{X_5}$(experiencia del usuario), $\mathbf{X_1}$ (web tecnología) y $\mathbf{X_3}$ (flexibilidad de precio) proporcionan las contribuciones sustantivas y, por lo tanto, son los predictores clave de la dimensio n de resultado. Éstos deberían ser los puntos de concentración en el desarrollo de cualquier estrategia que afecte a los resultados de **MKT Digital.**

Conclusión: El análisis de correlación canónica es una técnica útil y potente para:

1. Explorar las relaciones entre variables dependientes e independientes múltiples. La técnica es ante todo **descriptiva**, aunque puede ser empleada **con fines predictivos**.
2. Los resultados obtenidos a partir de un análisis canónico deben dar **respuestas a cuestiones relacionadas con el número de maneras en las que se relacionan dos conjuntos de múltiples variables, la validez de las relaciones y la naturaleza de las relaciones definidas.**
3. El análisis canónico posibilita al investigador **combinar en una medida compuesta, lo que de otra forma podría ser un gran número difícil de manejar de correlaciones bivariantes entre conjuntos de variables.**
4. Es útil para identificar **relaciones globales entre múltiples variables dependientes e independientes,** especialmente cuando el analista tiene poco conocimiento a priori sobre las relaciones entre los conjuntos de variables.
5. Fundamentalmente, **el investigador puede aplicar el análisis de correlación canónica a un conjunto de variables, seleccionar aquellas variables (tanto dependientes como independientes) que aparecen ser significativamente relacionadas, y llevar a cabo posteriores correlaciones canónicas con las restantes variables más significativas, o realizar regresiones con estas variables.**

11.11. Correlación canónica: Ejemplo 2.

Paso 1. Objetivos; Paso 2: Diseño; Paso 3: Condiciones de aplicabilidad

Problema 2.El análisis de **correlación canónica No lineal,** coincide con el análisis de correlación canónica categórica mediante escalamiento óptimo. El propósito de este procedimiento es **reducir con una mayor explicación en conjuntos de las variables categóricas**. El análisis de correlación canónica **no lineal**, se conoce también con el acrónimo de **OVERALS.**

El análisis de correlación canónica típico (**métrico**), es considerada como una extensión de la **regresión lineal múltiple**, en la que el segundo conjunto no contiene una única variable de respuesta sino varias.

En el caso de la correlación canónica no lineal, se desea abordar el estudio de un fenómeno de causalidad entre 2 fenómenos que vienen representados, tanto el causado como el explicativo, por un conjunto de variables y no se desea restringir el modo en que cada una de las variables explicativas incide en cada una de las explicadas.

Mediante el análisis de correlación canónica, Usted puede construir un modelo de causalidad entre ambos conjuntos de variables sin necesidad de especificar el detalle del modelo de causalidad.

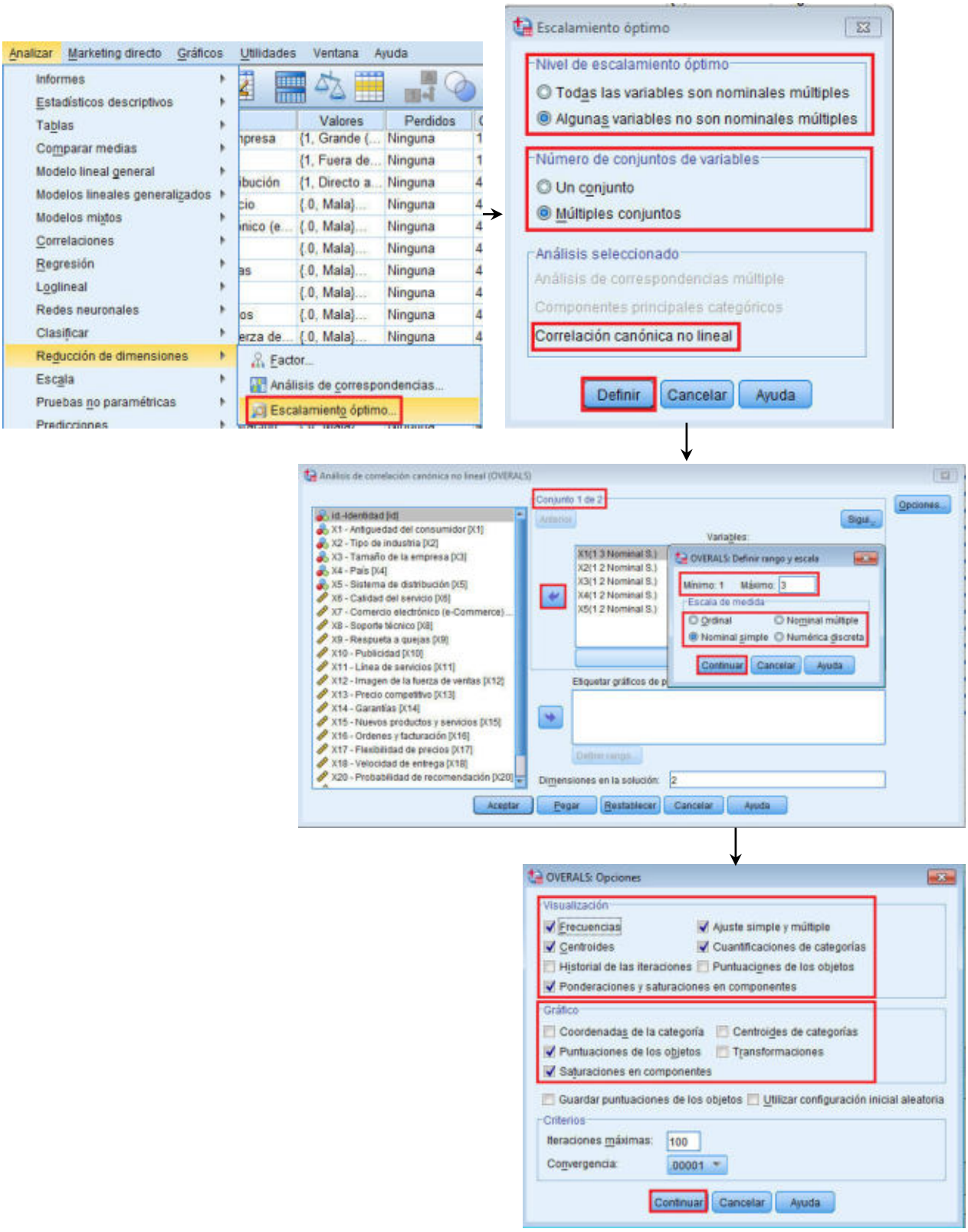
Para ilustrar la aplicación de la **correlación canónica no lineal,** se tomará el ejemplo de la base de datos **BM_MKT_Digital.sav** (esta base de datos se analiza con profundidad en el **Capítulo 6**). Recuerde que los datos consisten en una serie de medidas obtenidas a partir de una muestra de **200 clientes.** Las variables categóricas a trabajar son $\mathbf{X_1, X_2, X_3, X_4, X_5, X_{19}}$ **y** $\mathbf{X_{23}}$**.** Se **sugiere que el dato mínimo a registrar, codificado sea de valor 1.**

Problema: a nivel estadístico**, la empresa MKT Digital** requiere identificar la **reducción de variables categóricas** que permitan obtener una mayor explicación de las mismas, de la base de datos **BM_MKT_Digital.sav** . Se considera que **pasos 2 y 3**, se cumplen sólo para fines de aprendizaje.

Paso 4: Estimación y ajuste

Teclear: Analizar->Reducción de dimensiones->Escalamiento óptimo->Algunas variables no son nominales simples->Múltiples conjuntos->Definir rango y escala->Selección de variables categóricas a analizar como 1er grupo, del que se deberán indicar sus valores mínimos (siempre parten de 1) y máximos, por ejemplo $\mathbf{X_1}$ **Mínimo: 1, Máximo: 3 (en nuestro ejemplo, se escogerán las variables:** $\mathbf{X_1}$ **,** $\mathbf{X_2}$ **,** $\mathbf{X_3}$ **,** $\mathbf{X_4}$ **,** $\mathbf{X_5}$ **como grupo 1)->Escala de medida: Ordinal ; Nominal simple (ponderación indistinta, no hay sidefencias en la preferencia); Nominal múltiple (ponderación por pesos que hace diferencia y distancia, sí hay diferencia en la preferencia); Numérica discreta (distancias del 1-2 son iguales de 2-3)->Continuar->Siguiente->SE REALIZA LO MISMO, el ingreso y declaratoria de la definición de los rangos por categoría pero para grupo 2 (en nuestro caso** $\mathbf{X_{19}}$ **Ordinal;** $\mathbf{X_{23}}$ **Nominal simple)->Opciones-> frecuencias; controles; ajuste simple y múlltiple; cuantificaciones de categorías; ponderaciones y saturaciones en componentes; puntuaciones en los objetos->saturaciones en componentes->Continuar-Aceptar**

Figura 11.15. Proceso Correlación canónica no lineal



Fuente SPSS 20 IBM

Paso 5: Interpretación.

SPSS, genera diversas tablas, las cuales se dividen en los **2 conjuntos que generamos,** siendo las más relevantes (ver **Figura 11.16**):

Figura 11.16. Tablas que genera SPSS de correlación canónica no lineal

Lista de variables

Conjunto		Número de categorías	Nivel de escalamiento óptimo
1	X1 - Antiguedad del consumidor	3	Nominal simple
	X2 - Tipo de industria	2	Nominal simple
	X3 - Tamaño de la empresa	2	Nominal simple
	X4 - País	2	Nominal simple
	X5 - Sistema de distribución	2	Nominal simple
2	X19 - Satisfacción	10	Ordinal
	X23 - Consideración de alianza estratégica	2	Nominal simple

Conjunto 1

X1 - Antiguedad del consumidor

	Frecuencia marginal
< a 1 año	68
1 a 5 años	64
Más de 5 años	68
Perdidos	0
Perdidos dentro del conjunto	0

X2 - Tipo de industria

	Frecuencia marginal
Software para juegos	100
Software empresarial	100
Perdidos	0
Perdidos dentro del conjunto	0

Fuente SPSS 20 IBM

SPSS genera uno delos reportes más valiosos como la **tabla Resumen de análisis**. Ver **Figura 11.17**

Figura 11.17. Tabla Resumen del análisis

Resumen del análisis

		Dimensión		
		1	2	Suma
Pérdida	Conjunto 1	.090	.363	.453
	Conjunto 2	.090	.364	.454
	Media	.090	.364	.454
Autovalores		.910	.636	
Ajuste				1.546

Fuente SPSS 20 IBM

Se observa que el valor de **1.546** debe tender a ser **2** (75%, ya que son 2 dimensiones las que se analizan), para considerar que la agrupación para la definición de variables es correcta. La Dimensión 1 (**0.910**) explica más que la Dimensión 2 (**0.635**),

Por otro lado, **SPSS** genera la tabla **Ajuste. Ver Figura 11.18**

Figura 11. 18 . Tabla Ajuste

Ajuste

		Ajuste múltiple			Ajuste simple			Pérdida simple		
		Dimensión			Dimensión			Dimensión		
Conjunto		1	2	Suma	1	2	Suma	1	2	Suma
1	X1 - Antiguedad del consumidor[a]	.379	.088	.467	.379	.088	.467	.000	.000	.000
	X2 - Tipo de industria[a]	.021	.352	.373	.021	.352	.373	.000	.000	.000
	X3 - Tamaño de la empresa[a]	.063	.153	.216	.063	.153	.216	.000	.000	.000
	X4 - País[a]	.007	.015	.022	.007	.015	.022	.000	.000	.000
	X5 - Sistema de distribución[a]	.321	.065	.386	.321	.065	.386	.000	.000	.000
2	X19 - Satisfacción[b]	.787	.424	1.212	.778	.391	1.169	.009	.033	.042
	X23 - Consideración de alianza estratégica[a]	.014	.925	.939	.014	.925	.939	.000	.000	.000

a. Nivel de escalamiento óptimo: Nominal simple

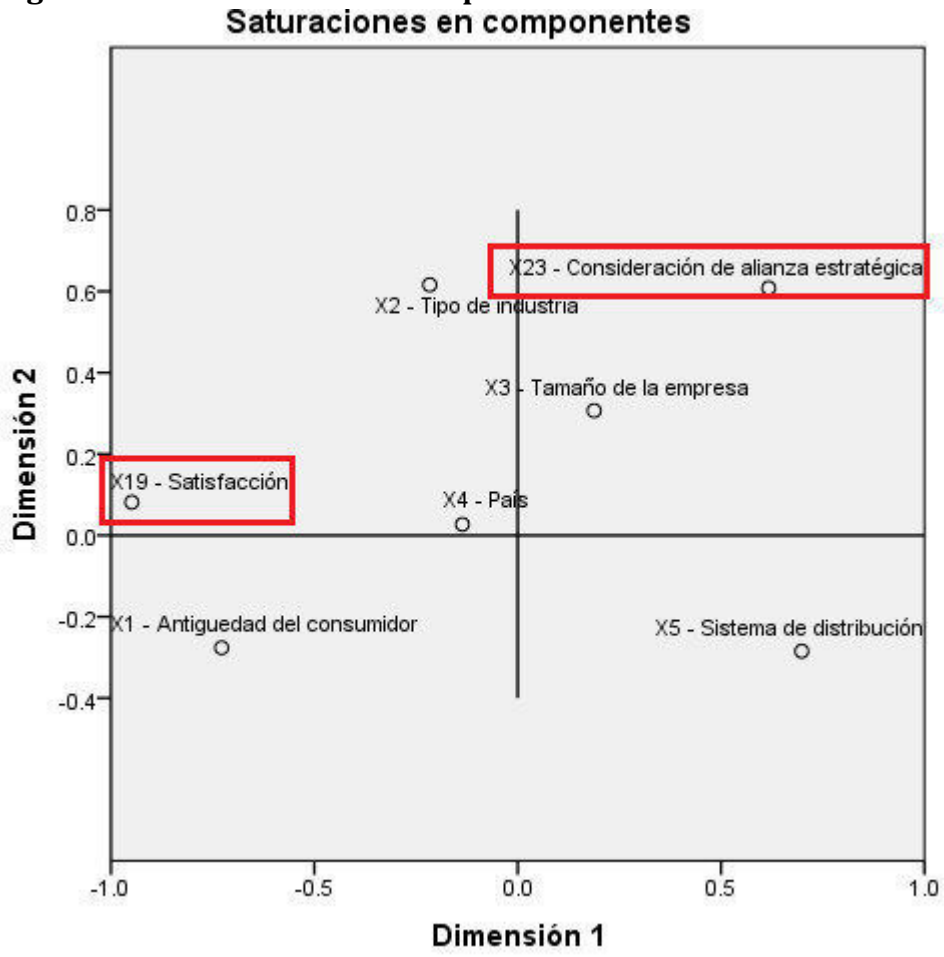
b. Nivel de escalamiento óptimo: Ordinal

Fuente SPSS 20 IBM

El cual nos indica la columna de **Pérdida simple**, que los valores cercanos a cero nos indican que las escalas de medición son adecuadas. Además, nos da indicativo de qué grupo está explicando más. Además, la variable $\mathbf{X_{19}}$ (Satisfacción) es la que más explica más y en general al conjunto 2, aunque a nivel de dimensiones $\mathbf{X_{23}}$ (Consideración de alianza estratégica) explica más a la dimensión 2, en el caso de Ajuste múltiple. Se sugiere cambiar agrupación de variables para encontrar aquellas que aportan más al modelo.

Finalmente, una de las tablas que **SPSS** contiene con mayor calidad de información gráfica es la llamada saturaciones en componentes, que habilita al investigador a explicarse el fenómeno de importancia de agrupación delas variables , para su óptima reducción. Ver **Figura 11.19**

Figura 11.19 Saturación de componentes



Fuente SPSS 20 IBM

Tomando en cuenta que la **Figura 11.18** donde la variable $\mathbf{X_{19}}$ (Satisfacción) es la que m +as explica, en general al **conjunto 2,** aunque a nivel de dimensiones $\mathbf{X_{23}}$ (Consideracio n de alianza estratégica) explica más a la **dimensión 2**; se **requiere asignar un nuevo nombre de variable** a este caso y asociarlas con las que se encuentren más cercanas a cada dimensión **.**

Referencias

Alpert, Mark 1., Robert A. Peterson, y Warren S. M. (1975), Testing the Significance of Canonical Correlations. Proceedings. *American Marketing Associtation 37: 117-19.*

Bartlett M. S. (1941), The Statistical Significance of Canonical Correlations. *Biometrika* 32: 29.

Dillon, W. R., y Goldstein, M. (1984), *Multivariate Analysis: Methods and Applications.* New York: Wiley.

Hair , J.F.; Anderson, R.E.; Tatham, R.L.; Black W.C. (1999).*Análisis Multivariante.*5a. Ed. España. Prentice Hall

IBM (2011a). IBM SPSS Statistics Base 20. EUA. .Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Base.pdf

IBM (2011b).Guía breve de IBM SPSS Statistics 20. EUA.Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Brief_Guide.pdf

IBM (2011c). IBM SPSS Missing Values 20. EUA. .Industrial Business Machines. Recuperado el 20161201 de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Missing_Values.pdf

Lambert, Z., y Durand R. (1975), Sorne Precautions in Using Canonica1 Analysis. *Journal of Marketing Research* 12 (November): 468-75.

Stewart, D., y William, L. (1968), A General Canonical Correlation Index. *Psychologica/ Bulletin 10:* 160-63.

Apéndice. Matriz de pruebas estadísticas sugeridas

¿Que desea realizar?	Número de variables /condiciones	Diseño	Paramétrico/No Paramétrico	Prueba Estadística Recomendada	Procedimiento Estadístico	
Búsqueda de diferencias entre condiciones	Una variable: 2 condiciones	Mediciones Independientes	Paramétrico	Prueba t de Muestras Independientes	Comparación de Medias	Prueba t de Muestras Independientes
		Mediciones repetidas	Paramétrico	Prueba t Relacionada	Comparación de Medias	Prueba t de Muestras Pareadas
		Mediciones Independientes	No Paramétrico	Mann-Whitney U	Pruebas No Paramétricas	Prueba Dos Muestras Independientes
		Mediciones repetidas	No Paramétrico	Wilcoxon	Pruebas No Paramétricas	Prueba de Dos Muestras Relacionadas
	Una variable: más de dos condiciones	Mediciones Independientes	Paramétrico	ANOVA de Un Factor de Mediciones Independientes	Modelo General Lineal	Univariado
		Mediciones repetidas	Paramétrico	ANOVA de Un Factor de Mediciones Repetidas	Modelo General Lineal	Medidas Repetidas
		Mediciones Independientes	No Paramétrico	Kruskal-Wallis	Pruebas No Paramétricas	K Muestras Independientes
		Mediciones repetidas	No Paramétrico	Friedman	Pruebas No Paramétricas	K Muestras Relacionadas
	Dos variables	Mediciones Independientes en ambas variables	Paramétrico	ANOVA de Dos Factores Independientes	Modelo General Lineal	Univariado
		Mediciones una Independiente y otra de Medidas Repetidas	Paramétrico	ANOVA de Dos Factores de Mediciones Repetidas	Modelo General Lineal	Mediciones Repetidas
		Mediciones Independientes en ambas variables	Paramétrico	ANOVA de Dos Factores de Diseño Combinado	Modelo General Lineal	Mediciones Repetidas
	Más que una variable dependiente	Mediciones Independientes	Paramétrico	MANOVA Independiente	Modelo General Lineal	Multivariado
		Mediciones Repetidas	Paramétrico	MANOVA Repetida	Modelo General Lineal	Mediciones Repetidas
Comparar conteo de frecuencias (Categorías)	NA	Asociación	No Paramétrico	Chi-Cuadrada	Descriptivo	Cruce-Tabular
Correlación de variables	Dos variables	Correlacional	Paramétrico	Pearson	Correlacionar	Bivariado
			No Paramétrico	Spearman	Correlacionar	Bivariado
			No Paramétrico	Kendall tau-b	Correlacionar	Bivariado
	Más de dos variables	Correlacional	Paramétrico	Regresión Múltiple	Regresión	Lineal
Reducción de datos	Muchas variables	Correlacional	Paramétrico	Análisis Factorial	Reducción de datos	Factor
		Correlacional	Paramétrico	Análisis de confiabilidad	Escala	Análisis de confiabilidad

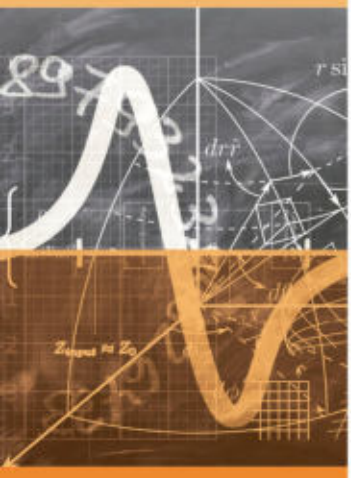
Fuente: Hinton, P.R.; Brownlow, Ch.; McMurray,I y Cozens, B. (2004). *SPSS Explained.* USA: Routledge,Taylor y Francis Group

Biografía del Dr. Juan Mejía Trejo

Nacido en la Ciudad de México en 1964, participó como Gerente de Explotación (1987-2008) de la planta interna en Teléfonos de México SAB (TELMEX), es Ingeniero en Comunicaciones y Electrónica (1998) por parte de la Escuela Superior de Ingeniería Mecánica y Eléctrica (ESIME), del Instituto Politécnico Nacional (IPN). Realizó sus estudios de posgrado en la Maestría en Administración de Empresas en Telecomunicaciones (2000), en el Instituto Tecnológico de Teléfonos de México (Inttelmex), logrando su Doctorado en Ciencias Administrativas (2010), en la Escuela Superior de Comercio y Administración (ESCA), del Instituto Politécnico Nacional (IPN). Ha sido profesor de asignatura en las instituciones de educación superior de la zona metropolitana de Guadalajara, Jal. México como UNITEC, UVM, ITESO, UP, UAG. Actualmente (2010 a la fecha) es profesor investigador titular en el Departamento de Mercadotecnia y Negocios Internacionales del Centro Universitario de Ciencias Económico Administrativas (CUCEA), de la Universidad de Guadalajara (UdG) desempeñándose como Coordinador del Doctorado en Ciencias de la Administración (2015 a la fecha) y como Presidente de la Academia de Negocios Electrónicos. Es miembro (2011 a la fecha) del Sistema Nacional de Investigadores (SNI) del Consejo Nacional de Ciencia y Tecnología (CONACYT) de México Nivel I, siendo su línea de investigación: los procesos de la innovación tecnológica así como los negocios electrónicos y la mercadotecnia digital, de los que ha escrito cerca de 50 obras entre artículos, capítulos de libros y libros, sobre estos temas, disponibles en el portal *Social Science Research Network* (https://papers.ssrn.com/sol3/results.cfm), además de poseer 5 solicitudes de patente sobre sistemas de información orientados al desarrollo de nuevos productos y servicios, toma de decisiones de alta gerencia para la planeación estratégica así como el desarrollo de la innovación.

Las Ciencias de la Administración y el Análisis Multivariante
bajo el enfoque de las Técnicas Dependientes
Tomo 1
se terminó de imprimir en junio de 2017
en los talleres de Ediciones de la Noche.
Madero 687, col. Centro
Guadalajara, Jalisco.

www.edicionesdelanoche.com



La obra *Las ciencias de la administración y el análisis multivariante. Proyectos de investigación, análisis y discusión de resultados. Tomo I. Las técnicas dependientes* tiene como principales objetivos presentar:

1. De forma teórica, las principales herramientas utilizadas en el análisis de datos y confiabilidad en cuestionarios, para entrar de lleno a las técnicas que el enfoque abarca, como son la descripción básica del software SPSS, la introducción a las técnicas multivariantes, el análisis de datos, la confiabilidad de los cuestionarios, la correlación y regresión lineal simple y múltiple, el análisis discriminante múltiple, pruebas no paramétricas de dos muestras, el análisis de la varianza univariante (ANOVA) y multivariante (MANOVA), el cruce-tabular y *chi*-cuadrada, el análisis de conjunto y el análisis de correlación canónica.

2. De forma práctica y demostrativa, problemas ejemplo que se resuelven a propósito de aplicar los seis pasos de solución de problemas multivariantes de Hair *et al.* (1999) utilizando los comandos del SPSS 20 de IBM.

3. Servir como libro base de las asignaturas de métodos cuantitativos I y II del Doctorado en Ciencias de la Administración del Centro Universitario de Ciencias Económico Administrativas de la Universidad de Guadalajara.

UNIVERSIDAD DE GUADALAJARA
CENTRO UNIVERSITARIO DE CIENCIAS
ECONÓMICO ADMINISTRATIVAS

ISBN 978-607-742-773-5
9 786077 427735