

TOMO I

Juan Mejía Trejo

EVALUACIÓN DE IMPACTO SOCIAL EN PROYECTOS DE INNOVACIÓN VÍA STATA

Métodos: Inferencia Causal, Aleatorización,
Propensión de Coincidencia de Puntaje y
Doble diferencia



CUCEA

El mejor lugar para el talento

TOMO I

Juan Mejía Trejo

EVALUACIÓN DE IMPACTO SOCIAL EN PROYECTOS DE INNOVACIÓN VÍA STATA

Métodos: Inferencia Causal, Aleatorización,
Propensión de Coincidencia de Puntaje y
Doble Diferencia



CUCEA

El mejor lugar para el talento

Este libro fue financiado con el fondo federal PROINPEP, Programa de Incorporación y Permanencia de Posgrado en el Programa Nacional de Posgrado de Calidad del CONACYT.

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 (CUCEA) de la Universidad de Guadalajara (UdeG).

Primera edición, 2021

D.R. © 2021, Universidad de Guadalajara
Centro Universitario de Ciencias Económico Administrativas
Periférico Norte No. 799
Núcleo Universitario, Los Belenes,
45100 Zapopan, Jalisco

ISBN Tomo: 978-607-571-166-9

ISBN Colección: 978-607-571-165-2

Hecho en México

Made in Mexico

Todos los Derechos son reservados. Esta publicación no puede ser reproducida ni en su totalidad o parcialidad, en español o cualquier otro idioma, ni registrada en, transmitida por, un sistema de recuperación de información, en ninguna forma ni por ningún medio, sea mecánico, foto-químico, electrónico, magnético, electroóptico, por fotocopia, o cualquier otro, inventado o por inventar, sin permiso expreso, previo y por escrito del autor.

ÍNDICE

Introducción	1
 CAPÍTULO 1.	
Conceptos Básicos	5
¿Qué son los impactos sociales?.....	5
Por qué evaluar	9
Monitoreo vs. Evaluación	11
Monitoreo.....	13
Los indicadores dentro de un marco de M&E	17
Monitoreo basado en resultados	20
Desafíos en la configuración de un sistema de monitoreo	23
Las evaluaciones de impacto	25
Estudios de eficacia y estudios de efectividad	29
Evaluación operativa y sus desafíos	31
Evaluación operativa vs. evaluación de impacto	31
Evaluación de impacto cualitativo vs. cuantitativo	33
Evaluación de impacto cuantitativo: ex post vs ex ante	35
Evaluación de impacto: prospectiva vs. retrospectiva	37
El problema de los contrafactuales.....	40
Búsqueda de contrafactuales con y sin comparaciones	41
Contrafactual antes y después de comparar	43
Evaluación de impacto y su relación con el sesgo.....	45
Describiendo un caso	50
El sesgo de selección	65
Evaluaciones de impacto ex post.....	73
Evaluaciones de procesos	75
Evaluación de impacto vs. evaluación de proyectos	77
Análisis de costo-beneficio y costo-efectividad	78

Ética y evaluación de impacto.....	80
Consideraciones finales.....	82

CAPÍTULO 2.

Preparación, Inferencia Causal

y Contrafactuales	85
Pasos por considerar	87
Construcción de una teoría del cambio.....	88
Desarrollo de una cadena de resultados.....	89
La especificación de las preguntas de la evaluación.....	92
La selección de indicadores de resultados y desempeño	94
Lista de verificación: datos para los indicadores	96
Inferencia causal	98
El contrafactual	100
La estimación del contrafactual	102
Dos estimaciones falsas del contrafactual	105
Contrafactual falso 1: comparación de resultados antes y después	106
Contrafactual falso 2: comparando grupos inscritos y no inscritos (autoseleccionados)	108

CAPÍTULO 3.

Aleatorización.....	111
Asignación aleatoria como excelente estimación del contrafactual.....	113
Validez externa e interna	117
Diseño estadístico aleatorio	118
Cuando aplicar la asignación aleatoria	122
Desventajas de la aleatorización.....	123
Cálculo de los efectos del tratamiento.....	126
Cómo asignar aleatoriamente el tratamiento	128
Efecto del tratamiento con aleatorización pura.....	131
Efecto del tratamiento con aleatorización parcial.....	132

Diferentes métodos de aleatorización.....	134
Intervención individual vs. conglomerados	135
Problemas con la aleatorización	137
Evaluación de impacto aleatorizada en la práctica	139
El modelo de diferencias.....	140
Cuestiones éticas.....	144
Calidad de la aleatorización.....	145
Efectos indirectos (<i>spillovers</i>)	147
Heterogeneidad en los impactos por aleatorización.....	149
Valor de un estudio de referencia	151
Dificultades en la aleatorización	152
Nivel al que se lleva a cabo una asignación aleatoria	154
Lista de verificación: la asignación aleatoria.....	156
Ejemplo utilizando STATA.....	158
Impacto de un programa de innovación de pbt por región	158
Evaluación de impacto por participación.....	163
Evaluación de impacto por participación y región	165
Evaluación de impacto por participación a nivel regional.....	167
Evaluando los efectos indirectos o de derrame (<i>spillover</i>)	169

CAPÍTULO 4.

Propensión de Coincidencia de Puntaje

(<i>Propensity Score Matching</i>)	171
¿Qué es PSM?	172
Lo que hace PSM	174
Coincidencia por propensión de puntaje	175
Verificación de coincidencia o pareamiento.....	177
Teoría PSM	178
El supuesto de la independencia condicional.....	185
El supuesto de soporte común.....	186
Breve ejemplo de soporte común por observación.....	192
TOT usando PSM.....	194
Reglas de aplicación PSM	194

Paso 1: Estimando un modelo de participación de programa.....	195
Paso 2: Definiendo la región de soporte común y pruebas de balanceo	197
Paso 3: Haciendo coincidentes participantes-no participantes	198
¿Qué algoritmo de emparejamiento elegir?	210
Cálculo del promedio de impacto en el grupo de tratamiento....	211
Calidad de la coincidencia o emparejamiento.....	211
Estimación de los errores estándar con PSM. Uso de Bootstrap	213
PSM y los métodos de regresión	215
La combinación de la coincidencia con otros métodos	216
Diferencias en diferencias pareadas.....	216
El método de control sintético	217
Crítica al PSM: ventajas y desventajas.....	217
Pruebas de falsificación	221
Ejemplo utilizando STATA.....	224
Ecuación PSM en la satisfacción de la propiedad de balance	225
Efecto de tratamiento promedio usando vecino más cercano (<i>nearest-neighbor</i>)	232
Efecto de tratamiento promedio usando estratificación de coincidencia (<i>stratification matching</i>)	233
Efecto de tratamiento promedio usando calibración por coincidencia (<i>radius matching</i>)	234
Efecto de tratamiento promedio usando núcleo y coincidencia local (<i>kernel matching</i>)	235
Efecto de tratamiento promedio y su chequeo de robustez	237
 CAPÍTULO 5.	
Doble Diferencia (DD. Double Difference).....	239
Método DD.....	242
Usos del modelo DD	246
¿Qué sucede en el modelo DD?	247
Modelo de panel de efectos fijos (<i>panel fixed-effects</i>)	251

Implementado DD	252
Ventajas y desventajas del uso de la DD	254
Utilidad del método DD	257
Tendencias iguales en el método DD	258
Comprobación del supuesto de igualdad de tendencias en el método DD	260
Modelos DD alternativos	262
Combinando PSM con DD	263
Método de la triple diferencia (<i>triple-difference method</i>)	264
Ajuste de las tendencias de diferencial de tiempo	264
Limitaciones del método de DD	265
Verificación DD	266
Ejemplo utilizando STATA	267
Comparación simple (uso de <i>ttest</i>).....	268
Implementación de regresión.....	271
Checando robustez de DD con regresión de efectos fijos.....	276
DD en datos de sección transversal	279
Tomando en cuenta las condiciones iniciales	285
DD combinada con PSM	289

ANEXO.

Introducción a STATA	295
Estructura de archivos (<i>File Structure</i>)	296
Descripción de archivos y ejemplos en STATA	296
Ejemplo 1. Abriendo el conjunto de datos (<i>Open data set</i>)	299
Ejemplo 2. Guardando el conjunto de datos (<i>Saving data set</i>)	301
Ejemplo 3. Saliendo de STATA (<i>Exiting STATA</i>).....	302
Ejemplo 4. Apoyo de STATA (<i>Help</i>)	302
Ejemplo 5. Apoyo de STATA (<i>Lookup/Search</i>).....	304
Ejemplo 6. Notas sobre los comandos STATA	305
Ejemplo 7. Listando las variables (<i>Listing variables</i>)	305
Ejemplo 8. Listando los datos (<i>Listing data</i>).....	307

Ejemplo 9. Listando datos específicos.....	309
Ejemplo 10. Sumando datos (<i>Summarizing data</i>)	310
Ejemplo 11. Ponderaciones (<i>Weight</i>)	311
Ejemplo 12. Ordenamiento con ponderación (<i>Sort/Weight</i>)	315
Ejemplo 13. Uso del comando <i>tabstat</i>	316
Ejemplo 14. Distribuciones de frecuencia (<i>Tabulations</i>).....	317
Ejemplo 15. Estadística descriptiva (<i>Table Command</i>)	319
Ejemplo 16. Despliegue de punto decimal	320
Ejemplo 17 . Comando contar (<i>Count</i>)	322
Ejemplo 18. Generando nuevas variables.....	323
Ejemplo 19. Etiquetando variables (<i>Labeling variables</i>)	326
Ejemplo 20. Etiquetando variables (<i>Labeling variables</i>) desde barra de herramientas.....	327
Ejemplo 21. Etiquetando datos (<i>labeling data</i>)	328
Ejemplo 22. Etiquetando valores de variables (<i>labeling values of variables</i>)	329
Ejemplo 23. Soltar variables así como observaciones.....	332
Ejemplo 24. Produciendo gráficos	335
Ejemplo 25. Combinando archivos de datos	338
Ejemplo 26. Haciendo apéndices en archivos (<i>Appending data sets</i>)	344
Ejemplo 27. Archivos .log (<i>log files</i>)	347
Ejemplo 28. Archivos .do (<i>.do files</i>)	350
Ejemplo 29. Archivos .ado (<i>.ado files</i>).....	354
Glosario	355
Bibliografía	423

Introducción

En el mundo empresarial e industrial, el diseño e implementación de proyectos normalmente toma en cuenta el impacto económico, financiero y hasta el político o ambiental en la introducción de innovaciones. Sin embargo, en los primeros veinte años del siglo XXI, han sucedido diversos acontecimientos que han demostrado que la evaluación de impacto social en la generación de bienestar, es de vital importancia.

De hecho, en los tiempos de la nueva normalidad que se avizora como la era PosCOVID-19, esto toma particular relevancia dado que todas las políticas y acciones que emitan empresas y gobiernos, deberán contar con el aval necesario de una evaluación de impacto social a la introducción de innovaciones.

Es por esta razón, que la obra: *EVALUACIÓN DE IMPACTO SOCIAL EN PROYECTOS DE INNOVACIÓN VÍA STATA. Métodos: Inferencia Causal, Aleatorización, Propensión de Coincidencia de Puntaje y Doble Diferencia, Tomo I* está orientada a describir tanto a propios como ajenos al tema, lo qué es el impacto social, sus características, condiciones e implicaciones, los principales métodos utilizados para calcularla así como las oportunidades que se vislumbran en la era PosCOVID-19, que demanda que los recursos y acciones de innovación a diseñar e implementar, reflejen altos estándares de impacto social que fomenten el bienestar, particularmente en los países emergentes.

Para lograrlo, esta obra está dividida en una colección de dos tomos, correspondiendo al Tomo I:

Capítulo 1. Conceptos básicos. Este capítulo está orientado a explicar qué son los impactos sociales, las razones de por qué evaluar, la diferenciación entre monitoreo vs. evaluación, los indicadores de ambos,

lo que implica un monitoreo basado en resultados y su configuración. Esto permite definir mejor la idea sobre lo que son las evaluaciones de impacto, su relación como estudios de eficacia y efectividad, las implicaciones sobre la evaluación operativa así como sus características de impacto cualitativo y cuantitativo además del punto de referencia en el que se requiera la evaluación de impacto ya sea con prospectiva o con retrospectiva. Realizar cualquier evaluación de impacto social para la introducción de innovaciones, requiere reconocer al grupo de tratamiento y al de control, lo que necesita resolver el problema de cómo determinar al grupo contrafactuales; con y sin comparaciones así como su relación con el sesgo. Se aporta una descripción de caso, explicando el sesgo de selección, en un escenario ex post y explicando a su vez las características que le acompañan como evaluaciones de procesos y de evaluación de proyectos, la importancia del análisis de costo-beneficio y costo-efectividad así como la ética en la evaluación de impacto social en la introducción de innovaciones.

Capítulo 2. Preparación, inferencia causal y contrafactuales.

Dada la preparación de los grupos de tratamiento y de control, este capítulo está diseñado para considerar los pasos que sirvan a la construcción de escenarios contrafactuales que posibiliten a los grupos, explicar la construcción de una teoría del cambio, explicar los eslabones de una cadena de resultados, las técnicas que facilitan especificación de las preguntas de la evaluación, la selección de indicadores de resultados y desempeño, realizar una lista de verificación basada en los datos para los indicadores. De esta manera, se equipa al investigador para realizar inferencia causal, la descripción y estimación de los contrafactuales, a través de dos casos de estudio con los que se cierra el capítulo.

Capítulo 3. Aleatorización. Es a partir de este capítulo hasta el 7, donde se inicia con las diferentes técnicas basadas en el software de STATA 14, por lo que se recomienda, si no está familiarizado con el software, iniciar con el Anexo: Introducción a STATA. Esto le permitirá

comprender las reglas de operación así como de acceder a los archivos del caso de ejemplo que se estará tratando y que se encuentran alojados en el enlace: [https://amidi.mx/en/publicaciones/archivo de soporte e-libros](https://amidi.mx/en/publicaciones/archivo%20de%20soporte%20e-libros). Así, el capítulo 3 le permitirá al lector conocer las características de la asignación aleatoria, ventajas y desventajas de su uso, las condiciones recomendables para su aplicación, las implicaciones éticas, los diferentes métodos de aleatorización que existen, listas de verificación para llevarlo a cabo así como un ejemplo con STATA para su comprensión.

Capítulo 4. Propensión de coincidencia de puntaje (PSM *Propensity Score Matching*). Explica al lector más asiduo, qué es y qué hace esta técnica, los supuestos y reglas que deben reunirse para su aplicación, pasos para lograrlo así como uso de técnicas complementarias. Se debaten sus ventajas y desventajas de uso y se completa con un ejemplo de STATA para comprobar su comprensión de uso.

Capítulo 5. Doble Diferencia (DD *Double-Difference*). El capítulo está diseñado para explicar el método qué es, cómo se usa, lo que sucede en el mismo a través del uso de panel de efectos fijos, cómo implementarlo, ventajas y desventajas, su utilidad, tendencias, modelos alternativos, la posibilidad de combinarlo con el **PSM**, ajustes, limitaciones, concluyendo con un ejemplo de STATA para comprobar su comprensión de uso.

CAPÍTULO 1.

Conceptos Básicos

El uso de métodos cuantitativos para medir el impacto de programas sociales ha cobrado un gran interés recientemente. En los últimos años han surgido organizaciones dedicadas a la elaboración y el financiamiento de evaluaciones de impacto. Las entidades multilaterales de crédito y las agencias de cooperación han enfatizado, cada vez con mayor fuerza, la necesidad de evaluar concienzudamente los proyectos de desarrollo. Muchos países han creado oficinas independientes de evaluación y monitoreo de programas públicos. Las evaluaciones de impacto han comenzado a desempeñar un papel preponderante en el diseño de diversas políticas como las de salud, educación ,innovación, etc. y, por ende, en el control político y la controversia democrática (**Bernal y Peña, 2011**).

¿Qué son los impactos sociales?

Los impactos sociales son cambios en uno o más de los siguientes ámbitos (AIEI, 2015):

- La forma de vida de las personas; es decir, cómo viven, trabajan, juegan e interactúan unas con otras en el quehacer cotidiano;
- Su cultura; esto es, sus creencias, costumbres, valores e idioma o dialecto;
- Su comunidad; su cohesión, estabilidad, carácter, servicios e instalaciones;

- Sus sistemas políticos; el grado al que las personas pueden participar en las decisiones que afectan sus vidas, el nivel de democratización que está teniendo lugar y los recursos suministrados para ese fin;
- Su entorno; la calidad del aire y el agua que utiliza la población, la disponibilidad y calidad de los alimentos que consume, el nivel de peligro o riesgo, polvo y ruido al que está expuesta, la idoneidad del saneamiento, su seguridad física; su acceso y control sobre los recursos;
- Su salud y bienestar; la salud es un estado de bienestar total desde el punto de vista físico, mental, social y espiritual, y no solamente la ausencia de enfermedad;
- Sus derechos tanto personales como a la propiedad; especialmente si las personas se ven económicamente afectadas o si sufren desventajas personales que pueden incluir la violación de sus libertades civiles;
- Sus temores y aspiraciones; sus percepciones acerca de su propia seguridad, sus temores acerca del futuro de su comunidad y sus aspiraciones tanto en lo que respecta a su propio futuro como al de sus hijos.

Los principios internacionales de la evaluación del impacto social la definen como:

los procesos de análisis, monitoreo y gestión de las consecuencias sociales voluntarias e involuntarias de intervenciones planeadas (políticas, programas, planes, proyectos) y todo proceso de cambio social invocado por dichas intervenciones (AIEI, 2015).

Los programas y políticas de desarrollo (por ejemplo: de innovación, educación o salud, entre otros), suelen estar diseñados para producir resultados planeados, como aumentar los ingresos, mejorar el bienestar o

reducir las enfermedades. Saber si estos cambios se logran o no, es una pregunta crucial para las diferentes autoridades que las mandan implementar, como las políticas públicas, educación, las de desarrollo industrial o las de salud, entre otras, aunque a menudo no suele abordarse. Es más habitual que los administradores de los programas y los responsables de las políticas se centren en medir e informar sobre los insumos y los productos inmediatos de un programa (cuánto dinero se gasta, cuántas personas participan en una innovación de procesos, cuántos empleos se generan, etc.) en lugar de evaluar si los programas han logrado sus objetivos de mejorar los resultados. (Gertler et al., 2017).

Khandker et al., (2017) refieren a que se pueden usar varios enfoques para evaluar los diversos programas que un gobierno emite, como lo son los *proyectos de innovación como tecnología, social, procesos, etc.* Por ejemplo:

- El *monitoreo*, que rastrea los *indicadores clave del progreso* en el curso de un programa como base para evaluar los resultados (*outcomes*) de la intervención.
- La *evaluación operativa*, que examina cuán efectivamente fueron los programas implementados y si hay *brechas* entre los *resultados (outcomes) planificados y los realizados*.
- La *evaluación de impacto*, que estudia si los cambios en el *bienestar* se deben realmente al programa intervención y no a otros factores.

El *monitoreo y la evaluación* son fundamentales en la formulación de políticas *basadas en evidencia*. Ofrecen un conjunto central de instrumentos que las partes interesadas pueden utilizar para verificar y mejorar la calidad, eficiencia y efectividad de las políticas y de los programas en diferentes etapas de implementación o, en otras palabras, para centrarse en los resultados. A nivel de la gestión del programa, es necesario saber cuáles son las opciones de diseño costo-efectivas, o demostrar ante los

responsables de la toma de decisiones que los programas están logrando sus resultados previstos con el fin de obtener asignaciones presupuestarias para continuarlos o ampliarlos. Los gobiernos deben rendir cuentas ante los ciudadanos para informales del resultado de los programas públicos. La evidencia puede constituir una base sólida para la transparencia y la rendición de cuentas. **(Gertler, et al., 2017).**

Dichos enfoques de evaluación pueden llevarse a cabo utilizando *métodos cuantitativos* (es decir, recopilación de datos de encuestas o simulaciones) *antes o después*, de la introducción de un programa. Así siguiendo a **Khandker et al., (2017)**:

- Una *evaluación ex ante* predice los impactos del programa utilizando datos antes de la intervención del programa, mientras que,
- La *evaluación ex post* examina los resultados (*outcomes*) después de la implementación de los programas.
- Las *comparaciones reflexivas* son un tipo de *evaluación ex post*; ellos examinan los impactos del programa a través de la *diferencia* en los *resultados (outcomes)* de los participantes *antes y después de la implementación del programa* (o entre participantes y no participantes). Se tratarán ejemplos al respecto.

El principal desafío, que atraviesan los diferentes tipos de *evaluación de impacto*, es encontrar un *buen contrafactual*, es decir, la situación que un sujeto participante habría experimentado si él o ella no hubiera estado expuesto al programa. Variantes de la *evaluación de impacto* discutidas en los siguientes capítulos que se incluyen son **(Khandker et al., 2012)**:

- Las evaluaciones aleatorias.
- Los métodos de coincidencia (*matching methods*), en especial, la técnica de propensión de coincidencia de puntaje (**PSM. Propensity Score Matching**).
- Los métodos de doble diferencia (**DD. Double-Difference methods**).

- El uso de variables instrumentales (**IV**. *Instrumental Variable method*).
- El diseño de regresión discontinuidad (**RD**. *Regression Discontinuity Design*) y los métodos de canalización (**PM**. *Pipeline Methods*).
- Impacto distributivo (**DI**. *Distributional Impacts*).

Cada uno de estos métodos, implica un conjunto diferente de supuestos en la contabilización, del posible *sesgo de selección* en la participación, que podría afectar la construcción de los efectos del tratamiento del programa.

Por qué evaluar

Gertler et al. (2017) afirman que las *evaluaciones de impacto* forman parte de una agenda mas amplia de formulación de políticas tales como: de innovación, educación, salud o públicas *basadas en evidencia*. Esta tendencia mundial creciente se caracteriza por un cambio de enfoque, *ya que en lugar de centrarse en los insumos lo hace en los productos y resultados*. Centrarse en los resultados no solo sirve para definir y hacer un seguimiento de los objetivos nacionales e internacionales, sino que además, los administradores de programas utilizan y necesitan cada vez más los resultados definir las asignaciones presupuestarias y orientar el diseño del programa y las decisiones de alto nivel del mismo.

La evidencia robusta generada por las *evaluaciones de impacto* está sirviendo cada vez más como fundamento para una mayor rendición de cuentas, innovación y aprendizaje. En un contexto, en que los responsables de las políticas y la sociedad civil exigen resultados y la rendición de cuentas de los programas públicos, la evaluación de impacto puede proporcionar evidencia robusta y creíble sobre el desempeño y ante todo, sobre si un programa concreto ha alcanzado o está alcanzando sus resul-

tados deseados. Las *evaluaciones de impacto* también son cada vez más utilizadas para *probar innovaciones* en el diseño de programas o en la prestación de servicios. A nivel mundial, estas evaluaciones son fundamentales para construir conocimientos acerca de la efectividad de los programas de desarrollo, iluminando sobre lo que funciona y no funciona para reducir la pobreza, incrementar la competitividad y mejorar el bienestar.

En pocas palabras, una *evaluación de impacto* mide los cambios en el bienestar de los individuos que se pueden atribuir a un proyecto, un programa o una política específicos. Este enfoque en la atribución es el sello distintivo de las evaluaciones de impacto. Por lo tanto, el reto fundamental en una evaluación de esta naturaleza consiste en identificar la relación causal entre el programa o la política y los resultados de interés. Las evaluaciones de impacto suelen medir el impacto promedio de un programa, las modalidades del programa o una innovación en el diseño. Por ejemplo, ¿el programa de introducción de innovaciones en procesos de agua y saneamiento, aumentó el acceso a agua potable y mejoró los resultados de salud?, ¿un programa innovador de estudios alternativo, mejoró las puntuaciones de las pruebas de los alumnos?, ¿la innovación de incluir destrezas cognitivas como parte de un programa de formación de jóvenes ha tenido éxito promoviendo la iniciativa empresarial e incrementando los ingresos?

En cada uno de estos casos, *la evaluación de impacto* proporciona información sobre si el programa provocó los cambios deseados en los resultados, al compararse con estudios de casos o anécdotas específicas, que solo pueden brindar información parcial y que quizá no sean representativos de los impactos generales del programa. En este sentido, las *evaluaciones de impacto* bien diseñadas e implementadas son capaces de proporcionar evidencia convincente y exhaustiva que puede ser utilizada para fundamentar las decisiones de las políticas, influir en la opinión pública y mejorar el funcionamiento de los programas. Las *evaluaciones*

de impacto clásicas abordan la efectividad de un programa en comparación con la ausencia del mismo.

Además de abordar la pregunta básica de si un programa es efectivo o no, las *evaluaciones de impacto* también se pueden utilizar para probar explícitamente modalidades de programas alternativos o innovaciones de diseño. A medida que los responsables de las políticas de implementación de programas innovadores, se centran cada vez más en entender mejor *cómo perfeccionar la implementación* y obtener más a cambio del dinero, los enfoques que prueban alternativas de diseño están ganando terreno rápidamente. Por ejemplo, una evaluación puede comparar el desempeño de un programa de formación con el de una campaña promocional para ver cuál es más efectivo para aumentar la alfabetización financiera. Una *evaluación de impacto* puede probar cuál es el enfoque de la combinación de innovaciones en nutrición y de estimulación del niño, que tiene el mayor impacto en su desarrollo. O la evaluación puede probar una innovación de diseño para mejorar el diseño de un programa existente, como utilizar mensajes de texto para mejorar el cumplimiento cuando se trata de tomar la medicación prescrita.

Monitoreo vs. Evaluación

El *monitoreo* es un proceso continuo mediante el cual se lleva a cabo un seguimiento de lo que ocurre con un programa y se utilizan los datos recopilados para fundamentar la implementación de los programas y la administración y las decisiones diarias. A partir, sobre todo de datos administrativos, el *monitoreo* realiza un seguimiento de los desembolsos financieros y del desempeño del programa en relación con los resultados esperados, y analiza las tendencias a lo largo del tiempo. El *monitoreo* es necesario en todos los programas y constituye una fuente crítica de información sobre el desempeño de los mismos, lo cual abarca también la implementación y los costos. Normalmente, el *monitoreo* se aplica a los

insumos, actividades y productos, aunque ocasionalmente también puede abarcar los resultados, como, por ejemplo, el progreso alcanzado en los objetivos de desarrollo nacional.

Un *sistema de monitoreo* su esencia central es la de establecer objetivos, indicadores y metas para los programas establecidos. La información y los datos resultantes se pueden utilizar para evaluar el rendimiento de los *programas de intervención* (Khandker, et al. 2017). Por ejemplo, el grupo de evaluación independiente del Banco Mundial evalúa el progreso de la estrategia de reducción de la pobreza (**PRS**. *Poverty Reduction Strategy*) del Banco Mundial y Fondo Monetario Internacional, contra sus objetivos a través de la supervisión; muchos países también han estado desarrollando sistemas de monitoreo para rastrear implementación de la iniciativa **PRS** y su impacto en la pobreza. Al *comparar* el programa de resultados (*outcomes*) con los objetivos específicos, el *monitoreo* puede ayudar a mejorar el diseño de políticas y su implementación, así como promover la rendición de cuentas y el diálogo entre políticas fabricantes y partes interesadas.

En contraste, la *evaluación* es un procesos *sistemático y objetivo* de los resultados (*outcomes*) logrado por el programa. En otras palabras, la *evaluación busca demostrar que los cambios en los objetivos se deben solo a las políticas específicas emprendidas*. Monitoreo y *evaluación* juntos se han denominado **M&E** (*Monitoring & Evaluation*). Por ejemplo, **M&E** pueden incluir (Khandker, et al., 2017):

- *Evaluación de procesos*, el cual examina cómo operan los programas y se enfoca en los problemas de entrega de servicio,
- *El análisis de costo-beneficio*, que compara los costos del programa vs. los beneficios de la entrega; y
- *Las evaluaciones de impacto*, que cuantifican los efectos de los programas en las personas, hogares y comunidades.

Todos estos aspectos, son considerados como parte de un buen sistema de **M&E** y generalmente son realizados por la agencia implementadora.

Monitoreo

El monitoreo de la *implementación de un programa*, las más de las veces mediante el uso de datos administrativos, es crítico en una *evaluación de impacto*. Permite al equipo de evaluación verificar si las actividades se están realizando según lo planificado, es decir, a qué participantes se les adjudicó el programa, con qué rapidez se amplió este último, y cómo se están gastando los recursos. Esta información es fundamental para implementar la evaluación, por ejemplo, para asegurar que los datos de *línea de base* se recopilen antes de que se introduzca el programa en la muestra de la evaluación y para verificar la integridad de los grupos de tratamiento y comparación. El monitoreo es esencial para verificar si un beneficiario realmente participa en el programa y para que no intervengan los no beneficiarios. Además, los datos administrativos pueden proporcionar información sobre el costo de implementación del programa, lo cual también es necesario para los análisis de costo-beneficio y costo-efectividad (Gertler et al., 2017).

Khandker et al. (2017) sugieren que los desafíos para monitorear el progreso de una intervención son:

1. *Identifique los objetivos* para los que está diseñado el programa o la estrategia sobre cómo incrementar la actividad de la innovación, reducir la pobreza o mejorar la escolarización de las niñas.
2. *Identifique indicadores clave* que se puedan usar para monitorear el progreso en relación con estos objetivos.
3. *Establecer objetivos*, que cuantifican el nivel de los indicadores que deben alcanzarse a una fecha dada.

4. *Establecer un sistema de monitoreo* para seguir el progreso hacia el logro de objetivos específicos e informar a los responsables políticos. Tal sistema fomentará una mejor gestión y rendición de cuentas de proyectos y programas.

Un ejemplo de los **puntos 1, 2, 3** anterior sería: *La Agenda 2030 y los Objetivos de Desarrollo Sostenible una oportunidad para América Latina y el Caribe (ONU CEPAL, 2015)*, donde se establecen **17** objetivos, metas e indicadores para alcanzarlos. Existen casos, como la innovación que son planteados a alcanzarse a través de varios de ellos. Ver **Tabla 1.1**.

Otro ejemplo de los mismos **puntos 1, 2, 3** es el *Global Innovation Index (GII, 2019: 371)*, el cual consta de **80** indicadores en tres categorías:

- a. **57** indicadores cuantitativos
- b. **18** indicadores compuestos
- c. **5** indicadores cualitativos

Ambos modelos, tienen oportunidad de generar el planteamiento del **punto 4** sobre el sistema de monitoreo.

Tabla 1.1. La innovación en los objetivos de la ONU CEPAL(2015)

Objetivo	Meta	Indicadores
8. Promover el crecimiento económico sostenido, inclusivo y sostenible, el empleo pleno y productivo y el trabajo decente para todos	8.2 Lograr niveles más elevados de productividad económica mediante la diversificación, la modernización tecnológica y la innovación , entre otras cosas centrándose en los sectores con gran valor añadido y un uso intensivo de la mano de obra	8.2.1 Tasa de crecimiento anual del PIB real por persona empleada
	8.3 Promover políticas orientadas al desarrollo que apoyen las actividades productivas, la creación de puestos de trabajo decentes, el emprendimiento, la creatividad y la innovación, y fomentar la formalización y el crecimiento de las microempresas y las pequeñas y medianas empresas, incluso mediante el acceso a servicios financieros	8.3.1 Proporción de empleo informal en el sector no agrícola, desglosada por sexo
9. Construir infraestructuras resilientes, promover la industrialización inclusiva y sostenible y fomentar la innovación	9.5. Aumentar la investigación científica y mejorar la capacidad tecnológica de los sectores industriales de todos los países, en particular los países en desarrollo, entre otras cosas fomentando la innovación y aumentando considerablemente, de aquí a 2030, el número de personas que trabajan en investigación y desarrollo por millón de habitantes y los gastos de los sectores público y privado en investigación y desarrollo	9.5.1 Gastos en investigación y desarrollo en proporción al PIB 9.5.2 Número de investigadores (en equivalente a tiempo completo) por cada millón de habitantes
	9.b Apoyar el desarrollo de tecnologías, la investigación y la innovación nacionales en los países en desarrollo, incluso garantizando un entorno normativo propicio a la diversificación industrial y la adición de valor a los productos básicos, entre otras cosas	9.b.1 Proporción del valor añadido por la industria de tecnología mediana y alta en el valor añadido total

17 Alianzas para lograr los objetivos	17.6 Mejorar la cooperación regional e internacional Norte-Sur, Sur-Sur y triangular en materia de ciencia, tecnología e innovación y el acceso a estas, y aumentar el intercambio de conocimientos en condiciones mutuamente convenidas, incluso mejorando la coordinación entre los mecanismos existentes, en particular a nivel de las Naciones Unidas, y mediante un mecanismo mundial de facilitación de la tecnología	17.6.1 Número de acuerdos y programas de cooperación en materia de ciencia o tecnología suscritos por los países, desglosado por tipo de cooperación 17.6.2 Número de abonados a Internet de banda ancha fija por cada 100 habitantes, desglosado por velocidad
	17.8 Poner en pleno funcionamiento, a más tardar en 2017, el banco de tecnología y el mecanismo de apoyo a la creación de capacidad en materia de ciencia, tecnología e innovación para los países menos adelantados y aumentar la utilización de tecnologías instrumentales, en particular la tecnología de la información y las comunicaciones	17.8.1 Proporción de personas que utilizan Internet

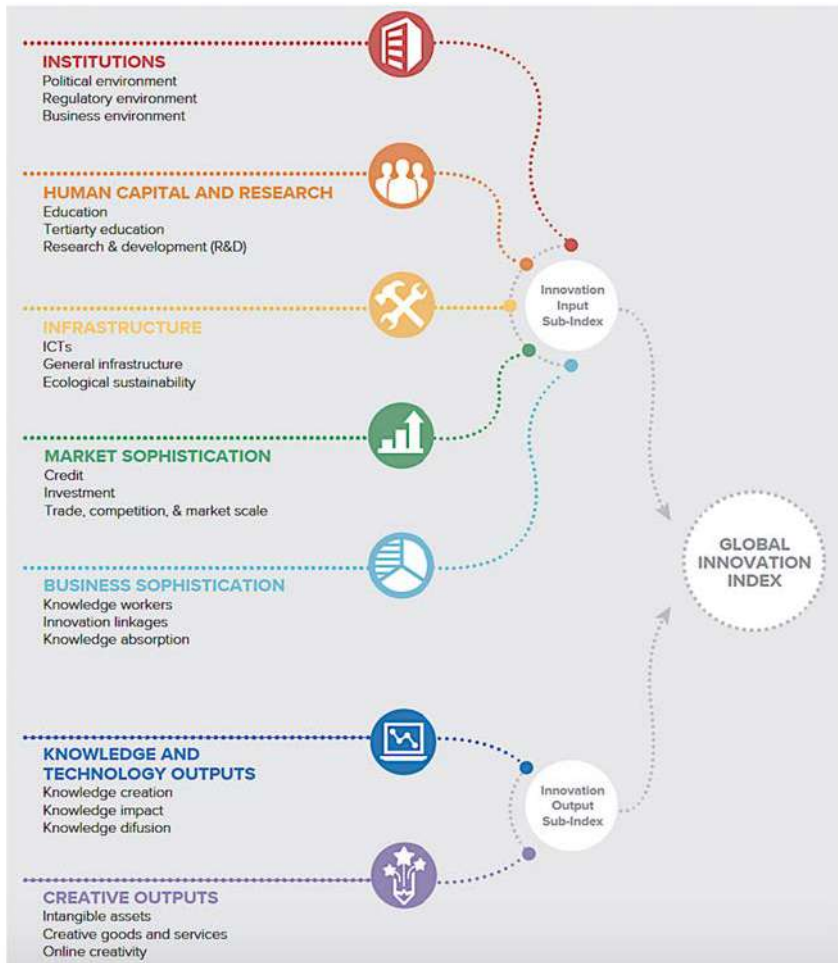
Fuente: ONU CEPAL (2015).

Los indicadores dentro de un marco de M&E

Los indicadores, generalmente se agrupan en factores a determinar:

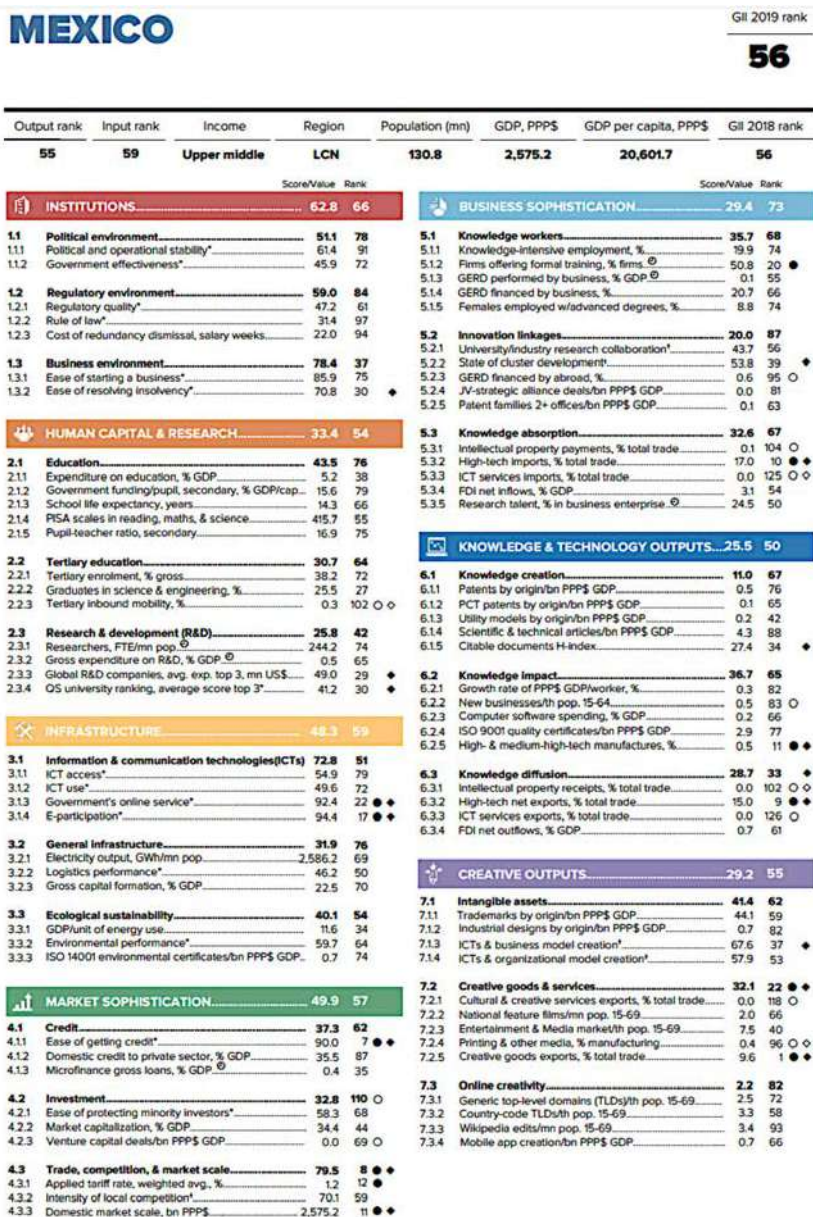
- Primero, los *factores finales* que miden los resultados (*outcomes*), como pro ejemplo, el Global Innovation Index (**GII, 2019**). Ver **Figura 1.1**

Figura 1.1. Modelo Global Innovation Index



Fuente: GII (2019).

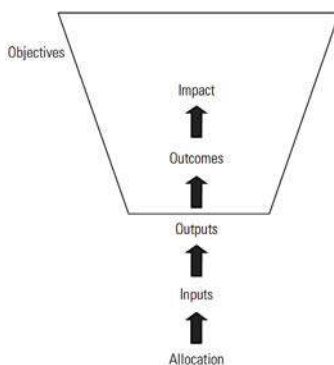
Figura 1.2. México en el Global innovation Index (2019)



NOTES: ● Indicates a strength; ○ a weakness; ● an income group strength; ○ an income group weakness; * an index; † a survey question. ○ indicates that the economy's data are older than the base year; see Appendix II for details, including the year of the data, at <http://globalinnovationindex.org>. Square brackets [] indicate that the data minimum coverage (DMC) requirements were not met at the sub-pillar or pillar level.

- Segundo, determinar los *factores intermedios* que miden las entradas (*inputs*) como el *innovation input sub-index* compuesto de **5 variables y 15 dimensiones**, o salidas (*outputs*) como *innovation output sub-index*, compuesto de **2 variables y 6 dimensiones** con del **GII (2019)**. Ver **Figura 1.2.** los mostrados como **x.y.**
- Los *indicadores objetivo*, que en el caso del modelo de la Global Innovation Index (GII, 2019) se encuentran los **80 indicadores** descritos líneas arriba. Ver **Figura 1.2.** los mostrados como **x.y.z.**
- Se puede apreciar, básicamente que un modelo como el **GII (2019)** o la **ONU (2015)**, pueden ser representados en cuatro grupos, como se presenta en la **Figura 1.3.** Este llamado marco lógico explica las entradas (*inputs*), salidas (*outputs*), resultados (*outcomes*) e impactos (*impacts*) en el sistema de **M&E**. La *evaluación de impacto*, que es el enfoque de esta obra, abarca las últimas etapas del marco de **M&E**.

Figura 1.3. Modelo de monitoreo y evaluación de impacto



Fuente: Khandker et al. (2017).

Como se observa, el *monitoreo* cubre tanto la *implementación* como el *monitoreo de rendimiento basado en resultados* (llamado también *results-based monitoring*). Los *indicadores* (variables) *intermedios* suelen variar más rápido que los *indicadores* (factores) *finales*, responden

más rápidamente a las intervenciones públicas y pueden ser medidos más fácilmente y de manera más oportuna. Seleccionar los *indicadores* (variables) para el *monitoreo* vs. *objetivos* y *metas* puede estar sujeto a limitaciones de recursos a los que se enfrenta el gestor del proyecto. Sin embargo, *es recomendable seleccionar solo unos pocos indicadores que puedan ser monitoreados adecuadamente en lugar de una gran cantidad de indicadores que no se puedan medir bien* (Khandker et al., 2017).

Monitoreo basado en resultados

La ejecución real de un sistema de monitoreo, a menudo se conoce como *monitoreo basado en resultados* (*results-based monitoring*). Kusek y Rist (2004), describen 10 pasos para el *monitoreo basado en resultados* (*outcomes*) como parte de un marco de M&E, como sigue:

1. Primero, *se debe realizar una evaluación de preparación*. La evaluación implica comprender las necesidades y características del área o región a ser objetivo, así como los actores clave (por ejemplo, el gobierno nacional o local y los donantes) que serán responsables de la implementación del programa. Es importante determinar, cómo responderá el esfuerzo a las presiones negativas y la información generada por el proceso de M&E.
2. En segundo lugar, los evaluadores del programa deben ponerse de acuerdo sobre *resultados* (*outcomes*) para *monitorear* y *evaluar*, así como *indicadores clave* de *desempeño* para *monitorear resultados* (*outcomes*). Hacerlo, implica la colaboración con los gobiernos y las comunidades receptoras, para llegar a un conjunto de *metas* y *objetivos* mutuamente acordados para el programa. Se establece la importancia de los resultados (*outcomes*) y los problemas relacionados para monitorearlos.
3. Tercero, *los evaluadores deben decidir los indicadores clave de desempeño* (KPI. *Key Performance Indicators*) *de cómo se medirán*

las tendencias en estos resultados (outcomes). Por ejemplo, de la **GII (2019)** sobre la *variable infraestructura* fuera un resultado (*outcomes*) importante para un programa, ¿la dimensión *acceso a internet* se medirá por la proporción de conexiones, puntajes de satisfacción de servicio, asistencia técnica u otra métrica? Las evaluaciones *cualitativas* y *cuantitativas* pueden ser realizadas para abordar este problema, como se discutirá más adelante en este capítulo. Los costos de la medición, también guiarán este proceso. Se establecen pros y contras del uso de indicadores prediseñados y su construcción.

4. Cuarto, *deben determinarse los instrumentos para recopilar información.* La línea base o los datos del preprograma, pueden ser muy útiles para evaluar el impacto del programa, ya sea utilizando los datos para *predecir* los resultados (*outcomes*) que podrían resultar del programa (como en las *evaluaciones ex ante*) o haciendo comparaciones de antes y después (también llamadas comparaciones reflexivas o *reflexive comparisons*). Los gerentes de programa, también pueden participar en discusiones frecuentes con los miembros del personal y comunidades seleccionadas. Incluye establecer las bases para datos en indicadores así como información; identificar fuentes de datos para los indicadores; diseño y comparación de métodos de colección de datos; la importancia de los indicadores pilotos.
5. Quinto, *se deben establecer objetivos de planeación de mejora; selección de resultados por objetivos.* Estos objetivos también se pueden utilizar para monitorear los resultados (*outcomes*). Este esfuerzo incluye establecer *objetivos periódicos a lo largo del tiempo* (por ejemplo, anualmente o cada dos años). Teniendo en cuenta la duración de los posibles efectos del programa, así como otros factores importantes, que pueden afectar la implementación del programa (como consideraciones políticas). El monitoreo de estos objetivos, en particular, representa el *sexto paso* en este

marco basado en resultados (outcomes) e implica la recopilación de datos de buena calidad. Incluye la definición de objetivos; factores a considerar cuando se selecciona objetivos de desempeño de indicadores;

6. Sexto. *Monitoreo de los resultados (outcomes)*. Se establecen tipos y niveles de monitoreo; principios clave en la construcción de un sistema de monitoreo; determinación del triángulo de la calidad de los datos: confiabilidad, validez y línea de tiempo; análisis del desempeño de los datos; pruebas previas de la colección de datos, instrumentos y procedimientos.
7. El séptimo paso, se relaciona con *usar la información de la evaluación para soportar los sistemas de administración basados en resultados (results-based management systems)* al momento del monitoreo, reconociendo que desde una perspectiva de la administración, el momento y la organización de las evaluaciones también impulsan el alcance a lo que las evaluaciones pueden ayudar a guiar las políticas. Si se encuentra que los indicadores reales son *divergentes* rápidamente de los objetivos iniciales, por ejemplo, evaluaciones realizadas en ese momento pueden ayudar a los administradores de programas a decidir rápidamente, si la implementación del programa u otros factores relacionados necesitan ser ajustados.
8. El octavo paso, *implica una cuidadosa consideración de los medios de presentación de los informes*, que incluyen el público al que se presentarán los resultados (*outcomes*). Incluye usos de la evaluación; el *timing* de las evaluaciones; los tipos y características de la calidad de las evaluaciones.
9. El noveno paso, implica el *uso de los hallazgos*; usar los resultados (*outcomes*) para crear vías para la retroalimentación (como la entrada de información independiente agencias, autoridades locales y comunidades específicas y no focalizadas). Tal retroalimentación puede ayudar a los evaluadores a aprender y actualizar

las reglas y procedimientos del programa para mejorar los resultados (*outcomes*). Incluye los usos de los hallazgos del monitoreo y la evaluación; conocer a la audiencia objetivo; presentación de los datos de desempeño en formas claras y entendibles. Incluye el uso de hallazgos de desempeño; beneficios adicionales del uso de hallazgos: retroalimentación, conocimientos y aprendizaje; estrategias para compartir información.

10. *Soportando el M&E dentro de la organización.* Los sistemas eficaces de **M&E** son duraderos y se basan, entre otras cosas, en la demanda continua (una función de incentivos para continuar el programa, así como el valor de la información creíble); transparencia y responsabilidad en procedimientos de evaluación; gestión efectiva de presupuestos; y responsabilidades bien definidas entre los miembros del personal del programa. Incluye la importancia de los incentivos; posibles problemas en los sistemas **M&E**; validación y evaluación de los sistemas **M&E** e información.

Desafíos en la configuración de un sistema de monitoreo

Los principales desafíos para un monitoreo efectivo, incluyen la *variación potencial en el programa implementación* debido a la falta de *habilidades* entre los funcionarios del programa, así como la *ambigüedad en los indicadores finales a evaluar*. En la mayoría de los programas de implementación de innovaciones y basados en el modelo marco de **M&E**, se han tenido uno o varios de los siguientes problemas:

- Los indicadores clave de rendimiento (**KPI**) no estaban bien definidos y, por lo tanto, no se capturaron exhaustivamente.
- Recursos humanos limitados disponibles para recopilar y registrar la información.

- El personal encargado de las actividades **M&E** con un nivel escaso de habilidades y capacidades limitadas; sus roles y responsabilidades no estaban bien definidos en los niveles de campo y oficina central.
- Carencia de herramientas y software sofisticados para analizar la información recopilada.

Las debilidades en estas áreas deben abordarse mediante diferentes enfoques. Por ejemplo, los **KPI** se deben definir con mayor precisión mediante:

- a. Una mejor comprensión las entradas y salidas en la etapa del proyecto,
- b. Especificando el nivel y la unidad de medición para indicadores,
- c. Con frecuencia, recolectando niveles comunes y datos adecuados para proporcionar actualizaciones periódicas sobre cómo evolucionan los resultados (*outcomes*) intermedios y si es necesario revisar los indicadores,
- d. Identificar claramente a las personas y entidades responsables del seguimiento.
- e. Para la recopilación de datos en particular, considerar: el *timing* de la encuesta (desde una *línea de base* previa al proyecto, por ejemplo, hasta el período actual); frecuencia (mensual o semestral, por ejemplo); instrumentos (como entrevistas o encuestas); y el nivel de recaudación (individual, hogar, comunidad o una administración más amplia unidades como el distrito) deben definirse y establecerse explícitamente dentro del marco de **M&E**.
- f. Proporcionar al personal, *capacitación y herramientas* para la recopilación y análisis de datos, así como la verificación en diferentes niveles de la estructura de monitoreo.

Los formuladores de políticas de implementación de la innovación, también necesitan establecer cómo impacta el programa a diversos niveles (sectores de la industria, pyme, mipyme) y tendencias a nivel de país como aumento del comercio, la inflación, el acceso a las tecnologías de información, etc. Una cuestión relacionada es la *heterogeneidad* en los impactos del programa en un grupo objetivo. Los efectos de un programa, por ejemplo, pueden variar durante su vida útil esperada. Las entradas (*inputs*) relevantes que afectan los resultados (*outcomes*) pueden también cambiar en este horizonte; por lo tanto, monitorear los resultados (*outcomes*) a largo y a corto plazo puede ser de interés para los responsables técnicos y políticos. Además, aunque los resultados (*outcomes*) de un programa de implementación de la innovación, a menudo se distinguen simplemente en áreas específicas y no específicas, monitoreando la variación en la implementación del programa (medidas de calidad, por ejemplo) puede ser extremadamente útil para comprender los efectos del programa. Con todas estas preocupaciones, un *monitoreo cuidadoso de áreas específicas y no específicas* (ya sea a nivel regional, industrial, sectorial, por tamaño de empresa o nivel individual mipyme, por ejemplo) será de gran ayuda para medir los efectos del programa (**Khandker et al., 2017**).

Las evaluaciones de impacto

Las evaluaciones son valoraciones periódicas y objetivas de un proyecto, programa o política planificado, en curso o terminado. Se utilizan para responder a preguntas específicas relacionadas con el diseño, la implementación y los resultados. En contraste con el monitoreo, que es permanente, las evaluaciones se llevan a cabo en momentos concretos en el tiempo y a menudo requieren una perspectiva externa de los técnicos expertos. Su diseño, método y costo varían considerablemente en función

del tipo de pregunta que la evaluación intente responder. En términos generales, las evaluaciones pueden abordar tres tipos de preguntas (**Imas y Rist, 2009**):

- *Preguntas descriptivas*, que apuntan a lo que está ocurriendo. Se centran en los procesos, las condiciones, las relaciones organizacionales y las opiniones de las partes interesadas.
- *Preguntas normativas*, que comparan lo que ocurre con lo que debería ocurrir. Evalúan las actividades e investigan si los objetivos se cumplen o no. Estas preguntas pueden aplicarse a los insumos, las actividades y los productos.
- *Preguntas de causa y efecto*, que se centran en la atribución. Investigan qué diferencia produce la intervención en los resultados.

Hay numerosos tipos de evaluación y de métodos de evaluación, basados en datos cuantitativos y cualitativos. Los datos cualitativos no se expresan en números sino más bien mediante un lenguaje o, a veces, imágenes. Los datos cuantitativos son mediciones numéricas y habitualmente se asocian con escalas o métricas. Tanto los unos como los otros se pueden utilizar para responder al tipo de preguntas planteado más arriba. En la práctica, numerosas evaluaciones trabajan con ambos tipos de datos. Hay múltiples fuentes de datos que se pueden emplear en las evaluaciones, tanto datos primarios recopilados para el objetivo de la evaluación como los *datos secundarios disponibles*.

Las *evaluaciones de impacto* son más completas si se sirven de datos cuantitativos pero subrayan el valor del monitoreo, de los métodos de evaluación complementarios y del uso tanto de datos cuantitativos como cualitativos. Las *evaluaciones de impacto* constituyen un tipo particular de evaluación que pretende responder a una pregunta específica de *causa y efecto*: *¿cuál es el impacto (o efecto causal) de un programa en un resultado de interés?* Esta pregunta básica incorpora una *dimensión causal* importante. Se centra únicamente en el impacto, es decir, en

los cambios directamente atribuibles a *un programa, una modalidad de programa o una innovación de diseño*.

La pregunta básica de la evaluación: *¿cuál es el impacto o efecto causal de un programa en un resultado de interés?* se puede aplicar en numerosos contextos. Por ejemplo, ¿cuál es el *efecto causal* de los estímulos a la innovación en la asistencia a pymes y sus logros de competitividad?, ¿cuál es el *impacto* en el acceso a la atención sanitaria de contratar la servicios de innovación en atención primaria con proveedores privados? Si la elaboración de agua de coco es reemplazado por un proceso innovador que aporta mayor productividad con menor costo, ¿cuál será el *impacto* económico en una comunidad cooperativista?, ¿la mejora de los caminos con tecnologías innovadoras aumenta el acceso a los mercados laborales e incrementa el ingreso de los hogares y, en caso afirmativo, en qué medida?, ¿influye el acceso de la clase por medios de internet, en los logros de los alumnos y, en caso afirmativo, en qué medida? Como muestran estos ejemplos, la pregunta de la evaluación básica se puede ampliar para analizar el impacto de una modalidad de programa o innovación de diseño, no solo de un programa.

El sello distintivo de las *evaluaciones de impacto* es centrarse en la *causalidad y la atribución*. Todos los métodos de evaluación de impacto plantean alguna forma de pregunta de causa y efecto. El enfoque para abordar la *causalidad* determina las metodologías que se pueden utilizar. Para estimar el *efecto causal* o el impacto de un programa en los resultados, cualquier método de *evaluación de impacto* elegido debe estimar el llamado *contrafactual*, es decir: *cuál habría sido el resultado de los participantes del programa si no hubieran participado en el mismo*. En la práctica, la *evaluación de impacto* requiere que el equipo de evaluación encuentre un *grupo de comparación* para estimar qué les habría ocurrido a los participantes del programa sin el programa, y posteriormente, efectuar comparaciones con el grupo de tratamiento que ha sido objeto del programa.

La elección de un método de *evaluación de impacto* depende de las características operativas del programa que se evalúa. Cuando las reglas de operación del programa son equitativas y transparentes y contemplan la rendición de cuentas, siempre se podrá encontrar un buen diseño de *evaluación de impacto*, ya sea que se planifique al comienzo, o durante el proceso de diseño o de implementación de un programa. El contar con *reglas de operación* claras y bien definidas para un programa no solo tiene un valor intrínseco en las políticas de implementación de innovaciones y en una gestión solvente de los programas: también es esencial para construir buenos grupos de comparación, lo cual constituye la base de las evaluaciones de impacto rigurosas. Concretamente, la elección de un método de *evaluación de impacto* está determinada por las características operativas del programa, en particular sus *recursos disponibles*, los *criterios de elegibilidad para seleccionar a los beneficiarios* y los *plazos para la implementación del programa*.

Así, se pueden formular tres preguntas acerca del *contexto operativo* de un determinado programa:

1. ¿El programa tiene recursos para servir a todos los beneficiarios elegibles?
2. ¿El programa está focalizado o es universal?
3. ¿El programa se ofrecerá a todos los beneficiarios de una sola vez o de manera secuencial?

La respuesta a estas tres preguntas determinará cuál de los métodos presentados *asignación aleatoria*, *variables instrumentales*, *regresión discontinua*, *diferencias en diferencias* o *pareamiento*, es el más adecuado para un determinado contexto operativo.

Estudios de eficacia y estudios de efectividad

Gertler et al. (2017) confirman que la función principal de la *evaluación de impacto* consiste en *producir evidencia* sobre el desempeño de un programa a fin de que sea utilizada por los *funcionarios públicos, los administradores del programa, la sociedad civil y otros actores relevantes*. Los resultados de las *evaluaciones de impacto* son particularmente útiles cuando las conclusiones se pueden aplicar a una población de interés más amplia. La cuestión de la generalización es clave para los responsables de las políticas, puesto que determina si los resultados identificados en la evaluación *pueden replicarse en grupos ajenos a los que han sido estudiados en la evaluación si aumenta la escala del programa*. En los primeros tiempos de las *evaluaciones de impacto* de los programas de desarrollo, se tenía que:

1. Una gran parte de la evidencia se basaba en *estudios de eficacia*, es decir, pruebas llevadas a cabo en un entorno específico en condiciones rigurosamente controladas para asegurar la consistencia entre el diseño de la evaluación y la implementación del programa. Dado que los *estudios de eficacia* suelen realizarse como *experiencias piloto* con una amplia participación técnica de los investigadores mientras el programa se está implementando, es posible que sus resultados, a menudo de pequeña escala, *no ofrezcan necesariamente mucha información acerca del impacto de un proyecto similar implementado a mayor escala en circunstancias normales*. Los *estudios de eficacia* analizan la *prueba de concepto*, a menudo para sondear la viabilidad de un nuevo programa o una teoría específica del cambio. Si el programa no genera impactos anticipados bajo estas condiciones cuidadosamente manejadas, es poco probable que funcione si se despliega en circunstancias normales. Por ejemplo, una *intervención piloto* que introduce nuevos protocolos de implementación de innovación tecnológica, puede funcionar en una empresa de software con excelentes administra-

dores y equipo de ingeniería, pero puede que la misma intervención no funcione en una empresa promedio con administradores menos esmerados y limitaciones de personal. Además, los cálculos de costo beneficio variarán, dado que los pequeños estudios de eficacia quizá no capturen los costos fijos ni las economías de escala. Como consecuencia, si bien la evidencia de los *estudios de eficacia* puede ser útil para probar un *enfoque innovador*, los resultados a menudo tienen una capacidad de *generalización limitada* y no siempre representan adecuadamente entornos más generales, que suelen ser la principal preocupación de los responsables de las *políticas de implementación de innovaciones*.

2. *Los estudios de efectividad* proporcionan evidencia a partir de las intervenciones que tienen lugar en *circunstancias normales*, utilizando vías de implementación regulares y con el objeto de *producir conclusiones que se pueden generalizar para una población grande*. Cuando las *evaluaciones de efectividad* están adecuadamente diseñadas e implementadas, los resultados pueden ser *generalizables* para los beneficiarios previstos fuera de la muestra de la evaluación, siempre y cuando la ampliación utilice las mismas estructuras de implementación y llegue a poblaciones similares a la de la muestra de la evaluación. Esta *validez externa* tiene una importancia crítica para los responsables de las políticas de implementación de innovaciones, porque les permite utilizar los resultados de la evaluación para fundamentar decisiones que afectan a todo el programa y que se aplican a los beneficiarios previstos más allá de la muestra de la evaluación.

Evaluación operativa y sus desafíos

Una evaluación operativa, busca comprender si la implementación de un programa se ha desplegado, según lo planeado. Específicamente, la evaluación operativa es una evaluación retrospectiva basado en los objetivos iniciales del proyecto, indicadores y metas del marco de **M&E (Khandker et al., 2017)**.

La evaluación operativa, puede basarse en entrevistas con los beneficiarios del programa y con funcionarios responsables de la implementación. El objetivo es comparar lo planeado con lo que realmente se entregó, para determinar si hay *brechas* entre lo planeado y resultados (*outcomes*) realizados, e identificar las lecciones que se aprenderán para el diseño de proyectos futuros e implementación.

Debido a que la evaluación operativa se relaciona en cómo son implementados los programas en última instancia, diseñar medidas apropiadas de calidad para lograrlo, es un desafío muy importante. Estos esfuerzos, incluyen monitorear cómo se gastó o asignó el dinero del proyecto en todos los sectores (en comparación con el objetivo), así como posibles efectos indirectos del programa, en áreas no consideradas inicialmente. Recopilar datos precisos sobre estos factores, puede ser difícil, pero es esencial para determinar el potencial de sesgos en la medición de los impactos del programa.

Evaluación operativa vs. evaluación de impacto

La justificación de un programa, para *atraer recursos públicos*, es mejorar un resultado (*outcomes*) seleccionado sobre lo que hubiera sido sin el programa. El principal problema de un evaluador, es medir el impacto o los efectos de una intervención y reportarlo, a fin de que los encargados de formular las políticas de implementación tengan los elementos suficientes de decisión, sobre si la intervención del programa, merece la pena continuarse, expandirse o extinguirse (**Khandker et al., 2017**).

La *evaluación operativa* se relaciona con asegurar la *implementación efectiva* de un programa de acuerdo con los objetivos iniciales del programa. La *evaluación de impacto* es un esfuerzo para comprender si los cambios esperados orientados al bienestar, se deben realmente a un proyecto o programa intervención. Específicamente, la *evaluación de impacto* intenta determinar si es posible identificar el efecto del programa y hasta qué punto el efecto medido se puede atribuir al programa y no a otras causas.

Como se sugiere en la **Figura 1.3**, la *evaluación de impacto* se basa en las últimas etapas del marco teórico del modelo **M&E**, que se centra en los resultados (*outcomes*) e impactos (*impacts*). La evaluación tanto operativa como de impacto, se consideran *complementarias* más que sustitutas. Una *evaluación operativa* debe ser parte del procedimiento normal dentro del agencia ejecutora. Pero la plantilla utilizada para una *evaluación operativa* puede ser muy útil para una *evaluación de impacto* más rigurosa. Realmente se necesita saber el contexto dentro del cual se generaron los datos y hacia dónde se dirigió el esfuerzo político. También la información generada a través de las oficinas de implementación del proyecto, es esencial para una *evaluación operativa* así como la interpretación de los resultados del impacto. Sin embargo, aunque la evaluación operativa y la práctica general de **M&E** son partes integrales de la implementación del proyecto, la evaluación de impacto no es imprescindible para cada proyecto ya que requiere mucho tiempo y recursos y, por lo tanto, debe ser aplicado selectivamente. Los responsables de alto nivel, pueden decidir si llevar a cabo una *evaluación de impacto*, sobre la base de los siguientes criterios:

- La intervención del programa, es innovador y de importancia estratégica.
- El ejercicio de la *evaluación de impacto*, contribuye a descubrir la brecha de conocimiento, de lo que funciona y lo que no. (La dispo-

nibilidad y la calidad de los datos son requisitos fundamentales para este ejercicio).

Evaluación de impacto cualitativo vs. cuantitativo

Los gobiernos, los donantes y otros profesionales de la comunidad de desarrollo están interesados para determinar la efectividad de los programas con objetivos de largo alcance como la introducción de innovaciones, la reducción de la pobreza o el aumento del empleo. Estas misiones de política, a menudo son posibles solo a través de *evaluaciones de impacto* basadas en evidencia sólida de datos de encuestas o a través de enfoques cuantitativos relacionados. Esta obra, se enfoca en los *métodos de evaluaciones de impacto cuantitativo* más que *cualitativos*.

La información cualitativa, como la comprensión del contexto local, sociocultural e institucional, así como los detalles del programa y de los participantes, es, sin embargo, esencial para una *evaluación cuantitativa sólida*. Por ejemplo, la información cualitativa puede ayudar a identificar mecanismos a través de los cuales los programas podrían tener impacto; dichas encuestas también pueden identificar a los responsables políticos locales o individuos que sean importantes para determinar el curso de cómo se implementan los programas, por lo tanto, ayudando a la *evaluación operacional*. Pero una *evaluación cualitativa* por sí sola, *no puede evaluar resultados (outcomes) contra alternativas relevantes o resultados (outcomes) contrafactuales*. Es decir, no puede realmente indicar lo que podría suceder en ausencia del programa. El *análisis cuantitativo* también es importante para abordar posibles sesgos estadísticas en los *impactos* del programa. Una mezcla de métodos cualitativos y cuantitativos (enfoque de métodos mixto) por lo tanto, podría ser útil para obtener una visión integral de la efectividad del programa.

Los enfoques de *métodos mixtos*, al combinar datos cuantitativos y cualitativos, constituyen un complemento clave en las *evaluaciones de impacto* que se basan únicamente en el uso de datos *cuantitativos*

(Gertler et al., 2017) sobre todo, para contribuir a generar hipótesis y enfocar las preguntas de la investigación antes de recopilar los datos cuantitativos, así como para presentar perspectivas y visiones novedosas del desempeño de un programa durante y después de su implementación. Hay numerosos *métodos cualitativos*, que componen su propio ámbito de investigación. Los métodos que generan datos *cualitativos* suelen basarse en enfoques abiertos, que no dependen de las respuestas predeterminadas de las personas entrevistadas. Los datos se generan a través de una gama de enfoques, incluidos grupos focales, historiales y entrevistas con beneficiarios seleccionados y otros informantes clave (Rao y Woolcock, 2003). También pueden incluir una gama de evaluaciones observacionales y etnográficas. A pesar de que las observaciones, ideas y opiniones recopiladas durante el trabajo cualitativo no suelen ser estadísticamente representativas de los beneficiarios del programa y, por lo tanto, no son generalizables, resultan útiles para entender por qué se han alcanzado o no ciertos resultados. Las evaluaciones que integran el análisis cuantitativo y cualitativo se caracterizan por utilizar *métodos mixtos* (Bamberger et al., 2010). En el desarrollo de un enfoque de *método mixto*, Creswell (2014) define tres aproximaciones básicas:

- a. *Convergente paralelo*. Se recopilan simultáneamente datos cuantitativos y cualitativos y se utilizan para triangular los hallazgos o para generar los primeros resultados sobre cómo se está implementando el programa y cómo lo perciben los beneficiarios.
- b. *Explicativo secuencial*. Los datos cualitativos proporcionan contexto y explicaciones para los resultados cuantitativos, para explorar casos *atípicos* de éxito y fracaso, y para desarrollar explicaciones sistemáticas del desempeño del programa, como se constató en los resultados cuantitativos. De esta manera, el trabajo cualitativo puede contribuir a determinar por qué en el análisis cuantitativo se observan ciertos resultados, y se pueden usar para entrar en

la *caja negra* de lo que ocurrió en el programa (**Bamberger et al., 2010**).

- c. *Exploratorio secuencial*. El equipo de evaluación puede utilizar grupos focales, listas, entrevistas con informantes clave y otros enfoques cualitativos para desarrollar hipótesis a propósito de cómo y por qué el programa funcionaría, y para clarificar preguntas acerca de la investigación que hay que abordar en el trabajo cuantitativo de evaluación de impacto, lo que incluye las alternativas más relevantes del diseño de programas que deben ser probadas a través de la evaluación de impacto.

Evaluación de impacto cuantitativo: ex post vs ex ante

Hay dos tipos de evaluaciones de impacto cuantitativas: *ex post* y *ex ante*. Una *evaluación de impacto ex ante* intenta medir los impactos previstos de futuros programas y políticas, dada la situación actual de un área potencialmente focalizada, y pueden involucrar *simulaciones basadas en suposiciones* sobre cómo funciona cierto factor como la economía (**Bourguignon y Ferreira, 2003; Todd y Wolpin, 2006**). En muchas ocasiones:

1. *Las evaluaciones ex ante*, se basan en *modelos estructurales* del entorno económico que enfrentan los participantes potenciales. Los supuestos subyacentes de los *modelos estructurales*, por ejemplo, implican identificar los principales agentes de innovación, sociales, económicos, etc. en el desarrollo del programa (individuos, comunidades, gobiernos locales o nacionales, pymes, mipymes, etc.), así como los vínculos entre los agentes y los diferentes mercados para determinar los resultados (*outcomes*) del programa. Estos modelos predicen los impactos del programa. **Gertler et al. (2017)** afirman sobre *las evaluaciones ex ante* que son evaluaciones que utilizan datos disponibles para *simular* los efectos esperados bajo las políticas de implementación de innovaciones, en los resultados de

interés. Pueden ser muy útiles para medir la *efectividad* esperada relativa de una gama de opciones de diseño de programas alternativos en los resultados. Se trata de métodos habitualmente usados que dependen de la disponibilidad de datos de gran alcance y calidad que se pueden utilizar para aplicar modelos de simulación adecuados a la pregunta en cuestión. Al contrario de las *evaluaciones de impacto*, estos métodos se emplean para simular futuros efectos potenciales, más que para medir los impactos reales de los programas implementados. Este tipo de métodos puede ser sumamente útil para establecer referencias para los probables efectos del programa y para instituir objetivos realistas, así como para estimar costos, tasas de retorno y otros parámetros económicos. Se suelen utilizar como la base de los análisis económicos de los proyectos, especialmente antes de que se introduzca una reforma o se implemente un proyecto.

2. Las *evaluaciones ex post*, por el contrario, miden los impactos reales acumulados por los beneficiarios, que son *atribuibles a la intervención del programa*. Una forma de este tipo de evaluación se encuentra en el *modelo de efectos del tratamiento* (**Heckman y Vytlačil, 2005**). Las *evaluaciones ex post* tienen *beneficios inmediatos y realidad reflejada*. Estas evaluaciones, sin embargo, en ocasiones carecen de los mecanismos subyacentes al impacto del programa en la población, del que los *modelos de análisis estructurales*, tienen como objetivo capturar y que pueden ser muy importantes para comprender la efectividad del programa (particularmente en entornos futuros). Las *evaluaciones ex post* también pueden ser mucho más costosas que las *evaluaciones ex ante* porque requieren recopilar datos sobre resultados para *grupos participantes y no participantes*, así como para otros factores sociales, políticos, técnicos, etc. que pueden haber determinado el curso de la intervención. Un costo adicional en la *configuración ex post* es el

fracaso de la intervención, que podría haber sido predicho a través del *análisis ex ante*.

3. Un enfoque propuesto a considerar, es combinar ambos análisis y *comparar las estimaciones ex post con las predicciones ante* (Ravallion, 2008) y hacerlos mixtos. Este enfoque puede ayudar a explicar cómo surgen los beneficios del programa, especialmente si el programa se lleva a cabo en diferentes fases y tiene la flexibilidad de ser refinada a partir del conocimiento adicional obtenido de la comparación. Sin embargo, un ejercicio de *impacto ex post* es más fácil de realizar si los investigadores tienen un *diseño ex ante de evaluación de impacto*. Es decir, se puede planificar un diseño, para una evaluación de impacto antes de implementar la intervención.

Evaluación de impacto: prospectiva vs. retrospectiva

De acuerdo con Gertler et al. (2017), las *evaluaciones de impacto* se pueden dividir en dos categorías: *prospectivas* y *retrospectivas*, describiéndose:

1. *Las evaluaciones prospectivas* se desarrollan simultáneamente con el *diseño del programa* y se incorporan en la implementación del mismo. Los datos de *línea de base* se recopilan antes de *implementar el programa*, tanto en el grupo que recibe la intervención (*denominado grupo de tratamiento*) como en el grupo utilizado como *comparación* y que *no es objeto de la intervención* (*denominado grupo de comparación*). Las *evaluaciones de impacto prospectivas* tienen más probabilidades de producir resultados solventes y creíbles, por tres motivos:
 - a. En primer lugar, se pueden recopilar *datos de línea de base* para establecer las medidas de los *resultados de interés* antes de que

el programa haya comenzado. Estos datos son importantes para *medir los resultados antes de la intervención*. Los datos de *línea de base* en los *grupos de tratamiento y comparación* se deben analizar para asegurar que los grupos sean *similares*. Las líneas de base también se pueden utilizar para evaluar la efectividad de la focalización, es decir, si un programa llega o no a sus beneficiarios previstos.

- b. En segundo lugar, definir la medida de éxito del programa en la etapa de planificación del mismo, centra tanto el programa como la evaluación en los resultados previstos. Como se verá, las *evaluaciones de impacto* se basan en la teoría del *cambio* de un programa o una cadena de resultados. El diseño de una evaluación de este tipo contribuye a clarificar los objetivos del programa, sobre todo porque requiere establecer medidas bien definidas de su éxito. Los responsables de las *políticas de implementación de programas de innovación*, deberían definir objetivos claros para el programa y formular preguntas claras que la evaluación debe contestar, para garantizar que los resultados sean relevantes para las políticas. En realidad, el pleno apoyo de los responsables de las políticas de implementación de programas de innovación, es un requisito necesario para el éxito de una evaluación; no se deberían emprender evaluaciones de impacto a menos que los responsables de dichas políticas, estén convencidos de la legitimidad de las mismas y de su valor para fundamentar decisiones clave de las políticas públicas.
- c. En tercer lugar, y lo que es aún más importante, en una *evaluación prospectiva* los grupos de tratamiento y comparación se definen antes de implementar la intervención que será evaluada. En resumen, las evaluaciones prospectivas son las que tienen más probabilidades de generar contrafactuales válidos. En la etapa de diseño, se pueden contemplar maneras alternativas de estimar

un contrafactual válido. El diseño de la evaluación de impacto también se puede alinear plenamente con las reglas operativas del programa, así como con el despliegue o el proceso de expansión de este último.

2. *Las evaluaciones retrospectivas*, evalúan el impacto del programa después de que se le haya implementado, y los *grupos de tratamiento y de comparación* se generan *ex post*. En las evaluaciones retrospectivas, el equipo que lleva a cabo la evaluación, a menudo tiene información tan limitada que resulta difícil analizar si el programa fue implementado con éxito y si sus participantes realmente se beneficiaron de él. Numerosos programas no recopilan datos de *línea de base* a menos que se haya incorporado la evaluación desde el principio, y una vez que el programa esté funcionando ya es demasiado tarde para hacerlo. Las *evaluaciones retrospectivas* que utilizan los datos existentes, son necesarias para evaluar los programas creados en el pasado. En estas situaciones, las opciones para obtener una estimación válida del *contrafactual* son mucho más limitadas. La evaluación depende de reglas claras de operación del programa en lo que respecta a la asignación de beneficios. También depende de la disponibilidad de datos con suficiente cobertura sobre los grupos de tratamiento y comparación, tanto antes como después de la implementación del programa. El resultado es que la viabilidad de una evaluación retrospectiva depende del contexto y nunca está garantizada. Incluso cuando son viables, las evaluaciones retrospectivas a menudo utilizan métodos cuasi experimentales y dependen de supuestos más fuertes y, por ende, pueden producir evidencia más discutible.

El problema de los contrafactuales

El principal desafío de una evaluación de impacto es determinar qué hubiera sucedido a los beneficiarios si el programa no hubiera existido (**Khandker et al., 2017**). Es decir, el investigador tiene que determinar por ejemplo, el ingreso por innovaciones de las empresas beneficiarias, en ausencia de la intervención de incentivos a la innovación. El resultado de un beneficiario, en ausencia de la intervención sería su *contrafactual*. Esto es, un programa o política de intervención busca *alterar* los cambios en el *bienestar* de los supuestos beneficiarios. En el análisis del impacto *cuantitativo ex post*, se observan los resultados (*outcomes*) de esta intervención en los beneficios previstos tales como el empleo o gastos. Así, cabe preguntarse: ¿este cambio se relaciona directamente con la intervención?, ¿esta intervención ha causado que crezcan los gastos o el empleo? No necesariamente. De hecho, con solo una observación puntual después del tratamiento, es *imposible* llegar a una conclusión sobre el impacto. En el mejor de los casos, se puede decir si el objetivo se cumplió en la intervención. *Pero el resultado (outcomes) después de la intervención no puede atribuirse a la programa en sí mismo.*

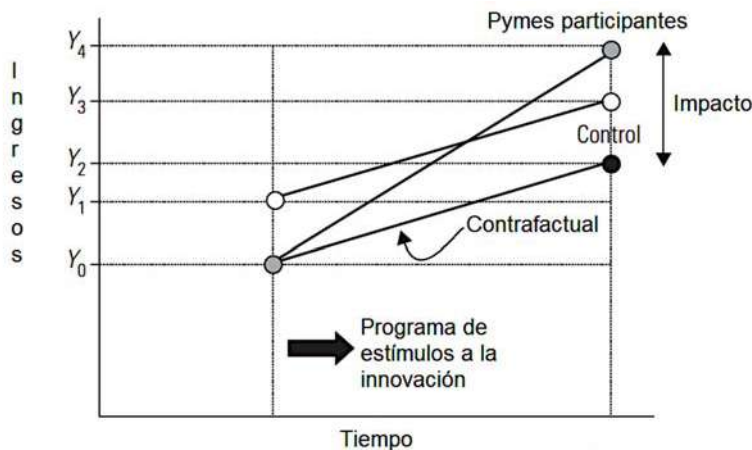
El problema de la evaluación, es que, si bien el impacto del programa (independiente de otros factores) puede evaluarse realmente solo comparando resultados (*outcomes*) reales y contrafactuales, *el contrafactual no se observa*. Entonces, el desafío de una evaluación de impacto es crear un grupo de comparación convincente y razonable para los beneficiarios a la luz de estos datos faltantes. Idealmente, el investigador desea comparar cómo la misma empresa, hogar o al individuo le hubiera ido bien y *sin intervención o tratamiento*. Pero no puede hacerlo porque en un momento dado, una empresa, un hogar o un individuo *no pueden tener dos existencias simultáneas*: una empresa, un hogar o un individuo no pueden estar en los grupos *con tratamiento y sin tratamiento* al mismo tiempo. Encontrar un contrafactual apropiado constituye el principal desafío de una evaluación de impacto. ¿Qué tal una comparación entre los *grupos*

tratados y no tratados, cuando ambos son elegibles para ser tratados? ¿Qué tal una comparación de los resultados (*outcomes*) de los *grupos tratados antes y después de que son tratados*? Estos grupos de comparación potenciales, pueden ser contrafactuales *falsificados*, como se discutirá en los ejemplos que siguen.

Búsqueda de contrafactuales con y sin comparaciones

Considere el caso de los beneficiarios, de un programa de gobierno sobre estímulos a la innovación para pymes que invierten en proyectos de investigación, desarrollo de tecnología e innovación dirigidos al desarrollo de nuevos productos, procesos o servicios, lo cual es considerado por las autoridades que brindan el apoyo económico como el camino para incentivar y promover el crecimiento y la competitividad de la empresa.

Los datos, sin embargo, muestran que la inversión per cápita (pymes) entre los *participantes* del programa *es menor* que el de los *no participantes* antes de la intervención del programa. ¿Es este un caso de quiebra potencial al programa? La respuesta es, no necesariamente. Normalmente, los gobiernos se dirigen a las pymes con ciertas desventajas, como la menor inversión per cápita en innovaciones para empezar, por lo que se considera que juzgar el impacto del programa, comparando las inversiones previas en innovación de los *participantes* del programa con los *no participantes*, es incorrecto. Lo que se necesita, *es comparar lo que habría sucedido con la inversión en innovación de las pymes participantes, de no haber existido el programa de intervención*. Así, es necesario determinar un grupo de comparación adecuado, como un *contrafactual* cercano a los beneficiarios del programa. La **Figura 1.4** da cuenta de lo anterior.

Figura 1.4. Evaluación de impacto *con y sin comparación*

Fuente: Elaboración propia.

Donde:

Y_4 Son los ingresos de las pymes *participantes* al programa de estímulos a la innovación, después de la intervención.

Y_3 Son los ingresos de las pymes *no participantes* o control de empresas.

Esta evaluación *con y sin comparación grupal*, mide el efecto del programa como $(Y_4 - Y_3)$. ¿Es esta medida una estimación correcta del efecto del programa? Sin saber ¿por qué algunas pymes participaron mientras que otras no, cuando las autoridades pusieron a disposición su programa de apoyo a la vista de todos los medios de publicidad, físicos y electrónicos posibles; esto lleva a pensar que tal comparación *podría ser engañosa*. Sin dicha información, un tercer observador no sabe si (Y_3) es el resultado (*outcomes*) *contrafactual correcto* para evaluar el efecto del programa. Por ejemplo, los ingresos son diferentes entre los participantes y los grupos de control antes del programa; este diferencial podría

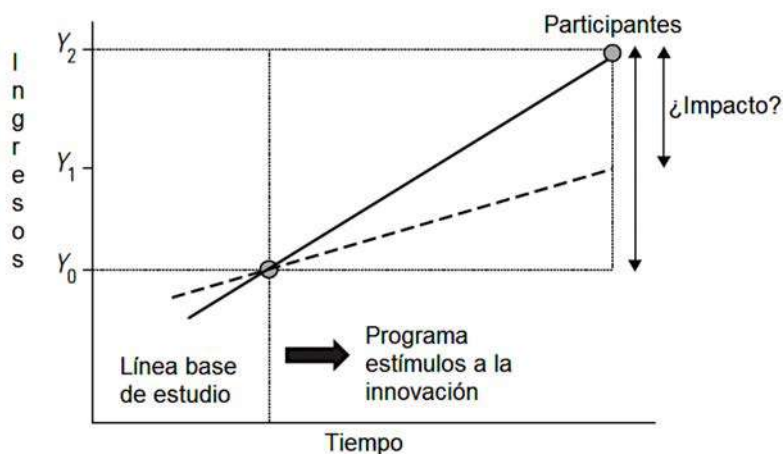
deberse a diferencias subyacentes que pueden sesgar la comparación a través de los dos grupos.

Si el investigador supiera los resultados (*outcomes*) contrafactuales (Y_0, Y_2), la estimación real del efecto del programa es $(Y_4 - Y_2)$, como lo indica la **Figura 1.4**, y no $(Y_4 - Y_3)$. En este ejemplo, el *falso contrafactual* produce una *subestimación del efecto del programa*. Tenga en cuenta, sin embargo, que dependiendo de las situaciones de pre-intervención de los *grupos tratados y grupos de control*, la *comparación del falso*, podría producir una *sobreestimación o subestimación* del efecto del programa.

Contrafactual antes y después de comparar

Otro *falso contrafactual* podría ser una comparación entre los resultados (*outcomes*) del *preprograma* y el *posprograma*, de los participantes. El investigador podría comparar los *resultados (outcomes) ex post* para las pymes beneficiarias, con datos sobre sus resultados (*outcomes*) *antes* de la intervención, ya sea con una encuesta de datos comparable antes de la introducción del programa o, en ausencia de un diseño de evaluación adecuado, con *datos retrospectivos*. Como se muestra en la **Figura 1.5**, se tienen dos puntos de observación para los beneficiarios pyme de la intervención: ingreso pre-intervención (Y_0) y el ingreso post-intervención (Y_2). En consecuencia, el efecto del programa podría estimarse como $(Y_2 - Y_0)$. La literatura (**Khandker et al., 2017**) se refiere a este enfoque, como el *método de impacto reflexivo*, donde los resultados (*outcomes*) de los participantes, antes de la intervención funcionan como resultados (*outcomes*) de comparación o control. ¿Este método ofrece una estimación realista del efecto del programa? Probablemente no.

**Figura 1.5. Evaluación de impacto
por comparación antes–después de**



Fuente: Elaboración propia.

La serie de tiempo, ciertamente hace que llegar a mejores conclusiones, sea más fácil, pero de ninguna manera es concluyente sobre el impacto de un programa. Observando la misma **Figura 1.5**, se aprecia, por ejemplo, que el impacto podría ser $(Y_2 - Y_1)$. De hecho, un método de diferencia tan simple no será una evaluación precisa porque hay muchos otros factores (fuera del programa) que pudieran haber cambiado durante el período. No controlar esos otros factores, significa que el investigador atribuiría falsamente el resultado (*outcomes*) del participante, en ausencia del programa como (Y_0) , cuando podría haber sido (Y_1) . Por ejemplo, los participantes en un programa de capacitación pueden tener mejores perspectivas de empleo después del programa. Aunque esta mejora puede ser debido al programa, también puede ser porque la economía se está recuperando de un pasado de crisis y el empleo están creciendo de nuevo. A menos que se hagan con cuidado, las comparaciones reflexivas existentes no pueden distinguir entre los efectos del programa y otros efectos externos, comprometiendo así la fiabilidad de los resultados (*outcomes*).

Las *comparaciones reflexivas* pueden ser útiles en evaluaciones de *intervenciones de cobertura total* tales como políticas y programas a nivel nacional en los que participa toda la población y *no hay margen para un grupo de control*. Incluso, *cuando el programa no sea tan amplio*, si se observan resultados (*outcomes*) para los participantes durante varios años, entonces los cambios estructurales en los resultados (*outcomes*) podrían ser probados (**Ravallion, 2008**).

En este contexto, por lo tanto, un estudio de *línea de base* amplio, que cubra múltiples características de preprogramas que consideren por ejemplo, la adopción de políticas de innovación, serían muy útiles para poder controlar tantos otros factores que pudieran estar cambiando con el tiempo. También, se necesitarían más datos detallados sobre la participación en programas existentes antes de que se implementara cualquier intervención.

Evaluación de impacto y su relación con el sesgo

Una *evaluación de impacto* es esencialmente un problema de falta de datos, porque el investigador no puede observar los resultados (*outcomes*) de los participantes del programa, si no hubieran sido beneficiarios. Sin información sobre el *contrafactual*, la siguiente mejor alternativa es *comparar los resultados (outcomes) de pymes o individuos tratados*, siguiendo nuestro ejemplo, con los de un grupo de comparación *que no ha sido tratado*. Al hacerlo, el investigador intenta elegir un grupo de comparación que sea muy similar al grupo tratado, de tal forma que, aquellos que recibieron tratamiento hubieran tenido resultados (*outcomes*) *similares a los del grupo de comparación en ausencia de tratamiento*.

Las *evaluaciones de impacto exitosas* dependen de encontrar un buen grupo de comparación. Allí se encuentran dos enfoques amplios a los que recurren los investigadores *para imitar el contrafactual de un grupo tratado* (**Khandker et al., 2017**):

- a. Crear un grupo de comparación a través de un diseño estadístico, o
- b. Modificar la estrategia de focalización del programa en sí mismo, para eliminar las diferencias que habrían existido entre los *grupos tratados y no tratados* antes de comparar resultados (*outcomes*) a través de los dos grupos.

La ecuación presenta el problema de evaluación básica que compara los resultados (*outcomes*) Y a través de individuos *tratados y no tratados* i :

$$Y_i = \alpha X_i + \beta T_i + \varepsilon_i$$

Donde:

T Es un *dummy* igual a (1) para los que participan y (0) para los que no participan.

X Es un conjunto de otras características observadas del individuo y quizás de sus características propias como pyme, hogar, individuo relacionadas a su entorno local.

ε Finalmente, es un término de error que refleja características no observadas que también afectan a Y .

La ecuación refleja, un enfoque comúnmente utilizado en las evaluaciones de impacto, el cual es *medir el efecto directo del programa* T sobre los resultados (*outcomes*) Y . Los *efectos indirectos* del programa (es decir, aquellos que no están directamente relacionados con participación) también pueden ser de interés, como los cambios en los precios dentro de las áreas del programa.

El problema con la estimación de la ecuación presentada, es que el tratamiento a la asignación frecuentemente, no aleatoria, debido a los siguientes factores:

- a. Colocación intencional del programa y
- b. Autoselección en el programa. Es decir, los programas se colocan de acuerdo con la necesidad de las comunidades empresariales,

sociales e individuos, quienes a su vez, seleccionan un programa dado de diseño y colocación. La *autoselección* podría basarse en las *características observadas*, *factores no observados*, o *ambos*. En el caso de *factores no observados*, el *término de error* en la ecuación de estimación, contendrá variables que también están correlacionadas con el *dummy de tratamiento*, T . El investigador puede medir, y por lo tanto explicar, estas *características no observadas* en la ecuación, que conduce a un *sesgo de selección no observado*. Es decir, $\text{cov}(T, \varepsilon) \neq 0$ implica la *violación* de uno de los supuestos clave de *ordinario mínimos cuadrados* en la obtención de *estimaciones imparciales*: independencia de los regresores de la término de perturbación ε . La correlación entre T y ε *sesga* naturalmente las otras estimaciones en la ecuación, incluida la estimación del efecto del programa β .

Este problema también se puede representar en un marco más conceptual. Suponga que el investigador, está evaluando un programa anti barrera de emprendimiento de innovación, como una *intervención de crédito*, destinado a aumentar ingresos de la pyme. Por ejemplo, suponga que (Y_i) represente el ingreso per cápita de la pyme i . Por *participantes*, $(T_i) = 1$, y el valor de (Y_i) *bajo tratamiento* se representa como $(Y_i)(1)$. Por *no participantes*, $(T_i) = 0$ e (Y_i) pueden representarse como $(Y_i)(0)$. Si $(Y_i)(0)$, se usa entre *no participantes* como resultado de comparación para los resultados (*outcomes*) de los participantes $(Y_i)(1)$, el efecto promedio del programa podría representarse de la siguiente manera:

$$D = E(Y_i(1) \mid T_i = 1) - E(Y_i(0) \mid T_i = 0)$$

El problema es que los *grupos tratados y no tratados pueden no ser los mismos antes de la intervención*, por lo que la diferencia esperada entre esos grupos puede, no deberse completamente, para programar

la intervención. Si, en la ecuación, se suma y resta el esperado resultado para los *no participantes* si hubieran participado en el programa: $E(Y_i(0) | T_i = 1)$ u otra forma de especificar el *contrafactual*, se obtiene:

$$D = E(Y_i(1) | T_i = 1) - E(Y_i(0) | T_i = 0) + [E(Y_i(0) | T_i = 1) - E(Y_i(0) | T_i = 0)]$$

$$D = ATE + B$$

En estas ecuaciones, *ATE* es el *efecto promedio del tratamiento*:

$$E(Y_i(1) | T_i = 1) - E(Y_i(0) | T_i = 0)$$

A saber, *la ganancia promedio en los resultados de los participantes en relación con los no participantes*, como si los sujetos o empresas pyme en nuestro caso, *no participantes*, también fueran tratados. El *ATE* (*Average Treatment Effect*) corresponde a una situación en la que se asigna un sujeto o pyme en nuestro caso, elegido al azar de la población a participar en el programa, por lo que los sujetos (pyme en nuestro caso) *participantes y no participantes* tienen una igual probabilidad de recibir el tratamiento *T*.

El término *B*:

$$[E(Y_i(0) | T_i = 1) - E(Y_i(0) | T_i = 0)]$$

Es el grado de *sesgo* de selección que surge en el uso de *D* como una estimación de la *ATE* (*Average Treatment Effect*). Dado que no se sabe $E(Y_i(0) | T_i = 1)$, *no se puede calcular la magnitud del sesgo de selección de D, por lo tanto, no se puede saber la diferencia exacta en los resultados entre los grupos tratados y de control.*

El objetivo básico de una *evaluación de impacto sólida*, es encontrar formas de deshacerse de *sesgo de selección* ($\mathbf{B} = \mathbf{0}$) o encontrar formas de explicarlo. Un enfoque, discutido más adelante, es *asignar aleatoriamente el programa*. También se ha argumentado, que el *sesgo de selección* desaparecería si se pudiera suponer que si los sujetos (empresas pyme en este caso) a *recibir tratamiento* (condicional en un conjunto de *covariables*, \mathbf{X}) fueran independientes de los resultados (*outcomes*) que tienen. Esta suposición se llama *suposición de falta de fundamento* (*assumption of unconfoundedness*), también conocido como el *supuesto de independencia condicional* (*conditional Independence assumption*) (Lechner, 1999; Rosenbaum y Rubin, 1983):

$$[Y_i(1), Y_i(0)] \perp T_i | X_i$$

También se puede hacer una suposición más débil de la *exogeneidad condicional de la colocación del programa* (*conditional exogeneity of program placement*). Estos diferentes enfoques y suposiciones se discutirán a en capítulos posteriores. La solidez de las estimaciones de impacto, depende de cuán justificables sean los supuestos están respecto de la *comparabilidad* de los participantes o grupos de comparación, así como del enfoque de la *exogeneidad* del programa en *áreas tratadas y no tratadas*. Sin embargo, si no se adopta de manera convincente alguno de los enfoques o suposiciones expuestos, no será posible evaluar el alcance del sesgo \mathbf{B} .

Describiendo un caso

El problema de evaluación consiste en medir el impacto del programa (*o tratamiento*) sobre un conjunto de variables de resultado en un conjunto de individuos (**Bernal y Peña, 2011**). Por ejemplo, en el efecto que tiene un programa de introducción de tecnología móvil para el aprendizaje, desarrollado en **Mejía-Trejo et al. (2015)** donde se espera la interacción de las variables:

- a. Tecnología (**TECH.** *Technology*)
- b. Contenidos de enseñanza aprendizaje y estilos (**CTLMS.** *Contents Teaching-Learning Management and Styles*)
- c. Rol profesor-estudiante (**PSR.** *Professor-Student Role*)
- d. Aprendizaje móvil (**ML.** *Mobile Learning*)

De los individuos participantes, nos arroja los siguientes resultados a partir de la observación:

1. Las variables de resultado son las variables sobre las cuales se espera que el programa tenga un efecto en los individuos beneficiarios del programa evaluado. En nuestro ejemplo, serían los indicadores del estado de adopción de tecnología móvil para el aprendizaje (**ML**) porque el programa trata de las variables que se incluyen para su adopción. El problema de *evaluación de impacto*, consiste entonces en establecer la *diferencia entre la variable de resultado del individuo participante en el programa en presencia del programa y la variable de resultado de ese individuo en ausencia del programa*. Esta diferencia es lo que se conoce como *efecto del tratamiento o programa*. El problema fundamental que se enfrenta en una *evaluación de impacto* es que para construir el *efecto del tratamiento* necesitaríamos conocer la diferencia entre la variable de resultado del individuo participante una vez se ha

implementado el programa y la variable de resultado que habría obtenido ese individuo en el caso hipotético de que no existiera el programa. Claramente, no se pueden observar ambos resultados para el mismo individuo al mismo tiempo.

2. El segundo resultado, es decir, el resultado del individuo participante si el programa no existiera, *es hipotético* y, por ende, *no se observa*. Este resultado hipotético se denomina *resultado contra-factual* en la literatura de *evaluación de impacto*.
3. Se debe tener un marco teórico de referencia para formalizar el problema de la *evaluación de impacto* (Mejía-Trejo et al., 2015). Formalmente, definimos el indicador del tratamiento como D_i . En el caso en que el programa que el *tratamiento es binario* (por ejemplo, el sujeto *participa en el aprendizaje móvil o no participa*) entonces $D_i = (1)$ si el individuo i recibe el tratamiento (*es tratado*) y (0) de lo contrario. Las variables de resultado se definen como $Y_i(D_i)$ para cada sujeto $i = (1) \dots N$ y N denota la población total. Es decir, $Y_i(1)$ es la variable del individuo i no es tratado. El efecto del tratamiento (o *impacto del programa*) para un individuo i se puede escribir como:

$$t_i = Y_i(1) - Y_i(0) \quad (1.1)$$

De nuevo, el problema fundamental de la *evaluación de impacto* es que en la *realidad*, solo se da uno de los dos resultados potenciales $Y_i(1)$ o $Y_i(0)$ para cada individuo i *pero no ambos*. Es decir, en los datos solo queda registrado $Y_i(1)$ si $D_i = (1)$ y $D_i = (0)$. En otras palabras, el investigador no dispone del resultado con tratamiento si el individuo no fue tratado, $Y_i(1)$ si $D_i = (0)$ ni dispone del resultado en ausencia del tratamiento si el individuo ha sido efectivamente tratado $Y_i(0)$ si $D_i = (1)$. Note que el impacto del programa medido por la diferencia de $t_i = Y_i(1) - Y_i(0)$, se refiere aun momento dado en el tiempo y, por lo tanto, no es equivalente

a comparar el mismo individuo en dos momentos distintos del tiempo (antes y después de la intervención). Por tanto, el resultado observado se puede escribir como:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0) = Y_i(0) \text{ si } D_i = (0) \text{ o } Y_i(1) \text{ si } D_i = (1) \quad (1.2)$$

En este caso, y en adelante, se entiende que una variable es observada si la información existe y está registrada en los datos a disposición del investigador. De manera análoga, una variable no observada es aquella que no existe o no quedó registrada en la base de datos disponible. Debido a que uno de los dos resultados en la ecuación:

$$t_i = Y_i(1) - Y_i(0)$$

No es observable para cada individuo i , *no es posible estimar el efecto individual promedio del tratamiento, del tratamiento, t_i* . El análisis se debe concentrar en el *impacto promedio* del programa en la población o en subconjuntos de la población (dependiendo del interés de política que se tenga).

4. En primera instancia, se puede estimar el impacto promedio del programa (o *efecto medio del tratamiento*) en la población (**ATE**):

$$Y_{ATE} = E_{it} = E [Y_i(1) - Y_i(0)] \quad (1.3)$$

Donde $E [\cdot]$ denota el operador de expectativas. Una representación simple de la variable de resultado con base en el modelo de regresión lineal está dada por:

$$Y_i = \beta_0 + t_i D_i + u_i \quad (1.4)$$

Donde:

$$t_i = Y_i(1) - Y_i(0) \text{ y } Y_i(0) = E[Y_i(0)] + u_i = \beta_0 + u_i$$

5. El resultado efecto t_{ATE} , se interpreta como el cambio promedio en la variable de resultado cuando un sujeto escogido al azar, pasa aleatoriamente de ser *participante a no participante*. Este parámetro es, particularmente relevante en el caso de la evaluación de un programa universal. En la mayoría de los casos, sin embargo, el tratamiento o programa no es universal sino que solo está disponible para un subconjunto de la población, generalmente porque el programa ha sido focalizado. En este caso, es posible utilizar un estimador únicamente que promedie el efecto sobre la población elegible.
6. Por un lado, se puede utilizar el *impacto promedio del programa sobre los tratados (ATT. Average Treatment on Treated)*, que es, por lo general, el *parámetro de mayor interés* en una evaluación de impacto. Es decir, el efecto promedio del *tratamiento* en el subconjunto de individuos que fueron efectivamente tratados. Éste corresponde a la diferencia entre la *media* de la variable de resultado, en el grupo de los *participantes* y la *media* que hubieran obtenido los *participantes si el programa no hubiera existido*:

$$t_{ATT} = E(t_i | D_i = (1)) = E[Y_i(1) | D_i = (1)] - E[Y_i(0) | D_i = (1)] \quad (1.5)$$

Donde: $E[| D]$ denota el operador de expectativas condicional

En este caso:

$$E[Y_i(1) | D_i = (1)]$$

Es el valor esperado de la variable de resultado en el grupo de tratamiento en *presencia del tratamiento* y:

$$E [Y_i (0) | D_i = (1)]$$

Se conoce como el *resultado contrafactual*, es el valor esperado de la variable de *resultado en el grupo de tratamiento en ausencia del tratamiento*. Evidentemente, el *promedio contrafactual*, es decir, el resultado promedio de los individuos tratados de no haber existido el programa o tratamiento, *es un resultado hipotético*, por lo cual no se observa en la realidad, y por tanto no queda registrado en los datos. *El efecto promedio del programa* sobre los tratados es particularmente *relevante* para definir si un programa existente debe *continuar* o, por el contrario, *debe eliminarse o modificarse*.

7. Por otra parte, se puede estimar el *impacto promedio del programa sobre los no participantes* (**ATU**. *Average Treatment on the Untreated*), que corresponde a la diferencia entre la *media* de la variable de resultado que habrían tenido los *no participantes si hubieran participado* en el programa y la media de la variable de resultado que efectivamente tuvieron los *no participantes al no haber participado*:

$$t_{ATU} = E (t_i | D_i = (0)) = E [Y_i (1) | D_i = (0)] - E [Y_i (0) | D_i = (0)] \quad (1.6)$$

En este caso, el resultado *contrafactual*:

$$E [Y_i (1) | D_i = (0)]$$

Corresponde al promedio de la variable de resultado de los *no participantes* si hubieran participado en el programa, dado que estos individuos *no han sido tratados*. Evidentemente, este *contrafactual* es hipotético, por lo cual no se observa en la realidad, y

por tanto no queda registrado en los datos. El parámetro t_{ATU} es relevante cuando la evaluación tiene por objetivo investigar si el programa se debe extender o no a otros grupos de la población. En cualquiera de los dos casos, t_{ATT} y t_{ATU} , es necesario escoger una aproximación apropiada (o sustituto) del *contrafactual* dado que este es un *resultado hipotético* que no se observa en la realidad, y por tanto no queda registrado en los datos. Por ejemplo, en el caso del t_{ATT} se requiere una aproximación de $E [Y_i (0) | D_i = (1)]$, es decir, el promedio de la variable de resultado entre los participantes en ausencia del programa. En principio, se podría utilizar el *promedio de la variable de resultado* entre los individuos *no participantes* pero elegibles para participar en el programa (conocido como el *grupo de control*, o *grupo de comparación*):

$$E [Y_i (0) | D_i = (0)], \text{ como una aproximación de } E [Y_i (0) | D_i = (1)]$$

Es decir, se podría utilizar el resultado de los *no participantes* (pero elegibles) como aproximación del resultado que *habrían tenido los participantes si el programa no hubiera existido*.

8. Sin embargo, esta comparación podría generar *estimaciones inexactas* del efecto del programa, dado que los *participantes* y los *no participantes generalmente son diferentes, aun en ausencia del programa*. Por ende, las variables de resultado del grupo de tratamiento y el grupo de control podrían ser diferentes, aun si el programa no existiera. Este problema se conoce como *sesgo de autoselección*. Por ejemplo, es posible que los participantes en el programa de aprendizaje móvil (**ML**) provengan de familias más vulnerables que los *participantes elegibles pero que no participan en el programa*. Las *variables de vulnerabilidad*, como el *ingreso del hogar* y la *accesibilidad a equipos móviles*, pueden tener efectos directos sobre los indicadores que estamos utilizando como variable de resultado (**ML**), aparte del efecto directo del programa

evaluado. Por ejemplo, en hogares con más ingreso, el *acceso a equipos móviles* puede ser mejor que en hogares con menos ingresos. De ser así, al comparar los resultados de *tecnología*, por ejemplo, del grupo de tratamiento con los del grupo de control, podríamos estar atribuyendo al programa un efecto negativo que en realidad se debe a que el *grupo de participantes tratados* está más restringido (por ingresos, por acceso a la tecnología móvil, etc.) que el grupo de *participantes usados como control*. En este caso, estaríamos subestimando el efecto del programa, debido a que no hemos tenido en cuenta las diferencias preexistentes entre los dos grupos que afectan tanto la probabilidad de participación en el programa como la variable de resultado (aprendizaje vía móvil **ML**). Además de posibles diferencias observadas entre los dos grupos (como los ingresos, el acceso a la tecnología móvil, el conocimiento de las apps, etc.), es también posible que existan diferencias no observadas o no medidas entre los dos grupos. Por ejemplo, puede ser que los padres de familia participantes en el programa, estén más motivados o atentos o sean más proactivas respecto al desarrollo de sus hijos, por lo cual se esmeran en lograr la participación en el programa. El problema de *autoselección* radica en que la motivación de los padres (*que no observamos y sería difícil de medir*) afecta no sólo la probabilidad de participar en el programa, sino también al estado de aprendizaje de los estudiantes. Esto es porque los padres más motivados podrían vigilar mejor el acceso tecnológico de sus hijos. Por tanto, la diferencia observada en el estado de aprendizaje de los participantes de los dos grupos se podría deber parcialmente a la diferencia en el nivel de motivación de las padres, y no exclusivamente a que un grupo participa en el programa y el otro no.

9. Note que, la comparación entre el promedio de la variable de resultado de los participantes y el promedio de la variable de resultado de los no participantes está dada por:

$$E [Y_i (1) | D_i = (1)] - E [Y_i (0) | D_i = (0)]$$

Adicionalmente, se sabe que:

$$\begin{aligned} t_{ATT} &= E [Y_i (1) | D_i = (1)] - E [Y_i (0) | D_i = (1)] \\ t_{ATT} + E [Y_i (0) | D_i = (1)] &= E [Y_i (1) | D_i = (1)] \end{aligned}$$

Restando $E [Y_i (0) | D_i = (0)]$ a ambos lados de la ecuación, se obtiene:

$$\begin{aligned} t_{ATT} + E [Y_i (0) | D_i = (1)] - E [Y_i (0) | D_i = (0)] &= E [Y_i (1) | D_i = (1)] \\ - E [Y_i (0) | D_i = (0)] & \end{aligned} \quad (1.7)$$

Donde el lado derecho de la ecuación, es la diferencia entre el promedio de la variable de resultado de *los participantes* y el promedio de la variable de resultado de *los no participantes*. De la ecuación, se deduce que utilizar:

$$\begin{aligned} E [Y_i (0) | D_i = (0)] & \text{aproximación del contrafactual,} \\ E [Y_i (0) | D_i = (1)] & \end{aligned}$$

Permite recuperar el t_{ATT} si y solo si:

$$E [Y_i (0) | D_i = (1)] - E [Y_i (0) | D_i = (0)] = 0 \quad (1.8)$$

Esta condición implica, que la *variable de resultado en ausencia del programa*, debería ser idéntica para el grupo de individuos tratados ($D = 1$) y el grupo de individuos no participantes o grupo de control ($D=0$), lo cual, como ya se ha expuesto, es un supuesto bastante fuerte. Por esta razón, la simple comparación de medias de las variables de resultado entre el grupo de tratamiento y el grupo de control generaría una aproximación inexacta del efecto

del programa sobre el grupo de participantes tratados, con excepción de un programa implementado a manera de *experimento aleatorio*.

- 10.** Si el supuesto anterior se cumple, es decir, si el valor promedio de la variable de resultado en *ausencia del programa* fuera idéntico para el grupo de individuos tratados ($D = 1$) que para el grupo de individuos no participantes ($D = 0$), entonces el efecto del programa estaría dado por:

$$t_{ATT} = E [Y_i(1) | D_i = (1)] - E [Y_i(0) | D_i = (0)] \quad (1.9)$$

En este caso, el estimador t_{ATT} estaría dado por el análogo muestral de la ecuación anterior, es decir:

$$t'_{ATT} = [Y' | D = (1)] - [Y' | D = (0)] \quad (1.10)$$

Donde $Y' | D$ es el promedio muestral de la variable de resultado, dado el estatus de participación (D). Es decir, el estimador del efecto del programa bajo el supuesto del **punto 9**:

$$E [Y_i(0) | D_i = (1)] - E [Y_i(0) | D_i = (0)] = 0$$

Resulta de comparar el promedio muestral de Y en el *grupo de tratamiento* con el promedio muestral de en el *grupo de control*. Esta comparación de *medias*, se puede implementar fácilmente por el método de regresión de la siguiente manera:

$$Y_i = \beta_0 + \beta_1 D_i + u_i \quad (1.11)$$

Donde Y_i es la variable de resultado para el individuo i , D_i es una variable binaria, también conocida como *variable dummy* o dicotómicas, que toma el valor de **(1)** si el individuo i participa en el

tratamiento (o programa) y (0) si el individuo i es elegible pero no participa en el programa, y u_i es el término de error de la regresión, que recoge las *variables observadas y no observadas* del individuo i , aparte de D_i , que afectan la variable de resultado. Como t_{ATT} es generalmente el parámetro de interés en evaluación de impacto, en adelante nos referimos simplemente a t .

11. Si el supuesto: $E [Y_i (0) | D_i = (1)] - E [Y_i (0) | D_i = (0)] = 0$ se cumple, entonces no existe correlación entre la participación en el programa (el indicador D_i) y todas las características del individuo contenidas en u_i porque el supuesto implica que los individuos participantes son idénticos a los *individuos no participantes*. En otras palabras, la participación es independiente de las características del individuo y, por lo tanto:

$$E [Y_i (0) | D_i = (0)]$$

Es una aproximación adecuada del contrafactual:

$$E [Y_i (0) | D_i = (1)]$$

Formalmente, esto implica que :

$$E (u_i | D_i) = (0) \quad (1.12)$$

Éste se conoce como el *supuesto de independencia condicional* e indica que conocer (D_i) no añade ninguna información adicional acerca de u_i . En otras palabras, los individuos que participan en el programa, *no son sistemáticamente distintos* de los individuos que no participan en el programa en características contenidas en u_i . Bajo el supuesto:

$$E (u_i | D_i) = (0)$$

El estimador de c por el método de mínimos cuadrados ordinarios (**OLS**. *Ordinary Least Squares*) es consistente e *insesgado*. Lo primero implica que el estimador de β_1 por **OLS** (*Ordinary Least Squares*), β_1' , converge en probabilidad a β_1 , es decir, el límite del estimador a medida que aumenta el tamaño de muestra es exactamente el parámetro que se va a estimar. Lo segundo significa que el valor esperado del estimador de **OLS** es igual al valor del parámetro, es decir:

$$E(\beta_1') = \beta_1$$

Independientemente del tamaño de la muestra. El parámetro β_i en la ecuación:

$$Y_i = \beta_0 + \beta_i D_i + u_i$$

Se obtiene:

$$\begin{aligned} E[Y_i(1) | D_i = (1)] &= E(\beta_0 + \beta_1 + u_i | D_i = (1)) = \beta_0 + \beta_1 + E(u_i / D_i = (1)) = \beta_0 + \beta_1 \\ E[Y_i(0) | D_i = (0)] &= E(\beta_0 + u_i / D_i = (0)) = \beta_0 + E(u_i / D_i = (1)) = \beta_0 \end{aligned}$$

Porque: $E(u_i / D_i = (0))$

Entonces:

$$t = E[Y_i(1) | D_i = (1)] - E[Y_i(0) | D_i = (0)] = (\beta_0 + \beta_1) - \beta_0 = \beta_1$$

El estimador de **OLS** (*Ordinary Least Squares*) de β_1 estaría dado por:

$$\beta_1' = [Y' | D = (1)] - [Y' | D = (0)] \quad (2.13)$$

Que coincide precisamente con la definición del t'_{ATT} previsto en la ecuación 2.10:

$$t'_{ATT} = [Y' | D = (1)] - [Y' | D = (0)] \quad (2.10)$$

12. Con las reflexiones anteriores y volviendo al caso descrito, suponga que debemos evaluar el impacto hipotético de un programa de aprendizaje móvil (**ML**) basado en el modelo de **Mejía-Trejo et al. (2015)** dirigido a niños entre los **6-15 años** de edad provenientes de familia con ingresos considerados de clase media-alta. El programa consiste, en la provisión de sujetos que cuentan con un mínimo de acceso, tanto de equipos móviles (Smartphone/ Tablet/LapTop) así como de internet de alta velocidad, con beneficios económicos en cuanto a descuentos en el servicio de internet como equipos móviles a utilizar. La composición de los sujetos participantes, es revisada por especialistas en métodos de tecnologías de aprendizaje, para asegurarse de que los contenidos de los programas educativos sean adecuados, dadas las edades de los niños beneficiarios. Por lo que el indicador de participación en el programa, se define como:

$$D_i = \begin{cases} (1) & \text{si el niño elegible } i \text{ participa en el programa (ML)} \\ (0) & \text{si el niño elegible } i \text{ no participa en el programa (ML)} \end{cases}$$

Así, los niños de familias entre los **6-15 años** de edad, que participan en el programa, constituyen el *grupo de tratamiento*, mientras que los niños de familias entre los **6-15 años** de edad, que no participan en el programa, constituyen el *grupo de control*. Como variable de resultado (*outcome*) de interés, (**ML**)= (Y_i), se identifica el puntaje **Z (Z-Score)**. El puntaje **Z (Z-Score)**, corresponde al nivel de aprendizaje, según la edad acordada por los especialistas en tecnologías del aprendizaje *estandarizada* de acuerdo con las

medias y varianzas poblacionales, por grupos de edad y sexo. El puntaje **Z** de nivel de aprendizaje según la edad resulta de *restarle el nivel de aprendizaje del niño la media poblacional de aprendizaje para su grupo de edad y sexo y dividir esta diferencia por la desviación estándar del nivel de aprendizaje de su grupo de edad y sexo en la población. El resultado indica el número de desviaciones estándar que el niño está por encima (o por debajo) de su media poblacional. Un puntaje **Z** de cero, indica que el niño se ubica exactamente en la media de su grupo.* Por lo que:

- $E [Y_i (1) | D_i = (1)]$ es en este caso, el *promedio* del puntaje **Z** de aprendizaje móvil (**ML**) en el grupo de tratamiento en presencia del programa.
- $E [Y_i (0) | D_i = (1)]$ es en este caso, el *promedio* del puntaje **Z** de aprendizaje móvil (**ML**) en el grupo de tratamiento en ausencia del programa. Es decir, cuál habría sido el promedio del **puntaje Z** de aprendizaje móvil (**ML**) según la edad en el escenario hipotético de que el programa no hubiera existido. Este *contrafactual* no se observa, y esto constituye la principal dificultad de la evaluación de impacto.
- $E [Y_i (0) | D_i = (0)]$ es el promedio del puntaje **Z** de aprendizaje móvil (**ML**) en el grupo de niños *no participantes* (o grupo de control).

Suponga que la variable de resultado en que estamos interesados, es la incidencia de *bajo rendimiento crónico del aprendizaje*, es decir, la probabilidad de que el aprendizaje del niño esté *dos desviaciones estándar por debajo de lo que debería estar*, dados su edad y su sexo (es decir, que el puntaje **Z** de aprendizaje móvil (**ML**) según la edad sea *menor que -2*). En este caso, $E [Y_i (1) | D_i = (1)]$ sería el porcentaje de niños participantes con *bajo rendimiento crónico del aprendizaje en presencia del*

programa. Si sucediera que los promedios del puntaje **Z** de aprendizaje móvil en el *grupo de tratamiento* y en el *grupo de control* son idénticos en ausencia del programa (es decir, se cumpliera el supuesto de la **ecuación 2.8**), entonces el efecto del programa se podría estimar simplemente *comparando la media de la variable de resultado en el grupo de tratamiento con la media de la variable de resultado en el grupo de control*. Esto se puede implementar con una regresión lineal del puntaje **Z** de aprendizaje móvil (**ML**) según la edad (Y_i) sobre la variable binaria de la participación, D_i , estimada por la **OLS** (*Ordinary Least Squares*).

$$Y_i = \beta_0 + \beta_i D_i + u_i$$

β_i sería el efecto del programa de aprendizaje móvil escogido si se cumple el supuesto de independencia condicional de la ecuación **1.12**.

En este apartado, se asume que el investigador sabe exactamente cuál es la variable de resultado Y_i que quiere medir y tiene los datos a disposición. En general, la elección de una variable de resultado Y_i con base en la cual se determinará la efectividad de un programa *no es una pregunta trivial*. La elección de Y_i debe estar guiada principalmente por *los objetivos y lineamientos del programa que se evalúa*. Por ejemplo, el programa hipotético anterior, basado en la introducción del aprendizaje móvil (**ML**), basado en el modelo de **Mejía-Trejo et al. (2015)**, tiene como objetivo mejorar el aprendizaje de los niños beneficiarios, con base a los requerimientos comentados. Por tanto, es razonable que la variable de resultado seleccionada sea un indicador del estado de aprendizaje de los niños (por edad, género, condiciones de alcance tecnológico, etc.). Existen casos como éste en que la selección de la variable de resultado

es relativamente obvia, pero también existen casos en los que no es elemental.

Por ejemplo, suponga un programa innovador para pymes/mipymes como grupos vulnerables económicamente. El programa consiste en la provisión de una suma de dinero mensual que corresponde a un determinado número de salarios mínimos diarios. En los objetivos y lineamientos se establece que el programa está diseñado para *proteger pymes/mipymes contra el riesgo económico* de la imposibilidad de generar ingresos, disminuir su vulnerabilidad por el tipo de actividad de bajos recursos y propiciar su inserción en la comunidad. En este caso, ¿cuál sería la variable de resultado Y_i ideal si le pidieran evaluar este programa? No es tan sencillo. En principio, podría ser un indicador los ingresos, puesto que las *pymes/mipymes* podría estar mejorando su estado financiero como resultado del aumento en el ingreso asociado a la participación en el programa. Sin embargo, las *pymes/mipymes* podrían estar gastando el dinero en cosas diferentes a sus insumos de operación totalmente ajenos al programa como alimentos, medicamentos u otros elementos asociados con un mejor estado de bienestar *aparente*, por lo cual la evaluación podría generar una conclusión muy negativa acerca del programa.

Por otra parte, la variable de resultado Y_i podría ser más ser más bien, una medición de la cantidad de proyectos a realizar con posibilidades reales de compra por parte de los interesados, a partir de asesorías dirigidas. Si las *pyme/mipyme* como personas morales, no perciben el suficiente reconocimiento de su actividad por los productos o servicios que ofrecen, a través de asesorías especializadas, se desarrollan con planeaciones estratégicas de deficientes, errática, poco realistas, etc. Entonces el hecho de que se les ofrezca participar en el programa, tendería a una mayor motivación que les impulse a generar mayor productividad, y el impacto podría verse más bien reflejado en una planeación estratégica más precisa, realista, alcanzable. Resulta que no es tan sencillo diseñar instrumentos que midan adecuadamente el nivel de planeación estratégica o el nivel de inserción de la *pyme/mipyme* en su comunidad.

Como se puede apreciar, *encontrar una variable de resultado adecuada en este caso es todo un reto para el evaluador*. Note, sin embargo, que en principio éste es el resultado de una política cuyo diseño es demasiado general y, por tanto, evaluar los logros de tales intervenciones es generalmente muy difícil. En suma, el evaluador debe tratar de relacionar, de la mejor manera posible, los objetivos, lineamientos y forma de operación del programa, con *variables de resultado* que se espera que puedan medir de manera relativamente razonable el desempeño del programa (Bernal y Peña, 2011).

El sesgo de selección

Como se explicó, la evaluación de impacto consiste en la estimación de:

$$t'_{ATT} = E(t_i | D_i = (1)) = E[Y_i(1) | D_i = (1)] - E[Y_i(0) | D_i = (0)] \quad (1.9)$$

Donde:

$E[Y_i(1) | D_i = (1)]$ es el valor esperado de la variable de resultado entre los *participantes en el programa en presencia del programa* a $E[Y_i(0) | D_i = (1)]$, o *resultado contrafactual*, es el valor esperado de la variable de resultado, entre los *participantes en ausencia del programa*. En otras palabras, evaluar la diferencia entre la *variable de resultado* entre el grupo de tratados si existe el programa y la variable de resultado entre el grupo de tratados si no se hubiera implementado el programa. Claramente, no es posible observar ambos resultados al mismo tiempo. Sin embargo, sí se puede observar la variable de resultado entre un grupo de individuos elegibles que no participan en el programa (o grupo de control), $E[Y_i(0) | D_i = (0)]$.

El principal reto de la *evaluación de impacto*, es determinar las condiciones bajo las cuales $E[Y_i(0) | D_i = (0)]$, se puede utilizar como una aproximación válida de $E[Y_i(0) | D_i = (1)]$, y por lo tanto, utilizarse en

la **ecuación 2.9** para obtener el efecto del programa t_{ATT} . Evidentemente, $E [(0) | = (0)]$, se podría utilizar como una aproximación adecuada del *contrafactual*, si:

$$E [Y_i (0) | D_i = (1)] = E [Y_i (0) | D_i = (0)] \quad (1.13)$$

Es decir, si la variable de resultado, en *ausencia del programa*, es idéntica para el *grupo de individuos tratados* ($D = 1$) que para el *grupo de individuos de control* ($D = 0$). El supuesto de la **ecuación 1.13** se viola toda vez que la *participación en el programa* es una elección del individuo elegible. La razón es que los *participantes y los no participantes generalmente son diferentes, aun en ausencia del programa*, y por tal motivo, es precisamente que se observa que *unos escogen participar y otros no*, aún si todos son elegibles para recibir el tratamiento. Es decir, existen *características (observadas y/o no observadas)* que causan que unos individuos participen y otros no. Probablemente, las diferencias en estas características entre *individuos participantes e individuos no participantes* también originen diferencias en la *variable de resultado* entre un grupo y el otro. Por ende, es muy probable que la *variable de resultado del grupo de tratamiento y la variable de resultado del grupo de control sean diferentes, aún si el programa no existiera*. Este hecho se conoce como *sesgo de selección*.

Recuerde que el t_{ATT} se puede escribir como:

$$E [Y_i (1) | D_i = (1)] - E [Y_i (0) | D_i = (1)] = t_{ATT} + E [Y_i (0) | D_i = (1)] - E [Y_i (0) | D_i = (1)] \quad (1.14)$$

Donde:

$E [Y_i (0) | D_i = (1)] - E [Y_i (0) | D_i = (1)]$ es el *sesgo de selección*.

En la **ecuación 1.14** es claro que si los individuos del *grupo de tratamiento* y el *grupo de control* son diferentes, aun en *ausencia del tratamiento* (la segunda parte al lado derecho de la ecuación), entonces la *diferencia entre la media del grupo de tratamiento y la media del grupo de control preexistente* (el lado izquierdo entre los de dos la grupos, ecuación) término será igual que al t_{ATT} más la diferencia preexistente entre los dos grupos, término que denomina *sesgo de selección*. Es decir, la *comparación de medias* entre el *grupo de tratamiento* y el *grupo de control* será una combinación del efecto directo del tratamiento, t_{ATT} y las diferencias preexistentes entre los dos grupos:

$$E [Y (0) | D = (1)] - E [Y (0) | D = (0)]$$

Y sin información adicional, el investigador no puede descifrar qué parte se debe a qué. En suma, los individuos que se *autoseleccionan en el grupo de tratamiento* son *sistemáticamente diferentes* (en formas que observamos y formas que no observamos) de *los individuos que se autoseleccionan en el grupo de control*, y precisamente por eso participan en el programa. Esas diferencias sistemáticas pueden estar relacionadas, a su vez, con la variable de resultado, objeto de la *evaluación de impacto*.

En la sección anterior, se explicó que si se cumple el supuesto de **la ecuación 1.13**, entonces, el efecto del programa sobre la variable de resultado, estimada por mínimos cuadrados ordinarios de la siguiente Y_i , es la pendiente de la regresión lineal:

$$Y_i = \beta_0 + \beta_1 D_i + u_i \quad (1.15)$$

Donde D_i es el *indicador del estatus del tratamiento*.

En este caso, β_i por **OLS** (*Ordinary Least Squares*). es un estimador *consistente e insesgado* del efecto del programa porque $E(u_i / D_i) = (0)$. Es decir, *no hay correlación* entre la decisión de participar o no en el programa, D_i , y las características en el término de error, u_i . Sin embargo, si la participación en el programa es una decisión de los sujetos, es improbable que se cumpla el supuesto:

$$E(u_i / D_i) = (0)$$

Esto se presenta, porque es posible que los sujetos que deciden *participar* en el programa lo hagan debido a que son sistemáticamente diferentes de los sujetos que deciden *no participar*. Algunas de estas diferencias pueden, a su vez, explicar la variable de resultado Y_i y por tanto, estar contenidas en el término de error u_i . Dicho de otra manera, saber si un sujeto *participa o no participa* en el programa informa sobre las variables no observadas del individuo contenidas en u_i por lo cual se *invalida el supuesto de independencia condicional*.

Suponga que el programa de aprendizaje móvil (**ML**) caso hipotético de **Mejía-Trejo et al. (2015)**, visto en la sección anterior, se ofrece y los *individuos elegibles deben decidir si quieren participar o no*. Además, participar tiene un costo en tiempo y trámites, debido a que los padres del niño elegible deben acercarse a la oficina de la institución encargada de administrar el programa, diligenciar un formulario de inscripción y llevar al niño potencialmente participante a un control en el que se registran sus condiciones iniciales (nombre, edad, género, equipo móvil, contenidos, acceso a internet, etc.). Aunque todas los padres de niños entre los **6-15 años** de edad que pertenecen al perfil de clase media-alta son elegibles para el programa, *es muy probable que un conjunto de padres considere que el costo de inscribirse para participar en el programa es muy alto* por motivos diversos como no ser atractivos los descuentos ofrecidos en la adquisición de equipos móviles, las tarifas de acceso a internet, las filas son muy largas, el examen de aptitudes de los niños es muy demorado o el

lugar de inscripción en el que se lleva a cabo, está muy lejos, por lo cual preferirán no participar. Es decir, se podría pensar que los padres que sí deciden participar son aquellas *más proactivos y motivados* y que, a pesar del *costo de participación*, se preocupan por sus hijos y su bienestar lo suficiente (o tienen el tiempo disponible) como para ir hasta la oficina, hacer la fila, sacar la cita en el sitio de examen, ir para asistir a la cita, volver, etcétera.

Note entonces que los *padres participantes* y los *padres no participantes* son diferentes en cuanto a su perfil de padre, la disponibilidad de tiempo, su disposición a ocuparse de los hijos, etc. Estas características, algunas de las cuales no son observadas por el evaluador porque no existen mediciones o porque no fueron registradas en los datos (como qué tan dedicado es un padre hacia sus hijos), pueden también afectar el nivel de aprendizaje de los *niños participantes* y *no participantes*. Por ejemplo, los padres más proactivos y motivados pueden ser también aquellas que se preocupan más por el aprendizaje de los niños y, por tanto, les ofrecen mayores atenciones, les procuran sus equipos móviles, les llevan controles de aprovechamiento, etc. Es decir, sería una *variable contenida* en u_i porque explica la medida de aprendizaje (o variable de resultado), Y_i .

Dado que existen características observadas y no observadas de los sujetos, contenidas en u_i , que explican tanto la decisión de participar en el programa como la variable de resultado, entonces:

$$E(u_i / D_i) \neq (0)$$

Es decir, la participación en el programa (D_i), sí está correlacionada con el término de error en la regresión el programa de la **ecuación (1.15)**. Esto constituye la *violación de uno de los supuestos fundamentales del modelo de regresión lineal (independencia condicional)* y, por ende, los estimadores de mínimos cuadrados ordinarios *no serán insesgados*. Esto se debe a que para demostrar que el estimador **OLS** (*Ordinary Least*

Squares) es insesgado, se requiere que cumpla el supuesto $E(u_i/D_i) = (0)$. Esto implica que el *valor esperado del estimador del efecto del programa estimado por OLS no corresponde al efecto verdadero del programa sino que está subestimado o sobrestimado*, según sean las características no observadas de los individuos. *En otras palabras, el efecto del programa no se puede obtener por la simple comparación de medias de la variable de resultado entre el grupo de tratamiento y el grupo de control.*

Intuitivamente, si hay variables que explican tanto la participación en el programa como la variable de resultado, la comparación de medias puede estar atribuyendo al programa un efecto que en realidad se debe a las diferencias preexistentes entre el grupo de tratamiento y el grupo de control. En nuestro ejemplo del programa de aprendizaje móvil (ML), si los padres *participantes* son más dedicados y están más motivados que las padres *no participantes*, entonces *podemos estar atribuyendo al programa un efecto positivo sobre el estado de aprovechamiento o aprendizaje de los niños*, cuando en realidad la diferencia en las medidas a favor de los niños tratados se debe a que tienen *padres más pendientes de su aprendizaje que los niños de control*, y no al programa en explica tanto la participación sí. Por lo que:

1. Si en la característica del individuo X_i que explica tanto la participación en el programa como la variable de resultado, es observable y está contenida en la base de datos disponible, entonces este problema se soluciona simplemente incluyéndola en la regresión de la **ecuación 1.15** como una variable explicativa adicional:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + u_i \quad (1.16)$$

Donde X_i es una característica observable del individuo, que explica tanto la participación en el programa como la variable resultado Y_i .

2. Por ejemplo, si los niños de padres más pobres son aquellos que deciden participar en el programa de aprendizaje móvil (ML),

mientras que los individuos más ricos eligen no hacerlo, entonces diferencia entre los X_i participantes sería el índice de riqueza del hogar. Si esta es la única diferencia entre los participantes y los no participantes, entonces el estimador de β_1 en la **ecuación 1.16** por **OLS** (*Ordinary Least Squares*) es un estimador consistente e insesgado del efecto del programa.

3. Si las diferencias entre los participantes y no participantes son todas *observables* (y la base de datos contiene información acerca de todas ellas), entonces la regresión de la **ecuación 1.15** se puede extender para incluir todas esas características. Si todas las diferencias entre el grupo de tratamiento y el grupo de control se incluyen en la regresión, entonces los factores restantes contenidos en u_i son efectivamente independientes de la decisión de participar, D_i y por lo tanto, el estimador de β_1 por **OLS** (*Ordinary Least Squares*), es un estimador insesgado y consistente del efecto del programa.
4. Sin embargo, en la mayoría de los casos, algunas de las diferencias entre los participantes y los no participantes no se observan o son características (en principio observables) que no están contenidas en la base de datos. Por ejemplo, los padres más dedicados a sus hijos pueden ser más propensos a participar en el programa de aprendizaje móvil (**ML**) que los padres menos dedicados. Pero ¿qué tan dedicado es un padre? no es una variable incluida en las encuestas. Por lo tanto, aunque sabemos que ésta es una diferencia entre los *beneficiarios* y los *no beneficiarios* por la que debemos controlar en la **ecuación 1.16**, no tenemos los datos a disposición. En este caso, el estimador de **OLS** (*Ordinary Least Squares*) de β_1 *está sesgado*, es decir:

$$E(\beta_1') > \beta_1 \text{ ó } E(\beta_1') < \beta_1$$

La dirección precisa del sesgo (llamado *sesgo de selección*, por las razones que se han expuesto anteriormente) depende de la rela-

ción existente entre la participación en el programa y la variable que diferencia a los participantes de los no participantes (llámese W), y de la dirección del efecto de la variable excluida W sobre la variable de resultado Y . Teniendo en mente el siguiente modelo:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + u_i$$

Donde W es una característica no observable (o no contenida en la encuesta), la dirección del sesgo de β_1 por **OLS** (*Ordinary Least Squares*). Sobre la regresión de la **ecuación 1.15**, se resume como se observa en la **Tabla 1.2**

Tabla 1.2. Dirección del sesgo de β_1 por OLS sobre la regresión

Ítems	$Corr(D_i, W_i) > 0$	$Corr(D_i, W_i) < 0$
$\beta_2 > 0$	$E(\beta_1') > \beta_1$	$E(\beta_1') < \beta_1$
$\beta_2 < 0$	$E(\beta_1') < \beta_1$	$E(\beta_1') > \beta_1$

Donde $Corr(D_i, W_i)$ es la correlación entre D_i y W_i

Fuente: Bernal y Peña (2011).

5. Por ejemplo, tomemos el caso de la *primera celda* (superior-izquierda). Si a mayor dedicación de los padres (*variable W no observada*):
 - a. Mayor es la probabilidad de participar en el programa aprendizaje móvil (**ML**), y entonces $Corr(D_i, W_i) > 0$.
 - b. Si, además, la dedicación de los padres aumenta el aprovechamiento del niño según su edad (variable de resultado Y_i) porque existe una mayor atención en la procuración de los medios tecnológicos y en el seguimiento de los contenidos, entonces $\beta_2 > 0$. En este caso, el estimador de **OLS** (*Ordinary Least Squares*) de β_1' en la regresión de la **ecuación 1.15** estaría sesgado hacia arriba:

$$E(\beta_1') > \beta_1$$

Es decir, el efecto estimado del programa sobre el peso según el aprovechamiento o aprendizaje, es mayor que el efecto verdadero del programa. Esto se presenta porque, al no poder incluir W en la regresión, le estamos atribuyendo al programa parte programa, (D_i) parte del efecto positivo que tiene W sobre Y . Es decir, se le atribuye al programa parte del efecto positivo de la mayor motivación de los padres participantes sobre el estado de aprovechamiento o aprendizaje vía móvil, de sus hijos. En otras palabras, (D_i) absorbe tanto sobre Y , como el efecto que tiene W directamente sobre Y , dando lugar a un efecto más grande de lo que en realidad es.

La gran mayoría de programas que se evalúan en la actualidad, están caracterizados por el hecho de que los individuos deben elegir si participan o no. Esto implica que las diferencias que surgen entre los participantes y no participantes son, en buena parte, no observables. Así, el gran reto de la evaluación de impacto, es encontrar metodologías que permitan obtener de un estimador consistente e insesgado de β_1 aun en presencia del sesgo de selección (Bernal y Peña, 2011).

Evaluaciones de impacto ex post

Se pueden usar varios métodos diferentes en la *teoría de la evaluación de impacto* para abordar la cuestión fundamental del *contrafactual faltante*. Cada uno de estos métodos, conlleva sus propios supuestos sobre la naturaleza del sesgo potencial de selección, en la participación del programa objetivo. Los supuestos son cruciales para desarrollar el modelo apropiado en determinar los impactos del programa. Existen así, métodos que varían según sus *supuestos subyacentes* con respecto a cómo resolver el *sesgo*

de selección en la estimación del efecto del tratamiento del programa, los cuales, en resumen, son:

1. Las evaluaciones aleatorias. Las *evaluaciones aleatorias* implican una iniciativa asignada al azar, a través de una muestra de sujetos (por ejemplo: comunidades, empresas o individuos); el progreso del tratamiento y control de los sujetos, que exhiban características similares en un preprograma, se rastrean con el tiempo. Los *experimentos aleatorizados* tienen el ventaja de evitar *sesgos de selección* a nivel de aleatorización.
2. Los métodos de coincidencia (*matching methods*), en especial, técnica de propensión de coincidencia de puntaje (**PSM**. *Propensity Score Matching*). En ausencia de un experimento, los métodos **PSM** comparan los *efectos del tratamiento* a través de *las unidades participantes y las unidades no participantes emparejadas*, con la coincidencia realizada en un rango de características observadas. Por lo tanto, los métodos de **PSM** suponen que el *sesgo de selección* se basa solo en la observación de características; *no pueden dar cuenta de factores no observados que afectan la participación*.
3. Los métodos de doble diferencia (**DD**. *Double-Difference Methods*). Los métodos **DD** suponen que *la selección no observada está presente y que es invariante en el tiempo*: el efecto del tratamiento, se determina tomando la diferencia en los resultados a través de las unidades de tratamiento y control, antes y después de la intervención del programa. Los métodos **DD** se pueden utilizar, tanto en *entornos experimentales como no experimentales*.
4. El uso de variables instrumentales (**IV**. *Instrumental Variable Methods*). Los modelos **IV**, se pueden usar con datos de *sección transversal o panel* y, en este último caso, permitir el *sesgo de selección* en las características no observadas para variar con el tiempo. En el enfoque **IV**, el *sesgo de selección* en las características no observadas se corrige encontrando una variable (o instru-

mento) que se correlaciona con la participación pero no con las características no observadas afectando el resultado; este instrumento se utiliza para *predecir la participación*.

5. El diseño de regresión discontinuidad (**RD**, *Regression Discontinuity Design*) y los métodos de tubería (**PM**, *Pipeline Methods*). El **RD** y los métodos **PM** son extensiones de los métodos **IV** y experimentales; explotan las reglas del programa exógenos (como los requisitos de elegibilidad) para *comparar participantes y no participantes* en un vecindario cercano alrededor del *límite de elegibilidad*. Los métodos **PM**, en particular, construyen un grupo de comparación de sujetos que son elegibles para el programa pero aún no lo han recibido.
6. Impacto distributivo (**DI**, *Distributional Impacts*).
7. Aproximación estructural y de otro modelaje (*Structural and other modeling approaches*).

Finalmente, el manual cubre los métodos para examinar los **DI**, así como los enfoques de modelado que pueden resaltar mecanismos (como las fuerzas intermedios del mercado) por las cuales los programas tienen un impacto. Estos enfoques cubren un mezcla de diferentes métodos cuantitativos discutidos en los siguientes capítulos, así como métodos *ex ante* y métodos *ex post*.

Evaluaciones de procesos

Gertler et al. (2017) consideran que las *evaluaciones de procesos*, se centran en cómo se implementa y funciona un programa, considerando si corresponde a su diseño original, y documentando su desarrollo y funcionamiento. Normalmente, estas evaluaciones pueden llevarse a cabo con relativa rapidez y a un costo razonable. En los proyectos piloto y en las etapas iniciales de un programa, pueden ser una valiosa fuente

de información sobre cómo mejorar la implementación del programa, y se suelen utilizar como primeros pasos para desarrollar un programa de modo que los ajustes operativos se puedan hacer antes de que se termine su diseño. Pueden probar si un programa funciona como estaba diseñado y si es consistente con la teoría del cambio del mismo. Una *evaluación de procesos* debería incluir los siguientes elementos, que a menudo se basan en una cadena de resultados o modelo lógico, complementados con documentos del programa y entrevistas con informantes clave y grupos focales beneficiarios:

- Objetivos del programa y contexto en el que funciona.
- Descripción del proceso utilizado para diseñar e implementar el programa.
- Descripción de las operaciones del programa, incluido cualquier cambio en las mismas.
- Datos básicos sobre las operaciones del programa, incluidos indicadores financieros y de cobertura.
- Identificación y descripción de eventos que escapen al control del programa que pueden haber influido en la implementación y los resultados.
- Documentación, como notas de concepto, manuales operativos, actas de las reuniones, informes y memorandos.

Aplicar una evaluación de impacto a un programa, cuyos procesos operativos no han sido validados, plantea el doble riesgo de que se malgasten los recursos de dicha evaluación, cuando en realidad podría bastar con una evaluación de proceso más sencilla, o bien el riesgo de que los ajustes necesarios en el diseño del programa se introduzcan una vez que la evaluación de impacto ya ha comenzado, lo cual cambia el carácter del programa que se evalúa y la utilidad de la mencionada evaluación.

Evaluación de impacto vs. evaluación de proyectos

Bernal y Peña (2011) afirman que la evaluación de proyectos (el análisis de rentabilidad de programas de infraestructura) *diffiere sustancialmente de la evaluación de impacto* (el análisis contrafactual de intervenciones sociales). La primera está basada en un análisis *ex ante* detallado de los beneficios y los costos esperados de un proyecto. En el caso de la construcción de una nueva vía, por ejemplo, es necesario estimar el número esperado de usuarios y los beneficios correspondientes: los ahorros de tiempo y dinero para un horizonte temporal previamente definido. Los beneficios deben compararse con los costos estimados, para así calcular la tasa de retorno del proyecto.

La evaluación de impacto, o *ex post*, está basada en un análisis contrafactual, en la comparación entre los resultados efectivamente observados en *presencia* del programa y los que habrían sido observados en su *ausencia*. Esta comparación permite, bajo algunos supuestos adicionales, calcular la tasa de retorno del programa o proyecto evaluado. En la *evaluación de proyectos*, la clave está en el cálculo correcto de los *beneficios y los costos esperados*; en la *evaluación de impacto*, en la estimación *adecuada del contrafactual*, para establecer una relación causal entre la política y los impactos generados. Ambos tipos de análisis son *complementarios* y ambos brindan una importante información para la toma de decisiones.

El *análisis de proyectos* no permite, tener en cuenta *los beneficios o efectos no esperados*. El *análisis de impacto*, en la mayoría de los casos, no permite medir efectos de largo plazo o efectos dinámicos. Hoy en día casi la totalidad de los proyectos de infraestructura cuentan con un análisis exhaustivo de costo-beneficio. De la misma manera, los programas sociales deberían contemplar, desde su mismo diseño, un análisis exhaustivo del impacto. El impacto estimado del programa es un insumo indispensable para la medición de los beneficios del programa

que se requieren para el *análisis costo-beneficio*. Este tipo de análisis, como se dijo previamente, contribuye a aumentar la calidad del gasto y la eficiencia de las iniciativas sociales. **Harberger (1972)**, economista estadounidense, fundador de la Escuela de Economía de Chicago, uno de los principales promotores de la evaluación en los países en desarrollo, ha dicho reiteradamente que:

“...para manejar la macroeconomía de un país basta con cinco o seis personas, pero que para gastar bien, se necesita un ejército de evaluadores de programas y proyectos...”

Sin embargo, para el caso de las políticas sociales, es difícil realizar un *análisis costo-beneficio* robusto. Con frecuencia, los resultados son muy sensibles a la manera como se valoran los beneficios de las intervenciones y a los supuestos que se hacen.

Análisis de costo-beneficio y costo-efectividad

Es sumamente importante que la evaluación de impacto pueda complementarse con información sobre el costo del proyecto, del programa o de la política que se evalúa. Una vez que están disponibles los resultados de la evaluación de impacto, estos pueden combinarse con información sobre los costos del programa para responder a otras dos preguntas. En primer lugar, en la forma básica de una evaluación de impacto, añadir información del costo permitirá llevar a cabo un análisis de costo-beneficio, a partir de lo cual se podrá responder a la pregunta: ¿cuáles son los beneficios de un programa con un determinado costo? El análisis de costo-beneficio estima los beneficios totales esperados de un programa, comparado con sus costos totales esperados. Busca cuantificar todos los costos y beneficios de un programa en términos monetarios, y evalúa si estos últimos superan a los costos (**Gertler et al., 2017**).

En un mundo ideal, *el análisis de costo basado en la evidencia de la evaluación de impacto* existiría no solo para un programa concreto sino también para una serie de programas o alternativas de programas, de modo que los responsables de las políticas pudieran valorar qué programa o alternativa es más efectivo en función de los costos para lograr un determinado objetivo. Cuando una evaluación de impacto ensaya alternativas de un programa, añadir información sobre costos le permite responder a la segunda pregunta: *¿cómo se comparan diversas alternativas de implementación en cuanto a su costo-efectividad?* Este *análisis de costo-efectividad* compara el costo relativo de dos o más programas o alternativas de programas para alcanzar un objetivo común, como la producción agrícola o las puntuaciones de los alumnos en las pruebas.

En un análisis de *costo-beneficio* o de *costo-efectividad*, la *evaluación de impacto* estima el lado del beneficio o el lado de la efectividad, mientras que examinar los costos proporciona la información sobre los mismos. Esta obra se centra en la *evaluación de impacto* y no aborda en detalle cómo recopilar datos sobre costos o llevar a cabo análisis de costo-beneficio o costo-efectividad. Sin embargo, *es fundamental que la evaluación de impacto se complemente con información sobre el costo del proyecto, del programa o de la política que se evalúa*. Una vez que se disponga de información sobre el *impacto y el costo de diversos programas*, el análisis de *costo-efectividad* puede identificar cuáles son las inversiones que arrojan la tasa más alta de retorno y permiten a los responsables de las políticas tomar decisiones bien fundamentadas sobre las intervenciones en las que hay que invertir.

Ética y evaluación de impacto

Cuando se toma la decisión de diseñar una evaluación de impacto, se deben considerar algunos asuntos éticos **Gertler et al. (2017)**. Se han formulado preguntas a propósito de si la *evaluación de impacto* es ética en sí misma y por sí sola. Un punto de partida de este debate, consiste en considerar *la ética de invertir recursos públicos en programas cuya efectividad se desconoce*. En este contexto, *la falta de evaluación puede en sí misma ser no ética*. La información sobre la efectividad del programa que generan las evaluaciones de impacto puede conducir a una inversión más efectiva y ética de los recursos públicos.

Otras consideraciones éticas tienen que ver con las reglas utilizadas para asignar los beneficios del programa, con los métodos con los que se estudia a los seres humanos y con la transparencia en la documentación de los planes de investigación, datos y resultados.

El principio ético más básico en una evaluación es que la *prestación de intervenciones con beneficios* conocidos no debería negarse o retrasarse únicamente en función de los objetivos de la evaluación. En esta obra, se sostiene que las evaluaciones no deberían dictar cómo se asignan los beneficios, sino más bien que deberían ajustarse a reglas de asignación del programa que sean equitativas y transparentes. En este contexto, cualquier preocupación ética a propósito de las reglas de asignación del programa no nace de la evaluación de impacto en sí misma sino directamente de las reglas de operación del programa. Planificar las evaluaciones puede ser útil para dilucidar las reglas de operación del programa, y contribuir a estudiar si son equitativas y transparentes, sobre la base de criterios claros de elegibilidad. La asignación aleatoria de los beneficios del programa a menudo suscita inquietudes éticas a propósito de la negación de dichos beneficios a destinatarios elegibles. Sin embargo, la mayoría de los programas funciona en contextos operativos con *recursos financieros y administrativos limitados*, lo cual impide llegar

a todos los beneficiarios elegibles de una sola vez. Desde una *perspectiva ética*, todos los sujetos que son igualmente elegibles para participar en cualquier tipo de programa social deberían tener la misma oportunidad de ser destinatarios del programa. *La asignación aleatoria cumple este requisito ético.*

En situaciones en las cuales un programa se aplicará en fases a lo largo del tiempo, la implementación se puede basar en *seleccionar aleatoriamente el orden en que los beneficiarios*, todos igualmente meritorios, serán receptores del programa. En estos casos, los destinatarios que ingresen más tarde en el programa pueden conformar el grupo de comparación para los primeros beneficiarios, generando un sólido diseño de evaluación, así como un método transparente e imparcial para asignar los escasos recursos.

La ética de la evaluación de impacto excede a la ética de las reglas de asignación del programa. También incluye la ética de la investigación en seres humanos, así como la ética de llevar a cabo investigaciones transparentes, objetivas y reproducibles. En numerosos países e instituciones internacionales, se han creado juntas de revisión institucional o comités éticos para regular las investigaciones que involucren a seres humanos. Estos organismos se encargan de asesorar, aprobar y monitorear los estudios de investigación, con los objetivos fundamentales de proteger los derechos y promover el bienestar de todos los sujetos. Aunque las evaluaciones de impacto son eminentemente empresas operativas, también constituyen estudios de investigación y, como tales, deberían adherir a las directrices de investigación para seres humanos.

Conseguir que una evaluación de impacto sea objetiva, transparente y reproducible es un componente ético igualmente importante de la investigación.

Para que la investigación sea transparente, los planes de evaluación de impacto pueden incluirse en un plan de análisis previo y ser sometidos a un registro de estudios. Una vez que la investigación se lleve a cabo,

los datos y códigos utilizados en el análisis pueden hacerse públicamente disponibles de modo que otras personas puedan replicar el trabajo, a la vez que se protege el anonimato.

Consideraciones finales

En resumen, se deben tomar varias medidas para garantizar que las *evaluaciones de impacto* sean efectivas y obtener comentarios útiles. Durante la identificación y preparación del proyecto, por ejemplo, *la importancia y los objetivos de la evaluación necesitan ser delineados claramente*. Las preocupaciones adicionales incluyen la *naturaleza y el momento de las evaluaciones*. Para aislar el efecto del programa sobre los resultados, independientemente de otros factores, el investigador debe medir el *tiempo* y estructurar *evaluaciones de impacto* de antemano para ayudar a los funcionarios del programa a *evaluar y actualizar la orientación*, así como otras pautas para la implementación, durante el curso de la intervención. La *disponibilidad y la calidad de los datos* también son esenciales para evaluar los efectos del programa; los requisitos de los datos, dependerán de si los evaluadores están aplicando un enfoque de *método cuantitativo o cualitativo*, o ambos, y sobre si el marco es *ex ante*, *ex post* o *ambos*.

Si se recopilan nuevos datos, es necesario abordar una serie de consideraciones adicionales, incluyendo *tiempos, diseño y selección de muestras*, así como la selección instrumentos para una encuesta apropiada. Además, se deberán realizar *encuestas piloto* en el campo para que las preguntas de la entrevista puedan ser revisadas y refinadas. La recopilación de datos, sobre *aspectos socioeconómicos* relevantes, las características tanto a nivel beneficiario como a nivel comunitario también deben ayudar a comprender mejor el comportamiento de los encuestados, dentro de su entorno económico y social, ambiental.

Ravallion (2003), también sugiere una serie de pautas para mejorar recopilación de datos en encuestas. Estas pautas, incluyen la comprensión de diferentes facetas y hechos estilizados del programa y del entorno económico de los *participantes* y *no participantes* para mejorar el diseño de muestreo y completar los módulos de encuesta a fin de obtener información adicional (por ejemplo, sobre la naturaleza de la participación o la focalización del programa) para comprender y abordar el *sesgo de selección* más adelante.

La contratación y capacitación de personal de campo, así como la implementación de un trabajo consistente al enfoque para gestionar y proporcionar acceso a los datos, también es esencial. Durante la implementación del proyecto, desde una perspectiva de gestión, el equipo de evaluación debe ser formado cuidadosamente para incluir suficiente experiencia técnica y administrativa para garantizar la precisión del informe de datos y resultados, así como transparencia en la implementación para que los datos pueden ser interpretados con precisión. La recopilación continua de datos, es importante para mantener actualizados a los responsables de las políticas de implementación de proyectos sobre el progreso del programa, así como, por ejemplo, parámetros del programa que deben adaptarse a las circunstancias o tendencias cambiantes con el acompañamiento de iniciativas. Los datos deben analizarse cuidadosamente y presentarse a los responsables de decisión y otras partes interesadas importantes en el programa para permitir potencialmente una valiosa retroalimentación. Esta entrada, además de los resultados de la evaluación en sí misma, puede ayudar a guiar el diseño de políticas futuras también.

CAPÍTULO 2.

Preparación, Inferencia Causal y Contrafactuales

No todos los programas justifican una evaluación de impacto. Las evaluaciones de impacto deberían utilizarse selectivamente cuando la pregunta que se plantea exige un exhaustivo análisis de la causalidad. Estas evaluaciones pueden ser costosas si uno tiene que recopilar sus propios datos, y el presupuesto con el que se cuenta para la evaluación debería utilizarse de manera estratégica a partir de realizar ciertos cuestionamientos base, tales como:

- a. La primera pregunta básica por realizar es: ¿qué está en juego?, ¿será que la evidencia del éxito del programa, o la modalidad del mismo o la innovación en el diseño, fundamentarán decisiones clave? Estas decisiones a menudo implican asignaciones presupuestarias y determinan la escala del programa. *Si el presupuesto es limitado o si los resultados afectarán solo a unas pocas personas, es posible que una evaluación de impacto no merezca la pena.*
- b. Si se decide que hay mucho en juego la siguiente pregunta es: ¿existe evidencia que demuestre que el programa funciona? Concretamente, ¿se sabe cuál sería el alcance del impacto del programa?, ¿hay evidencia disponible de programas similares en circunstancias similares? Si no hay evidencia disponible acerca del potencial del tipo de programa proyectado, puede que convenga comenzar con una prueba piloto que incorpore una evaluación de impacto. En cambio, si hay evidencia disponible de circunstancias similares, *el costo de una evaluación de impacto probablemente*

estaría justificado solo si puede abordar una nueva pregunta determinante para las políticas de introducción de innovaciones. Este sería el caso si el programa incluye innovaciones sustanciales que todavía no han sido probadas. Para justificar la movilización de recursos técnicos y financieros necesarios para llevar a cabo una evaluación de impacto de alta calidad, la intervención que será evaluada debe ser (**Gertler et al., 2017**):

- *Innovadora*. Probará un enfoque nuevo y prometedor.
 - *Aplicable*. Se puede aumentar la escala o se puede aplicar en un entorno diferente.
 - *Estratégicamente relevante*. La evidencia proporcionada por la evaluación de impacto fundamentará una decisión clave ligada a la intervención. Esto podría estar relacionado con la ampliación del programa, reformas o asignaciones presupuestarias.
 - *No probada*. Se sabe poco acerca de la efectividad del programa o de las alternativas de diseño, tanto a nivel global como en un contexto específico.
 - *Influyente*. Los resultados se utilizarán para fundamentar decisiones de políticas.
- c. Una pregunta final es la siguiente: ¿se cuenta con los recursos necesarios para una buena evaluación de impacto? Estos recursos atañen a elementos técnicos, como datos y el tiempo adecuado, recursos financieros para llevar a cabo la evaluación y recursos institucionales de los equipos que participan, más su interés y compromiso para construir y utilizar evidencia causal. *Un equipo de evaluación* es esencialmente una asociación entre dos grupos: un equipo de responsables de las políticas de introducción de innovaciones y un equipo de investigadores. Los equipos tienen que trabajar en aras del objetivo común de asegurar que una evaluación técnicamente robusta y bien diseñada se implemente de manera adecuada y arroje resultados relevantes para cuestiones clave de las políticas y del diseño del programa. Una clara comprensión de la

premisa y de la promesa de la evaluación de impacto por parte del equipo de evaluación contribuirá a asegurar su éxito.

Si usted decide que tiene sentido encarar una *evaluación de impacto*, en virtud de las preguntas planteadas y la necesidad relacionada de analizar la causalidad, más los elementos en juego asociados con los resultados y la necesidad de evidencia acerca del desempeño de su programa, entonces lo tendrá mayores probabilidades de continuar con la *evaluación de impacto*.

Pasos por considerar

Gertler (et al., 2017), recomiendan:

1. *Construir una teoría del cambio* que describa cómo se supone que el proyecto logrará los objetivos previstos;
2. *Elaborar una cadena de resultados* que sirva como instrumento útil para esbozar la teoría del cambio;
3. *Especificar las preguntas de la evaluación*; y
4. *Seleccionar los indicadores* para evaluar el desempeño.

Los cuatro pasos son necesarios y es preferible *implementarlos al inicio*, es decir, cuando comienza a diseñarse el proyecto de evaluación del programa o de las reformas. Esto requiere la participación de todas las partes interesadas, desde los responsables de las políticas hasta los implementadores del proyecto, con el fin de forjar una visión común de los objetivos y de cómo serán alcanzados. Esta participación permitirá crear un consenso sobre el enfoque de la evaluación y las principales preguntas a responder, y reforzará los vínculos entre la evaluación, la implementación del programa y el diseño de políticas de introducción de innovaciones. Cada uno de los pasos está claramente definido y está articulado en el modelo lógico incorporado en la cadena de resultados desde

la precisión en la determinación de los objetivos y las preguntas hasta la definición de las ideas integradas en la teoría del cambio, y los resultados esperados por la implementación del programa. Estos parámetros son esenciales para definir los elementos técnicos de la evaluación, incluyendo el tamaño de la muestra requerida para la evaluación y los cálculos de la potencia. En la mayoría de las *evaluaciones de impacto* será importante incluir una *evaluación de costo-beneficio, o costo-efectividad*, como se indica en el capítulo 1. Los formuladores de política de introducción de innovaciones, deberán estar atentos para saber qué programas o reformas son efectivos *pero también cuál es su costo*, es decir, precisar si es viable aumentar la escala de un programa y si es posible replicarlo.

Construcción de una teoría del cambio

La teoría del cambio es la descripción de cómo se supone que una intervención conseguirá los resultados deseados **Gertler et al. (2017)**. En ese sentido, expone la *lógica causal* de cómo y por qué un proyecto, una modalidad de programa o un diseño de innovación lograrán los resultados previstos. Debido al *enfoque causal* de la investigación:

- *Una teoría del cambio es la base de cualquier evaluación de impacto*. Su construcción es uno de los primeros requisitos para el diseño del proyecto, ya que contribuye a especificar las preguntas de la investigación.
- Las *teorías del cambio* describen una secuencia de eventos que generan resultados: analizan las condiciones y los supuestos necesarios para que se produzca el cambio, explicitan la lógica causal inscrita en el programa y trazan el mapa de las intervenciones del programa a lo largo de las vías lógicas causales.
- Configurar una *teoría del cambio* en conjunto con las partes interesadas puede clarificar y mejorar el diseño del programa. Esto es especialmente importante en los programas que pretenden influir

en las conductas, pues las *teorías del cambio* pueden ayudar a determinar los insumos y actividades de la intervención, qué productos se generan y cuáles son los resultados finales derivados de los cambios de comportamiento de los beneficiarios.

- El mejor momento para desarrollar una *teoría del cambio* es al comienzo de la fase de diseño, cuando es posible reunir a las partes interesadas con el fin de definir una visión colectiva del programa, sus objetivos y la ruta para alcanzar esos objetivos. Así, las partes interesadas podrán implementar el programa a partir de un entendimiento común del mismo, de sus objetivos y de su funcionamiento.
- Por otra parte, es útil que los diseñadores de programas revisen la literatura en *busca de evidencia que describa experiencias y programas similares*, y comprueben los contextos y los supuestos detrás de las vías causales de la *teoría del cambio* que configuran. Por ejemplo, en el trabajo de **Mejía-Trejo y Bravo (2019)**. Se tiene literatura de cómo las innovaciones, son basadas en distintos niveles de valores de la empresa que los impulsa y cómo éstas mejoran el bienestar a través de los productos y/o servicios que fabrica o entrega en consonancia con el propósito de la empresa.

Desarrollo de una cadena de resultados

Una cadena de resultados es una manera de describir una teoría del cambio. **Gertler et al. (2017)**. Otros enfoques, incluyen:

- *Modelos teóricos,*
- *Modelos lógicos,*
- *Marcos lógicos, y*
- *Modelos de resultados.*

Todos estos modelos integran los elementos básicos de una *teoría del cambio*, a saber: *una cadena causal*, una especificación de las condi-

ciones e influencias externas y la determinación de los supuestos clave. Aquí, se utilizará el *modelo de cadena de resultados* porque es el más sencillo y claro para describir la *teoría del cambio* en el contexto operativo de los programas de desarrollo.

Una *cadena de resultados*, establece la *lógica causal* desde el inicio del programa, empezando con los recursos disponibles, hasta el final, teniendo en cuenta los objetivos de largo plazo. Fija una definición lógica y plausible de cómo una secuencia de insumos, actividades y productos relacionados directamente con el proyecto interactúa con el comportamiento y define las vías para lograr los impactos. Ver **Gráfico 2.1**.

Gráfico 2.1. Elementos de una cadena de resultados



Fuente: Gertler et al. (2017).

Una cadena de resultados básica esquematizará *un mapa* con los siguientes elementos:

- *Insumos*. Los recursos de que dispone el proyecto, que incluyen el personal y el presupuesto.

- *Actividades*. Las acciones emprendidas o el trabajo realizado para transformar los insumos en productos.
- *Productos*. Los bienes y servicios tangibles que producen las actividades del programa (controlados de forma directa por la agencia ejecutora).
- *Resultados*. Los resultados que previsiblemente se lograrán cuando la población se beneficie de los productos del proyecto. En general, estos resultados se observan entre el corto y el mediano plazo y no suelen estar controlados de forma directa por la agencia ejecutora.
- *Resultados finales*. Los resultados finales alcanzados señalan si los objetivos del proyecto se cumplieron o no. Normalmente, los resultados finales dependen de múltiples factores y se producen después de un período más largo.

Tanto la *implementación* como los *resultados* forman parte de la *cadena de resultados*. La *ejecución* tiene que ver con el trabajo efectuado en el proyecto, que incluye *insumos*, *actividades* y *productos*. Estos ámbitos, que son responsabilidad directa del proyecto, suelen ser *monitoreados* para verificar si el proyecto está generando los bienes y servicios previstos. Los resultados comprenden los *resultados* y los *resultados finales*, que no son controlados de manera directa por el proyecto y dependen de cambios en el comportamiento de los beneficiarios del programa, es decir, dependen de las interacciones entre la oferta (*la implementación*) y la demanda (*los beneficiarios*). Por lo general, estos ámbitos son objeto de *evaluaciones de impacto* para *medir su efectividad*. Una buena *cadena de resultados* contribuirá a hacer aflorar los supuestos y riesgos implícitos en la *teoría del cambio*. Los formuladores de políticas de introducción de innovaciones están mejor situados para articular la lógica causal y los supuestos en los que descansa, así como los riesgos que pueden influir en el logro de los resultados previstos. El equipo, que dirige la evaluación deberá explicitar los supuestos y riesgos implícitos en consulta con los responsables de las políticas. *Una buena cadena de resul-*

tados, también incluirá evidencia provista por la literatura relacionada con los resultados de programas similares. Las cadenas de resultados son útiles para todos los proyectos, independientemente de que contemplen o no una evaluación de impacto, porque permiten a los responsables de las políticas y a los administradores del programa explicitar los objetivos del proyecto, lo que contribuye a clarificar la lógica causal y la secuencia de eventos que se encuentran detrás de un programa. Además, pueden identificar brechas y eslabones débiles en el diseño del programa y, por lo tanto, pueden ayudar a mejorar su diseño. Las cadenas de resultados también facilitan el monitoreo y la evaluación pues especifican cuál es la información que debe ser monitoreada en cada eslabón de la cadena para realizar un seguimiento de la implementación del programa, y definen qué indicadores de resultados hay que incluir cuando se evalúa el proyecto.

La especificación de las preguntas de la evaluación

La claridad de la pregunta de la evaluación es el punto de partida de cualquier evaluación efectiva (Gertler et al., 2017). La formulación de esta pregunta debe centrarse en la investigación, para asegurar que se ajusta al interés de las políticas en cuestión. En el caso de una *evaluación de impacto*, es necesario estructurarla como una hipótesis comprobable. La *evaluación de impacto* posteriormente, genera evidencia creíble para responder esa pregunta. Como se indicó antes, *la pregunta fundamental de la evaluación de impacto es:*

¿Cuál es el impacto (o el efecto causal) de un programa en un resultado de interés?

Se pone *énfasis en el impacto*, es decir, en los cambios directamente atribuibles a un programa, a una modalidad de programa o a una innovación de diseño. La pregunta de la evaluación debe orientarse según el interés central de las políticas de introducción de innovaciones, en cuestión. Como se señaló en el capítulo 1, las evaluaciones de impacto pueden

analizar toda una gama de preguntas. En ese sentido, antes de estudiar cómo se llevará a cabo el proyecto, el equipo de evaluación debería aclarar cuál es la pregunta que se analizará como primer paso, *sobre la base de la teoría del cambio*. Tradicionalmente, las *evaluaciones de impacto* se han centrado en el impacto que tiene un programa plenamente implementado, en los resultados finales y en contraste con los resultados observados en un grupo de comparación que no ha sido beneficiado por el programa. Sin embargo, el uso actual de las *evaluaciones de impacto* se está ampliando. El equipo de evaluación puede inquirir: *¿la pregunta de evaluación clave, es la pregunta clásica acerca de la efectividad de un programa para cambiar los resultados finales? O ¿se trata de probar si una modalidad de programa es más costo-efectiva que otra o bien de introducir una innovación en el diseño del programa que, de manera previsible, cambiará las conductas, como la matriculación?* En la actualidad la *evaluación de impacto* está incorporando *nuevos enfoques*, de manera creativa, para abordar las cuestiones de interés para el diseño de políticas, en una vasta gama de disciplinas. Tal es el caso del enfoque de *experimento de mecanismo*, que es una *evaluación de impacto* que prueba un mecanismo causal específico dentro de la teoría del cambio (Gertler, et al., 2017).

En una *evaluación de impacto*, la pregunta de la evaluación debe ser formulada como una *hipótesis bien definida y comprobable*, que pueda *cuantificar la diferencia* entre los resultados obtenidos *al contrastar los grupos de tratamiento y comparación*. La *cadena de resultados* puede usarse como base para formular la hipótesis que se busca probar a partir de la evaluación de impacto. Frecuentemente hay unas cuantas *hipótesis* asociadas con el programa, pero *no es necesario explorar* todas en una *evaluación de impacto*, y tampoco es posible hacerlo. Una *pregunta de la evaluación* deriva de elementos fundamentales de la *teoría del cambio* y se formula como una *hipótesis clara, comprobable y cuantificable*, siguiendo nuestro ejemplo de Mejía-Trejo et al. (2015) sobre el aprendi-

zaje móvil (ML): *¿cuál es el efecto del nuevo programa de aprendizaje móvil (ML) en el aprovechamiento de los niños de 6-15 años de edad?*

La selección de indicadores de resultados y desempeño

Una pregunta clara de la evaluación, debe:

- Ir acompañada de la *especificación acerca de qué medidas de resultado* se utilizarán para evaluar el desempeño, incluso en el caso de múltiples resultados. Las *medidas de resultado* seleccionadas se usarán para establecer si un programa o una reforma determinados tienen éxito o no. También son los indicadores que se pueden emplear como referencia al realizar los cálculos de la potencia con el fin de fijar los tamaños de la muestra necesarios para la evaluación.
- Posteriormente, de seleccionar los principales *indicadores* de interés, es preciso definir *objetivos claros* en lo relativo al éxito del programa. Este paso equivale a *determinar el efecto anticipado del programa sobre los principales indicadores de resultado que se han seleccionado*.
- Los tamaños del efecto son los cambios previstos como resultado del programa o de la reforma; por ejemplo, el cambio en las puntuaciones de las pruebas o en la tasa de adopción de una nueva tecnología. Los tamaños del efecto previstos son la base para llevar a cabo cálculos de la potencia.
- Es fundamental que los principales agentes interesados en el equipo de evaluación (tanto el equipo de investigación como el equipo de políticas de introducción de innovaciones) estén de acuerdo tanto en los *indicadores primarios de resultados de interés en la evaluación de impacto, como en los tamaños esperados de los efectos previstos como resultado del programa o de la innovación*, ya que

se usarán para juzgar el éxito del programa y formar la base de los *cálculos de la potencia*.

- Numerosas evaluaciones de impacto fracasan porque los *tamaños de muestra* no son lo bastante grandes para detectar los cambios generados por el programa y se dice que *tienen un déficit de potencia*.
- En este sentido, es *sustancial definir los tamaños mínimos previstos del efecto para establecer criterios básicos del éxito del programa o de la innovación*.
- Cuando existan datos disponibles, es posible realizar *simulaciones ex ante* para observar diferentes escenarios de resultados con el fin de establecer una referencia del tipo de *tamaños del efecto que se pueden esperar en una gama de indicadores*. Las *simulaciones ex ante* también se pueden utilizar para revisar indicadores iniciales de *costo-beneficio o costo-efectividad* y comparar intervenciones alternativas para provocar cambios en los resultados de interés.

Una cadena de resultados articulada con claridad proporciona un *mapa útil* para seleccionar los *indicadores* que se medirán a lo largo de la cadena, a fin de *monitorear* la implementación del programa y evaluar los resultados. Es útil contar con la participación de los equipos de políticas de introducción de innovaciones como de los de investigación, para seleccionar estos indicadores y asegurar que son medidas pertinentes del desempeño del programa. Los indicadores sugeridos como lista **EMARF** por **Gertler et al. (2017)**:

- *Específicos*: para medir la información requerida de la forma más rigurosa posible.
- *Medibles*: para garantizar que la información se puede obtener fácilmente.
- *Atribuibles*: para asegurar que cada medida está relacionada con los logros del proyecto.

- *Realistas*: para garantizar que los datos se pueden obtener de manera oportuna, con una frecuencia y un costo razonables.
- *Focalizados*: en la población objetivo.

Al elegir los *indicadores*, es importante identificarlos a lo largo de toda la *cadena de resultados*, y no solo en el nivel de los resultados, de modo que puedan seguir la *lógica causal* de cualquier resultado observado del programa. En las *evaluaciones de implementación* que se centran en probar dos o más alternativas de diseño, los resultados de interés pueden producirse antes en la *cadena de resultados*, como un resultado adelantado o como resultado de una fase temprana. Aun cuando el interés, solo esté puesto en las medidas de *resultados para la evaluación*, es sustancial realizar un seguimiento de los *indicadores de implementación*, de tal manera que sea capaz de determinar si las intervenciones:

- Se han llevado a cabo como estaban proyectadas,
- Si han sido recibidas por los beneficiarios previstos y
- Si han llegado a tiempo.

Si no se identifican estos indicadores en toda la *cadena de resultados*, se corre el riesgo de que la evaluación de impacto sea como una *caja negra* que podrá determinar si los resultados previstos se materializaron o no, pero no será capaz de explicar por qué.

Lista de verificación: datos para los indicadores

Una vez que se han seleccionado los indicadores, Como lista de verificación final es útil pensar en las disposiciones para *producir los datos con el fin de medir los indicadores*. Esta lista de verificación (PNUD, 2009) considera las disposiciones prácticas para producir todos los indicadores de manera fiable y oportuna:

- ¿Se han especificado con claridad los indicadores (productos y resultados)? Estos provienen de las preguntas clave de la evaluación y deberían ser consistentes con los documentos de diseño del programa y con la cadena de resultados.
- ¿Los indicadores son **EMARF**? específicos, medibles, atribuibles, realistas y focalizados.
- ¿Cuál es la fuente de los datos de cada indicador? Debe ser clara la identificación de la fuente de los datos, como una encuesta, un estudio o una reunión de las partes interesadas.
- ¿Con qué frecuencia se recopilarán los datos? Es preciso incluir un calendario.
- ¿Quién es el responsable de recopilar los datos? Se debe especificar quiénes responsable de organizar la recopilación de datos, verificar la calidad y la fuente de los datos y asegurar el cumplimiento de las normas éticas.
- ¿Quién es responsable del análisis y de los informes? Hay que fijar la frecuencia de los análisis, el método de análisis y el responsable de los informes.
- ¿Qué recursos se necesitan para producir los datos? Es fundamental que los recursos requeridos sean claros y que estén destinados a producir los datos, que a menudo es la parte más cara de una evaluación si se recopilan datos primarios.
- ¿La documentación es adecuada? Es útil diseñar planes para documentar los datos, incluir la utilización de un registro y asegurar el anonimato.
- ¿Qué riesgos implica? Al iniciar con el monitoreo planificado y las actividades de evaluación es preciso considerar los riesgos y los supuestos, así como la manera en que pueden influir en la puntualidad y la calidad de los datos y de los indicadores.

Inferencia causal

En las *evaluaciones de impacto precisas y fiables* existen dos conceptos esenciales:

- La inferencia causal y
- Los contrafactuales.

Muchas preguntas de política de introducción de innovaciones, tienen que ver con relaciones de *causa y efecto*: ¿la continuidad de los programas de estímulos a la innovación mejora la competitividad de las pymes? ¿Los programas de transferencias de tecnología, consiguen mejores resultados financieros para las pymes involucradas? ¿Los programas de formación profesional en innovación, aumentan los ingresos de quienes los han cursado? Las *evaluaciones de impacto* pretenden responder estas preguntas de *causa y efecto* con precisión. Evaluar el impacto de un programa en un conjunto de resultados equivale a evaluar el efecto causal del programa en esos resultados (**Gertler et al., 2017**)

A pesar de que las preguntas de causa y efecto son habituales, contestarlas no es un asunto sencillo. Por ejemplo, en el contexto de un programa de formación profesional en innovación, la sola observación de que los ingresos de una persona que ha recibido formación en la administración de la innovación, aumentan después de que ha completado ese programa *no basta para establecer la causalidad*. Tal vez los ingresos de esa persona se habrían incrementado aunque no hubiera sido objeto de la formación, sino de las condiciones cambiantes del mercado laboral o como consecuencia de sus esfuerzos, o de muchos otros factores que influyen en los ingresos. Las *evaluaciones de impacto* ayudan a superar el problema de atribuir causalidad pues *determinan, empíricamente, en qué medida un programa concreto, y solo ese programa, contribuye a cambiar un resultado*. Para establecer *causalidad* entre un programa y un

resultado se emplean métodos de *evaluación de impacto* a fin de descartar la posibilidad de la intervención de otros factores. Así:

- La respuesta a la pregunta básica de la *evaluación de impacto*:

¿Cuál es el impacto o efecto causal de un programa (P) en un resultado de interés (Y)?

Se obtiene mediante la *fórmula básica de la evaluación de impacto*, de la **ecuación 1.5**:

$$t_{ATT} = E(t_i | D_i = (1)) = E[Y_i(1) | D_i = (1)] - E[Y_i(0) | D_i = (1)]$$

- Según esta fórmula, el *impacto causal* (t_{ATT}) de un programa (D) en un resultado (Y) es la diferencia entre el resultado (Y) con el programa (es decir, cuando $D = 1$) y el mismo resultado (Y) sin el programa ($D = 0$). Por ejemplo, si (D) representa un programa de formación profesional en innovación y (Y) simboliza los ingresos, el impacto causal de un programa de formación profesional (t_{ATT}) es la diferencia entre los ingresos de una persona (Y) después de participar en el programa de formación profesional de innovación ($D = 1$) y los ingresos de la misma persona (Y) en el mismo momento en el tiempo, si no hubiera participado en el programa ($D = 0$).
- Dicho de otro modo, *se trata de medir el ingreso en el mismo momento en el tiempo para la misma unidad de observación (en este caso, una persona) pero en dos realidades diferentes*. Si esto fuera posible, se observaría cuál sería el ingreso de ese mismo individuo *en el mismo momento en el tiempo tanto con el programa como sin él*, de modo que la única explicación posible de cualquier diferencia en los ingresos de esa persona sería el programa. Al comparar al mismo individuo consigo, mismo en el mismo momento se conseguiría eliminar cualquier factor externo que

también pudiera explicar la diferencia de los ingresos. En ese caso, sería posible confiar en que la relación entre el programa de formación profesional y el cambio en los ingresos es causal.

La fórmula básica de la evaluación de impacto es válida para cualquier unidad analizada, ya sea una persona, un hogar, una comunidad, una empresa, un colegio, un hospital u otra unidad de observación que pueda beneficiarse del programa o verse afectada por él. La fórmula también es válida para cualquier resultado (Y) relacionado con el programa en cuestión. Una vez que se han medido los dos componentes clave de esta fórmula: el resultado (Y) con el programa y sin él, es posible responder cualquier pregunta acerca del impacto del programa.

El contrafactual

Es posible pensar en el impacto t_{ATT} de un programa como la diferencia en los resultados (Y) para la misma unidad (persona, hogar, comunidad, etc.) *cuando ha participado en el programa y cuando no lo ha hecho. Sin embargo, es imposible medir al mismo sujeto en dos realidades diferentes al mismo tiempo.* En cualquier momento del tiempo, un sujeto habrá participado en el programa o no lo habrá hecho. No se lo puede observar simultáneamente en dos realidades distintas (es decir, con el programa y sin él). Esto se denomina *problema contrafactual*, que genera el cuestionamiento:

¿Cómo medir lo que habría ocurrido si hubieran prevalecido las otras circunstancias?

Aunque se puede observar y medir el resultado (Y) para un participante del programa ($Y | D = 1$), no hay datos para establecer cuál habría sido su resultado en ausencia del programa ($Y | D = 0$). En la fórmula básica de

la evaluación de impacto, el término ($Y | D = 0$), representa el *contrafactual*, lo cual se puede expresar como:

¿Cuál habría sido el resultado si una persona no hubiera participado en el programa?

Por ejemplo, si la **pyme X** adopta una tecnología disruptiva y desaparece 1 año después, el solo hecho de que la **pyme X** haya desaparecido después de adoptará la tecnología disruptiva *no permite concluir que dicha tecnología disruptiva haya sido la causa de su muerte*. Quizá la **pyme X** ya estaba muy endeudada cuando adoptó la tecnología disruptiva y fue el endeudamiento la que provocó su desaparición en lugar de la tecnología disruptiva. Para *inferir la causalidad* habrá que descartar todos los demás factores potenciales capaces de haber influido en el resultado en cuestión. En este sencillo ejemplo, para saber si la adopción de la tecnología disruptiva causó la desaparición de la **pyme X**, un evaluador tendría que establecer qué le habría ocurrido a la **pyme X** si no hubiera adopción de la tecnología disruptiva. Como el la **pyme X** adoptó la tecnología disruptiva, no es posible observar de forma directa qué habría ocurrido si no lo hubiera hecho. Lo que le habría ocurrido si no hubiera adoptado la tecnología disruptiva es el *contrafactual*. Para definir el impacto de la adopción de la tecnología disruptiva, *el principal reto del evaluador consiste en determinar qué aspecto tiene el estado contrafactual de la realidad para la pyme X*.

Cuando se lleva a cabo una *evaluación de impacto*, es relativamente fácil obtener el primer término de la fórmula básica ($Y | D = 1$), el resultado con un programa, también conocido como el *resultado bajo tratamiento*, ya que basta con medir el resultado de interés para el participante del programa. Sin embargo, no es posible observar de forma directa el segundo término de la fórmula ($Y | D = 0$), para el participante. Es preciso obtener esta información estimando el *contrafactual*. Para entender este concepto clave de *estimación del contrafactual* es útil recurrir a otro caso hipotético. La solución del problema *contrafactual* sería posible si el

evaluador pudiera encontrar un *clon perfecto* de un participante **pyme X'** en el programa. Por ejemplo, si la **pyme X** comienza a recibir unidades de US\$ como ingreso por la adopción de la tecnología disruptiva y lo que se busca es medir el impacto de dicha adopción, la existencia del clon **pyme X'** perfecto de la **pyme X**, haría la evaluación muy fácil, pues se podría comparar el ingreso en unidades de US\$ de la **pyme X**, (por ejemplo, **6**) cuando adopta la tecnología disruptiva, con el ingreso en unidades de US\$ de la **pyme X'**, por ejemplo, **4** sin la adopción de la tecnología disruptiva. En este caso, el impacto en ingresos, sería de **2** unidades de US\$, es decir, la diferencia entre de *ingresos bajo el tratamiento* de adopción de tecnología disruptiva (**6**) y los *ingresos sin tratamiento* de adopción de tecnología disruptiva (**4**). En realidad, está claro que es imposible hallar clones perfectos, y que incluso entre gemelos genéticamente idénticos hay diferencias importantes.

La estimación del contrafactual

La clave para estimar el *contrafactual* para los participantes del programa, consiste en *desplazarse del nivel individual o de la persona al nivel del grupo*. A pesar de que no existe un clon perfecto de una persona única, es posible contar con *propiedades estadísticas* para generar *dos grupos de personas* que, si su número es lo bastante alto, *sean indistinguibles* una de otra desde el punto de vista estadístico en el *nivel del grupo*. El grupo que participa en el programa se denomina *grupo de tratamiento*, y su resultado es ($Y | D = 1$), después de que ha participado en el programa. El *grupo de comparación estadísticamente idéntico* (*grupo de control*) es el grupo que *no es objeto del programa* y permite estimar el resultado contrafactual ($Y | D = 0$), es decir, el resultado que se habría obtenido en el grupo de tratamiento *si no hubiera recibido el programa*. Por lo tanto, en la práctica, el reto de una *evaluación de impacto* es definir un *grupo de tratamiento* y un *grupo de comparación* que sean *estadísticamente idénticos*.

ticos, en promedio, en ausencia del programa. Si los dos grupos son idénticos (estadísticamente), con la única excepción de que un grupo participa en el programa y el otro no, es posible estar seguros de que cualquier diferencia en los resultados tendría que deberse al programa.

Encontrar esos grupos de comparación es la piedra angular de cualquier evaluación de impacto, al margen del tipo de programa que se evalúe. En pocas palabras, sin un grupo de comparación que produzca una estimación precisa del contrafactual, no se puede establecer el verdadero impacto de un programa. En ese sentido, el principal desafío para identificar los impactos es crear un grupo de comparación válido que tenga las mismas características que el grupo de tratamiento en ausencia del programa. Concretamente, los grupos de tratamiento y de comparación deben ser iguales en al menos tres aspectos:

- *En primer lugar, las características promedio del grupo de tratamiento y del grupo de comparación deben ser idénticas en ausencia del programa. Aunque no es necesario que las unidades individuales en el grupo de tratamiento tengan clones perfectos en el grupo de comparación, en promedio, las características de los grupos de tratamiento y de comparación deberían ser las mismas. Por ejemplo, la edad promedio de las pymes en el grupo de tratamiento debería ser la misma que en el grupo de comparación.*
- *En segundo lugar, el tratamiento no tendría que afectar al grupo de comparación de forma directa ni indirecta. En el caso del grupo semejante a la **pyme X** y su ingreso económico por la adopción de tecnología disruptiva, como grupo de tratamiento no debería transferir recursos al grupo de comparación (efecto directo) ni influir en los ingresos de los mercados locales (efecto indirecto). Por ejemplo, si lo que se busca es aislar el impacto de los ingresos por adopción de tecnología disruptiva, al grupo de tratamiento no se le deberían ofrecer más clientes o ventas que al grupo de compara-*

ción; de otra manera, no se podría distinguir si el ingreso adicional es consecuencia también, del mayor número de ventas.

- En tercer lugar, los resultados de las pymes en el *grupo de control*, deberían cambiar de la misma manera que los resultados en *el grupo de tratamiento*, si ambos grupos son objeto del tratamiento (o no). En este sentido, los grupos de tratamiento y de comparación tendrían que reaccionar al programa de igual modo. Por ejemplo, si los ingresos de las pymes de base tecnológica (**pbt**) del *grupo de tratamiento* aumentaran en **1** unidad de US\$ gracias a la adopción de tecnología disruptiva, los ingresos de las pymes en el *grupo de comparación* también tendrían que subir **1** unidad de US\$ si hubieran sido objeto de la formación.

Si se cumplen estas tres condiciones, solamente la existencia del programa de interés explicará cualquier diferencia en el resultado (Y) entre los dos grupos. Esto obedece a que la única diferencia entre los *grupos de tratamiento* y los de *comparación* es que los miembros del *grupo de tratamiento* recibieron el programa, mientras que los del *grupo de comparación* no lo recibieron. Cuando la diferencia en el resultado se puede atribuir totalmente al programa, se ha identificado el impacto causal del programa. En el caso de la **pyme X**, se observó que estimar el impacto de ingresos por adopción de tecnología disruptiva exigía la tarea impracticable de encontrar el clon perfecto de la **pyme X**. En vez de analizar el impacto en un solo individuo, es más realista evaluar el impacto promedio en un grupo de individuos. En este sentido, si se pudiera identificar otro grupo de **pymes** que comparten el mismo promedio de edad, composición por industria, sector, actividades principales, etc. con la salvedad de que no reciban ingresos adicionales, sería posible estimar el impacto de la adopción de tecnología disruptiva, pues este estaría conformado por la diferencia entre los ingresos económicos en ambos grupos. Por lo tanto, si el *grupo de tratamiento* tienen un ingreso de **6** unidades de US\$, mientras

que el *grupo de comparación* tienen un ingreso de **4** unidades de US\$, el impacto promedio de la adopción de tecnología disruptiva sería de **2** unidades de US\$.

Después de definir un *grupo de comparación* válido, es importante considerar qué ocurriría si la decisión fuera seguir adelante con una evaluación sin haber hallado ese grupo. Es evidente que un *grupo de comparación* no válido difiere del *grupo de tratamiento* de alguna manera distinta de la *ausencia* de tratamiento. Debido a esas otras diferencias, la *estimación de impacto* puede ser *no válida* o, en términos estadísticos, *puede ser sesgada*; en ese caso la *evaluación de impacto* no estimará el *verdadero impacto del programa*, sino el efecto del programa mezclado con esas otras diferencias existentes entre los grupos.

Dos estimaciones falsas del contrafactual

En lo que queda de la segunda parte de este libro se abordarán los diversos métodos disponibles para construir *grupos de comparación válidos* que permitirán estimar el *contrafactual*. Sin embargo, antes resulta útil analizar dos métodos de uso habitual, aunque sumamente riesgosos, para construir grupos de comparación que a menudo conducen a estimaciones inadecuadas (*falsas*) del *contrafactual*:

- *Las comparaciones antes-después* (también conocidas como *comparaciones pre-post* o *reflexivas*): cotejan los resultados del mismo grupo antes y después de participar en un programa.
- *Las comparaciones de inscritos y no inscritos* (o *autoseleccionados*): compara los resultados de un grupo que elige participar en un programa con otros de un grupo que elige no participar.

Contrafactual falso 1:

Comparación de resultados antes y después

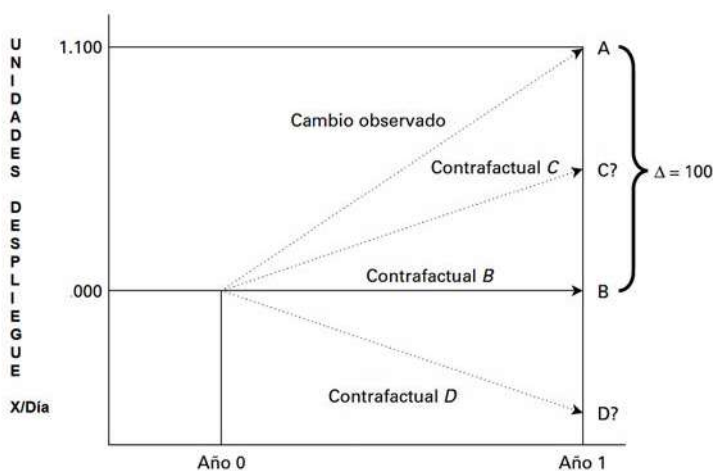
*Una comparación antes-después intenta establecer el impacto de un programa a partir de un seguimiento de los cambios en los resultados en los participantes del programa a lo largo del tiempo. De acuerdo con la fórmula básica de la evaluación de impacto, el resultado para el grupo de tratamiento ($Y | D = 1$), es, sin duda, el resultado después de participar en el programa. Sin embargo, las *comparaciones antes-después* consideran el *contrafactual* estimado ($Y | D = 0$), como el resultado para el grupo de tratamiento antes de que comience la intervención. Básicamente, esta comparación supone que si el programa no hubiera existido, el resultado (Y) para los participantes del programa habría sido igual a su situación antes del programa. *Lo cierto es que en la mayoría de los programas implementados durante meses o años este supuesto no puede sostenerse.**

Suponga el caso hipotético de un programa de adopción de innovación por mercadotecnia digital (**Mejía-Trejo, 2017**) para pymes en el 2019. El programa ofrece micropréstamos a las pymes, lo que les permite comprar dominios, diseñar páginas web, blog, chats, canales de video y redes sociales para aumentar sus ingresos y ser competitivos. Así:

- Se estableció como unidad de despliegue del programa de mercadotecnia digital, un **factor X** como **KPI** pactado sobre un producto innovador (compuesto por SEO/SEM/PPC/Engagement/Frecuencia/ monetización) a fin de medir la respuesta del programa. El año previo (2018) al inicio del programa, las pymes reportaron por la presentación de un producto innovador, un promedio de **1000X/Día** (punto **B** en el **Gráfica 2.1**).
- Un año después del lanzamiento de mejora de innovación por mercadotecnia digital, es decir, en 2019, las unidades de despliegue de mercadotecnia subieron a **1100X/Día** (el punto **A** en el **Gráfica 2.1**). Si se tratara de evaluar el impacto sobre la base de una *comparación antes-después*, se utilizaría el resultado de antes del

programa como una estimación del *contrafactual*. Al aplicar la misma fórmula básica de la *evaluación de impacto*, se llegaría a la conclusión de que el programa ha incrementado la producción de despliegues en **100X/Día (A-B)**.

Gráfica 2.1. Estimaciones antes-después de un programa hipotético de innovación por mercadotecnia digital



Fuente: Elaboración propia.

- Ahora bien, el crecimiento económico se sostuvo en el año del lanzamiento del programa (2019), sin embargo, en Marzo de 2020 con la presencia del COVID-19, se produjo una contracción económica, por lo que la cantidad de despliegues fue menor que **B**, y se sitúa, por ejemplo, en el nivel **D**. En ese caso, *el verdadero impacto del programa sería (A-D)*, que es mayor que los **100X/hr** estimados a partir de la comparación *antes-después*.
- En cambio, de haber continuado el crecimiento económico entre los dos años, el *contrafactual* de los despliegues podría haberse situado en el nivel **C**. Entonces el verdadero impacto del programa habría sido inferior a **100X/hr**.

- En otras palabras, si los análisis de impacto no pueden dar cuenta de eventos como COVID-19 y de todos los demás factores capaces de influir en la producción de despliegues a lo largo del tiempo, *es imposible calcular el verdadero impacto del programa mediante una comparación antes-después.*

En el ejemplo anterior de las innovación por mercadotecnia digital, fenómenos como el COVID-19 constituyen uno de varios factores externos que pueden influir en el resultado de interés del programa (los despliegues de mercadotecnia digital) durante su vigencia. De la misma manera, muchos resultados que los programas de desarrollo pretenden mejorar, como los ingresos, la productividad, la salud o la educación, *están afectados por un conjunto de factores en el transcurso del tiempo.* Por ese motivo, *el resultado de referencia casi nunca es una buena estimación del contrafactual y se lo considera un contrafactual falso.*

Contrafactual falso 2: Comparando grupos inscritos y no inscritos (autoseleccionados)

La comparación entre un grupo de individuos que se inscriben voluntariamente en un programa y un grupo de individuos que elige no participar es otro enfoque riesgoso de la *evaluación de impacto*. *Un grupo de comparación que se autoselecciona para no participar en un programa será otro contrafactual falso (Gertler et al., 2017).* La selección se produce cuando la participación en el programa se basa en las *preferencias, decisiones o características no observables de los participantes potenciales.*

A modo de ejemplo se puede pensar, en un programa hipotético de orientación profesional gestionar introducción de innovaciones para los jóvenes desempleados, por lo que:

- Dos años después de su lanzamiento, una evaluación intenta estimar su *impacto en los ingresos* a partir de la comparación de los

ingresos promedio de un grupo de jóvenes que decidieron *inscribirse en el programa* con los de un grupo de jóvenes que, a pesar de ser elegibles, *decidieron no inscribirse*.

- Si los resultados demostraran que los jóvenes que eligieron inscribirse en el programa *ganan el doble de los que decidieron no hacerlo*, *¿cómo debería interpretarse este hallazgo?* En este caso, el *contrafactual* se estima sobre la base de los *ingresos de quienes eligieron no inscribirse en el programa*.
- Sin embargo, es probable que los dos grupos sean, en esencia, diferentes. Aquellos individuos que decidieron participar pueden estar muy *motivados para mejorar sus vidas y quizá esperen un retorno alto de la formación*. Mientras que los que decidieron *no inscribirse* tal vez son *jóvenes desanimados* que no esperan beneficiarse de este tipo de programas. Es factible que estos dos grupos obtengan resultados bastante diferentes en el mercado laboral y consigan ingresos diferentes, incluso *sin el programa de formación profesional*.
- Los mismos problemas surgen cuando la admisión en un programa se basa en *preferencias no observadas* de los administradores del programa. Por ejemplo, si los administradores del programa basan la admisión y la inscripción *en una entrevista*.
 - a. Es posible que los admitidos en el programa sean aquellos en quienes los administradores ven una buena probabilidad de beneficiarse del programa.
 - b. Tal vez los no admitidos pueden mostrar menos motivación en la *entrevista*, tener calificaciones más bajas o sencillamente carecer de destrezas en una entrevista.
 - c. Como se señaló en el caso anterior, es probable que estos dos grupos de jóvenes, obtengan ingresos diferentes en el mercado laboral incluso *sin un programa de formación profesional*.
 - d. *Por lo tanto, el grupo que no se inscribió no proporciona una buena estimación del contrafactual* ya que la observación de una

diferencia en los ingresos entre los dos grupos no permite determinar *si se debe al programa de formación o a los contrastes subyacentes entre los dos grupos en motivación, destrezas y otros factores.*

- e. Así, el hecho de que *individuos menos motivados o cualificados no se hayan inscrito en el programa de formación genera un sesgo en la evaluación de impacto del programa.* Este sesgo se llama *sesgo de selección*, el cual se produce cuando los motivos por los que un individuo participa en un programa están *correlacionados con los resultados, incluso en ausencia del programa.*
- f. *Asegurarse de que el impacto estimado esté libre de sesgos de selección* es uno de los principales objetivos de cualquier evaluación de impacto, y plantea importantes dificultades.
- g. En este ejemplo, si los jóvenes que se inscribieron en la formación profesional, hubiesen tenido ingresos más altos incluso en *ausencia* del programa, *el sesgo de selección sería positivo*; en otras palabras, *se sobreestimaría el impacto del programa de formación profesional al atribuirle los ingresos más altos que los participantes habrían tenido de todas maneras.*

CAPÍTULO 3.

Aleatorización

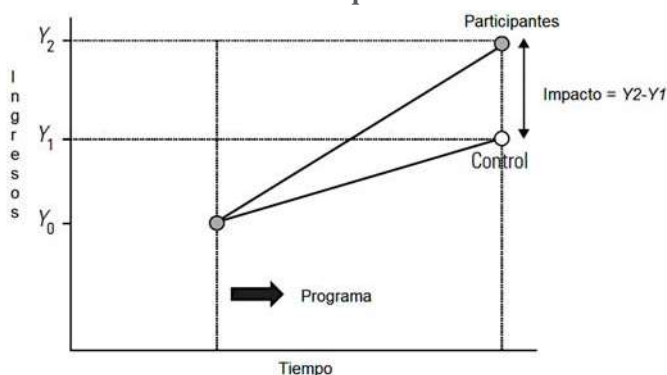
Después de analizar dos estimaciones *falsificadas del contrafactual* que se utilizan habitualmente, aunque con un alto riesgo de sesgo, las comparaciones antes-después y las comparaciones inscritos-no inscritos, a continuación se presentará un conjunto de métodos que se pueden aplicar para estimar con mayor precisión los impactos de un programa. Sin embargo, esas estimaciones no siempre son tan fáciles como puede parecer a primera vista. La mayoría de los programas se diseñan y posteriormente, se implementa en un entorno complejo y cambiante, donde diversos factores pueden influir en los resultados tanto de los participantes del programa como de aquellos que no participan. Imprevistos como sequías, recesiones, pandemias, los terremotos, los cambios de gobierno y los vaivenes nacionales e internacionales forman parte del mundo real. En una evaluación, se debe asegurar que la estimación del impacto del programa siga siendo válida a pesar de esta diversidad de factores. Las reglas de un programa para seleccionar a los participantes constituirán el parámetro clave para determinar el método de la evaluación de impacto.

Sin embargo, también se parte de la premisa de que todos los programas sociales deben tener reglas de asignación justas y transparentes. Una de las reglas más justas y transparentes para asignar recursos escasos entre poblaciones que los merecen de igual manera consiste en que todos aquellos que sean elegibles tengan la misma oportunidad de participar en el programa. Una manera sencilla de conseguirlo es mediante un *sorteo al azar o aleatorio* un método que decide quién participa en un programa en un determinado momento y quién no: el método de asignación aleatoria, también conocido como ensayo aleatorio controlado (**RCT**. *Randomized*

Control Trial). Este método no solo proporciona a los administradores del programa una regla imparcial y transparente para asignar recursos escasos entre poblaciones igualmente merecedoras de ellos, sino que también representa el método más sólido para evaluar el impacto de un programa. Por eso, la aplicación del mismo para evaluar los impactos de los programas ha aumentado de manera considerable en los últimos años (Gertler et al., 2017).

Como se ha visto, *encontrar un contrafactual* apropiado para el tratamiento, es el principal *desafío de evaluación de impacto* (Bernal y Peña, 2011; Gertler et al., 2017; Khandker et al., 2017). El *contrafactual* indica lo que le habría pasado a participantes de un programa si no hubieran participado. Sin embargo, *la misma persona no puede ser observada en dos situaciones distintas: ser tratado y no tratado al mismo tiempo*. El enigma principal, por lo tanto, es cómo los investigadores formulan *estados contrafactuales* del mundo en la práctica. En algunas disciplinas, como la ciencia médica, la evidencia sobre los *contrafactuales* se generan a través de *ensayos aleatorios*, que aseguran que los resultados en el *grupo de control* realmente capturan el *contrafactual* para un grupo de tratamiento. La **Figura 3.1** ilustra el caso de la aleatorización gráficamente.

Figura 3.1 Experimento ideal con un grupo de control equivalente



Fuente: Elaboración propia.

Considere una *distribución aleatoria* de dos grupos de sujetos o individuos *similares*: *un grupo recibe tratamiento y el otro grupo no recibe tratamiento*. Son *similares o equivalentes* en que ambos grupos, dado que se observa que tienen el mismo nivel de ingresos (en este caso, Y_0). Después de que se lleva a cabo el *tratamiento*, se descubre que los ingresos observados del *grupo tratado* son Y_2 mientras que el nivel de ingresos del *grupo de control* es Y_1 . Por lo tanto, el efecto de la intervención del programa, puede describirse como $(Y_2 - Y_1)$, como se indica en la **Figura 3.1**. Como se discutió en el capítulo 2, se debe tener *extremo cuidado al seleccionar el grupo de control* para garantizar la comparabilidad.

Asignación aleatoria como excelente estimación del contrafactual

El grupo de comparación ideal sería lo más similar posible al grupo de tratamiento en todos los sentidos, excepto con respecto a su participación en el programa que se evalúa (Gertler et al., 2017) Cuando se asignan unidades de manera aleatoria a los *grupos de tratamiento y de comparación*, ese proceso de asignación aleatoria producirá dos grupos que tienen una alta probabilidad de ser estadísticamente idénticos, siempre que el número de unidades potenciales a las que se aplica el proceso de asignación aleatoria sea suficientemente grande. Concretamente, con un gran número de unidades el proceso de asignación aleatoria producirá grupos que tienen *promedios estadísticamente equivalentes en todas sus características*. La **Figura 3.2** ilustra por qué la asignación aleatoria produce un *grupo de comparación* estadísticamente equivalente al *grupo de tratamiento*.

Supóngase que la población de unidades elegibles (el conjunto de participantes potenciales, o la población de interés para la evaluación) consiste en más de **1000 pbt** de desarrollo de software altamente inno-

vadores. Entonces, se asigna aleatoriamente la mitad al *grupo de tratamiento* y la otra mitad al *grupo de comparación* en una caja que contiene cada uno de sus nombres de los cuales se extraen los primeros **500**. Si los primeros **500** nombres constituyen el *grupo de tratamiento*, entonces tendríamos un grupo de tratamiento asignado de forma aleatoria (los primeros **500** números extraídos) y un grupo de comparación asignado también de manera aleatoria (los **500** nombres que quedaron en la caja). Ahora suponga, que el **40%** de las **1000** pymes originales tienen CEOs de género femenino; dado que los nombres se han asignado al azar, de los **500** nombres que se sacaron de la caja, alrededor **del 40%** serán CEOs femeninas. Si entre las **1000** pymes, el **20%** se dedica a desarrollo de software de apps, entonces casi el **20%** de ellas en los *grupos de tratamiento* y de *comparación* también deberían dedicarse al desarrollo de software de apps.

Figura 3.2. Características de los grupos bajo tratamiento con asignación aleatoria



Fuente: Gertler et al. (2017).

En general, si la población de unidades elegibles es lo suficientemente grande, el mecanismo de *asignación aleatoria* asegura que cualquier característica de la población *se transfiera tanto al grupo de tratamiento como al grupo de comparación*. Del mismo modo que las características observables, como el tamaño de la empresa o el giro al que se dedica la pyme, se transfieren tanto al *grupo de tratamiento como*

al de comparación, es lógico pensar que las características que son más difíciles de observar (*variables no observables*), como la motivación, el estilo de liderazgo, u otros rasgos de competitividad de una pyme, que son complejos de medir, también se aplicarán por igual al *grupo de tratamiento y al de comparación*. Así, los *grupos de tratamiento y comparación* generados a través de la *asignación aleatoria* serán similares no solo en cuanto a sus características observables sino también en relación con las no observables.

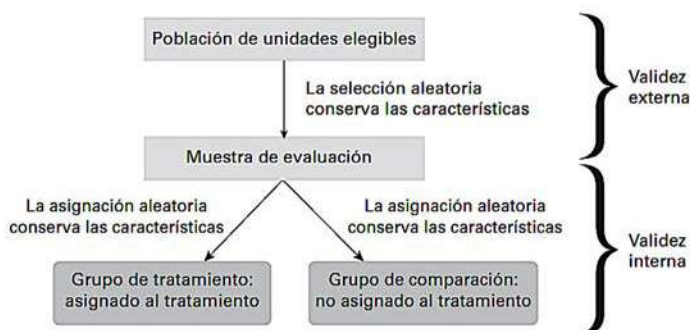
Tener dos grupos similares en todos los aspectos asegura que la estimación del contrafactual se aproxime al valor verdadero del resultado en ausencia de tratamiento, y que una vez que el programa se haya implementado, las estimaciones de impacto no sufrirán un sesgo de selección. Cuando una evaluación utiliza la asignación aleatoria para generar los grupos de tratamiento y de comparación, en teoría, el proceso debería producir dos grupos equivalentes, siempre que se cuente con un número de unidades lo suficientemente grande. Con los datos de *línea de base* de la muestra de evaluación con la que se cuente, se podrá comprobar empíricamente este supuesto y verificar que, de hecho, no hay diferencias sistemáticas en las características observables entre los grupos de tratamiento y de comparación antes del inicio del programa. Posteriormente, si después de lanzar el programa se observan diferencias en los resultados entre los grupos de tratamiento y comparación, sabremos que esas diferencias se deben únicamente a la incidencia del programa, dado que los dos grupos eran idénticos en la *línea de base*, antes del inicio del programa, y que están expuestos a los mismos factores externos a lo largo del tiempo. En este sentido, el grupo de comparación contiene todos los factores que también pueden explicar el resultado de interés.

Para estimar el impacto de un programa bajo la asignación aleatoria, se debe observar la diferencia entre el resultado bajo tratamiento (el resultado medio del grupo de tratamiento asignado de forma aleatoria) y nuestra estimación del contrafactual (el resultado medio del grupo de comparación asignado de manera aleatoria). Así, podemos confiar en que

nuestro impacto estimado constituye el verdadero impacto del programa, puesto que se han eliminado todos los factores observados y no observados que, de otra manera, podrían explicar la diferencia en los resultados.

En el **Gráfico 3.3** se presupone que todas las unidades de la población elegible serán asignadas ya sea al *grupo de tratamiento* o al *grupo de comparación*. Sin embargo, en algunos casos no es necesario incluir todas las unidades en la evaluación. Por ejemplo, si la población de unidades elegibles es de 1 millón de pymes de base tecnológica (**pbt**) y se quiere evaluar la efectividad de los bonos en efectivo en la probabilidad de que estas pymes adopten una tecnología disruptiva, bastará con seleccionar una muestra aleatoria representativa de, por ejemplo, **1000** pymes y asignarlas ya sea al *grupo de tratamiento* o de *comparación*. La **Figura 3.3** ilustra este proceso.

Figura 3.3. Muestra y asignación aleatorias de tratamiento



Fuente: Gertler et al. (2017).

Según la misma lógica que se detalló más arriba, la selección de una muestra aleatoria de la población de unidades elegibles para formar la muestra de evaluación conserva las características de la población de las unidades elegibles. Dentro de la muestra, la asignación aleatoria de individuos a los grupos de tratamiento y comparación también conserva dichas características.

Validez externa e interna

Los pasos de la *selección aleatoria* del tratamiento, que ya se han detallado, aseguran tanto la *validez interna como externa* de las evaluaciones de impacto (**Gráfico 3.3**), siendo (**Gertler et al., 2017**):

- *Validez interna* significa que el impacto estimado del programa es el impacto libre de todos los demás factores de confusión potenciales (o, en otras palabras, que el grupo de comparación represente una estimación precisa del contrafactual de modo que se estime el verdadero impacto del programa). Hay que recordar que la asignación aleatoria produce un grupo de comparación que es estadísticamente equivalente al grupo de tratamiento en la *línea de base*, antes de que empiece el programa. Una vez que el programa comienza, el grupo de comparación está expuesto al mismo conjunto de factores externos que el grupo de tratamiento a lo largo del tiempo, con la única excepción del propio programa. Por lo tanto, si aparece cualquier diferencia en los resultados entre los grupos de tratamiento y de comparación, solo puede deberse a la existencia del programa en el grupo de tratamiento. La validez interna de una evaluación de impacto se asegura a través del proceso de asignación aleatoria del tratamiento.
- *Validez externa* quiere decir que la muestra de la evaluación representa con precisión a la población de unidades elegibles. Los resultados de la evaluación se pueden entonces generalizar a la población de unidades elegibles. Se utiliza el muestreo aleatorio para asegurar que la muestra de la evaluación refleje adecuadamente la población de unidades elegibles, de modo que los impactos identificados en la muestra de la evaluación pueden extrapolarse a la población.

Nótese que se ha realizado un proceso de selección aleatoria con dos objetivos diferentes: selección aleatoria de una muestra (para la validez externa), y asignación aleatoria del tratamiento como método de evaluación de impacto (para la validez interna). Una evaluación de impacto puede producir estimaciones internamente válidas del impacto mediante una asignación aleatoria del tratamiento; sin embargo, si la evaluación se lleva a cabo con una muestra no aleatoria de la población, puede que los impactos estimados no sean generalizables para el conjunto de unidades elegibles.

Al contrario, si la evaluación utiliza una muestra aleatoria de la población de unidades elegibles, pero el tratamiento no se asigna de manera aleatoria, la muestra sería representativa pero el grupo de comparación puede no ser válido, lo cual pone en entredicho la validez interna. En algunos contextos, puede que los programas se enfrenten a limitaciones que exigen un equilibrio entre *validez interna* y *externa*.

Diseño estadístico aleatorio

En la práctica, sin embargo, puede ser muy difícil garantizar que un grupo de control sea muy similar para proyectar las áreas, para que los efectos del tratamiento observados en la muestra, sean generalizables, y que los efectos en sí mismos, son función del programa en sí (**Khandker et al., 2017**).

Cuando se asigna de forma aleatoria a los beneficiarios de un programa es decir, mediante sorteo, entre una población elegible numerosa, se puede generar una *estimación robusta del contrafactual*. La asignación aleatoria del tratamiento se considera la regla de oro de la evaluación de impacto. Utiliza un proceso aleatorio, o el azar, para decidir a quién se le concederá acceso al programa y a quién no. La *asignación aleatoria del tratamiento* también suele denominarse también:

- Ensayo aleatorio controlado,
- Evaluaciones aleatorias,
- Evaluaciones experimentales y
- Experimentos sociales, entre otras denominaciones.

Estrictamente hablando, *un experimento no tiene que identificar impactos mediante asignaciones aleatorias*, pero los evaluadores suelen utilizar el término *experimento* solo cuando la evaluación recurre a la asignación aleatoria (Gertler et al., 2017).

En la *asignación aleatoria*, todas las unidades elegibles (por ejemplo, una persona, un hogar, una empresa, un hospital, una escuela o una comunidad) tienen la misma probabilidad de ser seleccionadas para un programa. Nótese que esta probabilidad no necesariamente significa una probabilidad del 50% de ganar el sorteo. En la práctica, la mayoría de las evaluaciones con asignación aleatoria, darán a cada unidad elegible una probabilidad de selección determinada, de manera que el número de ganadores (*tratamientos*) sea igual al total de beneficios disponibles. Por ejemplo, si un programa tiene suficientes fondos para servir solo a **1000** pymes de base tecnológica (**pbt**), de un ecosistema de **10000** pymes elegibles, cada pyme tendrá una probabilidad de una entre **10** de ser seleccionada para el tratamiento. La *potencia estadística*, se maximizará cuando la muestra de evaluación se divida por igual, entre los grupos de tratamiento y comparación. En el ejemplo de este caso, para un tamaño total de la muestra de **2000** pymes, la potencia estadística se maximizará si se seleccionan las **1000 pbt** de tratamiento y una submuestra de **1000 pbt** de comparación, en lugar de tomar una muestra aleatoria simple del **20%** de las **10000 pbt** originales elegibles (lo que produciría una muestra de evaluación de alrededor de **200 pbt** de tratamiento y **1800 pbt** de comparación). (Gertler, et al., 2017).

Los estadísticos han propuesto, un *enfoque de aleatorización en dos etapas* que describe estas prioridades (Khandker et al., 2017):

- En la primera etapa, se selecciona aleatoriamente una muestra de posibles participantes de la *población relevante*. Esta muestra *debe ser representativa de la población*, dentro de un *cierto error de muestreo*. Esta etapa asegura la validez externa (*external validity*) del experimento.
- En la segunda etapa, los individuos en esta muestra son asignados aleatoriamente al tratamiento y a los *grupos de comparación*, asegurando la validez interna (*internal validity*) en aquellos cambios subsecuentes de los resultados de interés (*outcomes*) medidos, que sean debidos al programa en lugar de otros factores.
- Las condiciones para asegurar la validez externa e interna del diseño aleatorio, se discuten más adelante.

La *asignación aleatoria*, también es una manera justa y transparente de asignar los escasos recursos de un programa. Una vez que se ha definido una población objetivo (por ejemplo, pymes por debajo de la línea de rentabilidad), la *asignación aleatoria* es una regla de asignación justa porque permite que los administradores del programa *se aseguren de que todas las unidades elegibles tengan la misma probabilidad de participar en el programa* y de que el programa no sea asignado con criterios arbitrarios o subjetivos, ni por cuestiones de favoritismo u otras prácticas injustas. Cuando se produce un exceso de demanda de un programa, la asignación aleatoria es una regla que los administradores del mismo pueden explicar fácilmente, que todas las partes interesadas pueden entender y que se considera justa en numerosas circunstancias. Además, cuando el proceso de asignación se lleva a cabo de modo abierto y transparente, no es fácil manipularlo y, por lo tanto, protege a los administradores del programa de posibles acusaciones de favoritismo o corrupción. Por lo tanto, como mecanismo de asignación, la asignación aleatoria tiene sus propios méritos, que van mucho más allá de su utilidad como instrumento de evaluación de impacto. De hecho, diversos programas utilizan de manera rutinaria los sorteos como una forma de seleccionar a los parti-

cipantes del conjunto de individuos elegibles, sobre todo debido a sus ventajas administrativas y de gobernanza (**Gertler, et al., 2017**).

La decisión de participar en un programa depende de las características de los individuos (Bernal y Peña, 2011). En un experimento donde se asigna de manera aleatoria a los *potenciales beneficiarios* a los *grupos de tratamiento y control*, y no se les permite elegir si participan o no, se asegura que los individuos de los grupos de tratamiento y control tengan características idénticas. Aparte del *error de muestreo*, los *grupos de tratamiento y control deben ser idénticos ex ante*, es decir, no pueden diferir de manera sistemática en variables observables ni de las cuales no tenemos información. El efecto causal del programa es entonces la diferencia en las medias de las variables de interés entre los *grupos de tratamiento y control*. Los *experimentos* se usan con frecuencia en otras ciencias, por ejemplo, para determinar la efectividad de una nueva medicina. Los individuos, sin su conocimiento, son asignados aleatoriamente a un grupo de tratamiento, que recibe la medicina, y a un grupo de control, que recibe un placebo o la mejor medicina disponible en el mercado. La comercialización de la droga se aprueba si los resultados del experimento sugieren que es efectiva y segura. Los resultados de los experimentos, al estar *libres de sesgo de selección*, son muy deseables para determinar la efectividad de las políticas de introducción de innovaciones y programas. Además, pueden usarse como *punto de referencia* para comparar los efectos causales de otras intervenciones. Por ejemplo, antes de hacer una gran intervención por medio de un programa de introducción de nuevas tecnologías a pymes de base tecnológica (**pbt**), sería deseable estimar el impacto mediante la *evaluación de un piloto*, con el objeto de saber si funciona y si es *rentable* en términos sociales. Si, en cambio, hay un programa que lleva funcionando un tiempo y no ha sido objeto de evaluación, se puede medir el impacto para hacer ajustes a la política de introducción de la innovación. Por ejemplo, para decidir si se continúa con el programa, se reforma o se acaba completamente.

Cuando aplicar la asignación aleatoria

La asignación aleatoria puede utilizarse como regla de asignación de un programa en dos escenarios específicos (**Gertler et al., 2017**):

1. *Cuando la población elegible es mayor que el número de plazas disponibles del programa.* Cuando la demanda de un programa supera a la oferta, se puede utilizar un sorteo para seleccionar el grupo de tratamiento dentro de la población elegible. En este contexto, todas las unidades de la población tienen la misma probabilidad (o una probabilidad conocida superior a 0 e inferior a 1) de ser seleccionadas para el programa. El grupo que gana el sorteo es el grupo de tratamiento y el resto de la población a la que no se ha ofrecido el programa es el grupo de comparación. Siempre que exista una limitación que impida ampliar la escala del programa a toda la población, se pueden mantener los grupos de comparación para medir los impactos del programa a corto, mediano y largo plazo. En este contexto, no hay un dilema ético en mantener indefinidamente un grupo de comparación, ya que un subgrupo de la población quedaría necesariamente excluido del programa debido a problemas de capacidad.
2. *Cuando sea necesario ampliar un programa de manera progresiva hasta que cubra a toda la población elegible.* Cuando un programa se extiende por etapas, establecer de forma aleatoria el orden en el que los participantes se benefician del mismo ofrece a cada unidad elegible la misma posibilidad de recibir tratamiento en la primera fase o en una fase posterior. Siempre que no se haya sumado todavía el último grupo al programa, este sirve como grupo de comparación válido a partir del cual se podrá estimar el contrafactual para los que ya se han incorporado. Esta configuración también puede permitir que la evaluación recoja los efectos de una exposición diferencial al tratamiento, es decir, el efecto de recibir un programa durante un período más o menos prolongado.

Hay variantes del diseño básico del *experimento para introducción de innovaciones* que hacen que su adopción sea más fácil (**Bernal y Peña, 2011**). Por un lado, como algunas intervenciones son muy populares, la demanda excede los cupos disponibles. Así, se podría pensar en asignar los cupos aleatoriamente entre los solicitantes. La asignación aleatoria *entre tratamiento y control* podría hacerse a *nivel sujeto* (por ejemplo, a nivel individuo, CEO, familia de hogares), o a nivel de *conglomerado* (por ejemplo, cluster, comunidad, etc). Cuando no es posible asignar el tratamiento de manera aleatoria por *razones éticas o prácticas*, se puede incentivar a un subgrupo de la población objetivo, elegido de manera aleatoria a participar en el programa.

Desventajas de la aleatorización

Como las técnicas usadas para calcular el impacto del programa, usando datos experimentales son sencillas, sus resultados son transparentes para los políticos en introducción de innovaciones, diseñadores de las mismas y el público en general. Además, es difícil manipular los resultados de la evaluación, ya que el evaluador no puede elegir estratégicamente el método de estimación para generar los resultados deseados (**Bernal y Peña, 2011**). A pesar de las ventajas descritas, los *experimentos sociales* no son una solución fácil de todos los problemas evaluativos y sufren de diversas limitaciones. Hacer una correcta evaluación controlado es costoso monetaria y políticamente.

Hay además una injusticia fundamental: por pertenecer al *grupo de control*, se está excluyendo a un segmento de la población, igualmente vulnerable, de los beneficios de la intervención. Como no se pueden negar los beneficios del programa durante largos períodos, *eventualmente el grupo de control entra a formar parte del grupo de tratamiento*. Así, aunque se sabe que los efectos de corto y largo plazo con frecuencia difieren sustancialmente (en un experimento de este tipo, de introducción de innovaciones de corte social, no se pueden negar los bene-

ficios del programa en el largo plazo), entonces no se pueden estudiar dichos efectos. También puede haber externalidades o efectos de equilibrio general. Las externalidades son los efectos externos (esperados o no) generados por el programa, mientras que los efectos de equilibrio general son los efectos que puede tener la intervención sobre su entorno. Finalmente, por las características de los experimentos de introducción e innovaciones de corte social, solamente se puede evaluar el programa tal como está diseñado, y no se puede evaluar el impacto de variaciones en el programa existente (*análisis contrafactual*).

Si tanto el diseño como la *implementación de la aleatorización son adecuados*, los resultados son *internamente válidos*, es decir, el impacto estimado es generado por la intervención. Uno de los objetivos fundamentales de los experimentos de introducción de innovaciones de orientación social, es la posibilidad de generalizar los efectos estimados del programa a un grupo más grande de la población. Sin embargo, diversos aspectos pueden amenazar la validez externa del experimento. El experimento debe ser suficientemente similar al programa para poder generalizar los resultados a la población de interés; tanto la población estudiada como el programa implementado deben ser suficientemente similares a la población y al programa de interés.

Los efectos de equilibrio general y externalidades también pueden afectar la *validez externa*. Finalmente, dado que con frecuencia la *participación en los programas es voluntaria*, el efecto de un programa medido sobre una población elegida aleatoriamente puede no ser un estimador *insesgado* del efecto del programa real. Para aliviar este problema se puede aleatorizar entre los individuos que han manifestado su interés en participar, pero normalmente esto es más costoso desde un punto de vista político. Tres aspectos fundamentales afectan la posibilidad de generalizar los resultados de una evaluación aleatoria en otros contextos:

- Primero, los *experimentos modernos* son con frecuencia de pequeña escala y manejados con extremo cuidado, por ejemplo, operados

por los actores más experimentados de los programas. Por tanto, se teme que este estándar de implementación no se pueda replicar al escalar el programa. Para solucionar esto es necesario documentar adecuadamente la implementación del experimento para poder replicarlo.

- Segundo, dado que los *experimentos sociales* se realizan en regiones específicas, ¿se puede concluir que, dado que una población particular respondió de una manera al programa, otra población reaccionará de una manera similar a un programa parecido? No necesariamente. Para solucionar esto se puede, o bien realizar el mismo experimento en diversas regiones, o bien usar modelos comportamentales que determinen el canal mediante el cual se genera el efecto, para predecir si es posible replicar el éxito del programa.
- Tercero, una situación similar se da cuando nos preguntamos por el *efecto de variaciones* en el programa. Una posibilidad interesante es entonces combinar experimentos sociales con modelos económicos, bien sea para estimar un mayor número de parámetros, o para determinar los canales mediante los cuales se generan los efectos.

Si la aleatorización no es completamente exitosa, *el estimador de diferencias es sesgado* y necesitamos utilizar otras técnicas econométricas para estimar el efecto del programa. Pero, como se discutió anteriormente, *incluso si la aleatorización es exitosa, los resultados de los experimentos sociales sufren diversos problemas, tales como potencial falta de validez externa, efectos Hawthorne o John Henry*, no tienen en cuenta efectos de equilibrio general, comúnmente miden efectos de corto plazo, entre otros.

Cálculo de los efectos del tratamiento

La aleatorización, puede corregir el sesgo de selección (**B**), discutido en el capítulo 2, por asignación al azar, de individuos o grupos de los mismos a *grupos de tratamiento y control*. Volviendo para la configuración en el capítulo 2, considere el clásico problema de medir los efectos del tratamiento (**Imbens y Angrist, 1994**): deje que el tratamiento, $T_i = (1)$ si el sujeto i es tratado y (0) si no. Suponga que $Y_i(1)$ sea el resultado (*outcome*) bajo tratamiento y $Y_i(0)$ si no hay tratamiento.

Observe que Y_i y T_i , donde de acuerdo a **Heckman y Vytalil (2000)**, una caracterización de (Y) es identificada bajo diferentes aproximaciones, como la conocida como *modelo Neyman-Fisher-Cox-Rubin* de resultados de interés (*outcomes*) potenciales; también referida como modelo de switcheo de **Quandt** (*switching regression model of Quandt, Quandt, 1972*) y el modelo de distribución de ingreso de Roy (**Roy, 1951**). Así, se tiene:

$$Y_i = [T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0)]$$

Estrictamente hablando, el *efecto del tratamiento por unidad i es* $(1) Y_i - (0)$ y el $ATE = E[Y_i(1) - Y_i(0)]$ o la diferencia en los resultados (*outcomes*) de estar en un proyecto en relación con el área de control para una persona o unidad i extraída al azar de la población. Esta formulación supone, por ejemplo, que todos en la población tienen la misma probabilidad de ser tomados en cuenta.

Generalmente, sin embargo, solo $E[Y_i(1) | T_i = 1]$ los resultados (*outcomes*) promedio de los *tratados*, condicional a estar en un *área de tratamiento*, y $E[Y_i(0) | T_i = 0]$ los resultados (*outcomes*) promedio de los *no tratados*, condicionados a no estar en un área tratada, son observados. Con un *enfoque no aleatorio* y observaciones solo en una submuestra de la población $E[Y_i(1)]$, no es necesariamente igual a $E[Y_i(1) | T_i = 1]$, y $E[Y_i(0)]$ no es necesariamente igual a $E[Y_i(0) | T_i = 0]$.

Por lo general, se observan *efectos de tratamiento* alternativos en forma de $TOT = E [Y_i(1) - Y_i(0) / T_i = 1]$, o la diferencia en los resultados (*outcomes*) de recibir el programa en comparación con estar en un área de control para una persona o sujeto i al azar extraído de la muestra tratada. Es decir, el **TOT** refleja las ganancias promedio de los participantes, condicional a que estos participantes reciban el programa. Supongamos que el área de interés es el **TOT**, $E [Y_i(1) - Y_i(0) / T_i = 1]$. Si T_i es no aleatorio una simple diferencia entre áreas tratadas y de control $D = E [Y_i(1) / T_i = 1] - E [Y_i(0) / T_i = 0]$, no será igual a **TOT**. La discrepancia entre **TOT** y **D**, será $E [Y_i(0) / T_i = 1] - E [Y_i(0) / T_i = 0]$, el cual es igual al sesgo **B** al estimar el *efecto del tratamiento*:

$$TOT = E [Y_i(1) - Y_i(0) / T_i = 0] = \quad (3.1)$$

$$E [Y_i(1) / T_i = 1] - E [Y_i(0) / T_i = 1] = \quad (3.2)$$

$$D = E [Y_i(1) / T_i = 1] - E [Y_i(0) / T_i = 0] \text{ Si } E [Y_i(0) / T_i = 0] =$$

$$E [Y_i(0) / T_i = 1] \quad (3.3.)$$

$$TOT = D \text{ Si } B=0 \quad (3.4)$$

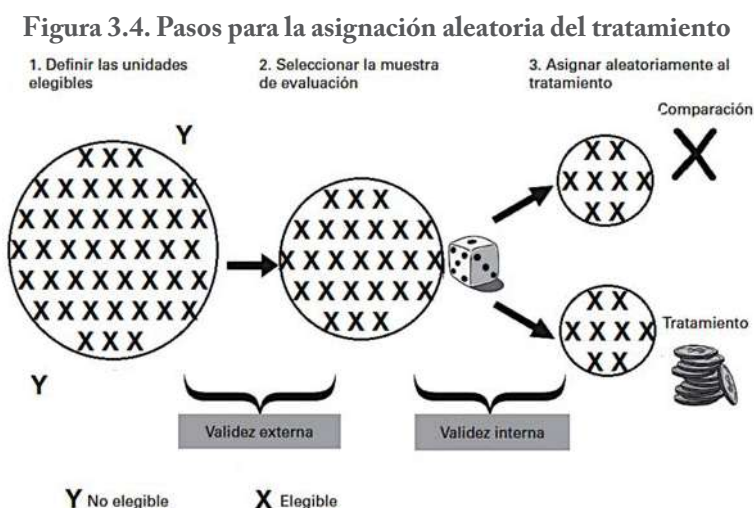
Aunque en principio el resultado (*outcome*) *contrafactual* $E [Y_i(0) / T_i = 1]$ en la **ecuación 3.2** no se puede observar directamente para comprender el alcance del *sesgo*, todavía hay algo de intuición al respecto que pudiera existir.

En un esfuerzo por unificar la literatura de los *efectos del tratamiento*, **Heckman y Vytlačil (2005)**, también describen un parámetro llamado *efecto de tratamiento marginal* (**MTE**. *Marginal Treatment Effect*), del cual el **ATE** y **TOT** pueden ser derivados. (**Björklund y Moffitt, 1987**). El **MTE** es el cambio promedio en los resultados (*outcomes*) Y_i para individuos que están al margen de participar en el programa, dado un conjunto de características observadas X_i y condicionamiento en un conjunto de características no observadas U_i en la ecuación de participación: **MTE** = $E [Y_i(1) - Y_i(0) | X_i = x, U_i = u]$. Es decir, el **MTE** es el *efecto promedio*

del programa para individuos que simplemente son indiferentes entre participantes y no participantes.

Cómo asignar aleatoriamente el tratamiento

De acuerdo a **Gertler et al. (2017)**, se puede verificar el proceso como se observa en la **Figura 3.4**.



Fuente: Gertler et al. (2017) con adaptación propia.

Y se enlista a continuación:

1. *Definir las unidades elegibles para el programa.* Cabe recordar que, dependiendo del programa concreto, una unidad podría ser una persona, una pyme, un centro de salud, una empresa, pueblo, municipalidad, etc. La población de unidades elegibles está compuesta por aquellos para los cuales interesa conocer el impacto de un programa.

2. *Comparar el tamaño del grupo con el número de observaciones requeridas para la evaluación.* El tamaño de la muestra de la evaluación se establece mediante cálculos de la potencia y se basa en el tipo de preguntas a las que el evaluador desearía que se respondiera. Si la población elegible es pequeña, quizás haya que incluir todas las unidades elegibles en la evaluación. Por el contrario, si hay más unidades elegibles de las que se requiere para la evaluación, entonces el segundo paso consiste en seleccionar una muestra de unidades a partir de la población que se incluirá en la muestra de evaluación. Este segundo paso responde sobre todo a la necesidad de limitar los costos de la recopilación de datos. Si se observa que los datos de los sistemas de monitoreo existentes se pueden usar para la evaluación, y que esos sistemas abarcan al conjunto de unidades elegibles, tal vez no sea necesario elaborar una muestra distinta de la evaluación.
3. Configurar los grupos de tratamiento y de comparación a partir de las unidades de la muestra de la evaluación, mediante la asignación aleatoria. En los casos en que la asignación aleatoria tenga que llevarse a cabo en un foro público, por ejemplo en la televisión, es posible que sea necesario utilizar una técnica sencilla, sugiriéndose:
 - a. Si se quiere asignar el **50%** de los individuos al grupo de tratamiento y **50%** al grupo de comparación, hay que lanzar la moneda para cada persona. Hay que decidir con antelación cuál cara de la moneda asignará una persona al grupo de tratamiento.
 - b. Si se quiere asignar una *tercera parte* de la muestra de la evaluación al grupo de tratamiento, se puede tirar un dado para cada persona. Antes, hay que decidir una regla, por ejemplo, si el dado muestra **1 o 2**, el individuo será asignado al grupo de tratamiento, mientras que si arroja un **3, 4, 5 o 6** el individuo será derivado al grupo de comparación. El dado se tiraría una vez para cada

persona en la muestra de evaluación, y se la asignaría sobre la base del número del dado.

- c. Escribir los nombres de todos los individuos en trozos de papel de igual tamaño y forma. Plegar los papeles de modo que no se puedan leer los nombres y mezclarlos de manera conveniente en una caja o en algún otro recipiente. Antes de empezar a sacar los nombres, debe decidirse la regla, es decir, cuántos trozos de papel se extraerán, y que extraer un nombre significa asignar a esa persona al grupo de tratamiento. Una vez que la regla esté clara, se debe solicitar a alguien del público (una persona imparcial, como un niño) que extraiga tantos trozos de papel como participantes se requiera en el grupo de tratamiento.

Si tienen que asignarse muchas unidades (por ejemplo, más de **100**), utilizar enfoques sencillos como los descritos requerirá demasiado tiempo y habrá que utilizar un proceso automatizado. Para ello, primero habrá que decidir una regla de modo que se asignen los participantes sobre la base de números aleatorios. Para implementar la aleatoriedad, se asignará un número al azar a cada unidad en la muestra de evaluación, utilizando un generador aleatorio de números en una hoja de cálculo, *o en un programa estadístico especializado* y se utilizará la regla ya definida para formar los *grupos de tratamiento y comparación*. Es importante decidir la regla antes de generar los números al azar. De otra manera, puede que el evaluador se vea tentado de usar una regla basada en los números aleatorios que ve, lo que invalidaría la asignación aleatoria. La lógica en que se fundamenta el proceso automatizado no es diferente de la asignación aleatoria basada en lanzar una moneda o extraer nombres de un sombrero. Se trata de un mecanismo que asigna al azar si cada unidad pertenece *al grupo de tratamiento o de comparación*. Al utilizar un sorteo público, dados o números al azar generados por computador, es importante documentar el proceso para asegurar que sea transparente, así:

- En primer lugar, eso significa que la regla de asignación debe decidirse con antelación y comunicarse al público.
- En segundo lugar, el evaluador debe ceñirse a la regla una vez que se extraen los números al azar.
- En tercer lugar, debe demostrarse que el proceso era realmente aleatorio. En el caso de los sorteos y el lanzamiento de dados, se puede grabar el proceso en video; la asignación mediante números al azar generados por computador requiere que se presente un registro de los cálculos, de modo que el proceso pueda ser auditado.

Efecto del tratamiento con aleatorización pura

La aleatorización se puede configurar de dos maneras: *aleatorización pura* y *aleatorización parcial*. Si el tratamiento se realizó de forma *puramente aleatoria* siguiendo el *procedimiento en dos etapas* descrito anteriormente, para después, los sujetos *tratados* y *no tratados* tendrían el mismo resultado (*outcome*) esperado en ausencia del programa. Entonces $E[Y_i(0) | T_i = 1]$ es igual a $E[Y_i(0) | T_i = 0]$. Debido a que el tratamiento debería ser aleatorio y no una función de características *no observadas* (como la personalidad, el carácter o los gustos del sujeto) entre los individuos, los resultados (*outcomes*) no se espera que haya variaciones para los dos grupos, si la intervención no hubiera existido. *Por lo tanto, el sesgo de selección se convierte en cero en el caso de la aleatorización.*

Considere el caso de la *aleatorización pura*, donde una muestra de individuos o empresas, se extrae al azar de la población de interés. La muestra experimental posteriormente, se divide al azar en dos grupos:

- El grupo de tratamiento que está expuesto a la intervención del programa y
- El grupo de control que no recibe el programa. En términos de una regresión, este ejercicio se puede expresar como:

$$Y_i = \alpha X_i + \beta T_i + \varepsilon_i \quad (3.5)$$

Donde T_i es el tratamiento ficticio (*dummy*) igual a **(1)** si la unidad i se trata aleatoriamente y **(0)** en caso contrario. Como se dijo, Y_i se define como:

$$Y_i \cong [Y_i(1) \cdot T_i] + [Y_i(0) \cdot (1 - T_i)] \quad (3.6)$$

Si el *tratamiento es aleatorio* (entonces T y ε son independientes), la **ecuación 3.5** puede estimarse mediante el uso de mínimos cuadrados ordinarios (**OLS. Ordinary Least Squares**) y el efecto del tratamiento estimados β_{OLS} estima la diferencia en los resultados del grupo tratado y el grupo control. Si una evaluación aleatorizada está correctamente diseñada e implementada, entonces podrá ser determinada una estimación de la evaluación de impacto de programa sin sesgo, es decir, imparcial.

Efecto del tratamiento con aleatorización parcial

*Sin embargo, una aleatorización pura es extremadamente rara de realizar. Más bien, aleatorización parcial es la que se utiliza, donde las muestras de tratamiento y control se eligen al azar, condicional en algunas características observables de X (por ejemplo, edad o ingresos). Si el investigador puede hacer una suposición llamada *exogeneidad condicional de la colocación del programa* (conditional exogeneity of program placement), el investigador puede encontrar una estimación *sin sesgo* de la estimación del programa.*

Aquí, se sigue el modelo de **Ravaillon (2008)**. Denotando por simplicidad $Y_i(1)$ como Y_i^T y $Y_i(0)$ como Y_i^C , la **ecuación 3.5** podría aplicarse a una submuestra de participantes y no participantes como sigue:

$$Y_i^T = \alpha^T + X_i \beta^T + \mu_i^T \text{ if } T_i = 1, i = 1, \dots, n \quad (3.7)$$

$$Y_i^C = \alpha^C + X_i \beta^C + \mu_i^C \text{ if } T_i = 0, i = 1, \dots, n^* \quad (3.8)$$

Es una práctica común, estimar lo anterior como una regresión única al agrupar los datos para los *grupos de control y tratamiento*. Se puede multiplicar la **ecuación 3.7** por T_i , multiplique la **ecuación 3.8** por $(1 - T_i)$ y usar la identidad en la **ecuación 3.6** para obtener:

$$Y_i = \alpha^C + (\alpha^T - \alpha^C)T_i + X_i\beta^C + X_i(\beta^T - \beta^C)T_i + \varepsilon_i, \quad (3.9)$$

Donde:

$$\varepsilon_i = T_i(\mu_i^T - \mu_i^C) + \mu_i^C$$

El efecto del tratamiento de la **ecuación 3.9** se puede escribir como:

$$A^{TT} = E(Y_i | T_i = 1, X) = E[\alpha^T - \alpha^C + X_i(\beta^T - \beta^C)]$$

Que es el *efecto justo del tratamiento* del grupo tratado, **TOT**, discutido anteriormente. Para la **ecuación 3.9**, se puede obtener una estimación consistente del efecto del programa con **OLS**:

$$E(\mu_i^T | X, T = t) = E(\mu_i^C | X, T = t) = 0, \quad t = \{0, 1\}$$

Es decir, *no hay sesgo de selección* debido a la aleatorización. En la práctica, un modelo de impacto común, es a menudo usado que supone:

$$\beta^T = \beta^C$$

El **ATE** es entonces, simplemente:

$$\alpha^T - \alpha^C$$

Diferentes métodos de aleatorización

Si la aleatorización fuera posible, habría que tomar una decisión sobre qué tipo de aleatorización se utilizaría. Estos enfoques, detallados en **Duflo et al. (2008)** se analizan a continuación:

- *Oversubscription*. Si los recursos limitados, repercuten en el programa, la implementación puede ser asignados aleatoriamente a través de un subconjunto de participantes elegibles, y el resto de los sujetos elegibles quienes no reciben el programa, pueden considerarse *grupo de control*. Se debe hacer un examen del presupuesto, evaluando cuántos sujetos podrían ser encuestado vs. aquellos realmente dirigidos, para obtener un control lo suficientemente grande de grupo como muestra de beneficiarios potenciales.
- *Randomized phase-in*. Este enfoque pasa gradualmente por el programa a través de un conjunto de áreas elegibles, de modo que los *grupos de control*, representan áreas elegibles que aún esperan recibir el programa. Este método ayuda a aliviar los problemas de equidad y aumenta la probabilidad que las áreas de programa y grupos de control sean similares en características observadas.
- *Within-group randomization*. En un enfoque de introducción gradual aleatorizado, sin embargo, si el retraso entre la génesis del programa y la recepción real de los beneficios es grande, una gran controversia puede surgir acerca de qué área o áreas deben recibir primero el programa. En ese caso, todavía se puede introducir un elemento de aleatorización proporcionando el programa, a algunos subgrupos en cada área objetivo. Este enfoque, es por lo tanto similar a la *randomization phased-in* en una escala más pequeña. Existen problemas de efectos indirectos.
- *Encouragement design*. En lugar de aleatorizar el tratamiento, los investigadores al azar, asignan un anuncio o incentivo a los sujetos para participar en el programa. Algunos anuncios del programa se

dan por adelantado (ya sea antes de que se implemente el programa o durante el tiempo que dura este, conservar recursos) a un subconjunto aleatorio de beneficiarios elegibles. Este aviso puede ser utilizado como un instrumento para incorporarse al programa. Los efectos indirectos, también se pueden medir en este contexto, si se recopilan datos en las redes sociales de los sujetos que reciben el aviso, para ver cómo la aceptación puede diferir entre los sujetos que están conectados o no. Tal caso, requeriría una recopilación más intensiva de datos.

Intervención individual vs. conglomerados

Suponga que un programa de estímulos a la innovación fuera a implementarse en el país, como un experimento social. Para **Bernal y Peña (2011)** en este caso, considera elegir de uno de dos diseños básicos:

- *Aleatorización a nivel individual.* Se conforma un listado de los sujetos elegibles. De la lista de elegibles, se escogen aleatoriamente unos sujetos que harán parte del programa (por ejemplo pymes, hogares, individuos, etc.) y el resto de los sujetos, no serán beneficiarios del programa por lo cual constituirán el *grupo de control*.
- *A nivel de conglomerados (clusters).* Se asigna de manera aleatoria la participación en el programa, pero no a *nivel de sujeto o individuo*, sino desde barrios o comunidades hasta grupos especializados (clústers). Es posible aprovechar las restricciones logísticas o presupuestales típicas en los gobiernos, o las restricciones en la capacidad operativa del programa para aleatorizar el orden de entrada de las diferentes regiones, estados, municipios al programa. En una asignación a nivel de *conglomerados*, se pospone la entrada de algunos grupos, de tal manera que actúen como *grupo de control*. La *aleatorización por conglomerados* es una buena alter-

nativa, cuando es difícil negar los beneficios de una intervención a una fracción de los miembros de una comunidad, pues pueden ser tan pobres y vulnerables como los beneficiarios. La intervención a *nivel de conglomerados* es también preferible en el caso en que (parte de) los beneficios se transfieren de los grupos tratados a los de control, simplemente porque conviven en la misma comunidad por lo que se les llama efectos indirectos o de derrame (*spillovers*). Por ejemplo, una pyme de base tecnológica que recibe el beneficio de la capacitación que se pone en contacto indirecto contra que no lo recibe pero que lo adopta de buen grado y se pudiera concluir erróneamente que la intervención no tuvo efectos sobre la prevalencia de la baja capacitación.

En muchas intervenciones es posible elegir el nivel al que se hace la aleatorización. No existe una clara ventaja de alguno de los diseños básicos sobre el otro; *a veces será más apropiado hacer la aleatorización a nivel individual y a veces a nivel de conglomerado*. La deseabilidad de uno u otro depende por ejemplo del tipo de preguntas de interés en la evaluación, consideraciones políticas, restricciones logísticas, y existencia de externalidades, entre otras. La *deseabilidad* de uno u otro depende por ejemplo del tipo de preguntas de interés en la evaluación, consideraciones políticas de introducción de innovaciones, restricciones logísticas, y existencia de externalidades, entre otras. Si por su naturaleza las intervenciones pretenden afectar a una comunidad completa, es claro que la aleatorización debe hacerse a este nivel.

Para un nivel de impactos y de confianza determinados, el *tamaño de muestra* en el caso en que la unidad de observación es el *conglomerado* debe ser mayor que en el caso en que la unidad de observación es el individuo, para poder estimar efectos significativos de la intervención. Además, el *tamaño de muestra* necesario crece a medida que aumenta el tamaño de los conglomerados (es decir, a medida que aumenta el número de observaciones individuales por conglomerado). Por tanto, el *nivel de*

aleatorización puede tener efectos importantes en el presupuesto del experimento ya que con frecuencia el levantamiento de información es el rubro más importante en los costos de las evaluaciones.

Problemas con la aleatorización

Varias inquietudes merecen consideración con un diseño de aleatorización, que incluye problemas éticos, validez externa, incumplimiento parcial o total, desgaste selectivo y efectos indirectos, entre otros, tales como (**Khandker et al., 2017**):

- *Retener un tratamiento particular* de un grupo aleatorio de sujetos y proporcionar acceso, a otro grupo de sujetos de forma, puede ser simplemente, poco ético. La realización aleatorizada el diseño, a menudo es políticamente inviable, porque es difícil justificar dicho diseño para las personas que podrían beneficiarse del programa.
- *La validez externa* es otra preocupación. Un proyecto de introducción de innovación tecnológica a pequeña escala puede no afectar las tasas salariales generales, mientras que un proyecto a gran escala podría serlo a nivel nacional. Es decir, el impacto medido de un proyecto piloto puede no ser una guía precisa del impacto del proyecto llevado a escala, en un país. El problema es, cómo generalizar y replicar los resultados obtenidos mediante la aleatorización evaluaciones.
- *El cumplimiento* también puede ser un problema con la aleatorización, que surge cuando una fracción de las personas a las que se les ofrece el tratamiento, no lo toman. Por el contrario, algunos los miembros del grupo de comparación pueden recibir el tratamiento. Esta situación es referida como *cumplimiento parcial (o imperfecto)*. Para ser válido y evitar el sesgo de selección, un análisis debe centrarse en los grupos creados por la aleatorización inicial.

El análisis no puede excluir sujetos o cortar la muestra, de acuerdo con el comportamiento que pudo haber sido afectado por la asignación aleatoria. En términos más generales, el *interés a menudo radica en el efecto de un tratamiento dado*, pero la aleatorización afecta solo la *probabilidad* de que el individuo está expuesto al tratamiento, en lugar del tratamiento en sí.

- Además, los posibles *efectos indirectos* surgen cuando el *grupo de tratamiento* también ayuda al *grupo de control* como participantes de la muestra, confundiendo así las estimaciones del impacto del programa. Por ejemplo, las pymes fuera de la muestra que pueden mudarse a una región donde la capacitación de innovación tecnológica, se tiene establecido al azar, contaminando así los efectos del programa.

A pesar de las bondades evaluativas que ofrecen los *experimentos de introducción de innovaciones*, considerados como experimentos sociales, en la práctica, *pueden desviarse del diseño ideal* (Bernal y Peña, 2011). Problemas en la aleatorización pueden implicar que la *inferencia estadística* acerca de los efectos causales calculados no es válida para la población bajo estudio. Por ejemplo, los individuos actúan de manera diferente al conscientemente participar en un experimento, que si participaran en un programa fuera del marco experimental. El participar en un experimento puede alterar el comportamiento del grupo de tratamiento (*efecto experimental o Hawthorne*) o el del grupo de control (*efecto John Henry*). Otro problema potencial es que, dados los altos costos de los experimentos de introducción de innovaciones, es posible que el tamaño de la muestra sea pequeño y, por tanto, la medición de los efectos no sea suficientemente precisa. Además, si la *muestra es pequeña no se garantiza el supuesto de independencia*, porque la probabilidad de que todas las pymes de base tecnológica (**pbt**), por ejemplo, con ciertas características particulares caigan en un grupo particular (*tratamiento o control*), no es despreciable.

Otros factores, como una *mala aleatorización*, el *no cumplir con el protocolo de tratamiento y la pérdida de muestra*, introducen correlación entre el diferencias tratamiento, D_i , y el término de error u_i . Así, *el estimador de el tratamiento no es asignado de manera completamente aleatoria*, sino que se basa en parte en las características o preferencias de los individuos. Los resultados de las variables objetivo reflejarán tanto *el impacto del programa como el efecto de la falla en la aleatorización*.

Evaluación de impacto aleatorizada en la práctica

La aleatorización ha ido creciendo en popularidad en algunas partes del mundo, en parte porque si se implementa correctamente, la aleatorización es un indicador sólida del impacto del programa (**Khandker, et al., 2017**). Además, una vez que la encuesta ha sido diseñada y los datos recopilados, los ejercicios empíricos para inferir los impactos de los *experimentos aleatorios* son bastante sencillos. Por lo general, justificar o iniciar un experimento aleatorio es más fácil en el inicio de un programa, durante la *fase piloto*. Esta fase ofrece una oportunidad natural para introducir la aleatorización antes de que el programa se amplíe. Presenta una ocasión para que los participantes de la implementación, *evalúen* rigurosamente la efectividad del programa. Eso también puede brindar la oportunidad de *mejorar el diseño* del programa. También se puede introducir un elemento de aleatorización en programas existentes de muchas maneras diferentes con un mínimo de disrupción o ruptura. Mientras que las secciones anteriores de este capítulo han discutido en teoría las preocupaciones con la aleatorización, las siguientes secciones discuten varios temas prácticos y estudios de caso en la implementación de estudios aleatorizados.

El modelo de diferencias

Lo más relevante de una aleatorización exitosa es el tipo de datos que genera, más que las técnicas econométricas utilizadas (Bernal y Peña, 2011) Esto se debe a que el tratamiento está distribuido de manera independiente de otros determinantes de los impactos. Un buen experimento social asegura que las condiciones iniciales, tanto en la variable de resultado como en otras características, sean idénticas entre los grupos de tratamiento y control. Por tanto, la variable de resultado en ausencia del programa debería ser idéntica para los tratados $D_i = (1)$ y para el grupo de control $D_i = (0)$, es decir, se cumple el principio de independencia condicional: $E[u_i | D_i] = 0$, y se puede utilizar por tanto el promedio de la variable de resultado para el grupo de control:

$E[Y_i(0) | D_i = (0)]$, como aproximación del contrafactual $E[Y_i(0) | D_i = (1)]$

Si D_i es binario dadas las características de los datos generados por una aleatorización, las técnicas necesarias para establecer el efecto del tratamiento son extremadamente sencillas:

- Se puede medir como la diferencia de medias en la variable de interés entre los grupos de tratamiento y control después de la intervención.
- Esta diferencia de medias se puede implementar de manera sencilla, con base en el modelo de regresión lineal.
- Como los datos generados en una aleatorización exitosa no sufren de contaminación por sesgo de selección, la asignación aleatoria del tratamiento *implica que el supuesto de independencia condicional*, $E[u_i | D_i] = 0$, se cumple automáticamente. El modelo que permite la comparación, es como se expresó en la **ecuación 1.11**:

$$Y_i = \beta_0 + \beta_i D_i + u_i + \quad (1.11)$$

Donde Y_i es la variable de resultado, D_i es el indicador de tratamiento y u_i es el error. El estimador de **OLS** (*Ordinary Least Squares*) en la **ecuación (1.11)**, β_i 's, se llama *estimador de diferencias*, dado que es igual a la diferencia en medias entre los *grupos de tratamiento y control*. Dado que se cumple el supuesto de independencia condicional $E [u_i | D_i] = 0$, el estimador de diferencias es *insesgado y consistente*. Así:

$$t' = \beta_i = E [Y_i(1) | D_i = (1)] - E [Y_i(0) | D_i = 0]$$

Donde $[Y' / D]$ es el promedio muestral de cada grupo. A medida que el tamaño de la muestra aumenta, esta diferencia converge a:

$$t = E [Y_i(1) | D_i = (1)] - E [Y_i(0) | D_i = (0)]$$

Cuando un experimento por introducción de innovaciones de corte social se diseña e implementa exitosamente, genera *estimaciones insesgadas* del efecto del programa para la población objetivo. Es decir, los resultados tienen validez interna. Sin embargo, el estimador de diferencias no necesariamente es eficiente. La eficiencia es un concepto asociado a la *varianza* del estimador. *Mayor eficiencia significa menor varianza* o dispersión del estimador con respecto a un estimador de la misma clase (ejemplo: lineal e insesgado). Esto implica que el efecto del programa es estimado con mayor precisión (*exactitud*). Evidentemente, es una propiedad deseada que el estimador del programa sea lo más preciso posible.

Suponga que la participación de pymes de base tecnológica (**pbt**), en un programa de estímulos a la innovación, cuyo objetivo es lograr la adopción de una tecnología disruptiva de manufactura, fue asignado el apoyo económico de manera aleatoria. La variable de resultado rele-

vante, es la cantidad de productos manufacturados, según la cantidad de equipo disruptivo instalado. En particular, se asume que se tiene a disposición el puntaje **Z** de la cantidad de productos manufacturados, según la cantidad de equipo disruptivo instalado, es decir, *el número de desviaciones estándar que el pyme de base tecnológica está por encima o por debajo de la media de su grupo relevante*. La **Tabla 3.1**

**Tabla 3.1. Caso hipotético de aleatorización
por el modelo de diferencia**

Nivel de ingresos de producto innovador en la pyme de base tecnológica	Promedio puntaje Z, tratamiento	Promedio puntaje Z, control	Nivel de ingresos por innovación en la pyme de base tecnológica, tratamiento y control
Bajo	0.45	0.45	0.20
Medio	0.70	0.50	0.40
Alto	0.90	0.70	0.40

Fuente: Elaboración propia.

Al mostrar los datos relacionados con la evaluación del programa mencionado, se tiene:

1. La primera columna, presenta el porcentaje de pymes de base tecnológica (**pbt**) con infraestructura instalada para adaptar tecnología disruptiva de manufactura, es decir, está dentro del programa de estímulos a la innovación, o *grupo de tratamiento*:

$$E [Y_i (1) | D_i = (1)]$$

Dado el nivel de ingresos de producto innovador en la pyme de base tecnológica (según la fila).

2. La segunda columna, contiene los mismos datos de pymes de base tecnológica (**pbt**) con infraestructura instalada para adaptar tecnología disruptiva de manufactura *de un grupo de control* elegido de manera experimental:

$$E [Y_i (0) | D_i = (0)] - E [Y_i (0) | D_i = (1)]$$

Recuerde que en un experimento aleatorio, el grupo de control es exactamente el *contrafactual*, que en un experimento dado que se elige de manera aleatoria.

3. La tercera columna, presenta la distribución de pymes de base tecnológica (**pbt**) con infraestructura instalada para adaptar tecnología disruptiva de manufactura de los *grupos de tratamiento y control*, que es idéntica. Por ejemplo, el **20%** de las pymes de base tecnológica (**pbt**) (tanto de *tratamiento como de control*) tienen bajo nivel de manufactura del producto innovador con la tecnología disruptiva de adopción.
4. Con los datos de la tabla podemos calcular el **ATT**, es decir, el efecto del programa sobre los participantes por el método de diferencias, dado que el *grupo de control fue* elegido de manera aleatoria:

$$\begin{aligned} t_{ATT} &= E (t | D_i = (1)) = E [Y_i (1) | D_i = (1)] - E [Y_i (0) | D_i = (0)] \\ E' [Y_i (1) | D_i = (1)] &= (0.45 * 0.2) + (0.70 * 0.4) + (0.90 * 0.4) = \mathbf{0.73} \\ E' [Y_i (0) | D_i = (0)] &= (0.45 * 0.2) + (0.50 * 0.4) + (0.70 * 0.4) = \mathbf{0.57} \\ t_{ATT} &= 0.73 - 0.57 = \mathbf{0.16} \end{aligned}$$

El programa, aumentó el puntaje **Z** de las pymes de base tecnológica (**pbt**) con infraestructura instalada para adaptar tecnología disruptiva de manufactura beneficiarias en **0.16** desviaciones estándar. La idea detrás de un experimento aleatorio es que al

asignar aleatoriamente las familias a los *grupos de tratamiento y control*, se garantiza que:

$$E[Y_i(0) | D_i = (1)] - E[Y_i(0) | D_i = (0)]$$

Si la aleatorización fue exitosa, entre los individuos de *tratamiento y control* sólo se deberían observar diferencias en variables de resultado; no debería haber diferencias en ninguna de las variables que no se ven afectadas por el programa.

Cuestiones éticas

La implementación de *experimentos aleatorios* en los países en desarrollo, a menudo plantea problemas de cuestiones éticas. Por ejemplo, convencer a los funcionarios de gobierno para que retengan un programa en particular de un contingente seleccionado al azar, que comparte el mismo estado de limitaciones o pobreza así como limitaciones para ganar oportunidades, puede ser difícil.

Un contra-argumento es que la aleatorización es una forma científica de determinar el impacto del programa, por lo tanto, en última instancia, ayudaría a decidir, entre un conjunto de diferentes programas disponibles para los formuladores de políticas, de introducción de innovaciones de cuáles funcionan realmente y, por lo tanto, merecen inversión. Por lo tanto, a largo plazo, la aleatorización tiende a ayudar a un mayor número de personas, además de aquellos que fueron inicialmente abordados.

Además, considerando los recursos limitados, no todas los sujetos pueden ser objetivo de un programa, *ya sea experimental o no experimental*. En ese caso, la orientación aleatoria, *no es poco ética*. La conclusión es que, en la práctica el potencial convincente para llevar a cabo diseños aleatorios a menudo es difíciles; por lo tanto, el primer desafío es encontrar *habilitadores* adecuados para llevar a cabo tal diseño. Por ejemplo, representantes de gobiernos, organizaciones no gubernamen-

tales, y algunas veces las empresas del sector privado pueden ser dichos *habilitadores potenciales* (Khandker, et al., 2017).

Calidad de la aleatorización

Los diferentes enfoques en la implementación de estudios aleatorios, reflejan la necesidad de adaptar el programa de *intervención y encuesta* adecuadamente dentro de la muestra objetivo (Khandker, et al., 2017). Estas preocupaciones, están integradas en un proceso más amplio de dos etapas *que guía la calidad de la experimentación diseño*:

- En la primera etapa, los responsables de la política de introducción de las innovaciones, deben definir claramente no solo la muestra aleatoria que será seleccionada para el análisis, sino también la población de la cual esa muestra será analizada. Específicamente, el experimento tendría *validez externa*, lo que significa que los resultados obtenidos podrían generalizarse a otros grupos o entornos (por ejemplo, a través de otras intervenciones del programa,). Usando la notación discutida anteriormente, este enfoque correspondería a las condiciones:

$$E[Y_i(0) | T_i = 1] = E[Y_i(0) | T_i = 0]$$

y

$$E[Y_i(1) | T_i = 1] = E[Y_i(1) | T_i = 0]$$

- En segundo lugar, se deben tomar medidas al asignar aleatoriamente esta muestra a través de las condiciones de los *grupos de tratamiento y control*, para asegurar que el efecto del tratamiento solo sea una función de la intervención y no causado por otros factores. Este criterio se conoce como *validez interna* (*internal validity*) y refleja la *capacidad de controlar problemas* que afecta-

rían la interpretación causal del impacto del *grupo de tratamiento*. El *sesgo sistemático* (asociado con la selección de grupos que no son equivalentes, el desgaste selectivo de la muestra, la contaminación del área objetivo por la muestra del *grupo de control* y los cambios en los instrumentos utilizados para medir el progreso y los resultados del curso del experimento), así como el efecto de *enfocarse en las elecciones relacionadas y los resultados de los participantes* dentro de la muestra objetivo, proporcionan un ejemplo de problemas para la interpretación causal. La variación aleatoria en otros eventos, que ocurren mientras el experimento está en progreso, aunque no representa una amenaza directa a la *validez interna*, también debe ser monitoreado dentro de la recopilación de datos, porque *una variación aleatoria muy grande puede representar una amenaza para la previsibilidad de medición de datos*. La siguiente sección analiza algunos enfoques que, junto con una metodología aleatorizada, pueden ayudar a explicar estos factores potenciales de intervención.

Aunque seguir el *enfoque de dos etapas* conducirá a una medida consistente de la **ATE** (Kish, 1987), los investigadores en la implementación, casi nunca han considerado este enfoque en la práctica. Más específicamente, la única suposición que puede hacerse, dada la aleatorización, es que:

$$E[Y_i(0) | T_i = 1] = E[Y_i(0) | T_i = 0]$$

Incluso, manteniendo el criterio de *validez interna* en un entorno económico es muy difícil. En el mejor de los casos, por lo tanto, los formuladores de políticas de introducción de innovaciones, examinan el efecto de las intervenciones del programa de forma aleatoria, que puedan estimarse consistentemente con el **TOT** o el efecto en una subpoblación dada:

$TOT = E[Y_i(1) - Y_i(0) | T_i = 1]$ en oposición a $ATE = E[Y_i(1) - Y_i(0)]$

Efectos indirectos (*spillovers*)

Asegurarse de que los grupos de control y tratamiento, no se mezclen, es crucial para medir un impacto sin sesgo o imparcial del programa (Khandker, et al., 2017). En el *diseño experimental*, varios enfoques pueden ayudar a reducir la probabilidad de contaminación de los grupos de tratamiento. Los grupos de tratamiento y control, que se encuentren suficientemente separados, por ejemplo, se pueden seleccionar para que la migración a través de las dos áreas, sea poco probable. Como resultado, la contaminación de las áreas de tratamiento es más probablemente que ocurra, con grupos de tratamiento realizados a mayor escala.

Sin embargo, a pesar de los esfuerzos por aleatorizar la intervención del programa *ex ante*, la participación en el programa *puede no ser completamente al azar*, por lo que:

- Los sujetos en el *grupo de control*, pueden moverse a *grupos de tratamiento*, afectando en última instancia, sus resultados (*outcomes*) de la exposición al programa.
- Del mismo modo, los sujetos seleccionados en los grupos de tratamiento, pueden no serlo en última instancia a participar, ya que también, pueden verse afectados indirectamente por el programa. Si un programa para apuntar al *grupo de tratamiento* también ayuda al *grupo de control*, confundiría las *estimaciones de impacto* del programa, en algunos casos, los *grupos de tratamiento* no se pueden ampliar sin crear efectos de equilibrio.
- A menudo, el *efecto Hawthorne* puede afectar los resultados de una aleatorización experimento, donde el simple hecho de ser incluido en un experimento puede alterar comportamiento no aleatorio. Específicamente, *el efecto Hawthorne* se relaciona con beneficia-

rios que se sienten de manera diferente, porque saben simplemente que están bajo un tratamiento. Esta simple sensación, cambiar sus elecciones y comportamiento. Factores distintos por lo tanto, al funcionamiento real del programa, pueden cambiar los resultados de los participantes.

- Específicamente, en casos donde el tratamiento real es distinto de la variable que se manipula aleatoriamente, llame a (**Z**) la variable que se asigna aleatoriamente (por ejemplo, la carta invitando a los CEOs de empresas de base tecnológica a participar en un programa y ofreciéndoles estímulo por asistir), (**T**) es el tratamiento de interés (por ejemplo, participar en el programa). Utilizando la misma notación que antes, el investigador sabe, por asignación aleatoria que:

$$E[Y_i(0) | Z_i = (1)] - E[Y_i(0) | Z_i = (0)] = (0)$$

Es igual a cero y que la diferencia:

$$E[Y_i(1) | Z_i = (1)] - E[Y_i(0) | Z_i = (0)] = \text{Efecto causal de } Z.$$

- Sin embargo, no es igual al efecto del tratamiento, (**T**), porque (**Z**) no es igual a (**T**). Debido a que (**Z**) ha sido elegido para influenciar mínimamente el tratamiento, esta diferencia es, el impacto de **ITT**. Debido también, a que el **ITT** es en principio aleatorio, también puede actuar como una *variable instrumental válida* para identificar el impacto del tratamiento. La estimación de **ITT**, sería el coeficiente estimado en la variable que describe la asignación inicial. El impacto, en aquellos cuyo estado de tratamiento es cambiado por el instrumento, también se conoce como el *efecto de tratamiento promedio local (the local average treatment effect)* (**Abadie et al., 2002**).
- *El desgaste selectivo* también es un problema potencial: es el que los sujetos abandonen un programa, donde aquellos que no

se encuentren debidamente motivados, que muestren vulnerabilidades evidentes, podrían sesgar el efecto del programa.

- Si la medición de la extensión de los *efectos indirectos* es de interés para los responsables políticos en la introducción de innovaciones, la *aleatorización* puede permitir que este fenómeno se mida con mayor precisión. La precisión, por supuesto, depende del nivel de los *efectos indirectos*. Si se producen *efectos indirectos* en la economía global, por ejemplo, cualquier metodología, ya sea *aleatorización* o un *enfoque no experimental*, tendrá dificultades para capturar el impacto del programa. Sin embargo, las repercusiones locales pueden ser medidas con una *metodología aleatoria* (Miguel y Kremer, 2004).
- *Seleccionar el nivel de aleatorización*, sobre la base del nivel en el cual los *efectos indirectos* se espera que ocurran (es decir, ya sea sobre individuos, comunidades o unidades más grandes), es por lo tanto, crucial para comprender el impacto del programa. Una cantidad sustancial de factores de medición de datos que pueden conducir a *contaminación y efectos secundarios* (por ejemplo, desde falta de motivación, interés hasta migración, cambio de paridad, etc.) también necesitaría ser examinado durante el curso de la evaluación, para ser capaz de estimar el impacto del programa con precisión.

Heterogeneidad en los impactos por aleatorización

El nivel al que se produce la *intervención aleatoria* (por ejemplo, el nacional, regional o comunitario), por lo tanto, afecta de múltiples maneras a los *efectos del tratamiento* que van a ser estimados. La *aleatorización* a nivel agregado (digamos, regional) no puede necesariamente, tener en cuenta la *heterogeneidad* individual en la participación y los resultados (*outcomes*) del programa.

Una implicación de este problema, es que el programa final o el impacto del tratamiento, en el nivel *individual no necesariamente se puede medir con precisión como una variable binaria (es decir, $T = 1$ para un participante individual y $T = 0$ para un individuo en un área de control)*. Aunque cierto programa puede ser aleatorizado a un nivel más amplio, la selección individual todavía puede existir en la respuesta al tratamiento. Se puede usar una mezcla de métodos, incluyendo *variables instrumentales*, para tener en cuenta la selección no observada en el individuo nivel. Las interacciones entre los criterios de focalización y el indicador de tratamiento también pueden ser introducidos en la regresión. Por ejemplo, los efectos del *tratamiento por cuartil* también se pueden estimar para medir los impactos distributivos de programas aleatorizados sobre resultados como el consumo y gasto per cápita (**Abadie, et al., 2002**).

Una desviación relacionada de la *aleatorización perfecta*, es cuando la aleatorización es una función de algún conjunto de observables (clima, densidad de población y similares) que afectan la probabilidades de que ciertas áreas sean seleccionadas. Por lo tanto, el *estado del tratamiento es aleatorio, condicionado a un conjunto de características observadas*. Dentro de cada área tratada, sin embargo, el tratamiento es aleatorio entre sujetos, individuos o comunidades. Por lo tanto, se pueden hacer observaciones al tratamiento y comparación, dentro de cada área, y se puede hacer un *promedio ponderado* tomado en todas las áreas, para dar el efecto promedio del programa, en las muestras tratadas.

Valor de un estudio de referencia

La realización de *encuestas de referencia* en un entorno aleatorio conlleva varias ventajas:

- Primero, las *encuestas de referencia*, permiten examinar las interacciones entre las *condiciones iniciales* y *el impacto del programa*. En muchos casos, esta comparación será de considerable importancia para evaluar la *validez externa*.
- Los *datos de referencia* también son útiles, al realizar experimentos de política de introducción de innovaciones, porque las áreas tratadas podrían haber tenido acceso a programas similares o iniciativas antes de la implementación de la nueva iniciativa. Comparar la captación de los de actividades de los participantes, como la introducción de procesos innovadores, de innovación por modelo de negocios, acceso a créditos, etc. *antes y después* de la intervención aleatoria, también pueden ser útil para evaluar las respuestas al experimento.
- Otros valores de un estudio de referencia, incluyen la oportunidad de verificar que la *aleatorización se realizó de manera adecuada*. Los gobiernos que participan en esquemas aleatorizados pueden percibir la necesidad, por ejemplo, de compensar las áreas de control, por no recibir el programa mediante la introducción de otros esquemas, al mismo tiempo.
- Datos recopilados provenientes de las intervenciones del programa en áreas de control antes y durante el curso de la encuesta, ayudarán en la contabilidad de estas fuentes adicionales de *efectos indirectos*.
- La recopilación de los *datos de referencia*, también ofrece la oportunidad de probar y refinar los procedimientos correspondientes. Sin embargo, las *encuestas de referencia* pueden ser costosas y deben realizarse con cuidado.

- El problema con la realización de una referencia o *línea de base*, es que puede conducir a *sesgos* en los impactos del programa al *alterar el contrafactual*.
- La decisión de realizar una *encuesta de referencia*, se reduce para comparar el *costo de la intervención*, *el costo de la recopilación de datos* y *el impacto* de las variables en una encuesta de referencia y se reflejan necesariamente en el resultado (*outcomes*).

Dificultades en la aleatorización

Debido a que minimizan el sesgo de selección en los impactos del programa, las evaluaciones aleatorias pueden ser muy atractivas en los países en desarrollo (**Khandker et al., 2017**). Desafortunadamente, los factores contextuales en tales configuraciones, están *plagadas de situaciones que pueden conjuntar otros factores la implementación aleatoria* y de ahí la calidad de los efectos del programa. La recopilación de datos detallada, sobre estos factores de conjunción y el uso de una combinación de métodos, además de examinar los **ATE**, por lo tanto, pueden ayudar a explicar la *heterogeneidad* individual resultante en el tratamiento impactos.

Incluso en el contexto de los países industrializados, **Moffitt (2003)**, defiende un enfoque integral que compara estudios y programas, *experimentales y no experimentales*, sobre políticas de introducción de innovaciones. Tales comparaciones, pueden revelar posibles mecanismos que afectan la participación, los resultados (*outcomes*) y otros comportamientos de los participantes, lo que ayuda a los evaluadores a comprender las posibles implicaciones de dichos programas, cuando se aplican a diferentes contextos:

- En los estudios *no experimentales* a discutir, se intenta dar cuenta del problema de *sesgo de selección* de diferentes maneras. Básicamente, los *estudios no experimentales*, intentan replicar un experimento natural o aleatorización, tanto como sea posible.
- A diferencia de la *aleatorización*, donde *el sesgo de selección* se puede corregir directamente (aunque también existen problemas en esta área), en las *evaluaciones no experimentales*, es necesario un enfoque diferente, que generalmente implica suposiciones sobre la *forma del sesgo*.
- *Un enfoque, es hacer el caso para asumir la falta de fundamento*, o de condicional de exogeneidad de la colocación del programa (*conditional exogeneity of program placement*), que es una versión más débil de la falta de fundamento. Los métodos de coincidencia (*matching methods*), como el **PSM** (*Propensity Score Matching*) o técnica de propensión de coincidencia de puntaje y los métodos **DD** de doble diferencia (*Double-Difference Methods*), caen bajo esta categoría.
- El enfoque de la *variable instrumental IV* (*Instrumental Variable Methods*) no necesita hacer esta suposición. Intenta encontrar instrumentos que estén correlacionados con la participación en la decisión, pero no correlacionada con la variable de resultado condicional a la participación.
- Finalmente, otros métodos, como el *diseño de regresión discontinua* (**RD**. *Regression Discontinuity Design*), explotan las características del diseño del programa, para evaluar el impacto.

Nivel al que se lleva a cabo una asignación aleatoria

La asignación aleatoria puede llevarse a cabo en diversos niveles: individual, hogares, empresas, comunidades o regiones. En general, el nivel en el que se asignan aleatoriamente las unidades a los grupos de tratamiento y de comparación dependerá en gran medida de dónde y cómo se implemente el programa. Por ejemplo, si se aplica un programa sanitario a nivel de las clínicas de salud, primero se elegirá una muestra aleatoria de dichas clínicas y después se asignará algunas de ellas al grupo de tratamiento y otras al grupo de comparación.

Cuando el nivel de asignación aleatoria es superior o más agregado, como el nivel regional o provincial, puede resultar difícil realizar una evaluación de impacto, porque el número de regiones o provincias en la mayoría de los países no es suficientemente grande para obtener grupos equilibrados de tratamiento y de comparación. Si un país tiene únicamente seis provincias, solo podrá haber tres de ellas en el grupo de tratamiento y tres en el grupo de comparación, lo cual es insuficiente para asegurar que las características de referencia de los grupos de tratamiento y comparación estén equilibradas. Además, para que la asignación aleatoria genere estimaciones de impacto no sesgadas, es importante garantizar que los factores externos dependientes del tiempo (como el clima o los ciclos de las elecciones locales) sean en promedio los mismos en los grupos de comparación y de tratamiento. A medida que el nivel de asignación aumenta, se vuelve cada vez más improbable que estos factores estén equilibrados entre ambos grupos (**Gertler et al., 2017**).

Asimismo, a medida que el nivel de la asignación aleatoria disminuye, por ejemplo, a nivel del sujeto, aumentan las probabilidades de que el grupo de comparación se vea afectado de forma involuntaria por el programa. Hay dos tipos particulares de riesgos que se deben tener en cuenta cuando se escoge el nivel de asignación, a saber:

- Los efectos indirectos llamados también, *de derrame (spillover)* se produce cuando el grupo de tratamiento influye de forma directa o indirecta en los resultados del grupo de comparación (o viceversa).
- y
- El *cumplimiento imperfecto*. Por su parte, el cumplimiento imperfecto tiene lugar cuando algunos miembros del grupo de comparación participan en el programa o algunos miembros del grupo de tratamiento no lo hacen.

Tener en cuenta el nivel de asignación aleatoria de manera rigurosa puede minimizar el riesgo de *derrame* y de *cumplimiento imperfecto*. Los sujetos pueden asignarse a grupos, para minimizar los flujos de información y los contactos entre los *grupos de tratamiento y comparación*. Para reducir la contaminación, el nivel de asignación también debería escogerse según la capacidad del programa para mantener una clara diferencia entre grupos de tratamiento y comparación a lo largo de la intervención. Si el programa comprende actividades entre los grupos (como de tipo comunitario), puede que sea difícil evitar exponer a todos los individuos de esa comunidad al programa.

Un ejemplo de *efecto de derrame* es la administración de recursos técnicos (*capacitación*) y financiero a pymes de base tecnológica (**pbt**). Si en el grupo de tratamiento hay pymes situados cerca de una pyme del *grupo de comparación*, las pymes pueden verse afectadas positivamente por un *efecto de derrame* del *grupo de tratamiento*, porque se reducirán sus probabilidades de contraer los recursos técnicos (capacitación) y financieros de sus vecinos (**Miguel y Kremer, 2004**). Para aislar el impacto del programa, los pymes de base tecnológica (**pbt**) de *tratamiento y comparación* deben estar situados suficientemente lejos unos de otros de modo de evitar ese tipo de derrames. Sin embargo, a medida que la distancia entre las pymes aumente, se volverá más oneroso implementar el programa y, a la vez, administrar las encuestas. Como regla general, si se pueden descartar los efectos de derrame de forma razonable,

es preferible llevar a cabo una asignación aleatoria del tratamiento en el nivel más bajo posible de implementación del programa, lo cual garantizará que el número de unidades de los grupos de tratamiento y comparación sea el mayor posible.

Una vez que se haya seleccionado una muestra de evaluación aleatoria y asignado el tratamiento de manera aleatoria, es bastante sencillo estimar el impacto del programa. Después de que el programa ha funcionado durante un tiempo, tendrán que medirse los resultados de las unidades de tratamiento y de comparación. El impacto del programa es sencillamente la diferencia entre el resultado promedio (\bar{Y}) para el grupo de tratamiento y el resultado promedio (\bar{Y}) para el grupo de comparación.

Ver Figura 3.5.

Figura 3.5. Estimación del impacto con la asignación aleatoria

	Tratamiento	Comparación	Impacto
	Media (\bar{Y}) para el grupo de tratamiento= 100	Media (\bar{Y}) para el grupo de tratamiento= 80	Impacto= $\bar{Y} = 20$
Se inscribe si y solo si está asignado al grupo de tratamiento			

Fuente: Gertler et al. (2017) con adaptación propia.

Lista de verificación: la asignación aleatoria

La asignación aleatoria es el método más robusto para estimar los contrafactuales; se considera el sello de oro de la evaluación de impacto. Para estimar la validez de esta estrategia de evaluación en un determinado contexto, deben contemplarse algunas pruebas básicas (**Gertler et al., 2017**):

- ¿Están equilibradas las características de referencia o *línea de base*? Deben compararse las características de *línea de base* del

grupo de tratamiento y del grupo de comparación. Como se ha mencionado, por motivos estadísticos no todas las características observables deben ser similares en los *grupos de tratamiento y de comparación* para que la aleatorización sea exitosa. Incluso cuando las características de los dos grupos son verdaderamente idénticas, se puede esperar que el **5%** de las mismas aparecerán con una diferencia estadísticamente significativa, cuando se utiliza un intervalo de confianza de **95%** para la prueba. Las variables, en cuyo caso se presenta una diferencia grande entre los *grupos de tratamiento y de comparación*, son especialmente preocupantes.

- ¿Se ha producido algún incumplimiento con la asignación? Se debe verificar, si todas las unidades elegibles han recibido tratamiento y que no haya unidades no elegibles que hayan recibido tratamiento. Si ha habido incumplimiento, tendrá que utilizarse el método de variable instrumental.
- ¿Son suficientemente numerosas las unidades en los grupos de tratamiento y comparación? Si no, sería necesario combinar la asignación aleatoria con diferencias en diferencias.
- ¿Hay algún motivo para creer que los resultados en algunas unidades de alguna manera dependen de la asignación de otras unidades? ¿Podría haber un impacto del tratamiento en las unidades del grupo de comparación?

Por lo anterior, se analizará un ejemplo detallado del empleo de esta técnica, por lo que antes de iniciar Antes de iniciar, se recomienda la lectura previa del anexo: Introducción a STATA.

Ejemplo utilizando STATA

La *aleatorización*, funciona en el *escenario ideal* donde se asignan individuos o en nuestro caso, pymes de base tecnológica (**pbt**) con tratamiento, al azar, *eliminando el sesgo de selección*. Es un intento de obtener un estimado del impacto del programa, comparando las mismas **pbt** tratadas a lo largo del tiempo. Sin embargo, no proporciona una estimación consistente del impacto del programa, porque existen otros factores que afectan los resultados. Así, al comparar el resultado de las **pbt tratadas** con el grupo de *control* similar, esta comparación, es la base de la estimación de impacto del programa que funciona bien con la aleatorización, porque la asignación de individuos o grupos (**pbt**) al *tratamiento* y grupos de *comparación*, es al azar. Por lo tanto, se espera una estimación no sesgada, o imparcial, del impacto del programa en la muestra que será obtenida cuando el diseño y la implementación de la evaluación aleatoria, son apropiados. El ejercicio descrito más adelante, demuestra la estimación de impacto aleatoria, con diferentes escenarios. En esta sección, la evaluación del impacto aleatoria se demuestra de arriba abajo: es decir, desde la colocación del programa hasta la participación de la **pbt** en el mismo.

Impacto de un programa de innovación de pbt por región

Suponga que es un programa hipotético de entrega de microcréditos, como programa de estímulos a la innovación (**pei**) que se asignan aleatoriamente a ciertas regiones de un país, y además sin diferencias entre las regiones tanto *tratadas como de control*. Así, se desea determinar el impacto de la colocación del **pei** en los gastos anuales totales per cápita de las **pbt**. Para este ejercicio, se utilizará el archivo de datos **pei_08.dta**. Los siguientes comandos, abren y crean un archivo **log** de dos variables

de salida: *exptot* y *smefac* (unidad que se deberá dividir entre **100**) antes de unirse al programa de microcréditos **pei** para las **pbt**:

```
use "/Users/DCA/Desktop/STATA log files/pei_08.dta"
gen lexptot=ln(1+exptot)
gen lnsmeffac=ln(1+smefac/100)
```

Así, se crea una variable *dummy* (ficticia), para la colocación del **pei** en las **pbt** de microcrédito en las regiones del país. Se crean dos variables para la colocación del **pei**: una como programas para CEOs masculinos y otra como programa para CEOs femeninos:

```
gen clsterng=cluster*10+region
egen pclsternm=max(dmmfd), by(clsterng)
egen pclsternf=max(dfmfd), by(clsterng)
```

Así que:

- Primero, use el método más simple para calcular el efecto de *tratamiento* promedio de la colocación del **pei** para las **pbt** por región. Se realiza mediante el comando **STATA** *ttest*, que compara el resultado entre las **pbt** por región *tratadas* y las de *control*.
- El siguiente comando, muestra los efectos de la colocación del **pei** a los CEO femeninos de las **pbt** por región. **Ver Tabla 3.2.**

```
ttest lexptot, by(pclsternf)
```

Tabla 3.2. Comandos y tabla de resultados colocación del *pei* a los CEO femeninos de las *pbt* por región

```
. use "/Users/DCA/Desktop/STATA log files/pei_98.dta"

. gen lexptot=ln(1+exptot)

. gen lnsmefac=ln(1+smefac/100)

.

. gen clstergn=cluster*10+region

. egen pclstergnm=max(dmmfd), by(clstergn)

. egen pclstergnf=max(dfmfd), by(clstergn)

.

. ttest lexptot, by(pclstergnf)
```

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
0	67	8.328525	.0644093	.5272125	8.199927	8.457122
1	1,062	8.458371	.0157201	.5122923	8.427525	8.489217
combined	1,129	8.450665	.0152934	.5138679	8.420659	8.480672
diff		-.1298466	.0646421		-.2566789	-.0030142

diff = mean(0) - mean(1) t = -2.0087
 Ho: diff = 0 degrees of freedom = 1127

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0
 Pr(T < t) = 0.0224 Pr(|T| > |t|) = 0.0448 Pr(T > t) = 0.9776

Command

Fuente: STATA con datos propios.

- El resultado, muestra que la diferencia entre las **pbt** por región *tratadas y de control* es significativa. Esto es, la colocación del **pei** para las CEO femeninas por región mejora el gasto per cápita de las **pbt**. Aunque la diferencia es negativa en el resultado, el impacto se interpreta como positivo. El signo negativo, simplemente significa que el resultado del **pei** regional de las **pbt** ($pclstergnf = 1$) es más que eso, en **pei** no regionales **pbt** no programados

(*pclsternf* = 0), lo que implica que el impacto de la participación, es de hecho positivo.

- Alternativamente, puede ejecutar la ecuación más simple que ejecuta la regresión del **pei** por gasto per cápita **pbt** vs. el **pei dummy** (ficticio) **pbt** de la región. Ver **Tabla 3.3**.

reg lexptot *pclsternf*

Tabla 3.3. Comandos y tabla de resultados regresión del *pei* por gasto per cápita *pbt* vs. el *pei dummy* (ficticio) *pbt* de la región

. reg lexptot *pclsternf*

Source	SS	df	MS	Number of obs	=	1,129
Model	1.06259118	1	1.06259118	F(1, 1127)	=	4.03
Residual	296.797338	1,127	.263351676	Prob > F	=	0.0448
				R-squared	=	0.0036
				Adj R-squared	=	0.0027
Total	297.85993	1,128	.264060221	Root MSE	=	.51318

lexptot	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
<i>pclsternf</i>	.1298466	.0646421	2.01	0.045	.0030142 .2566789
_cons	8.328525	.0626947	132.84	0.000	8.205513 8.451536

.
Command

Fuente: STATA con datos propios.

El resultado (**0.130**) provoca el mismo efecto, el cual es significativo.

- La regresión anterior, estima el impacto general de los **pei** de la región sobre el gasto per cápita de las **pbt**. Es posible que sea diferente del impacto en el gasto, después de mantener constantes otros factores, esto es, especificando el modelo ajustado para las covariables que afectan los resultados de interés. Ahora, realice la

regresión del mismo resultado (log del gasto per cápita de las **pbt**) contra el *dummy* de las **pbt** de la región más otros factores que pueden influir en el gasto. **Ver Tabla 3.4.**

reg lexptot pclstergnf genceo ageceo educeo lnsme fac bdbaccess
pcirr raw1 raw2 raw3 raw4 raw5 [pw=weight]

Ajustando para los otros factores covariantes, se encuentran impactos no significantes por colocación del programa en la variable de resultado:

Tabla 3.4. Comandos y tabla de resultados regresión del pei por gasto per cápita pbt vs. el pei *dummy* (ficticio) pbt de la región con covariantes adicionales

```

. reg lexptot pclstergnf genceo ageceo educeo lnsme fac bdbaccess pcirr raw1 raw2 raw3 raw4 raw5 [pw=weight]
(sum of wgt is 1.1260e+03)
note: raw4 omitted because of collinearity

```

Linear regression	Number of obs	=	1,129
	F(11, 1117)	=	21.34
	Prob > F	=	0.0000
	R-squared	=	0.2411
	Root MSE	=	.4628

lexptot	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
pclstergnf	-.0476228	.1048081	-0.45	0.650	-.2532657	.1580201
genceo	-.0378989	.0638106	-0.59	0.553	-.1631011	.0873033
ageceo	.0030443	.0012857	2.37	0.018	.0005216	.0055669
educeo	.0490643	.0057238	8.57	0.000	.0378336	.060295
lnsme fac	.187556	.0388141	4.83	0.000	.1113992	.2637128
bdbaccess	-.0401021	.0494552	-0.81	0.418	-.1371376	.0569334
pcirr	.1102857	.0605069	1.82	0.069	-.0084342	.2290057
raw1	.0108086	.0109051	0.99	0.322	-.0105881	.0322054
raw2	-.0261441	.0197685	-1.32	0.186	-.0649316	.0126435
raw3	.0126889	.0071662	1.77	0.077	-.0013718	.0267496
raw4	0	(omitted)				
raw5	.1001298	.0610757	1.64	0.101	-.0197062	.2199657
_cons	7.925905	.2181363	36.33	0.000	7.497902	8.353908

Command

Fuente: STATA con datos propios.

Evaluación de impacto por participación

Aunque la asignación del **pei** de microcrédito a **pbt** es aleatoria en todas las regiones, la participación puede no serlo, por lo que:

- Solo aquellas **pbt** que tengan menos de **50** décimos de activos como facilidades, pueden participar en los programas de microcrédito (llamados también, *grupos objetivo*). Como se hizo anteriormente, comience con el método más simple para calcular el efecto promedio del tratamiento de **pei** de participación para CEOs femeninas de las **pbt**, a través del comando **STATA** *ttest*, que compara el resultado entre las **pbt** regionales tanto *tratadas como de control*:

```
ttest lexptot, by(dfmfd)
```

El resultado, muestra que la diferencia entre *participantes* y *no participantes*, no es significativa. **Ver Tabla 3.5.**

Tabla 3.5. Comandos y tabla de resultados prueba t del peipor gasto per cápita pbt vs. participantes y no participantes CEO femenina

```
. ttest lexptot, by(dfmfd)
```

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
0	534	8.447977	.023202	.5361619	8.402398	8.493555
1	595	8.453079	.0202292	.4934441	8.413349	8.492808
combined	1,129	8.450665	.0152934	.5138679	8.420659	8.480672
diff		-.005102	.0306448		-.0652292	.0550253

diff = mean(0) - mean(1) t = -0.1665
 Ho: diff = 0 degrees of freedom = 1127

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0
 Pr(T < t) = 0.4339 Pr(|T| > |t|) = 0.8678 Pr(T > t) = 0.5661

Fuente: STATA con datos propios.

- De nuevo, puede ejecutar el modelo de regresión simple contra la participación en el **pei** de las CEO femeninas:

```
reg lexptot dfmfd
```

La regresión, ilustra que el efecto de la participación de las CEO femeninas en los **pei** de microcrédito no es diferente de cero. **Ver Tabla 3.6.**

Tabla 3.6. Comandos y tabla de resultados de regresión del pei por gasto per cápita pbt vs. CEO femenina

```
. reg lexptot dfmfd
```

Source	SS	df	MS	Number of obs	=	1,129
Model	.007325582	1	.007325582	F(1, 1127)	=	0.03
Residual	297.852604	1,127	.264288025	Prob > F	=	0.8678
				R-squared	=	0.0000
				Adj R-squared	=	-0.0009
Total	297.85993	1,128	.264060221	Root MSE	=	.51409

lexptot	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
dfmfd	.005102	.0306448	0.17	0.868	-.0550253	.0652292
_cons	8.447977	.0222468	379.74	0.000	8.404327	8.491626

```
*
```

Command

Fuente: STATA con datos propios.

- Ahora, de manera similar a la regresión de la colocación del programa por región, incluya otras **pbt** y covariables a nivel regional en la ecuación de participación CEOs femenina:

```
reg lexptot dfmfd genceo ageceo educeo lnsme fac bdbaccess pcirr  
raw1 raw2 raw3 raw4 raw5 [pw=weight]
```

El impacto de la participación de las CEO femenina en el gasto de las **pbt**, ahora ha cambiado de no significativa a significativa (nivel del 10 por ciento). Ver **Tabla 3.7**.

Tabla 3.7. Comandos y tabla de resultados de regresión del pei por gasto per cápita pbt vs. CEO femenina y covariantes adicionales

```
. reg lexptot dfmf gencen agecen educen lnsmefac hdbacces pcirr raw1 raw2 raw3 raw4 raw5 [pennslght]
      (sum of wgt is 1.1368e+03)
      note: raw4 omitted because of collinearity
```

```
Linear regression               Number of obs   =    1,119
                                # of variables   =    10
                                Prob > F           =    0.0000
                                R-squared          =    0.2443
                                Root MSE      =    .46182
```

lexptot	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]
dfmf	.0097051	.0346748	2.80	0.006	.0013577 .1382325
gencen	-.0314738	.0542991	-0.52	0.603	-.1598345 .0924669
agecen	.0010966	.0013114	2.36	0.019	.0005186 .005675
educen	.0497782	.0068587	8.22	0.000	.0378884 .061668
lnsmefac	.2093298	.0421298	4.82	0.000	.1265933 .2931184
hdbacces	-.0332328	.0497227	-0.67	0.504	-.1307933 .0643275
pcirr	.0981767	.0687408	1.43	0.153	-.0268189 .2137732
raw1	.009243	.0111588	0.83	0.407	-.012636 .031122
raw2	-.0245718	.0196224	-1.25	0.176	-.0646723 .011928
raw3	.012282	.0072423	1.70	0.090	-.001926 .026492
raw4	(omitted)				
raw5	.1023195	.0611801	1.67	0.095	-.0177281 .2227781
_cons	7.950205	.2346789	34.07	0.000	7.405592 8.114918

Command:

Fuente: STATA con datos propios.

Evaluación de impacto por participación y región

Los dos ejercicios anteriores, mostraron en regresiones separadas, los efectos de la colocación y participación en el programa. Sin embargo, estos dos efectos se pueden combinar en la misma regresión, lo que da una estimación más imparcial, sin sesgos:

```
reg lexptot dfmfd pclsternf genceo ageceo educeo lnsmefac
bdbaccess pcirr raw1 raw2 raw3 raw4 raw5 [pw=weight]
```

Los resultados muestran un efecto no significativo de la colocación del programa, con un efecto positivo (**7.3%**) de las CEO femeninas de las **pbt** participantes en el programa (**t = 2.05**).

Tabla 3.8. Comandos y tabla de resultados de regresión del pei por gasto per cápita pbt vs. CEO femenina por región y covariantes adicionales

```
. reg lexptot dfmfd pclsternf genceo ageceo educeo lnsmefac bdbaccess pcirr raw1 raw2 raw3 raw4 raw5 [pwwweight]
(sum of wgt is 1,1290e+03)
note: raw4 omitted because of collinearity
```

```
Linear regression               Number of obs   =    1,129
                              F(12, 1116)       =    19.10
                              Prob > F           =    0.0000
                              R-squared          =    0.2455
                              Root MSE      =    .46165
```

lexptot	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]
dfmfd	.0783778	.0360089	2.18	0.030	.0077251 .1490185
pclsternf	-.0704618	.107272	-0.73	0.465	-.2809394 .1329158
genceo	-.0382663	.0637062	-0.60	0.548	-.1632638 .0867312
ageceo	.0029962	.0012832	2.33	0.020	.0004685 .005524
educeo	.0304366	.0057791	5.23	0.000	.0190976 .0417756
lnsmefac	.2814613	.040454	6.95	0.000	.1220071 .4409159
bdbaccess	-.0387826	.0490283	-0.79	0.429	-.1349005 .0574154
pcirr	.098372	.0046197	2.13	0.035	.0095095 .187234
raw1	.0008528	.0109168	0.08	0.438	-.0125671 .0142726
raw2	-.0257747	.0197685	-1.30	0.193	-.0645623 .0130129
raw3	.0120047	.0071891	1.70	0.093	-.0027000 .0267093
raw4	0 (omitted)				
raw5	.1033218	.0613179	1.69	0.092	-.0160095 .2226532
_cons	7.936712	.2172097	36.54	0.000	7.510526 8.362907

Command

Fuente: STATA con datos propios.

Evaluación de impacto por participación a nivel regional

Ahora, vea si la participación en el programa es importante, para las **pbt** de la región. Comience con el modelo simple y restrinja la muestra para programar a nivel región:

```
reg lexptot dfmfd if pclstergnf ==1 [pw=weight]
```

El resultado muestra, que el impacto de la participación de CEOs femeninos en el **pei** de microcrédito en las **pbt** individuales, es de hecho, negativo. La participación en el **pei** de CEOs femeninos reduce el gasto per cápita de las **pbt** del en un **7.0** por ciento. **Ver Tabla 3.9.**

Tabla 3.9. Comandos y tabla de resultados de regresión del pei por gasto per cápita pbt vs. CEO femenina participante por región

```
. reg lexptot dfmfd if pclstergnf ==1 [pw=weight]
(sum of wgt is 1.0517e+03)
```

Linear regression	Number of obs	=	1,062
	F(1, 1060)	=	3.57
	Prob > F	=	0.0590
	R-squared	=	0.0044
	Root MSE	=	.51788

lexptot	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
dfmfd	-.0700156	.0370416	-1.89	0.059	-.1426987	.0026675
_cons	8.519383	.0294207	289.57	0.000	8.461653	8.577112

```
.
Command
```

Fuente: STATA con datos propios.

Ahora, realice la regresión del modelo extendido (es decir, que incluya otras variables que influyan en los gastos totales):

```
reg lexptot dfmfd genceo ageceo educeo lnsmejac bdbaccess pcirr
raw1 raw2
raw3 raw4 raw5 if pclsternf ==1 [pw=weight]
```

Al mantener constantes todas las demás variables, se observa que la participación de las CEOs femeninas se vuelve positivo y es significativo al nivel del **10%. Ver Tabla 3.10.**

Tabla 3.10. Comandos y tabla de resultados de regresión del pei por gasto per cápita pbt vs. CEO femenina y región participante

```
. reg lexptot dfmfd genceo ageceo educeo lnsmejac bdbaccess pcirr raw1 raw2 raw3 raw4 raw5 if pclsternf ==1 [pw=weight]
(sum of wgt is 1.2517e+03)
note: raw4 omitted because of collinearity
```

Linear regression	Number of obs	=	1,862
	F(13, 1850)	=	19.75
	Prob > F	=	0.0000
	R-squared	=	0.2931
	Root MSE	=	.4087

lexptot	Robust		t	P> t	[95% Conf. Interval]	
	Coeff.	Std. Err.				
dfmfd	.0715208	.0351238	2.01	0.044	.0010447	.1412127
genceo	-.0512014	.0651407	-0.79	0.432	-.1790242	.0766213
ageceo	.0025424	.0012743	2.00	0.046	.0000419	.0050429
educeo	.0547913	.0050044	0.81	0.000	.0435575	.065985
lnsmejac	.2613056	.0330327	4.77	0.000	.0949104	.2276934
bdbaccess	-.0427308	.0400058	-0.87	0.387	-.1246042	.0412066
pcirr	.1147208	.0507036	1.95	0.051	-.0060729	.2300754
raw1	.0040494	.0306694	0.40	0.688	-.057981	.0230079
raw2	-.0271331	.0215043	-1.26	0.207	-.0693294	.0150933
raw3	.0135604	.0067754	2.00	0.046	.0002555	.0268553
raw4	# (omitted)					
raw5	.0050010	.0007051	1.00	0.300	-.0014051	.0124066
_cons	7.957003	.2263049	35.16	0.000	7.513021	8.401944

Fuente: STATA con datos propios.

A continuación, ejecute la regresión del modelo extendido.

```
. reg lexptot pclstergnf genceo ageceo educeo lnsmfac bdbaccess  
pcirr raw1 raw2 raw3 raw4 raw5 if dfmfd ==0 [pw=weight]
```

Como se puede ver en el resultado que sigue, la colocación del programa en las **pbt** muestra que no hay efectos indirectos (*spillover*) sin efecto de desbordamiento después de controlar las otras variables. **Ver Tabla 3.12.**

Tabla 3.12. Comandos y tabla de resultados de regresión del pei por gasto per cápita pbt vs. CEO femenina no participante por región con covariantes adicionales

```
. reg lexptot pclstergnf genceo ageceo educeo lnsmfac bdbaccess pcirr raw1 raw2 raw3 raw4 raw5 if dfmfd ==0 [pweight]  
[coef of wgt is 6.9634e+01]  
note: raw4 omitted because of collinearity
```

Linear regression	Number of obs	=	134
	F(11, 112)	=	28.88
	Prob > F	=	0.0000
	R-squared	=	0.3122
	Root MSE	=	.46284

lexptot	Robust				
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
pclstergnf	-.0002184	.0000525	-0.40	0.693	-.0002832 .0003368
genceo	-.0000001	.0011234	-0.04	0.737	-.0000042 .0000040
ageceo	.0000001	.0017643	2.13	0.032	.0000000 .0013645
educeo	.0012046	.0000005	7.73	0.000	.0007107 .0016985
lnsmfac	.0000001	.0000000	0.10	0.918	-.0000000 .0000000
bdbaccess	-.0000001	.0000000	-0.04	0.967	-.0000000 .0000000
pcirr	.0000001	.0000000	1.10	0.272	-.0000000 .0000000
raw1	.0000001	.0000000	0.00	0.999	-.0000000 .0000000
raw2	-.0000001	.0000000	-0.00	0.999	-.0000000 .0000000
raw3	.0000001	.0000000	0.00	0.999	-.0000000 .0000000
raw4	* (omitted)				
raw5	.0000001	.0000000	1.10	0.272	-.0000000 .0000000
_cons	7.6155001	.2941247	25.88	0.000	7.046001 8.204999

Command

Fuente: STATA con datos propios.

CAPÍTULO 4.

Propensión de Coincidencia de Puntaje (PSM. Propensity Score Matching)

Dadas las preocupaciones con la implementación de evaluaciones aleatorias, el enfoque de evaluación de impacto perfecto sigue siendo un método en teoría. Por lo tanto, cuando un tratamiento *no puede ser aleatorizado*, lo mejor que puede hacer es intentar *imitar la aleatorización*, es decir, intentar tener un análogo observacional de un experimento aleatorio. Con los *métodos coincidentes*, se intenta desarrollar un *grupo contrafactual* o de *control* que sea tan similar al *grupo de tratamiento* como sea posible, en términos de *características observadas*. La idea es encontrar, desde un gran grupo de *no participantes*, individuos que son observacionalmente *similares a los participantes* en términos de características no afectadas por el programa (pueden incluir, características de preprograma, por ejemplo, porque esos claramente no se ven afectados por la participación subsecuente del programa). Cada *participante* es *emparejado* con un *no participante* observacionalmente similar, y posteriormente, se compara la diferencia promedio en los resultados entre los dos grupos, para obtener el *efecto del tratamiento del programa*. Si se supone que las diferencias en la participación son basados únicamente en las diferencias en las características observadas, y si hay suficientes *no participantes disponibles* para coincidir con los *participantes*, se puede medir el efecto del tratamiento correspondiente, incluso si el tratamiento no es al azar.

El problema es identificar, con credibilidad, los grupos que se parecen. La identificación, es un problema porque incluso si las **pbt** coinciden a lo

largo de un vector, X , de características diferentes, raramente se encontrarían dos **pbt** que sean exactamente similares entre sí en términos de las muchas características que compartan. Debido a que existen muchas características posibles, una forma común de hacer coincidir las **pbt** es hacer coincidir la propensión del puntaje. En **PSM**, cada *participante* se *compara* con un *no participante* sobre la base de un *propensión de puntaje* único, reflejando el probabilidad de participar condicionalmente en las diferentes características observadas de X (**Rosenbaum y Rubin, 1983**). **PSM** por lo tanto, evita la *maldición de la dimensionalidad*, asociado con tratar de unir a *participantes* y *no participantes* en todas las posibilidades característica cuando X es muy grande.

¿Qué es PSM?

Gertler et al. (2017) afirman que los métodos de coincidencia (*matching*) se pueden aplicar en el contexto de casi todas las reglas de asignación de un programa, siempre que se cuente con un grupo que no haya participado en el mismo. La coincidencia, utiliza técnicas estadísticas para construir un grupo de comparación. Para cada unidad posible, el *tratamiento* intenta encontrar una unidad de *no tratamiento* (o conjunto de unidades de no tratamiento) que tengan características lo más parecidas posible. Suponga, un caso en el que se propone evaluar el impacto de un programa de capacitación de innovación eso y se cuenta con una base de datos, con los registros de ingreso, domicilio, declaraciones tributarias, que contiene tanto a los individuos que se inscribieron en el programa como a los individuos que no lo hicieron.

El programa que se intenta evaluar, no tiene reglas de asignación claras (*como asignación aleatoria o un índice de elegibilidad*) que explique por qué ciertos individuos se inscribieron en el programa y otros no lo hicieron. En este contexto, *los métodos de coincidencia o pareamiento (matching)*, permitirán identificar el conjunto de individuos no

inscritos que más se parece a los individuos tratados, a partir de las características que ya se tienen en la base de datos. Estos individuos *no inscritos* coincidentes posteriormente, se convierten en el grupo de comparación que se emplea para estimar el *contrafactual*. La búsqueda de una buena pareja para cada participante del programa, requiere aproximarse todo lo posible a las características que explican la decisión del individuo de inscribirse en el programa. Desafortunadamente, *en la práctica esto es más difícil*. Si la lista de características observables relevantes es muy grande, o si cada característica adopta muchos valores, es posible que sea complicado identificar una pareja para cada una de las unidades del grupo de tratamiento. A medida que aumenta el número de características o dimensiones con las que se quiere coincidir o parear las unidades que se inscribieron en el programa, es posible encontrarse con lo que se denomina *la maldición de la dimensionalidad*.

Por ejemplo, si solo se consideran tres características importantes para identificar el grupo de comparación del pareamiento, como la edad, el género y si la persona tiene un diploma de estudios de pregrado, es probable que se encuentren parejas para todos los participantes que se *inscribieron* en el programa entre el conjunto de aquellos que *no se inscribieron* (los no inscritos), pero se corre el riesgo de dejar al margen otras características potencialmente importantes. Sin embargo, si se aumenta la lista de características, por ejemplo, para incluir el número de años de estudios, el número de meses que el individuo lleva desempleado, el número de hijos, el número de años de experiencia, etc., es posible que la base de datos no contenga una buena pareja coincidente para la mayoría de los participantes del programa que están inscritos, a menos que abarque un número muy grande de observaciones. El **Gráfico 4.1** ilustra la coincidencia (*matching*) sobre la base de cuatro características: edad, género, meses de desempleo, y diploma de estudios de pregrado.

Gráfico 4.1. Coincidencia exacta en cuatro características

Unidades tratadas				Unidades no tratadas			
Edad	Género	Meses desempleado	Diploma de pregrado	Edad	Género	Meses desempleado	Diploma de pregrado
19	1	3	0	24	1	8	1
35	1	12	1	38	0	1	0
41	0	17	1	58	1	7	1
23	1	6	0	21	0	2	1
55	0	21	1	34	1	20	0
27	0	4	1	41	0	17	1
24	1	8	1	46	0	9	0
46	0	3	0	41	0	11	1
33	0	12	1	19	1	3	0
40	1	2	0	27	0	4	0

Fuente: Gertler et al. (2017).

Lo que hace PSM

PSM construye un grupo de comparación estadística modelando la probabilidad de participar en el programa, sobre la base de las características observadas no afectadas por el programa. Posteriormente, los *participantes* se comparan en función de esta probabilidad o propensión de puntaje, a los *no participantes*, utilizando diferentes métodos descritos más adelante en el capítulo. La media del efecto del tratamiento (**ATE**, *Average Treatment Effect*) del programa, se calcula como la diferencia de la media en los resultados a través de estos dos grupos. Por sí solo, el **PSM** es útil cuando solo se observan características que se cree, afectan la participación en el programa. Esta suposición depende de las reglas que rigen la focalización del programa, así como cualquier factor que impulse la autoselección de individuos o **pbt** en este caso. Idealmente, si está disponible, la *línea de base* previa al programa, los datos sobre *participantes* y *no participantes* pueden usarse para calcular la propensión de puntaje y hacer coincidir los dos grupos.

La selección de características observadas, también puede ayudar en el diseño de *experimentos de ondas múltiples* (*multiwave experiments*). **Hahn et al. (2008)** muestran que los datos disponibles sobre covariables para individuos que son blanco de un experimento, digamos en la primera etapa de una intervención de dos etapas, se pueden usar para elegir una regla de asignación de tratamiento para la segunda etapa, condicionada en características observadas. Esto equivale, a elegir el propensión de puntaje en la segunda etapa y permite una estimación más eficiente de los efectos causales.

Coincidencia por propensión de puntaje

La llamada *maldición de la dimensionalidad*, es posible resolverla fácilmente utilizando el método denominado *propensión de coincidencia de puntaje* (*propensity score-matching*, **Rosenbaum y Rubin, 1983**). Con este enfoque, *ya no se requiere que se intente coincidir o parear a cada unidad inscrita con una unidad no inscrita que tenga exactamente el mismo valor para todas las características de control observables*. En cambio, para cada unidad del grupo de *tratamiento* y del conjunto de *no inscritos*, se computa la probabilidad de que esta unidad se inscriba en el programa (el denominado propensión de puntaje) sobre la base de los valores observados de sus características (las variables explicativas). Esta puntuación, es un número real entre **0** y **1** que resume la influencia de todas las características observables en la probabilidad de inscribirse en el programa.

Deberían utilizarse solo las características observables en la *línea de base* para calcular la *propensión de puntaje*. Esto se debe a que las características post tratamiento, pueden haberse visto afectadas por el propio programa, y el uso de dichas características para identificar a un grupo de comparación coincidente o pareado, sesgaría los resultados. Cuando el tratamiento influye en las características del individuo y se usan aquellas

características para parear, se escoge un grupo de comparación que se parece al grupo de tratamiento debido al propio tratamiento.

Sin el tratamiento, esas características tendrían un aspecto muy diferente. Esto incumple el requisito básico de una buena estimación del contrafactual, a saber: que el grupo de comparación debe ser similar en todos los aspectos, excepto en el hecho de que el grupo de tratamiento recibe el tratamiento y el grupo de comparación no lo recibe. Una vez que se ha computado la *propensión de puntaje* de todas las unidades, aquellas del grupo de tratamiento pueden parearse con unidades en el conjunto de no inscritos que tienen los puntajes de propensión más cercanos. En la práctica, se utilizan muchas definiciones de lo que constituye la unidad más próxima o cercana para llevar a cabo un la coincidencia o pareamiento. Las unidades de control más cercanas, se pueden definir sobre la base de una estratificación de la propensión de puntaje, la identificación de los vecinos más próximos de la unidad de tratamiento, considerando la distancia, dentro de un determinado radio, o utilizando técnicas de núcleo (*kernel*). Se considera una buena práctica verificar la robustez de los resultados de la coincidencia o pareamiento, empleando diversos algoritmos de que lo realizan (**Rosenbaum, 2002**).

Estas unidades próximas se convierten en el grupo de comparación y se utilizan para producir una estimación del contrafactual. El método de coincidencias o pareamiento por *propensión de puntajes*, intenta imitar la asignación aleatoria a los grupos de tratamiento y comparación escogiendo para el grupo de comparación aquellas unidades que tienen propensiones similares a las unidades del grupo de tratamiento. Dado que el pareamiento de puntajes de propensión no es un método de asignación aleatoria pero intenta imitarlo, pertenece a la categoría de *métodos cuasi-experimentales*. La diferencia promedio en los resultados entre las unidades de *tratamiento*, o inscritas, y sus unidades de *comparación* correspondientes genera la estimación del impacto del programa. En resumen, el impacto del programa se estima comparando los resultados promedio de un grupo de tratamiento, o inscrito, y el resultado promedio

del subgrupo de unidades estadísticamente pareadas, donde la coincidencia o pareamiento se basa en características observables en los datos disponibles. Para que la coincidencia o pareamiento por propensión de puntajes, produzca estimaciones del impacto de un programa para todas las observaciones tratadas, cada unidad de tratamiento o inscrita debe parearse con una unidad no inscrita. En este libro, el análisis de coincidencias o pareamiento, se centra en un *pareamiento de uno a uno*. No se analizarán otros tipos de pareamiento, como el de *uno a varios* o el de *reemplazo/sin reemplazo*. Sin embargo, en todos los casos, el marco conceptual descrito aquí seguiría vigente.

En la práctica, puede ocurrir que para algunas unidades inscritas, no haya unidades en el conjunto de no inscritos que tengan puntajes de propensión similares. En términos técnicos, puede que se produzca una falta de rango común, o falta de superposición, entre la propensión de puntajes del grupo de tratamiento o *inscrito* y los del conjunto de *no inscritos*.

Verificación de coincidencia o pareamiento

El pareamiento se basa en el supuesto de que las unidades inscritas y no inscritas son similares en términos de cualquier variable no observable que podría influir tanto en la probabilidad de participar en el programa como en el resultado, por lo que se debe cuestionar **(Gertler et al., 2017)**:

- ¿La participación en el programa está determinada por variables que no se pueden observar? Esto no se puede comprobar directamente, de modo que para orientarse habrá que fiarse de la teoría, del sentido común y del conocimiento adecuado del contexto de la evaluación de impacto.
- ¿Las características observables están bien equilibradas entre los subgrupos pareados? Compárense las características observables

de cada grupo de tratamiento y su grupo de unidades de comparación pareados en la *línea de base*.

- ¿Se puede encontrar una unidad de comparación pareada para cada unidad de tratamiento? Verifíquese si hay un rango común suficiente en la distribución de los puntajes de propensión. Las pequeñas zonas de rango común o superposición señalan que las personas inscritas y no inscritas son muy diferentes, y aquello arroja dudas sobre si el pareamiento es un método creíble.

Teoría PSM

PSM es un *método flexible y oportunista*, pues en principio se puede aplicar al tener una sola observación en el tiempo, siempre que existan observaciones de grupo de *tratamiento* y *control*. Por ejemplo, no es necesario contar con información de *línea de base*. Como **PSM** se puede aplicar cuando ninguno de los demás métodos no experimentales se puede usar, entonces se utiliza con mucha frecuencia. Por tanto, es necesario tener cuidado de asegurarse de que se cumplen los supuestos para poder aplicar la técnica y, por tanto, dar una interpretación acertada de los resultados. ¿Qué necesitamos en cuanto a los datos para asegurarnos de que **PSM** funcione? No se puede determinar si se cumplen los supuestos, pues no podemos verificar si lo no observado afecta o no la decisión de participación. Lo que en la práctica se hace es calcular las diferencias en las variables observables X ; si los dos grupos son demasiado diferentes en las características observadas, esto podría ser evidencia de que es probable que también existan diferencias entre los dos grupos en características *no observadas*. Los resultados de los métodos de coincidencia o emparejamiento son confiables siempre y cuando existan razones para pensar que las *variables no observables o no disponibles* en la base de datos, que pueden incluir variables intrínsecamente difíciles de medir, no son un determinante fundamental tanto de la participación en el programa como

de las variables de resultado potenciales. Por ejemplo, en un programa en el cual el *costo de participación es nulo*, es más difícil justificar que unos individuos elegibles participen y otros no, debido a la existencia de diferencias sistemáticas observables entre unos y otros. Si la participación no era costosa, la *motivación* (Variable que usualmente no es posible medir) posiblemente no será un factor importante tanto en la participación del programa como en la determinación de los resultados potenciales (Bernal y Peña, 2011).

Una alternativa a **PSM**, sería estimar el *modelo de diferencias*, controlando por todo el vector X de variables independientes relevantes. De hecho, **PSM** puede interpretarse como un *estimador de mínimos cuadrados* ponderado por la probabilidad de participación en el programa. Sin embargo, **PSM** es preferible al simple modelo de diferencias con regresores adicionales por tres razones:

- Primero, dado que la coincidencia o emparejamiento, es un *método no paramétrico*, no debemos hacer los supuestos implícitos en cuanto a la forma funcional que se hacen al usar las regresiones lineales.
- Segundo, al restringir la muestra al soporte común, aseguramos que estamos *comparando grupos comparables* (aunque esto puede implicar que los impactos medidos no correspondan al total de la población, sino a una submuestra).
- Tercero, cuando existen dudas sobre la posibilidad de extrapolación de resultados: si los grupos de tratamiento y control son bastante disímiles, el método de coincidencia o pareo, disminuye la extrapolación al reducir la comparación a la *región de soporte común*.

No existe un método ideal para resolver el *sesgo de selección* fuera del ámbito experimental. En la práctica, lo mejor que podemos lograr es un método que funcione bajo circunstancias o supuestos que se pueden verificar, y que provea señales de alarma si no está funcionando bien. El

método de emparejamiento funciona cuando el sesgo de selección está determinado únicamente por variables observables. Con buenos datos y decisiones acertadas por parte del investigador, **PSM** puede ser un método poderoso.

En la práctica, la participación en el programa puede depender tanto de *variables observables* como *no observables o no disponibles* en la base de datos. Las características de cada programa, en particular su focalización y la respuesta de los agentes, afectan la importancia de las características observables y no observables en la determinación de la selección en el programa. Entonces, para determinar si las variables observadas determinan exclusivamente la participación en el programa, y, por tanto, asegurar que **PSM** sea una aproximación válida, es necesario estudiar las características particulares del programa en cuestión. Dada la disponibilidad de datos representativos y altamente comparables para los grupos de tratamiento y control, el proceso de estimación de **PSM** se puede resumir en el siguiente *algoritmo* (Bernal y Peña, 2011):

1. Estimar la probabilidad de participar en el programa usando las muestras de tratamiento y control. Hay dos preguntas fundamentales al estimar la probabilidad de participación: *qué modelo usar* y *qué variables incluir*. En cuanto a la selección del modelo, la probabilidad de participación se puede estimar de varias maneras, incluidas la *probabilidad lineal*, *probit* o *logit*. La idea es especificar un modelo $P(D_i = 1 | X) = f(X)$, es decir, la probabilidad de participación como una función $f(\cdot)$, que puede ser lineal o no lineal en las características observables de los individuos, X . Con frecuencia se prefieren *logit* o *probit*, y no los modelos de probabilidad lineal, por las deficiencias de ese último, en particular las predicciones fuera del intervalo $[0, 1]$ y la presencia de *heterocedasticidad*. Cuando el tratamiento es binario y se estima la probabilidad de participación vs. la de no participación, los modelos *logit* y *probit* usualmente generan resultados similares. Por tanto,

la decisión de qué modelo escoger *no es crítica*. Sin embargo, cuando existen varias categorías de tratamiento, el caso es distinto. Se pueden usar los *modelos logit o probit multinomiales*. El *probit multinomial* está basado en supuestos menos fuertes que el *logit multinomial* y es, por tanto, preferido. En particular, el *logit multinomial* asume la independencia de alternativas irrelevantes, es decir, que la deseabilidad relativa de dos alternativas (*odds ratio*) es independiente de otras alternativas. Este supuesto, aunque conveniente para la estimación, es poco probable que se cumpla en la práctica. Una vez seleccionado el modelo, ¿qué variables incluir en la estimación de la probabilidad de participación? **Heckman et al., (1998)** muestran que las estimaciones de **PSM** son muy sensibles a la selección de variables usadas para estimar $P(X)$. La estrategia de coincidencia o emparejamiento, se basa en el supuesto de independencia condicional, que implica que las variables de resultado deben ser independientes del tratamiento, una vez se condiciona por la probabilidad de participación. Por tanto, en la estimación de $P(X)$ se deben incluir únicamente variables que afecten la decisión de participación y la variable de resultado de manera simultánea.

Los investigadores se pueden guiar por los modelos que describan el fenómeno bajo estudio, investigaciones previas y conocimiento del diseño institucional. *Candidatos naturales son entonces, variables que no cambien en el tiempo*, o aquellas que hayan sido medidas antes de participar en el programa (asegurándose de que el efecto anticipación del programa no haya afectado dicha variable).

2. Predecir las probabilidades de participación de cada individuo tanto para los tratados como para los controles. En el cálculo de la probabilidad de participación *se incluye siempre una constante*. Si no se tiene claridad acerca de qué variables incluir en la estimación, ¿es mejor incluir más o menos variables? *Hay varias razones para evitar incluir demasiadas variables*. Por un lado, *incluir variables no relacionadas puede exacerbar el problema de soporte común*.

Por el otro, aunque el incluir variables irrelevantes en la estimación de la participación no sesgará los resultados, ni los hará inconsistentes, *sí tiende a aumentar su varianza*. También hay razones que sugieren que *no es buena idea estimar modelos demasiado estilizados*. Solo se debe excluir una variable si hay consenso acerca de que no está relacionada con la variable de resultado o que no es una variable explicativa apropiada (por ejemplo, es una variable que puede verse afectada por el tratamiento). La decisión de qué variables, interacciones y términos de segundo orden incluir en la estimación de la probabilidad de participación tiene que ver con la *significancia estadística*. Una alternativa es estimar la especificación preferida y revisar los estadísticos de la estimación para asegurarse del ajuste de ésta. **Imbens y Rubin (2010)** proponen un algoritmo secuencial para determinar la especificación de la probabilidad de participación, basado en la significancia estadística de cada término. A grandes rasgos, proponen comenzar con una especificación muy parsimoniosa: incluir una constante y las variables que se consideran vitales en la coincidencia o emparejamiento, o que se piensa a priori que están altamente correlacionadas con la variable de resultado. Se define un nivel de significancia esperado para las variables que entren de manera lineal en la especificación. Partiendo del conjunto de variables seleccionadas a priori, se estima repetidamente la probabilidad de participación a través de un probit usando máxima verosimilitud, incluyendo términos adicionales.

Se deja el nuevo término solo si la significancia estadística está por encima del valor predeterminado (usando el estadístico de la prueba de relación de verosimilitud con la hipótesis nula de que el coeficiente del nuevo término es igual a cero, *likelihood ratio test statistic*). Después de elegir las variables que entran la especificación de manera lineal, se eligen de manera similar uno a uno

los términos de segundo orden (cuadráticos e interacciones). Se incluyen términos de segundo orden solamente si las variables incluídas de manera lineal y se dejan solo si la significancia estadística está por encima de unos valores predeterminados de nuevo usando el estadístico de relación de verosimilitud.

3. Restringir la muestra al soporte común. Explicado con mayor detalle en *el supuesto del soporte común*, más adelante
4. Seleccionar un algoritmo de emparejamiento. Para cada individuo tratado se busca un individuo o grupo de individuos de control que tenga una probabilidad de participación similar:
 - Si se encuentran individuos con una probabilidad similar de participación en el programa, siga al **paso 6**.
 - Si no se encuentran individuos con probabilidad de participación en el programa similar, se vuelve al **paso 1** y se especifica nuevamente la ecuación de participación, por ejemplo, incluyendo más interacciones. Si después de varios intentos los indicadores de calidad del emparejamiento no mejoran, esto puede implicar que no se cumple el supuesto de independencia condicional. Por tanto, no es posible estimar los efectos y se debe considerar una estrategia de estimación distinta a **PSM**.
5. Revisar que las variables observables entre los grupos de tratamiento y control sean parecidas o estén balanceadas (*balancing property*) por grupos de probabilidad predicha. Para realizar esta comparación es necesario conocer las ponderaciones asignadas a los individuos del grupo de control, que están determinadas por el algoritmo de emparejamiento seleccionado.
6. Se calculan los impactos del programa como el promedio apropiadamente ponderado de la diferencia entre la variable de resultado de los tratados y los no tratados.
7. Se calculan los errores estándar e intervalos de confianza para determinar si los impactos son estadísticamente significativos.

El enfoque **PSM**, intenta capturar los efectos de diferentes covariables X observadas en el grupo de participación y capturarlas en *un solo puntaje o índice de propensión (propensity score index)*. Posteriormente, en nuestro caso, los resultados de **pbt participantes** y los *no participantes* con puntajes de propensión similares se comparan, para obtener el efecto del **pei** en las **pbt**. Los **pbt** que no se encuentren en ninguna coincidencia, se descartan porque no hay base existente para la comparación. **PSM** construye un grupo de comparación estadístico, que se basa en un modelo de probabilidad de participar en el tratamiento T condicional a las características observadas X , o la propensión de puntaje: $P(X) = Pr(T = 1 / X)$. **Rosenbaum y Rubin (1983)** muestran que, bajo ciertos supuestos, emparejar en $P(X)$ es tan bueno como emparejar en X . Los supuestos necesarios para la identificación del efecto del programa, son (**Khandker et al., 2017**):

- Condicionales independencia, y
- Presencia de un soporte común.

Además, como se discutió en los capítulos previos, el efecto del tratamiento del programa, usando estos métodos pueden representarse como el efecto del tratamiento promedio (**ATE. Average Treatment Effect**) o el efecto del tratamiento sobre el tratado (**TOT. Treatment effect On the Treated**). Por lo general, los investigadores y evaluadores pueden garantizar solo la *validez interna* de la muestra en *oposición a la externa*, por lo que solo se puede estimar el **TOT**. Los supuestos más débiles de *independencia condicional*, así como *apoyo común* se aplican a la estimación del **TOT** y también se analizan en este capítulo.

El supuesto de la independencia condicional

La *independencia condicional*, establece que dado un conjunto de covariables observables X que no son afectados por el tratamiento, los resultados potenciales Y son independientes de la asignación del tratamiento T . Si Y_i^t representa los resultados para los participantes y Y_i^c son los resultados para *no participantes*, la independencia condicional implica:

$$(Y_i^t, Y_i^c) \perp T_i \mid X_i \dots \dots \dots (4.1)$$

Este supuesto también se llama *falta de fundamento* (*uncounfoundedness*), (**Rosenbaum y Rubin, 1983**), e implica que la adopción del programa se basa completamente en las características observadas. Para estimar el **TOT** en lugar del **ATE**, se necesita una suposición más débil:

$$Y_i^c \perp T_i \mid X_i \dots \dots \dots (4.2)$$

La *independencia condicional*, es un supuesto fuerte y *no es un criterio directamente comprobable*; depende de características específicas del programa en sí. *Si las características no observadas determinan la participación en el programa, se violaría la independencia condicional y el PSM no es un método apropiado*. Si las variables no observadas, afectan de hecho la participación y los resultados, esta situación produce lo que se llama un *sesgo oculto* (*hidden bias*) (**Rosenbaum, 2002**). Aunque el supuesto de independencia condicional, o *falta de fundamento* (*uncounfoundedness*), no se puede verificar, la sensibilidad de los resultados estimados del método **PSM** pueden verificarse con respecto a las desviaciones de este supuesto de identificación. En otras palabras, incluso si no se puede estimar el alcance de la selección o el *hidden bias* (sesgo oculto), el grado para lo cual los resultados de **PSM** son sensibles a esta suposición de *falta de fundamento* (*uncounfoundedness*) pueden ser probados.

En los siguientes capítulos, se discuten enfoques cuando la selección de no observados, está presente. Tener un amplio conjunto de datos de preprograma, ayudará a soportar la asunción *independencia condicional* al permitir que el investigador controle, tantas características observadas como podría estar afectando la participación en el programa (suponiendo que una *selección no observada*, esté limitada). Las alternativas generadas, al existir la *selección de características no observadas*, violan la independencia condicional, por lo que se discute en los siguientes capítulos, incluyendo la *variable instrumental* y los *métodos de doble diferencia*.

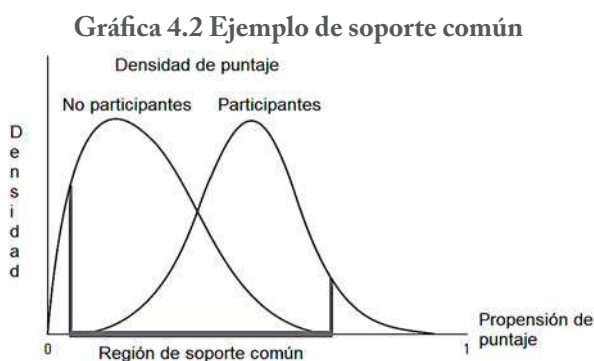
El supuesto de soporte común

Una segunda suposición, es el soporte común (*common support*) o la condición de superposición (*overlap condition*), donde: $0 < P(T_i = 1 | X_i) < 1$. Esta condición asegura que las observaciones de tratamiento tengan observaciones de comparación cercanas en la distribución de la propensión de puntaje (Heckman et al., 1999). Específicamente, la efectividad de **PSM** también depende de tener un gran y aproximadamente igual número de observaciones *participantes* y *no participantes* para que una región sustancial de apoyo común se pueda encontrar. Para calcular el estimado de **TOT**, esta suposición puede ser expresada como: $P(T_i = 1 | X_i) < 1$.

Por lo tanto, las *unidades de tratamiento tendrán que ser similares a las unidades sin tratamiento* en términos de características observadas no afectadas por la participación; por lo tanto, algunas *unidades sin tratamiento* es posible que deban descartarse, para garantizar la comparabilidad. Sin embargo, en ocasiones un subconjunto no aleatorio de la muestra de tratamiento puede tener que descartarse, si las unidades de comparación similares no existen (Ravallion, 2008). Esta situación es más problemática, porque crea un posible sesgo de muestreo en el efecto

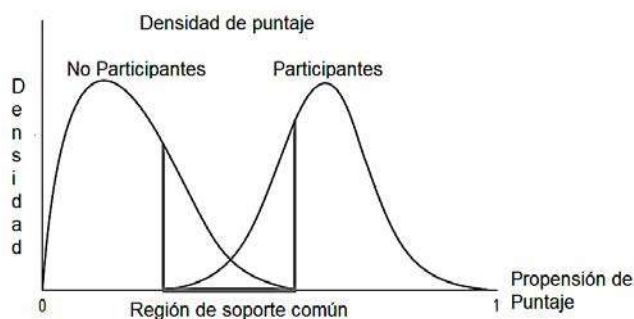
del tratamiento. Examinar las características de las unidades retiradas, puede ser útiles para interpretar el sesgo potencial en los efectos estimados del tratamiento.

Heckman et al., (1997), alientan a abandonar las observaciones del tratamiento con débil soporte común ya que sólo en ésta pueden hacerse inferencias sobre la causalidad, como se refleja en la **Gráfica 4.2**.



Fuente: Khandker et al. (2017).

Gráfica 4.3. Ejemplo de pobre balanceo y soporte común débil



Fuente: Khandker et al. (2017).

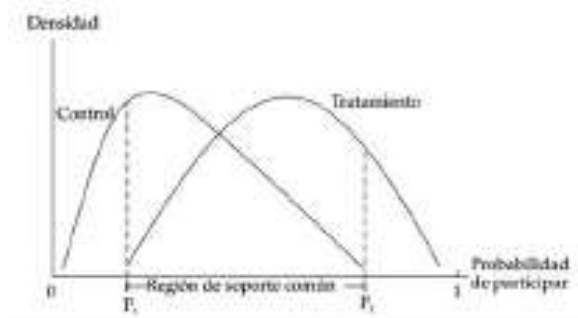
La **Gráfica 4.3**, representa un ejemplo de la falta de rango común, con las siguientes observaciones (**Gertler et al., 2017**):

- En primer lugar, se estima la probabilidad de que cada unidad de la muestra se inscriba en el programa a partir de las características observables de esa unidad, es decir, la propensión de puntaje. El gráfico muestra la distribución de los puntajes de propensión por separado para los inscritos y no inscritos. El problema es que estas distribuciones no se superponen perfectamente. En el medio de la distribución, es relativamente fácil encontrar las parejas porque hay tanto *inscritos* como *no inscritos* con estos niveles de puntajes de propensión. Sin embargo, los *inscritos* con puntajes de propensión cercanos a **1** no se pueden aparear con ningún *no inscrito* porque no hay *no inscritos* con propensión de puntajes tan altos. Hay tan poca similitud entre las unidades que tienen muchas probabilidades de inscribirse en el programa y las unidades no inscritas que no se puede encontrar una buena pareja para ellas.
- De la misma manera, los *no inscritos* con puntajes de propensión cercanos a **0** no pueden aparearse con ningún inscrito porque no hay inscritos que tengan puntajes de propensión tan bajos. Por lo tanto, en los extremos, o colas, de la distribución de la propensión de puntaje *aparece una falta de rango común*. En este caso, el procedimiento de pareamiento estima el *efecto local promedio del tratamiento* (**LATE**. *Local Average Treatment Effect*) para las observaciones sobre el rango común.
- Los pasos que hay que seguir cuando se aplica un pareamiento por propensión de puntajes, se resumen en **Jalan y Ravallion (2003)** y **Rosenbaum (2002)**:
 - a. Se necesitarán encuestas representativas y altamente comparables en las que se puedan identificar las unidades que se inscribieron en el programa y las que no lo hicieron.
 - b. Se reúnen las dos muestras y se estima la probabilidad de que cada individuo se inscriba en el programa, a partir de las características individuales observables en la encuesta. Este paso produce la propensión de puntaje.

- c. Se limita la muestra a unidades para las que aparece un rango común en la distribución de la propensión de puntaje.
- d. Para cada unidad inscrita, se identifica un subgrupo de unidades con puntajes de propensión similares.
- e. Se comparan los resultados de las unidades de tratamiento, o inscritas, y las parejas de las unidades de comparación, o no inscritas. La diferencia de los resultados promedio de estos dos subgrupos es la medida del impacto que se puede atribuir al programa para esa observación específica tratada.
- f. La media de estos impactos individuales arroja una estimación del efecto local promedio del tratamiento. En la práctica, los programas estadísticos habitualmente usados incluyen comandos que realizan los pasos 2 a 6 de manera automática.

Por ejemplo, suponga que las dos funciones de la **Gráfica 4.4**

Gráfica 4.4. Otro ejemplo de soporte común



Fuente: Bernal y Peña (2011).

Son distribuciones de probabilidad de participar para los grupos de *control* y *tratamiento*. Observe que para probabilidades mayores a P_2 no existen observaciones del grupo de control; esto implica que no es posible encontrar sujetos en el grupo de *control* que sean suficientemente pare-

cidas a las de *tratamiento*, ubicadas después de P_2 y, por lo tanto, *no se tiene un buen grupo de control* para estas observaciones. Así mismo, para valores inferiores a P_1 , no existen observaciones del grupo de *tratamiento*. Entonces, la región de soporte común está entre P_1 y P_2 siendo estos puntos de corte se determinan por inspección visual.

Hay otras maneras de determinar el soporte común con más precisión. Una de ellas es el *criterio del mínimo y el máximo*, que sugiere eliminar las observaciones inferiores al mínimo y mayores que el máximo del otro grupo. Esto se hace tanto para el grupo de *tratamiento* como para el de *control*. Estos dos criterios, de inspección visual y del máximo y el mínimo, presentan fallas si hay áreas dentro del soporte común, donde no haya observaciones para uno de los dos grupos.

Otra forma de definir la zona de soporte común, es la de elegir los valores de probabilidad estimada de participación para los cuales hay densidad positiva tanto para el grupo de *tratamiento* como para el de *control*. En el aseguramiento de que las densidades son positivas estrictamente, se excluyen las regiones que tienen densidad positiva pero muy pequeña. Formalmente, aquellas que no excedan una densidad mínima de q (Smith y Todd, 2005). Este criterio se conoce como de *trimming*, en la que después de realizarlo se requiere estimar de nueva cuenta la probabilidad de participación. Mientras más alto sea el porcentaje de individuos coincidentes o emparejados, lo cual es un parámetro clave de la evaluación, los resultados probablemente serán calculados con mayor precisión. Independientemente del criterio usado para determinar el soporte común, si la proporción de individuos eliminados es alta, esto genera dudas sobre la posibilidad de generalizar el efecto estimado con los individuos restantes al total de la población, por lo que se pone en juego la validez externa de las estimaciones.

En general, es importante recordar tres cuestiones esenciales acerca de la coincidencia o pareamiento:

- En primer lugar, los métodos de coincidencia o pareamiento solo pueden utilizar características observables para construir grupos de comparación, dado que las características no observables no se pueden considerar. Si hay alguna característica no observable que influye en la inscripción o no inscripción de la unidad en el programa, y que también influye en el resultado, las estimaciones de impacto obtenidas con el grupo de comparación pareado estarían sesgadas. Para que el resultado del pareamiento no esté sesgado, requiere un supuesto de mucho peso, a saber: que no hay diferencias no observables en el grupo de tratamiento y de comparación que también estén asociadas con los resultados de interés.
- Segundo, la coincidencia o pareamiento debe realizarse utilizando solo características que no estén afectadas por el programa. La mayoría de las características que se miden después del comienzo del programa no pertenecerían a esta categoría. Si los datos de *línea de base* (antes de la intervención) no están disponibles y los únicos datos son los existentes después de que la intervención ha comenzado, las únicas características que se podrán utilizar para construir una muestra pareada serán aquellas (normalmente pocas) características que no se ven afectadas por un programa, como la edad y el sexo. Aunque se quisiera parear utilizando un conjunto mucho más rico de características, entre ellas los resultados de interés, no se podrá hacerlo porque aquellas están potencialmente afectadas por la intervención. No se recomienda el pareamiento basado únicamente en características posteriores a la intervención. Si hay datos de *línea de base* disponibles, se puede realizar el pareamiento sobre la base de un conjunto más rico de características, entre ellas, los resultados de interés. Dado que los datos se recopilan antes de la intervención, el programa no puede haber afectado aquellas variables anteriores a la misma. Sin embargo, si hay datos de *línea de base* sobre los resultados disponibles, no se debería utilizar el método de pareamiento solo, sino que habría que

combinarlo con diferencias en diferencias para reducir el riesgo de sesgo. Este procedimiento se detallará en la próxima sección.

- Tercero, los resultados de la estimación del método de pareamiento solo son tan buenos como las características que se utilizan para el pareamiento.

Si bien es importante poder parear utilizando un gran número de características, lo es aún más poder parear sobre la base de características que determinan la inscripción. Cuanto más se comprenda acerca de los criterios utilizados para la selección de los participantes, en mejores condiciones se estará de construir el grupo de comparación.

Breve ejemplo de soporte común por observación

Suponga que toma una submuestra de **10 pbt** del grupo de tratamiento y **10 pbt** del grupo de **control** de del programa **pei**, con las siguientes probabilidades estimadas de *participación*. Ver **Tabla 4.1**.

**Tabla 4.1. Ejemplo determinación zona de soporte común
con datos hipotéticos**

	1	2	3	4	5	6	7	8	9	10
Tratamiento	0.40	0.48	0.54	0.65	0.75	0.76	0.77	0.8	0.86	0.9
Control	0.21	0.30	0.38	0.42	0.52	0.55	0.59	0.62	0.65	0.77

Fuente: Bernal y Peña (2011) con adaptación propia.

Si éste fuera el universo de estudio, la zona de *soporte común* está entre el *mínimo* del grupo de tratamiento y el *máximo* del grupo de control. Así, se restringe entonces el análisis a los grupos con probabilidades $P'(X_i) = p'E [0.40, 0.77]$. Los grupos de tratamiento y control restringidos son (ver **Tabla 4.2**).

Tabla 4.2. Ejemplo determinación zona de soporte común por grupos de tratamiento y control con datos hipotéticos

	1	2	3	4	5	6	7
Tratamiento	0.40	0.48	0.54	0.65	0.75	0.76	0.77
Control	0.21	0.30	0.38	0.42	0.52	0.55	0.77

Fuente: Bernal y Peña (2011) con adaptación propia.

Note que, dada la restricción al soporte común, perdimos tres observaciones del grupo de *tratamiento* y tres del grupo de *control*. Otra manera de restringir el soporte común consiste en tomar las zonas para las cuales existe probabilidad positiva en ambos grupos, es decir, elegir el rango de probabilidades de *participación* para el cual existen observaciones tanto del grupo de tratamiento como el de *control*. En este caso, también se eliminan las observaciones con probabilidades menores que **0.40** y mayores que **0.77**. Además, note que entre **0.54** y **0.65** no existen observaciones del grupo de tratamiento, y se eliminan además dos observaciones del grupo de control (ver **Tabla 4.3.**).

Tabla 4.3. Ejemplo determinación zona de soporte por zonas de probabilidad positiva en ambos grupos con datos hipotéticos

	1	2	3	4	5	6	7
Tratamiento	0.40	0.48	0.54	0.65	0.75	0.76	0.77
Control	0.21	0.30	0.38	0.42	0.52	0.55	0.77

Fuente: Bernal y Peña (2011) con adaptación propia.

TOT usando PSM

Si la independencia condicional se mantiene, y si hay una superposición considerable (*sizable overlap*) en $\mathbf{P}(\mathbf{X})$ entre los *participantes* y *no participantes*, el estimador de PSM para el TOT puede especificarse como la media de la diferencia en \mathbf{Y} sobre el *soporte común* (*common support*), ponderando las unidades de comparación por la distribución de la propensión de puntaje de los participantes. Un *estimador de sección transversal* típico, se especifica de la siguiente manera:

$$\text{TOT}_{\text{PSM}} = E_{P(\mathbf{X})|T=1} \{E[Y^T | T=1, P(\mathbf{X})] - E[Y^C | T=0, P(\mathbf{X})]\}. \quad (4.3)$$

Más explícitamente, con datos de *sección transversal* y dentro del *soporte común*, el efecto de tratamiento, puede escribir de la siguiente manera (ver Heckman et al., (1997) y Smith y Todd (2005):

$$\Rightarrow \text{TOT}_{\text{PSM}} = \frac{1}{N_T} \left[\sum_{i \in T} Y_i^T - \sum_{j \in C} \omega(i, j) Y_j^C \right] \dots\dots\dots (4.4)$$

Donde:

N_T es el número de participantes; i y $\omega(i, j)$ es la ponderación o peso usado para agregar las respuestas para coincidir a los no participantes j . Hay varios esquemas de ponderación disponibles para calcular resultados ponderados de los comparadores coincidentes.

Reglas de aplicación PSM

Para calcular el efecto del tratamiento del programa, primero se debe calcular la *propensión de puntaje* $\mathbf{P}(\mathbf{X})$ sobre la base de todas las covariables observadas \mathbf{X} que afectan conjuntamente la participación y el resultado de interés. El objetivo de la coincidencia, es encontrar el grupo de comparación más cercano de una muestra *de no participantes*

a la muestra de *participantes* del programa. Se mide el *más cercano* en términos de características observables no afectadas por la participación en el programa. Por lo que se sugiere seguir los siguientes pasos (Khandker et al., 2017):

Paso 1: Estimando un modelo de participación de programa

Las muestras de *participantes* y *no participantes* deben agruparse, y posteriormente, la participación T debe estimarse en todas las covariables observadas X en los datos que son probables, para determinar la participación. Cuando uno está interesado solo en comparar resultados para aquellos que participan ($T = 1$) con los que *no participan* ($T = 0$), esta estimación puede ser construida a partir de un modelo *probit* o *logit* de participación en el programa. **Caliendo y Kopeinig (2008)** también proporcionan ejemplos de estimaciones de la ecuación de participación, con una variable de *tratamiento* no binario. En esta situación, es posible usar un *multinomio probit* (que es computacionalmente intensivo, pero basado en supuestos más débiles que el *logit multinomial*) o una serie de modelos binomiales.

Después de estimar la *ecuación de participación*, los valores de predicción de T , se pueden calcular. *El resultado de predicción, representa la probabilidad estimada de participación o propensión de puntaje.* Cada *participante* y *no participante* de la muestra tendrá una *propensión de puntaje estimado*, $P'(X | T = 1) = P'(X)$. Tenga en cuenta que la *ecuación de participación*, no es un modelo determinante, por lo que los resultados de la estimación, como las estadísticas t y el R^2 ajustado no son muy informativos y pueden ser engañosos. Para esta etapa de **PSM**, la causalidad no es tan interesante como la correlación de X con T .

En cuanto a las covariables X relevantes, **PSM** estará sesgado si las covariables que determinan la participación no están incluidas en la ecuación de participación, por otros motivos. Estas razones podrían

ser, por ejemplo, datos de baja calidad o una comprensión deficiente del contexto local en el que se está introduciendo el programa. Como resultado, una orientación limitada existe sobre cómo seleccionar variables X usando pruebas estadísticas, porque las características observadas que tienen más probabilidades de determinar la participación, tienen probabilidad de que estén basados en datos y contexto específico (Dehejia, 2005 para mayores sugerencias de selección de covariables). Heckman et al., (1997, 1998) muestran que el sesgo en PSM, en las estimaciones del programa, pueden ser bajas, dadas tres disposiciones generales (Khandker et al., 2017):

- Primero, si es posible, se debe utilizar el mismo instrumento de encuesta o fuente de datos para los *participantes* y *no participantes*. El uso de la misma fuente de datos, ayuda a garantizar que las características observadas ingresen al *modelo logit o probit* de participación para que sean medidas de manera similar, a través de los dos grupos y por lo tanto, reflejan los mismos conceptos.
- Segundo, una encuesta de muestra representativa tanto de los *no participantes* elegibles como de los *participantes*, pueden mejorar en gran medida, la precisión de la *propensión de puntaje*. Además, cuanto mayor sea la muestra de *no participantes* elegibles, se facilitarán más las coincidencias. Si las dos muestras provienen de encuestas diferentes, entonces deberían ser encuestas altamente comparables (mismo cuestionario, mismos entrevistadores o capacitación de entrevistadores, mismo período de encuesta, etc.). Un punto relacionado, es que los *participantes* y los *no participantes*, deben enfrentar los mismos incentivos económicos que podrían impulsar opciones como la participación en el programa (Ravallion, 2008). El investigador podría explicar este factor al elegir *participantes* y *no participantes* de la misma área geográfica.
- Sin embargo, incluir demasiadas variables X en la ecuación de participación, debería evitarse; la sobreespecificación del modelo

puede generar errores estándar más altos para la *propensión de puntaje* estimado $P'(X)$ y también puede resultar en una predicción perfecta participación para muchas **pbt** ($P'(X)$). En el último caso, tales observaciones abandonarían el *soporte común* (como se analiza más adelante). Como se mencionó previamente, determinar la participación es un problema menor en la *ecuación participante*, que obtener una distribución de las probabilidades de participación.

Paso 2: Definiendo la región de soporte común y pruebas de balanceo

Posteriormente, la región de *soporte común*, necesita ser definida donde las distribuciones de propensión de puntaje al *tratamiento* y la superposición del grupo de *comparación*. Como se mencionó anteriormente, algunas de las observaciones *no participantes* pueden tener que descartarse porque caen fuera del área de *soporte común*. Sin embargo, el *sesgo de muestreo* aún puede ocurrir si las observaciones *no participantes*, son sistemáticamente diferentes en términos de las características observadas de la muestra *no participante* retenida; estas diferencias deben ser monitoreadas cuidadosamente para ayudar a interpretar el efecto del tratamiento. También, se pueden realizar *pruebas de equilibrio* para verificar si, dentro de cada cuartil de la distribución de la *propensión de puntaje*, la *propensión de puntaje promedio* y la *media de X* son iguales.

Para que la **PSM** funcione, los grupos de *tratamiento* y *comparación* deben estar equilibrados en los puntajes de propensión que se basan en una X similar observada. Aunque, un grupo tratado y su comparador no tratado coincidente, podrían tener los mismos puntajes de propensión, no son a nivel observación, similares si existe una especificación errónea en la ecuación de participación. Las distribuciones del *grupo tratado* y el *comparador* deben ser similares, lo que implica el equilibrio. Formalmente, el investigador necesita verificar que $P'(X | T = 1) = P'(X | T = 0)$.

Paso 3: Haciendo coincidentes participantes-no participantes

La metodología de **PSM** se puede entender como una manera de ponderar las observaciones del grupo de control para que la distribución de características observables X sea lo más parecida posible a la del grupo de tratamiento. Todos los estimadores por coincidencia o emparejamiento, contrastan la variable de resultado de un individuo *tratado* con los resultados de uno o más miembros del grupo de *control* que más se parezcan al individuo *tratado*, con base en la medida $P(X)$. Sin embargo, los estimadores **PSM** difieren tanto en la manera en que se encuentra este grupo de individuos de control parecidos al tratamiento, como en la manera como estos vecinos son ponderados a la hora de hacer la comparación. Algunos de los estimadores más usados son el **PSM** por vecino más cercano (**NN** *Nearest-Neighbor matching*) y el **PSM** por kernel.

Se pueden utilizar *diferentes criterios de coincidencia* para asignar *participantes a no participantes* en la base de *propensión de puntaje*. Hacerlo, implica calcular un peso para cada conjunto coincidente *participante-no participante*. Como se discute a continuación, la elección de una técnica de emparejamiento particular por lo tanto, puede afectar la estimación del programa resultante a través de los pesos asignados con varias técnicas a saber, para selección del algoritmo de coincidencia o emparejamiento más idóneo (**Khandker et al., 2017; Bernal y Peña, 2011**):

1. *Vecino más cercano (NN. Nearest-Neighbor matching)*. Una de las técnicas de coincidencia o emparejamiento más utilizadas, es ésta también llamada **NN** coincidente, donde cada unidad de tratamiento, se hace coincidente a la unidad de comparación, con la propensión de puntaje más cercano. También se pueden elegir n vecinos más cercanos y hacer coincidencia (generalmente se usa $n = 5$). La coincidencia, se puede hacer con o sin reemplazo. Hacer coincidir con reemplazo, por ejemplo, significa que el mismo *no participante* se puede usar como una coincidencia para diferentes

participantes. Así, el individuo en el grupo de *control*, es tal que la distancia entre la probabilidad predicha de participación y la probabilidad predicha de participación del sujeto en el grupo de tratamiento, *sea mínima*. Una vez se ha coincidido o emparejado a cada sujeto *participante* en el programa con uno o más individuos del grupo de *control*, se calcula la diferencia entre la variable de resultado de los sujetos de *tratamiento* y la variable de resultado de los sujetos de *control*. El **ATT** se obtiene promediando estas diferencias de forma lineal. Si tomamos solo al vecino más cercano, el impacto del programa se calcula simplemente como la diferencia entre las variables de resultado. Si en vez de tomar al individuo con la probabilidad de participación más cercana al tratamiento decidimos tomar a los dos, tres o de manera general los n vecinos más cercanos, se debe decidir cuántos vecinos cercanos elegir para cada observación del grupo de tratamiento y qué peso darle a cada una de estas observaciones al calcular el impacto. Otra alternativa es hacer un promedio ponderado, según qué tan comparable es cada vecino. Es decir, la diferencia entre la variable objetivo del tratamiento con la de cada control se pondera por la distancia en la probabilidad de participar. Así, al impacto medido con un vecino muy cercano se le da más peso que al impacto medido con un vecino no tan cercano. La decisión de cuántos vecinos más cercanos usar genera una tensión (*trade-off*) entre el sesgo del estimador y su varianza. Al tener más de una observación en el grupo de control el emparejamiento es peor, pues a mayor número de vecinos, estaremos comparando individuos con probabilidades de participación más lejanas y, por tanto, *menos parecidos* entre sí. Esto genera *sesgo* en las estimaciones. Sin embargo, como se usan más observaciones diferentes para calcular el contrafactual de cada participante, *se reduce la varianza*. Este método se puede aplicar *con o sin reemplazo*. En caso de permitir *reemplazo*, un individuo del grupo de control puede ser el vecino más cercano para más de un individuo

de tratamiento; se usa solo una vez, en el caso *sin reemplazo*. Permitir el reemplazo genera de nuevo tensión (*trade-off*) entre el *sesgo y la varianza*. Al permitir el *reemplazo*, la calidad promedio del emparejamiento sube y, por tanto, el sesgo disminuye. Pero se reduce el número de individuos diferentes que se usan para calcular el contrafactual y, por tanto, *las estimaciones son menos precisas*. *Por lo general, este método se aplica con reemplazo*. En el caso del **PSM** por vecino más cercano (**NN**), todos los individuos del grupo de tratamiento en el *soporte común* son emparejados con algún miembro del grupo de control para el cálculo del **ATT**. Algunas de estas parejas pueden ser cuestionables, dado que para algunos individuos de tratamiento el vecino más cercano podría tener una probabilidad de participación $P(X)$ bastante distinta, 43 y aún contribuir al cálculo del **ATT**. Recuerde que el vecino más cercano de un individuo tratado i es aquel que tiene la diferencia en la probabilidad de participación más baja con i que cualquier otro individuo del grupo de control. Esta diferencia, sin embargo, puede no ser lo suficientemente pequeña como para estar seguros de que los dos individuos son muy parecidos, porque esta distancia mínima no tiene ninguna restricción de tamaño. Existe por otro lado, una variante que de le llega a conocer como emparejamiento de distancia máxima (**FN**. *Farthest-Neighbor*) que ocurre cuando si el vecino más cercano es lejano a la observación, el estimador de **PSM** por vecino más cercano puede emparejar observaciones que no son muy parecidas. Una manera de solucionar esto consiste en imponer un nivel de tolerancia en la distancia entre probabilidades de *participación*. Coincidir o emparejar a cada individuo del grupo de *tratamiento* con la observación más cercana, siempre y cuando la distancia entre las probabilidades de participación sea menor o igual a un número determinado, mejora la calidad de la coincidencia o emparejamiento. También se puede hacer coincidir o emparejar a cada individuo del grupo de *tratamiento* con todos

los individuos de *control* dentro del nivel de tolerancia. Bajo este enfoque, se compara cada observación con tantas parejas como existan dentro del nivel de tolerancia. Así, el grupo de *control* crece si hay varias parejas cercanas y decrece sino. Por tanto, la comparabilidad de las parejas se asegura y se usa eficientemente la información disponible. Para calcular el **ATT** se sigue un procedimiento similar al caso de n vecinos cercano. Si solo existe una observación del grupo de *control*, el impacto del programa sobre ese tratado es la diferencia entre la variable objetivo entre el *tratado* y el *control*. Si hay más de un *control* en la distancia máxima, se hace un promedio (ponderado o no) de las diferencias en la variable objetivo entre el *tratamiento* y cada uno de los controles. No es claro, sin embargo, cómo determinar *ex ante*, el nivel de tolerancia, es decir, qué distancia entre las probabilidades de participación, es razonable.

2. *Calibración por coincidencia. (Caliper or radius matching)*. Un problema con la técnica **NN** coincidente, es que la diferencia entre la propensión de puntajes de un *participante* y su vecino *no participante* más cercano más cercano, todavía puede ser muy alto. Esta situación da como resultado, pobres cálculos de coincidencia que se es posible evitar al imponer un umbral o *tolerancia* sobre la distancia máxima de propensión de puntaje, a manera de calibración (*caliper*). Por lo tanto, este procedimiento implica hacer coincidir solo la propensión de puntajes, dentro de cierto rango, y reemplazar. Un gran número de no participantes no considerados, sin embargo, es probable que aumente la posibilidad de sesgo de muestreo.
3. *Estratificación o intervalo de coincidencia. (Stratification or interval matching)*. Este procedimiento, divide *el soporte común* en diferentes estratos (o intervalos) y calcula el impacto del programa dentro de cada intervalo específicamente, dentro de cada intervalo, el efecto del programa es la diferencia de la media en los resultados

entre las *observaciones tratadas* y *las de control*. Un promedio ponderado de estas estimaciones de impacto de intervalo, produce el impacto general del programa, tomando la proporción de participantes en cada intervalo como los pesos. También se le conoce como estimador de **PSM** por estratificación (*subclassification o blocking*) y hace una partición en el espacio de las probabilidades estimadas, y lo divide en *estratos* o rangos de probabilidad de participación. Posteriormente, se calcula el impacto en cada estrato, por ejemplo en el estrato *j*, como la diferencia promedio en la variable de resultado entre los grupos de *tratamiento* y *control*. El **ATT** se obtiene como la media ponderada de las diferencias por estrato, siendo los pesos la proporción de individuos tratados en el estrato sobre el total de individuos tratados en el soporte común. Surge una cuestión: ¿cómo determinar cuántos estratos armar? como se mencionó anteriormente, dentro de cada estrato tanto la probabilidad de participación promedio como la media de las variables de control debe ser igual entre los grupos de *tratamiento* y *control*. A esto se le llama que los grupos de *tratamiento* y *control* estén *balanceados*. Si la probabilidad no está balanceada, el intervalo es demasiado grande y es necesario partirle. Una vez balanceada la probabilidad de participación, es necesario checar que las variables observadas, *X*, estén balanceadas también. Si no lo están, es necesario reespecificar la probabilidad de participación, por ejemplo, incluyendo términos de interacción. Basándose en **Cochran (1968)**, en la práctica con frecuencia se usan *cinco estratos*. Tomemos el siguiente caso como ejemplo. Una vez que se restringe la muestra al *soporte común*, comparamos las implicaciones de elegir diferentes métodos de emparejamiento entre observaciones del grupo de *tratamiento* y *control*. Usando estos datos ficticios se ejemplificarán los diferentes algoritmos de emparejamiento y, suponiendo que la calidad del emparejamiento es adecuada, se estimará el **ATT**. Volviendo al del **pei** para **pbt**, la siguiente tabla contiene

información restringida al soporte común tanto de la probabilidad de participación como de la variable resultado, que en este caso es el gasto percápita de las **pbt** anual medida en un factor convenido monetario (ver **Tabla 4.4**):

Tabla 4.4. Ejemplo estimador PSM por estratificación

	Observación	1	2	3	4	5	6	7
Tratamiento	Probabilidad de participación P'	0.4	0.48	0.54	0.65	0.75	0.76	0.77
	Gasto percápita de la pbt anual Y_i	69	74	73	74	77	70	80
Control	Probabilidad de participación p'	0.42	0.52	0.55	0.59	0.62	0.65	0.77
	Gasto percápita de la pbt anual Y_i	71	69	70	72	78	70	74

Fuente: Bernal y Peña (2011) con adaptación propia.

Al realizar las coincidencias o emparejamientos, la observación de *tratamiento* con probabilidad de participación **0.40** por *vecino más cercano* se hace coincidente o empareja con el de *control* con probabilidad **0.42**, y su contribución al **ATT** es **69-71 = -2**. El *tratamiento* con probabilidad **0.48** se empareja con el control **0.52** y la contribución es **74-69 = 5**. Ya habíamos descrito el emparejamiento del *tratamiento* **0.54** y su contribución al **ATT** es **73-70 = 3**. El *tratamiento* con probabilidad **0.65** se empareja con el control **0.62** y la contribución es **74-70 = 4**. Los *tratamientos* con probabilidad **0.75**, **0.76** y **0.77** se emparejan con el control **0.77** y las contribuciones son: **77-74 = 3**, **70-74 = -4** y **80-74 = 6** respectivamente. En este caso el **ATT** es el promedio simple de las contribuciones individuales: **$(-2+5+3+4+3-4+6)/7 = 2.1429$** . Si en cambio, se usa el algoritmo de *tres vecinos más cercanos*, el grupo de control de la **pbt** del grupo de *tratamiento* con probabilidad de participación

de **0.54** serían las observaciones con probabilidad de participación **0.52, 0.55 y 0.59**. El promedio de la variable de resultado para estas tres observaciones es **70.33**. La contribución de este individuo al **ATT**, es: $73 - 70.33 : 2.67$; el **ATT** es, de nuevo, el promedio simple de la contribución de cada observación del grupo de *tratamiento*. También se puede usar la coincidencia o emparejamiento de *distancia máxima*, tomando todas las observaciones cuya distancia no exceda una diferencia en probabilidades de participación de **0.03**. Así, el grupo de control para la observación **0.54** tendría dos observaciones (**0.52 y 0.55**), mientras que la observación de **0.65** del grupo de tratamiento no tendría grupo de control. Para la coincidencia o emparejamiento por estratificación, suponga que la manera óptima de partir los grupos es en dos estratos, y que el punto de corte es **0.6**. En este caso, el impacto en el primer estrato, con probabilidades de participación por debajo de **0.6**, es el promedio de la variable resultado entre los tratamientos, menos el promedio en el grupo de control: $72 - 70.5 = 1.5$. En el segundo estrato, con probabilidades de participación por encima de **0.6**, el impacto del programa es $75.25 - 74 = 1.25$. El **ATT** es el promedio del efecto sobre cada estrato ponderado por la proporción de tratados en el soporte común. En este ejemplo los pesos serían **3/7** para el primer estrato y **4/7** para el segundo. Es decir, el **ATT** sería $(1.5 \times 3/7) + (1.25 \times 4/7) = 1.3571$. La estimación por estratificación es menos sensible a la especificación de la probabilidad de participación dado que estamos volviendo discreta la distribución de los pesos asignados a cada observación. En vez de darle a cada observación un peso igual al inverso de la probabilidad de participación, le asigna el mismo peso a todas las observaciones en un mismo estrato. Otra ventaja de este algoritmo es que, dado que se comparan observaciones dentro del mismo estrato, no se hace extrapolación en los resultados. es decir, no usan los parámetros

estimados de una población y se los aplica a otra muy diferente para estimar los impactos.

4. *Núcleo y coincidencia local (Kernel and local matching)*. Hasta ahora hemos usado información de un subconjunto de sujetos del grupo de *control* para construir el resultado contrafactual de un individuo *tratado*. Los PSM por *kernel* y regresión local lineal (**RLL**. *Regresion Local Linear*) son estimadores *no paramétricos* que hacen coincidente o emparejan a cada sujeto del grupo de tratamiento con un promedio ponderado de (potencialmente) todos los individuos del grupo de control. El *kernel* es un procedimiento estadístico que se basa exclusivamente en los datos, sin hacer supuestos paramétricos a priori sobre los modelos económicos. La función de *kernel* pondera los datos del grupo de *control*, dando más peso a las observaciones con probabilidades de participación cercanas, y menos a aquellas con probabilidades lejanas. En la práctica las funciones *kernel* son continuas por partes (*piecewise continuous*), acotadas, simétricas alrededor de cero, cóncavas en cero y, por conveniencia, generalmente se integran a uno. Es necesario determinar la vecindad a cada lado de la probabilidad de participación del individuo i que se quiere utilizar para determinar los individuos que se comparan con i . Esta elección es crítica en los métodos de estimación *no paramétrica*. En principio, con estos estimadores se puede comparar a cada individuo de tratamiento con todos los individuos de control, y darles menos peso en la comparación a aquellos controles con probabilidades de participación más lejanas. Tomar una vecindad más amplia implica un mayor número de observaciones y, por tanto, estimadores que pueden ser más eficientes. Sin embargo, estaremos comparando al individuo de tratamiento con controles que no son muy parecidos a él, de tal manera que el sesgo se incrementa. En la elección de la vecindad hay una tensión entre sesgo potencial del estimador del efecto del programa y su varianza. Si la vecindad es pequeña, se reduce el sesgo porque los

individuos son todos más parecidos, pero se aumenta la varianza porque hay pocas observaciones. En la coincidencia o emparejamiento por *kernel* se compara al individuo tratado i con su respectivo grupo de *control*, apropiadamente ponderados por una función de qué tan lejos se encuentra cada individuo de *control* de i con respecto a la probabilidad de *participación*. Dependiendo de la función de *kernel* que se elija, se estará comparando a cada individuo del grupo de *tratamiento* con todos los individuos del grupos de *control*, como es el caso del *kernel gaussiano* o el uniforme, o con un subconjunto de grupo de *control*, como en el caso de los *kernels triangular* o el *Epanechnikov*. El emparejamiento por *kernel* se puede interpretar como una *regresión* de la variable objetivo del grupo de control en el vecindario de cada observación del grupo de tratamiento sobre un intercepto, usando los pesos del kernel (Smith y Todd, 2005). El **RLL** es esta misma regresión sobre un intercepto, pero se le adiciona un término de pendiente, que es la probabilidad de participación (de allí el nombre de *lineal*, porque se asume que la variable de resultado cambia linealmente con la probabilidad de participación). En ambos casos el intercepto estimado es la contribución de la observación del individuo del grupo de tratamiento al **ATT**. Como se mencionó anteriormente, el *kernel Gaussiano* utiliza todos los individuos del grupo de *control* para construir la contribución de cada individuo del grupo de tratamiento al **ATT**. En este caso, el ancho de banda determina el peso relativo de las observaciones lejanas vs. cercanas. Los *kernels triangular* y *Epanechnikov* solo utilizan algunas observaciones (las más cercanas) del grupo de control. En este caso, el ancho de banda desempeña un doble papel. No solo determina el peso relativo de las observaciones lejanas vs. cercanas, sino que además determina la vecindad a cada lado de la probabilidad de participación que se quiere utilizar para incluir las observaciones del grupo de control en la *estimación del contrafactual*. Los pesos usados para calcular

el contrafactual construidos con base en un *kernel* dependen del tipo de *kernel* elegido. Por ejemplo, en el *kernel* uniforme, se les da el mismo peso a todas las observaciones en el grupo de control elegido. Otros *kernels* usados con frecuencia asignan pesos inversamente proporcionales a la distancia entre la probabilidad de participación $P(X)$ de cada individuo del grupo de control y la probabilidad de participación del individuo *tratado* para el cual se está estimando el *contrafactual*. Una ventaja importante de los estimadores de *kernel* y **RLL** es que tienen *menor varianza* (por ejemplo, que **PSM** por vecino más cercano,) porque usan más información **RLL**, al incluir el término lineal de la probabilidad de participación, aventaja al emparejamiento por *kernel* cuando los controles están distribuidos de manera asimétrica alrededor del tratamiento (por ejemplo, en los extremos de la distribución de $P(X)$). No existe un método para determinar el ancho de banda pero en general éste debe ser menor cuanto mayor sea la muestra. Un riesgo con los métodos que hemos descrito, es que solo un *pequeño subconjunto de no participantes finalmente satisfará los criterios* para caer dentro del soporte común y así construir el resultado *contrafactual*. Los estimadores de coincidencia no paramétricos, como la coincidencia de *kernel* (*kernel matching*) y los métodos lineales locales (**LLM**. *Lineal Local Matching*), utilizan un *promedio ponderado* de todos los *no participantes* para construir la coincidencia contrafactual para cada participante. Si P_i es la propensión de puntaje propensión para el *participante i* y P_j es la propensión de puntaje para el *no participante j*, y si se sigue la notación en la **ecuación 4.4**, se infiere que las ponderaciones para la *coincidencia de núcleo* (*kernel matching*), sea:

$$\omega(i, j)_{LM} = \frac{K\left(\frac{P_j - P_i}{a_n}\right)}{\sum_{k \in C} K\left(\frac{P_k - P_i}{a_n}\right)} \dots\dots\dots (4.5)$$

Donde:

$K(\cdot)$ es la *función de núcleo (kernel)* y a_n es el parámetro de ancho de banda. **LLM** en contraste, estima una regresión no paramétrica ponderada localmente (lowess) del grupo de comparación, resultado en la vecindad de cada observación de tratamiento (**Heckman et al., 1997**). La coincidencia de núcleo (*kernel matching*) es análoga, a la regresión en un término constante, mientras que **LLM** utiliza un término *constante* y un término de *pendiente*, por lo que es *lineal*. **LLM** puede incluir una tasa más rápida de convergencia cerca de los puntos de frontera (**Fan 1992, 1993**). El estimador **LLM** tiene la misma forma que el estimador de coincidencia de núcleo (*kernel matching*), excepto por la función de ponderación:

$$E[\hat{\tau}_{LLM}] = \frac{K_e \sum_{i \in C} K_a(P_i - P_t)^2 - [K_e(P_t - P_t)] \sum_{i \in C} K_a(P_i - P_t)}{\sum_{j \in C} K_a \sum_{i \in C} K_a(P_i - P_t)^2 - \left(\sum_{i \in C} K_a(P_i - P_t) \right)^2} \dots\dots\dots(4.6)$$

5. *Coincidencia de diferencia en diferencia. (DD. Difference-in-Difference Matching)*. Con datos sobre las observaciones de *control* y *participantes*, *antes* y *después* de la intervención del programa, una *diferencia en la diferencia (DD. Difference-in-Difference)* se construye como estimador de coincidencia. La **DD**, se discute en mayor detalle en el capítulo 5; importante, pero de lo más importante a destacar, es que permite que características no observadas afecten la adopción del programa, suponiendo que estos rasgos *no observados no varían con el tiempo*. Para presentar al estimador **DD**, tenemos que volver a la configuración de la sección transversal del estimador **PSM** de la **ecuación 4.4**. Suponga datos de panel de dos períodos $t = \{1,2\}$, el estimador de **DD** lineal local (*local linear DD estimador*) para la diferencia de medias en los

resultados Y_{it} entre los *participantes* i y los *no participantes* j con soporte común, se tiene:

$$TOT_{PSM}^{DD} = \frac{1}{N_T} \left[\sum_{i \in T} (Y_{it}^T - Y_{it}^C) - \sum_{j \in C} \omega(i, j) (Y_{jt}^C - Y_{jt}^T) \right] \dots \dots \dots (4.7)$$

Con solo las secciones transversales en el tiempo, o en lugar de datos de panel (**Todd, 2007**) donde TOT_{PSM}^{DD} , se reescribe:

$$TOT_{PSM}^{DD} = \frac{1}{N_{T_2}} \left[\sum_{i \in T_2} Y_{i2}^T - \sum_{j \in C_2} \omega(i, j) Y_{j2}^C \right] - \frac{1}{N_{T_1}} \left[\sum_{i \in T_1} Y_{i1}^T - \sum_{j \in C_1} \omega(i, j) Y_{j1}^C \right] \dots (4.8)$$

Donde:

Y_{it}^T y Y_{it}^C , $t = \{1, 2\}$, son los resultados de las observaciones de diferentes *participantes* y *no participantes* en cada período t . El enfoque **DD** combina la el enfoque tradicional de **PSM** y **DD** que se analizan en el próximo capítulo. Las características de lo *observado* así como también, de lo *no observado* que afectan la participación, pueden explicarse si suponemos que los *factores no observados* que afectan la participación, son *constantes a lo largo del tiempo*. Tomando la diferencia en los resultados a lo largo del tiempo, también debería diferenciar las características no observadas del tiempo invariante y, por lo tanto, el *sesgo de selección* potencial no observada. De nuevo, el capítulo 5 trata este tema en detalle. También, se puede usar un *estimador ajustado por regresión* (descrito con más detalle en este capítulo y en el capítulo 5). Este método, supone el uso de un *modelo lineal estándar* para resultados para estimar el **TOT** (como $Y_i = \alpha + \beta T_i + \gamma X_i + \varepsilon_i$) aplicando pesos sobre la base de la propensión de puntaje al grupo de comparación coincidente. También puede permitir que el investigador controle la selección de *características no observadas*,

suponiendo nuevamente que estas *características no varían con el tiempo*. Por lo tanto, se pueden usar varios pasos para *unir a los participantes con los no participantes*. Una de las ventajas del método de coincidencia o emparejamiento, es que se puede aplicar a un único levantamiento de información, siempre y cuando existan observaciones de tratamientos y controles. Sin embargo, si hay información longitudinal, o cortes transversales repetidos, se pueden combinar los métodos de emparejamiento con el de *diferencias-en-diferencias*, de tal manera que se relajan los supuestos de cada uno de estos métodos, y, por tanto, los resultados se vuelven más robustos. En adelante, supondremos que contamos con información longitudinal, para facilitar la exposición. Recuerde que el **PSM** asume que la selección en el programa, se basa única y exclusivamente en características observables de los individuos. Al contar con información longitudinal y combinar **PSM** con *diferencias-en-diferencias*, la selección en el programa puede también depender de *variables no observables*, siempre y cuando éstas permanezcan *constantes* en el tiempo.

¿Qué algoritmo de emparejamiento elegir?

En muestras grandes, asintóticamente todos los algoritmos de coincidencia o emparejamiento, deben generar los mismos resultados, puesto que tienden a comparar sólo parejas idénticas. En muestras pequeñas, sin embargo, la selección de algoritmo puede afectar los resultados estimados, pues con frecuencia existe una tensión entre dos objetivos deseables: *tener un estimador insesgado y minimizar la varianza*. La selección del estimador depende de las características de cada caso, pues ningún algoritmo es siempre mejor que los demás. Es mejor, entonces, ser pragmáticos e intentar diversos algoritmos. Si los resultados son muy similares, entonces la selección del algoritmo es poco importante. Pero si

los resultados difieren, es necesario determinar qué genera la disparidad, antes de elegir un algoritmo (**Bryson et al., 2002**).

Todos los métodos de emparejamiento descritos anteriormente se pueden entender como una asignación de pesos para promediar las observaciones de los individuos de control, de tal manera que el grupo de control sea idéntico al grupo de tratamiento. Estos mismos pesos se utilizan para calcular la calidad del emparejamiento, como se describe en la siguiente sección (**Bernal y Peña, 2011**).

Cálculo del promedio de impacto en el grupo de tratamiento

Como se discutió anteriormente, si la *independencia condicional* (*conditional independence*) y una *superposición considerable* (*seizable overlap*) en la propensión de puntaje, entre *participantes* y *no participantes* coincidentes, puede ser supuesta, entonces el promedio del efecto por tratamiento **PSM** es igual a la diferencia de la media (*mean difference*) de los resultados sobre el *soporte común*, ponderando las unidades de comparación por la distribución de la propensión de puntaje de los participantes. Para comprender el potencial de los mecanismos observados que conducen el efecto estimado del programa, es posible examinar el impacto del tratamiento, en diferentes características observables, como la posición en la distribución muestral de ingresos, edad, etc.

Calidad de la coincidencia o emparejamiento

Como el emparejamiento no se hace condicional en todas las variables observadas sino únicamente en la probabilidad de participación, es necesario determinar si la coincidencia o emparejamiento, balancea la distribución de las variables relevantes entre los grupos de tratamiento

y control. Para que **PSM** funcione, condicional en la probabilidad de participación, los grupos de tratamiento y control deben ser similares o estar balanceados. Esto implica que tanto la probabilidad de participación promedio como las medias de las variables contenidas en el vector X sean idénticas entre los grupos de tratamiento y control. Así, se quiere determinar si después de condicionar por la probabilidad de participación aún existen diferencias en las características observables entre los grupos de tratamiento y control. Note que este análisis se realiza usando únicamente información de las variables de control, y no las de resultado, y por tanto hacen parte de la etapa de diseño. Sin embargo, se presentan después de la elección del método de emparejamiento pues para realizar los ejercicios que permitirán determinar la calidad del emparejamiento necesitamos los pesos (*weights*) para ponderar las variables del grupo de control; estos pesos dependen del método de coincidencia o emparejamiento.

¿Cómo evaluar si las coincidencias o emparejamiento fueron correctas? Intuitivamente, después de controlar por el efecto de la probabilidad de participación, condicionar en las variables observables no generar nueva información acerca del tratamiento se pueden calcular estadísticos t en las muestras de *tratamiento* y *control* para determinar si hay diferencias significativas en el promedio de las variables de control entre los grupos de *tratamiento* y *control*. Las variables del grupo de control deben estar apropiadamente ponderadas por los pesos, que dependen del método de emparejamiento. Es de esperar que haya diferencias antes del emparejamiento, pero después de éste la distribución de variables observables debe estar balanceada entre los dos grupos y no debe haber diferencias estadísticamente significativas. Otra manera de determinar la calidad del emparejamiento consiste en usar la idea de estratificación: se divide el espacio de las probabilidades estimadas de los grupos de tratamiento y control en *estratos* o rangos de probabilidad de participación. Los estratos se eligen de tal manera que esté balanceada la probabilidad de participación entre ellos, es decir, que sea igual el promedio de la probabilidad

de participación estimada de los grupos de tratamiento y control. Posteriormente, se usan estadísticos t dentro de cada estrato para determinar si los primeros dos momentos de la distribución de las variables en X son idénticos entre grupos. Si la muestra está balanceada, se prosigue con el cálculo del impacto del programa, descrito en la sección de selección del algoritmo de emparejamiento. De lo contrario, es necesario reestimar la probabilidad de participación incluyendo términos que añadan flexibilidad, como términos cuadráticos o cúbicos, interacciones o diferentes variables en la estimación de la probabilidad de participación, hasta que estas diferencias desaparezcan (Bernal y Peña, 2011).

Estimación de los errores estándar con PSM. Uso de Bootstrap

Una vez calculado el **ATT**, es necesario calcular los *errores estándar* y con ellos los *intervalos de confianza de las estimaciones*, para determinar si el impacto del programa es estadísticamente significativo. Una de las desventajas del **PSM** es que calcular los errores estándar no es fácil, pues deben incorporar la varianza debida a la estimación de la probabilidad de participación. Si el emparejamiento se hizo *sin reemplazo*, también se debe incluir la varianza asociada a este método de emparejamiento. Los errores estándar, en este caso de los estimadores de **PSM**, se pueden calcular de dos maneras. Por un lado, se pueden calcular analíticamente, es decir, encontrando una expresión explícita de la varianza. Por ejemplo, al usar el método de vecino más cercano, es posible calcular los errores de manera analítica (Bernal y Peña, 2011).

En comparación con los métodos de regresión tradicionales, la varianza estimada del efecto del tratamiento en **PSM**, debe incluir la varianza atribuible a la derivación de la propensión de puntaje, la determinación del soporte común y (si hace la coincidencia sin reemplazo) el

orden en que se hacen coincidentes los individuos tratados (**Caliendo y Kopeinig, 2008**). No tener en cuenta esta variación adicional, más allá de la variación normal de muestreo, provocará que los errores estándar se estimen incorrectamente (**Heckman et al., 1998**).

Sin embargo, cuando las fórmulas para la varianza son tan complicadas que no se conoce la distribución asintótica del estimador, ni se tiene una fórmula para el cálculo de errores estándar, los errores se pueden calcular usando *bootstrapping*. Éste es el caso, por ejemplo, del **PSM** por *kernel*. El *bootstrapping* consiste en estimar propiedades de un estimador, por ejemplo, su varianza, midiendo dichas propiedades en la distribución empírica de los datos. Cuando se puede asumir que las observaciones vienen de una población independiente e idénticamente distribuida, el *bootstrapping* se puede implementar construyendo una serie de muestras (*con reemplazo*) proveniente de la base de datos original. El cálculo de los errores estándar por ambos métodos debe generar resultados similares. La técnica del *bootstrapping* (**Efron et al., 2003**), se extraen muestras repetidas de la muestra original y los estimados de las propiedades (tales como el error estándar y sesgo) se vuelven a calcular con cada muestra. Cada estimación (*bootstrap*) de muestra, incluye los primeros pasos de la estimación que derivan la propensión del puntaje, el soporte común, etc. La justificación formal para los estimadores (*bootstrap*), son limitados; sin embargo, debido a que los estimadores son asintóticamente lineales, el *bootstrapping* probablemente conducirá a errores estándar válidos e intervalos de confianza (**Imbens, 2004**). **Abadie e Imbens (2008)** muestran que el *bootstrapping* no funciona para los estimadores de n vecinos cercanos, ya que en este caso no genera bandas de confianza asintóticamente válidas. Por tanto, para el estimador de n vecinos cercanos se sugiere calcular los errores estándar de manera analítica.

PSM y los métodos de regresión

Dado que las coincidencias o emparejamiento resultante, produce estimaciones consistentes bajo condiciones *débiles*, una ventaja práctica de **PSM** sobre los mínimos cuadrados ordinarios (**OLS**. *Ordinary Least Squares*) es que reduce el número de dimensiones en las cuales hacer coincidentes (emparejar) a las unidades de participación y de comparación. Sin embargo, las estimaciones consistentes de **OLS** del **ATE** se pueden calcular bajo el supuesto de condicional de exogeneidad (*conditional exogeneity*). Un enfoque sugerido por **Hirano et al. (2003)** es estimar una *regresión ponderada de mínimos cuadrados* de resultado en el tratamiento T y otras covariables observadas X no afectadas por la participación, *utilizando el inverso de una estimación no paramétrica* de la propensión de puntaje. Este enfoque lleva a un estimador eficiente, y el efecto del tratamiento se estima mediante:

$$Y_{it} = \alpha + \beta T_{it} + \gamma X_{it} + \varepsilon_{it}$$

Con pesos de 1 para los participantes y pesos de $P'(X) / (1 - P'(X))$ para el control observaciones T_{it} es el indicador de tratamiento, y los intentos de especificación anteriores, para tener en cuenta las diferencias latentes entre las unidades de tratamiento y comparación que afectan la selección en el programa, así como los resultados. Para una estimación de **ATE** para la población, los pesos serían $1 / P'(X)$ para los *participantes* y $1 / (1 - P'(X))$ para las unidades de *control*.

La combinación de la coincidencia con otros métodos

Aunque la técnica de coincidencia o pareamiento, *requiere un volumen importante de datos y tiene un riesgo significativo de sesgo*, ha sido utilizada para evaluar programas de desarrollo en una amplia gama de contextos (**Gertler, et al., 2017**). Los usos más convincentes de la coincidencia o pareamiento son aquellos que lo combinan con otros *métodos* y aquellos que utilizan el *método de control sintético*. En esta sección, se analizarán las diferencias en *diferenciascoincidentes o pareadas* y el método de control sintético.

Diferencias en diferencias pareadas

Cuando dispone de datos de *línea de base* sobre los resultados, el pareamiento se puede combinar con diferencias en diferencias para reducir el riesgo de sesgo en la estimación. Como se ha analizado, el simple pareamiento con propensión de puntajes, no puede dar cuenta de características no observables que podrían explicar por qué un grupo decide inscribirse en un programa, y eso también podría afectar los resultados. La *coincidencia o pareamiento combinado con diferencias en diferencias*, al menos tiene en cuenta cualquier característica no observable *que sea constante* a lo largo del tiempo entre ambos grupos. Se implementa en 5 pasos, de la siguiente manera:

1. El pareamiento debe realizarse a partir de características observables de la *línea de base* (como se ha señalado).
2. Para cada unidad inscrita, se debe calcular el cambio en los resultados entre los períodos antes y después (primera diferencia).
3. Para cada unidad inscrita, calcúlese el cambio en los resultados entre los periodos antes y después para la comparación pareada de esta unidad (segunda diferencia).

4. Réstese la segunda diferencia de la primera diferencia, es decir, aplíquese el método de diferencias en diferencias.
5. Por último, calcúlese un promedio de esas dobles diferencias.

El método de control sintético

El método de control sintético permite utilizar la estimación del impacto en contextos donde una única unidad (como un país, una empresa o un hospital) es objeto de una intervención o se expone a un evento. En lugar de comparar esta unidad tratada con un grupo de unidades no tratadas, el método usa información sobre las características de la unidad tratada y las unidades no tratadas para construir una unidad de comparación “sintética” o artificial, ponderando cada unidad no tratada de tal manera que la unidad de comparación sintética se asemeje lo más posible a la unidad tratada. Esto requiere una extensa serie de observaciones de las características de la unidad tratada y de las unidades no tratadas a lo largo del tiempo. Esta combinación de unidades de comparación en una unidad sintética proporciona una mejor comparación para la unidad tratada que cualquier unidad no tratada individualmente.

Crítica al PSM: ventajas y desventajas

Aunque los procedimientos de coincidencia o pareamiento se pueden aplicar en numerosos contextos, independientemente de las reglas de asignación de un programa. Así, las principales ventajas de **PSM**, son (**Khandker et al., 2017**):

- Se basa en el grado en que las características observadas, impulsan la participación en el programa. Si el *sesgo de selección* de las *características no observadas* es con toda probabilidad insig-

nificante, entonces **PSM** puede ser una técnica adecuada para la comparación con *estimaciones aleatorias*.

- En la medida en que las variables de participación se encuentren incompletas, los resultados **PSM** pueden ser *sospechosos*. Esta condición, como se mencionó anteriormente, no es comprobable directamente con criterios; requiere de un examen cuidadoso de los factores que impulsan la participación en el programa (por ejemplo, a través de encuestas).
- No requiere necesariamente una *línea de base* o *encuesta de panel*, aunque en la sección transversal resultante, las covariables observadas que ingresan al *modelo logit* para la propensión de puntaje, tendrían que satisfacer el supuesto de la *independencia condicional* reflejando las *características observadas X* que no son afectadas por la participación. Una línea base previa al programa, es más útil a este respecto, ya que cubre la observación de variables *X* que son independientes del estado del tratamiento.
- Como se discutió anteriormente, los datos sobre *los participantes* y *los no participantes* a lo largo del tiempo, también pueden ayudar a contabilizar para algunos *no observados* con sesgo de selección, combinando enfoques tradicionales de **PSM** con supuestos **DD** (detallado en el capítulo 5).
- **PSM** también es un *método semiparamétrico*, que impone menos restricciones en la forma funcional del *modelo de tratamiento*, así como menos suposiciones sobre la distribución del término de error. Aunque las observaciones se descartan para lograr el *soporte común*, **PSM** aumenta la probabilidad de comparaciones razonables entre las unidades en tratamiento y de control, disminuyendo potencialmente el sesgo en el impacto del programa. Este resultado es cierto sin embargo, solo si el soporte común es robusto en cuanto a datos suficientes sobre los *no participantes* ya que es esencial para garantizar una muestra lo suficientemente grande, de la cual extraer coincidencias.

- El sesgo puede también resultar de la eliminación de observaciones *no participantes* que son sistemáticamente diferentes de los retenidos; este problema también se puede aliviar mediante la recopilación de datos en de una gran muestra de *no participantes*, con suficiente variación para permitir una muestra representativa. De lo contrario, examinar las características de la muestra *no participante* eliminada puede refinar la interpretación del efecto del tratamiento.

Por otro lado, se considera que tienen varias limitaciones o desventajas importantes (Gertler, et al. 2017):

- En primer lugar, requieren conjuntos de datos amplios sobre grandes muestras de unidades, e incluso cuando estos están disponibles, puede que se produzca una falta de rango común entre el grupo de tratamiento, o inscrito, y el conjunto de no participantes.
- En segundo lugar, solo se puede aplicar el pareamiento basándose en características observables; por definición, no se pueden incorporar las características no observables en el cálculo de la propensión de puntaje. Por lo tanto, para que el procedimiento de pareamiento identifique un grupo de comparación válido, no deben existir diferencias sistemáticas en las características no observables entre las unidades de tratamiento y las unidades de comparación pareadas que podrían influir en el resultado (**Y**). Dado que no se puede demostrar que existen esas características no observables que influyen en la participación y en los resultados, se debe suponer que no existen. Normalmente se trata de un supuesto de mucho peso. A pesar de que el pareamiento contribuye a controlar por características básicas observables, nunca se puede descartar el sesgo que nace de las características no observables.
- En resumen, el supuesto de que no se ha producido un sesgo de selección debido a las características no observables es de mucho

peso y, lo que es más problemático, no puede comprobarse. El pareamiento por sí solo suele ser menos robusto que los otros métodos de evaluación analizados, dado que requiere el fuerte supuesto de que no hay características no observables que influyan simultáneamente en la participación en el programa y en sus resultados. Por otro lado, la asignación aleatoria, la variable instrumental y el diseño de regresión discontinua no requieren el supuesto inde demostrable de que no hay tales variables no observables. Tampoco requieren muestras tan grandes o características básicas tan amplias como el pareamiento por puntajes de propensión.

En la práctica, los métodos de pareamiento suelen usarse cuando no es posible recurrir a las opciones de asignación aleatoria, variable instrumental y diseño de regresión discontinua. El denominado pareamiento *ex post* es muy riesgoso cuando no hay datos de *línea de base* disponibles sobre el resultado de interés o de las características básicas. Si una evaluación utiliza datos de encuestas que fueron recopilados después del comienzo del programa (es decir, *ex post*) para deducir las características básicas de las unidades de la *línea de base* y posteriormente, emparejar el grupo tratado con un grupo de comparación empleando esas características deducidas, puede emparejar involuntariamente basándose en características que también fueron afectadas por el programa; en ese caso, el resultado de estimación sería inválido o estaría sesgado.

Por el contrario, cuando se dispone de datos de *línea de base*, el pareamiento basado en las características básicas puede ser muy útil si se combina con otras técnicas, como el método de diferencias en diferencias, lo que permite corregir por las diferencias entre los grupos que son fijas a lo largo del tiempo. El pareamiento también es más fiable cuando se conocen las reglas de asignación del programa y las variables fundamentales, en cuyo caso el pareamiento se puede llevar a cabo con esas variables.

A estas alturas, es probable que quede claro que las evaluaciones de impacto se diseñan mejor antes de que un programa comience a ser implementado. Una vez que el programa ha comenzado, si hay que influir en cómo se asigna y no se han recopilado datos de *línea de base*, habrá pocas o ninguna opción rigurosa para la evaluación de impacto.

Pruebas de falsificación

El supuesto de independencia condicional, permite la identificación de los impactos de interés en el método de pareo. Lo que este supuesto implica es que la selección depende únicamente de variables observables. Por tanto, para eliminar el sesgo de selección es suficiente controlar por el conjunto de variables observables relevantes.

El supuesto de independencia condicional, sin embargo, no tiene implicaciones que se puedan comprobar directamente en los datos y, por tanto, no es posible determinar de manera directa si se cumple o no. Sin embargo, es posible realizar algunos ejercicios para evaluar si la estrategia de identificación tiene validez, comúnmente llamados *pruebas de falsificación*.

Imbens y Rubin (2010) sugieren dos tipos de pruebas de falsificación:

1. Estimar el efecto del programa sobre una variable de resultado ficticia, de la cual tenemos la seguridad de que no puede ser afectada por el tratamiento (*pseudoresultado*). Consiste en estimar el efecto causal de la variable de *tratamiento* sobre una variable que claramente no puede haber sido afectada por éste, o *pseudoresultado*. Típicamente, se eligen las variables de resultado rezagadas, medidas en uno o más periodos de tiempo anteriores a la intervención. La selección de cuál variable es adecuada para hacer la prueba de falsificación es específica al contexto. Por ejemplo, si nos inte-

resa el efecto de un programa de una capacitación en el periodo t sobre los ingresos futuros Y_{i+1} , una prueba de falsificación podría ser estimar el efecto del entrenamiento sobre los ingresos de los individuos dos años antes de entrar al programa Y_{i-2} , controlando por todo el historial anterior de ingresos que tengamos disponible (Y_{i-3} , Y_{i-4} , etc.). Note que una vez se controla por este historial de ingresos pasados, la intervención en t no debería tener efectos sobre Y_{i-2} . Si no se controla apropiadamente por el historial de ingresos, es posible que si se encuentre un efecto de la intervención sobre los ingresos rezagados Y_{i-2} simplemente porque estaría capturando trayectorias de ingresos sistemáticamente diferentes entre individuos participantes e individuos no participantes que, a su vez, determinaron que unos participaran en el programa y otros no. Por otra parte, si se encuentra que la intervención no tiene efectos sobre Y_{i-2} , controlando por el historial de ingresos, es plausible que se cumpla el supuesto de selección en observables. Si la intervención si afecta Y_{i-2} , entonces esto sería evidencia que sugiere que no se cumple el supuesto de selección por características observables y, por tanto, el método de coincidencia o emparejamiento no sería adecuado para estimar el impacto del programa de entrenamiento. Note que al igual que la *evaluación de la calidad* del emparejamiento, esta prueba de falsificación se realiza usando únicamente información de las *variables de control* y no las de *resultado*. Si la prueba de falsificación falla, la coincidencia o emparejamiento no es un buen método para calcular el impacto, pues la selección seguramente no se da solo en características observables. *Sin embargo, esta prueba desafortunadamente no genera sugerencias acerca de qué métodos serían más adecuados.* Seguramente se tendrían que usar métodos que controlen por el *sesgo de selección*, debido a variables observables y también a variables no observables o no disponibles en los datos, tales como variables instrumentales,

regresión discontinua o los demás descritos en los capítulos posteriores.

2. Estimar el efecto causal de una intervención ficticia que se sabe no debería tener efectos sobre nuestra variable de resultado (*pseudotrataamiento*). La segunda prueba de falsificación, tiene que ver con estimar el efecto causal de una intervención *ficticia*, que se sabe que *no tiene efectos*, o *pseudotrataamiento*. Para esto necesitaríamos, por ejemplo, tener información de *múltiples grupos de control*. Como ejemplo, para estimar los efectos de un programa de capacitación en innovación para una ciudad en particular, podemos tener dos grupos de control:
 - a. El primero está compuesto por individuos elegibles (vivían en la ciudad) pero decidieron no tomar el entrenamiento (control a).
 - b. El segundo está compuesto por aquellos que *vivían justo por fuera del límite de la ciudad*, y por tanto no eran elegibles para beneficiarse del tratamiento, pero que en otras dimensiones eran comparables a los beneficiarios (control b).

En este caso estimaríamos el impacto del *pseudotrataamiento* *pertenecer al grupo de control a* con respecto al grupo de comparación *pertenecer al grupo de control b* sobre los ingresos en el periodo posterior al entrenamiento. Sabemos que este pseudotrataamiento no debe tener impacto alguno. Si los cálculos sugieren que el pseudotrataamiento no tiene efectos, es plausible que se cumpla el supuesto de selección en observables. Sino, el método de coincidencia o emparejamiento no es adecuado para estimar el impacto del programa.

Estas pruebas de falsificación deberían ser parte de la etapa de diseño. Este análisis, debe realizarse incluso antes de calcular los impactos del programa. Si los análisis sugieren que la estrategia de identificación no es creíble, esto implica que el aplicar la metodología propuesta, no tiene

sentido pues no estaremos identificando el impacto deseado. Las pruebas, sin embargo, no sugieren qué camino tomar; solo hacen explícito que el método elegido no es apropiado para estudiar los impactos, ya que no se cumplen los supuestos de identificación. Las *pruebas de falsificación* se pueden realizar no solo en el contexto de coincidencia o emparejamiento, sino también, para los métodos discutidos en varios de los capítulos posteriores.

Ejemplo utilizando STATA

La idea básica detrás de la propensión de coincidencia de puntaje (**PSM**, *Propensity Score Matching*) es hacer coincidente o emparejar a cada *participante* con un *no participante* idéntico y posteriormente, medir la diferencia promedio en la variable de resultado de interés, entre los *participantes* y los *no participantes*. Este ejercicio ilustra cómo implementar **PSM** en el programa **STATA**.

El comando que lo ejecuta en **STATA** es **pscore.ado**, desarrollado por **Becker e Ichino (2002a)**. El comando **pscore**, estima el puntaje de propensión, que es la probabilidad de obtener un tratamiento para cada sujeto (en nuestro caso, **pbt**) y probar la propiedad de equilibrio, es decir, las observaciones con la misma propensión de puntaje, deben tener la misma distribución de características observables independientes, del estado del tratamiento. Después de hacer el equilibrio, es posible usar diferentes comandos para llevar a cabo diferentes tipos de correspondencia y posteriormente, derivar el efecto promedio del tratamiento.

Ecuación PSM en la satisfacción de la propiedad de balance

Se proponen realizar, los siguientes pasos para su resolución:

- El primer paso en **PSM**, es determinar la propensión de puntaje y la propiedad de equilibrio. Se realiza utilizando el comando **pscore** en **STATA**.
- Utilice los datos del periodo 08 de **pei_08.dta**. Comience con la variable de participación masculina del programa **dmmfd** como la variable de tratamiento.
- Se utiliza el comando *pscore.do* de **Becker e Ischino (2002b)**. Se destaca que si dicho comando no se encuentra disponible, deberá teclear:

```
search pscore
```

Seleccionar la actualización de dicho comando en la versión de la cual disponga, en nuestro caso **STATA 14.1**, como se ve en la imagen:

```
Web resources from Stata and other users
[connecting http://www.stata.com]
18 packages found (Stata Journal and STB listed first)
-----
r00328 from http://www.stata-journal.com/software/sj14-1
3214=1 s00328: Estimation of the generalized... / Estimation of the
generalized propensity score / through generalized linear models / by
Barbara Guardabascio, Istat, Italian National / Institute of Statistics,
Rome, Italy / Marco Venturi, Istat, Italian National / Institute of
```

- El siguiente comando muestra la aplicación del comando **pscore**:

```
. use "/Users/DCA/Desktop/STATA log files/pei_08.dta"
. gen lnsmejac=ln(1+smefac/100)
```

```
. pscore dmmfd genceo ageceo educeo lnsme fac bdbaccess pcirr
raw1 raw2 raw3 raw5 raw6 [pw=weight], pscore (ps98) blockid
(blockf1) comsup level (0.001)
```

- Los resultados incluyen los de la regresión *probit*, la estimación y la descripción de la propensión de puntajes, el número de bloques y estratificación usando propensión de puntajes, y las prueba de propiedad de equilibrio. El área de *soporte común*, la constituyen los puntajes de propensión dentro del rango de los valores estimados más bajos y más altos para las **pbt** en el grupo de tratamiento.
- El siguiente resultado, muestra que la región identificada de soporte común es [.00180123, .50022341], el número final de bloques es **4** y la propiedad de equilibrio no está satisfecho.
- El elemento más importante a buscar en la salida, es la lista de variables, que hace que, la propiedad de equilibrio no se satisfaga. La salida muestra la variable raw5 no está equilibrada en el bloque 2. La solución a este problema es usar un conjunto diferente de covariables y vuelva a ejecutar el comando *pscore*. **Ver Tabla 4.5.**

Tabla 4.5. PSM para prueba 1 de balanceo

 Algorithm to estimate the propensity score

The treatment is dmmfd

innovation male microcredit participant : 1=Y, 0=N	Freq.	Percent	Cum.
0	909	80.51	80.51
1	220	19.49	100.00
Total	1,129	100.00	

Estimation of the propensity score

(sum of wgt is 1.1260e+03)
 Iteration 0: log pseudolikelihood = -424.61883
 Iteration 1: log pseudolikelihood = -390.85321
 Iteration 2: log pseudolikelihood = -389.10243
 Iteration 3: log pseudolikelihood = -389.05511
 Iteration 4: log pseudolikelihood = -389.05501

Probit regression	Number of obs	=	1129
	Wald chi2(11)	=	64.36
	Prob > chi2	=	0.0000
Log pseudolikelihood = -389.05501	Pseudo R2	=	0.0838

dmmfd	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
genceo	.915108	.2432905	3.76	0.000	.4382675	1.391949
ageceo	-.0036952	.0046186	-0.80	0.424	-.0127475	.005357
educeo	.0161662	.0170125	0.95	0.342	-.0171777	.04951
lnsmefac	-.3341691	.1113146	-3.00	0.003	-.5523417	-.1159965
bdbaccess	-.0752904	.1770457	-0.43	0.671	-.4222935	.2717128
pcirr	.2088394	.1753383	1.19	0.234	-.1348174	.5524961
raw1	.145771	.0384417	3.79	0.000	.0704268	.2211153
raw2	.0465751	.0648087	0.72	0.472	-.0804475	.1735977
raw3	-.0017358	.023861	-0.07	0.942	-.0485026	.045031
raw5	-.7687454	.2311995	-3.33	0.001	-1.221888	-.3156028
raw6	-.0249797	.0135856	-1.84	0.066	-.051607	.0016476
_cons	-1.188481	.8358266	-1.42	0.155	-2.826671	.4497088

Note: the common support option has been selected
 The region of common support is [.00180123, .50022341]

Description of the estimated propensity score
 in region of common support

Estimated propensity score

	Percentiles	Smallest		
1%	.0055359	.0018012		
5%	.0170022	.0020871		
10%	.0346036	.0026732	Obs	1,127
25%	.069733	.0028227	Sum of Wgt.	1,127
50%	.1206795		Mean	.1339801
		Largest	Std. Dev.	.0850809
75%	.1811405	.4698302		
90%	.2527064	.472444	Variance	.0072388
95%	.2965199	.4735467	Skewness	.8931864
99%	.3903884	.5002234	Kurtosis	3.942122

Step 1: Identification of the optimal number of blocks

Use option detail if you want more detailed output

The final number of blocks is 4

This number of blocks ensures that the mean propensity score is not different for treated and controls in each blocks

Step 2: Test of balancing property of the propensity score

Use option detail if you want more detailed output

Variable raw5 is not balanced in block 2

The balancing property is not satisfied

Try a different specification of the propensity score

Inferior of block of pscore	innovation male microcredit participant: 1=Y, 0=N		Total
	0	1	
0	380	49	429
.1	382	97	479
.2	140	70	210
.4	5	4	9
Total	907	220	1,127

Note: the common support option has been selected

End of the algorithm to estimate the pscore

Fuente: STATA con datos propios.

- Después de algunas iteraciones, se encontrará que con omitir *raw5* y *lnsmefac* permitirá que el comando *pscore* se vuelva a ejecutar, con la propiedad de equilibrio satisfecha.
- Antes de volver a ejecutar el comando *pscore*, también deben retirarse las variables *ps98* y *blockf1*, que fueron creadas como última ejecución del programa, dado que la participación de CEOs femeninas es de mayor interés. Así, se deberá cambiar a *dfmfd* así como teclear:

gen lnsmefac=ln(1+smefac/100)

pscore dfmfd genceo ageceo educeo lnsmefac bdbaccess pcirr
raw1 raw2 raw3 raw5 raw6 [pw=weight], pscore (ps98) blockid
(blockf1) comsup level (0.001). **Ver Tabla 4.6.**

Tabla 4.6. PSM para prueba 1 de balanceo

Algorithm to estimate the propensity score

The treatment is dfmfd

innovation female microcredit participant : 1=Y, 0=N	Freq.	Percent	Cum.
0	534	47.30	47.30
1	595	52.70	100.00
Total	1,129	100.00	

Estimation of the propensity score

(sum of wgt is 1.1260e+03)
Iteration 0: log pseudolikelihood = -750.38718
Iteration 1: log pseudolikelihood = -682.82636
Iteration 2: log pseudolikelihood = -680.63459
Iteration 3: log pseudolikelihood = -680.62452
Iteration 4: log pseudolikelihood = -680.62452

Probit regression

Log pseudolikelihood = -680.62452

Number of obs = 1129
Wald chi2(11) = 85.21
Prob > chi2 = 0.0000
Pseudo R2 = 0.0930

dfmfd	Robust		z	P> z	[95% Conf. Interval]	
	Coef.	Std. Err.				
genceo	-.037986	.1662857	-0.23	0.819	-.3639	.287928
ageceo	.0013931	.0037305	0.37	0.709	-.0059185	.0087047
educeo	-.0465567	.0151559	-3.07	0.002	-.0762618	-.0168516
lnsmefac	-.6662184	.101586	-6.56	0.000	-.8653232	-.4671136
bdbaccess	-.1173796	.13358	-0.88	0.380	-.3791916	.1444323
pcirr	.4304416	.154365	2.79	0.005	.1278917	.7329915
raw1	.0571981	.0307982	1.86	0.063	-.0031652	.1175615
raw2	-.0055393	.056959	-0.10	0.923	-.1171769	.1060982
raw3	.015395	.0184184	0.84	0.403	-.0207044	.0514944
raw5	-.1114687	.1647319	-0.68	0.499	-.4343373	.2113999
raw6	.0235048	.01239	1.90	0.058	-.000779	.0477887
_cons	-1.483823	.7367316	-2.01	0.044	-2.927791	-.0398558

Note: the common support option has been selected

The region of common support is [.02576077, .71555996]

Description of the estimated propensity score
in region of common support

Estimated propensity score				
	Percentiles	Smallest		
1%	.0515941	.0257608		
5%	.1326484	.0301603		
10%	.1918711	.034091	Obs	1,124
25%	.3237772	.0343378	Sum of Wgt.	1,124
50%	.4419659		Mean	.4174089
		Largest	Std. Dev.	.1488984
75%	.5230785	.7051936		
90%	.6033309	.70725	Variance	.0221707
95%	.6323751	.71107	Skewness	-.501193
99%	.6918163	.71556	Kurtosis	2.723861

Step 1: Identification of the optimal number of blocks
 Use option detail if you want more detailed output

The final number of blocks is 4

This number of blocks ensures that the mean propensity score
is not different for treated and controls in each blocks


```
*****
Step 2: Test of balancing property of the propensity score
Use option detail if you want more detailed output
*****
```

The balancing property is satisfied

This table shows the inferior bound, the number of treated and the number of controls for each block

Inferior of block of pscore	innovation female microcredit participant: 1=Y, 0=N		Total
	0	1	
.0257608	91	31	122
.2	180	146	326
.4	222	333	555
.6	36	85	121
Total	529	595	1,124

Note: the common support option has been selected

```
*****
End of the algorithm to estimate the pscore
*****
```

Fuente: STATA con datos propios.

Con la propensión de puntajes generados, los resultados de interés (tales como, el gasto total por **pbt**) tales como la coincidencia o emparejamiento entre el grupo de *tratamiento* y el grupo de *control* combinado, ahora son verificados, para ver si los programas de microcrédito, afectan al resultado de interés. Las siguientes secciones, estiman el efecto del tratamiento de la participación en el programa de microcrédito, utilizando diferentes técnicas de correspondencia que están disponibles.

Efecto de tratamiento promedio usando vecino más cercano (*nearest-neighbor*)

El comando para estimar el efecto de tratamiento promedio en en los tratados, usando la coincidencia o emparejamiento por vecino más cercano (**NN**. *Nearest-Neighbor*) es *attnd*. A continuación se muestra un ejemplo de su aplicación, para estimar el efecto del tratamiento promedio de la participación de CEOs femeninas de las **pbt** en el pei por micro-crédito, considerando el gasto total per cápita de las **pbt** y utilizando la correspondencia del vecino más cercano, por lo que se deberá realizar. Ver **Tabla 4.7**.

```
gen lexptot=ln(1+exptot)
attnd lexptot dfmfd [pweight=weight], pscore(ps98) consup
```

Tabla 4.7 Comandos y resultados del ATT por vecino más cercano
(*nearest-neighbor*)

```
. gen lexptot=ln(1+exptot)
. attnd lexptot dfmfd [pweight=weight], pscore(ps98) consup
```

The program is searching the nearest neighbor of each treated unit.
This operation may take a while.

ATT estimation with Nearest Neighbor Matching method
(random draw version)
Analytical standard errors

n. Treat.	n. contr.	ATT	Std. Err.	t
595	293	0.136	0.042	3.256

Note: the numbers of treated and controls refer to actual nearest neighbour matches

Command

Fuente: STATA con datos propios.

Al usar el comando *attnd* con o sin pesos, no afecta los resultados. Solo para lo propósitos de este ejercicio, *attnd* se muestra con la estimación de los pesos. Como muestra el siguiente resultado, la participación de las CEOs femeninas de las **pbt** en el **pei** con microcrédito, tiene como significado de que no puede afectar el gasto per cápita de los **pbt** con la técnica del vecino más cercano ($t = 3.256$). El *tratamiento promedio de los tratados (ATT)* del gasto per cápita de las **pbt** para la participación de CEOs femeninas en el programa es del **13.6 %**.

Efecto de tratamiento promedio usando estratificación de coincidencia (*stratification matching*)

El comando *att*, es el que calcula el efecto de tratamiento promedio en los tratados, usando estratificación de coincidencia (*stratification matching*), sobre las CEOs femeninas de las **pbt** que participan en el **pei** por microcréditos, por lo que:

- Para calcularlo sobre los gastos per capita de las **pbt**, se teclea lo siguiente:
`atts lexptot dfmfd, pscore(ps98) blockid(blockf1) comsup`
- El resultado muestra, **9.9%** de incremento de gastos per capita **pbt**, debido a la participación de las CEO femeninas de las **pbt** en el **pei** por microcréditos. El impacto es significativo en el nivel del **5%** ($t=3.320$). Ver **Tabla 4.8**.

Tabla 4.8. Comandos y resultados del ATT por estratificación de coincidencia (*stratification matching*)

. atts lexptot dfmfd, pscore(ps98) blockid(blockf1) comsup

ATT estimation with the Stratification method
Analytical standard errors

n. treat.	n. contr.	ATT	Std. Err.	t
595	529	0.099	0.030	3.320

Command

Fuente: STATA con datos propios.

Efecto de tratamiento promedio usando calibración por coincidencia (*radius matching*)

El comando *attr*, es el que calcula el efecto de tratamiento promedio en los tratados, usando calibración por coincidencia (*radius matching*), sobre las CEOs femeninas de las **pbt** que participan en el **pei** por micro-créditos, por lo que:

- Para calcularlo sobre los gastos per capita de las **pbt**, se teclea lo siguiente:

attr lexptot dfmfd, pscore(ps98) radius(0.001) comsup

- El resultado muestra, un incremento en el impacto (**14.6%**) con alta significancia (**t=3.793**) de CEOs femeninas de las **pbt** que

participan en el **pei** sobre el gasto per capita de las **pbt**. Ver **Tabla 4.9**.

Tabla 4.9. Comandos y resultados del ATT por calibración por coincidencia (*radius matching*)

```

. att lexptot dfmfd, pscore(pscore) radius(0.001) consup

```

The program is searching for matches of treated units within radius.
This operation may take a while.

ATT estimation with the Radius Matching method
Analytical standard errors

n. treat.	n. contr.	ATT	Std. Err.	t
478	386	0.146	0.039	3.793

Note: the numbers of treated and controls refer to actual matches within radius

Command

Fuente: STATA con datos propios.

Efecto de tratamiento promedio usando núcleo y coincidencia local (*kernel matching*)

El comando *attk*, es el que calcula el efecto de tratamiento promedio en los tratados, usando núcleo y coincidencia local (*kernel matching*), sobre las CEOs femeninas de las **pbt** que participan en el **pei** por microcréditos, por lo que:

- Se consideran *reps* que actúa con el uso de bootstrapping en 50 veces.

- Para calcularlo sobre los gastos per capita de las **pbt**, se teclea lo siguiente:

```
attk lexptot dfmfd, pscore(ps98) comsup bootstrap reps(50)
```

- Los resultados se consideran consistentes con los hallazgos encontrados hasta ahora. La participación de las CEOs femeninas de las **pbt** que participan en el **pei** con microcréditos incrementan su participación por gasto per capita de la **pbt** con una significancia del **4% al 5%. Ver Tabla 4.10**

Tabla 4.10. Comandos y resultados del ATT por núcleo y coincidencia local (*kernel matching*)

```
. attk lexptot dfmfd, pscore(ps98) comsup bootstrap reps(50)
```

```
The program is searching for matches of each treated unit.
This operation may take a while.
```

ATT estimation with the Kernel Matching method

n. treat.	n. contr.	ATT	Std. Err.	t
595	529	0.107	.	.

Note: Analytical standard errors cannot be computed. Use the bootstrap option to get bootstrapped standard errors.

Bootstrapping of standard errors

```
command:      attk lexptot dfmfd , pscore(ps98) comsup bwidth(.06)
statistic:    attk          = r(attk)
```

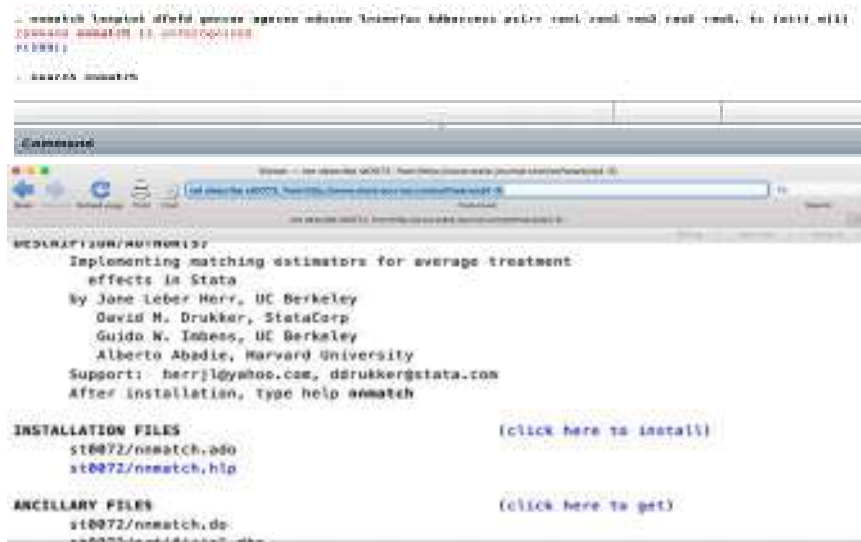
Command

Fuente: STATA con datos propios.

Efecto de tratamiento promedio y su chequeo de robustez

Hay varias formas de verificar la robustez de los resultados:

- Un enfoque, es estimar la ecuación de puntaje de propensión y luego usar los diferentes métodos de coincidencia previamente discutidos para comparar los resultados. Los hallazgos con diferentes técnicas de emparejamiento son bastante consistentes
- Otra forma de verificar la robustez, es aplicar la *coincidencia directa del vecino más cercano* (NN) en lugar de estimar primero la ecuación de propensión de puntaje. **STATA** tiene un comando *nnmatch* para hacer eso. Si ambos métodos dan resultados similares, entonces se supone, que son más confiables. Observe que si *nnmatch* ni se encuentra como comando disponible, con teclear *search nnmatch*, tendrá opción de actualizarlo, como se muestra:



- El siguiente comando **STATA** estimará el efecto promedio del tratamiento en el resultado de interés, utilizando la *coincidencia directa del vecino más cercano* (NN) con una coincidencia (*match=1*) por

tratamiento, la opción *m* especifica el número de coincidencias (*matches*) más cercanas a las observaciones tratadas.

```
. nnmatch lexptot dfmfd genceo ageceo educeo lnsmejac bdbac-  
cess pcirr raw1 raw2 raw3 raw5 raw6, tc (att) m(1)
```

- Los resultados de nueva cuenta son consistentes con los hallazgos previos. Un **13.6%** de impacto positivo de participación de las CEOs femeninas de las **pbt** en el **pei** por microcrédito sobre los gastos per capita de la **pbt**, los cuales se aprecian con un **5%** de nivel de significancia. Ver **Tabla 4.11**.

Tabla 4.11. Comandos y resultados del ATT para checar su robustez

```
. nnmatch lexptot dfmfd genceo ageceo educeo lnsmejac bdbaccess pcirr raw1 raw2 raw3 raw5 raw6, tc (att) m(1)
```

Matching estimator: Average Treatment Effect for the Treated

Weighting matrix: inverse variance Number of obs = 1129
Number of matches (m) = 1

Lexptot	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
SATT	.1368467	.0377988	3.62	0.000	.0619672 .2117384

Matching variables: genceo ageceo educeo lnsmejac bdbaccess pcirr raw1 raw2 raw3 raw5 raw6

Command

Fuente: STATA con datos propios.

Para mayor información sobre el contenido de los archivos **ado** componentes de las rutinas expuestas, consulte:

Becker, S.O.e Ichino (A). (2002a). *Stata programs for ATT estimation based on propensity score matching*, en: <http://sobecker.userweb.mwn.de/pscore.html>

Becker, S. e Ichino, A. (2002b). Estimation of Average Treatment Effects Based on Propensity Scores. *Stata Journal* 2 (4): 358–77.

CAPÍTULO 5.

Doble Diferencia (DD. *Double Difference*)

Los dos métodos discutidos en los capítulos anteriores: *evaluación aleatoria* y **PSM** se concentran en varios estimadores de *diferencia única* que a menudo requieren solo una apropiada *encuesta transversal*. Este capítulo, analizará la *técnica de estimación de doble diferencia* que generalmente utiliza el *panel datos*. Tenga en cuenta, sin embargo, que **DD** puede usarse también, en datos de sección transversal repetidos, siempre que la composición del participante y los grupos de control, sean claramente estables en el tiempo (**Khandker et al., 2017**). De hecho, asignación aleatoria, variables instrumentales (**IV**) y diseño de regresión discontinua (**DRD**) (estos dos a analizar más adelante), estiman el contrafactual mediante reglas explícitas de asignación del programa que el equipo de evaluación conoce y entiende. Dichos métodos ofrecen estimaciones creíbles del contrafactual haciendo relativamente pocas suposiciones e imponiendo pocas condiciones. Los dos próximos tipos de métodos, doble diferencia (**DD**) y coincidencia o pareamiento (**PSM**), ofrecen al equipo de evaluación, herramientas adicionales que pueden aplicarse cuando las reglas de asignación del programa son menos claras o cuando no es factible ninguno de los tres métodos antes descritos. En casos como este, se suele utilizar tanto el método de **DD** como el de **PSM**. Sin embargo, ambos también requieren supuestos más fuertes que la *asignación aleatoria*, o los métodos de **IV** o **DRD**. Se entiende que si no se conoce la regla de asignación del programa, hay una incógnita más en la evaluación, acerca de la cual se deben formular supuestos. Dado que los supuestos no son necesariamente verdaderos, puede que el método **DD** o

el **PSM** *no siempre proporcionen estimaciones fiables de los impactos de los programas* (Gertler et al., 2017).

Siguiendo a **Peña y Benal (2011)**, el modelo de **DD**, es comúnmente utilizado para estimar el efecto del programa tanto en experimentos aleatorios, cuando existen varios levantamientos de información, como en experimentos naturales, se describe a continuación. Esta metodología es común en el análisis de cuasi experimentos, debido a que es posible que, aun en presencia del evento fortuito que genera una asignación del tratamiento que parece aleatoria, se observen diferencias preexistentes entre el grupo de tratamiento y el grupo de control. Por tanto, es necesaria una metodología que corrija por esa diferencia preexistente entre los dos grupos.

Por otra parte, si el evento fortuito determina solamente de manera parcial el tratamiento, entonces el método de evaluación no puede ser el mismo que se aplica a los experimentos aleatorios controlados. La razón es que el tratamiento será parcialmente endógeno (el *sesgo de selección* está presente en cierta medida) y, por ende, la comparación de medias entre el grupo de tratamiento y el grupo de control no es válida. En este caso, se utilizan las metodologías que se discuten en los capítulos que siguen, y que también se aplican en los estudios no experimentales. Típicamente, en experimentos naturales la aleatorización que origina el evento fortuito no es perfecta, como podría serlo en un experimento aleatorio controlado. Por ende, es muy probable que existan diferencias sistemáticas entre el grupo de tratamiento y el grupo de control, aun antes de la aplicación del tratamiento. Por tal razón, es importante tener en cuenta estas diferencias preexistentes a la hora de estimar el efecto del programa sobre la variable de resultado. La razón es que la diferencia entre el grupo de tratamiento y el grupo de control en el período posterior al tratamiento estaría asociada tanto al tratamiento en sí como a diferencias que ya estaban presentes antes de la implementación del programa. El modelo de **DD** es una manera de controlar por estas posibles diferencias preexistentes entre los dos grupos.

El modelo de **DD** es simplemente el cambio esperado en Y entre el periodo posterior y el periodo anterior a la implementación del tratamiento en el grupo de tratamiento, menos la diferencia esperada en Y en el grupo de control durante el mismo periodo. Como su definición lo indica, el estimador de **DD** diferencias requiere de la existencia de datos panel, es decir, observaciones de los mismos individuos antes y después de la implementación del tratamiento (en un experimento aleatorio controlado o en un experimento natural). De hecho, en el caso de un experimento natural y en ausencia de datos tipo panel, es posible implementar el estimador **DD** utilizando datos de corte transversal repetidos. (Bernal y Peña, 2011). En el siguiente cuadro, se describe la información requerida:

Items	Tratamiento	Control
$t = 1$ (línea de base)	$Y_1 \mid D=1$	$Y_1 \mid D=0$
$t = 2$ (seguimiento)	$Y_2 \mid D=1$	$Y_2 \mid D=0$

donde $t = 1$ es el periodo anterior a la implementación del experimento aleatorio o la ocurrencia del evento fortuito que da origen al experimento natural (comúnmente conocido como la línea de base), y $t = 2$ es el periodo posterior a la implementación del tratamiento o periodo de seguimiento. El subíndice de Y indica el período al que corresponde esa observación de la variable de resultado.

En una configuración de panel, la estimación **DD** resuelve el problema de la falta de datos al medir resultados y covariantes tanto para los *participantes* como para los *no participantes* en los periodos previos y posteriores a la intervención. **DD** esencialmente, compara el *tratamiento* y los *grupos de comparación* en términos de cambios en los resultados a lo largo del tiempo, en relación con los resultados observados para una preintervención de *línea de base*. Es decir, dada una configuración de dos periodos, donde $t = 0$ antes del programa y $t = 1$ después de la imple-

mentación del programa, dejando Y_t^T y Y_t^C como los resultados respectivos para un programa beneficiario y unidades no tratadas en el tiempo t , el método **DD** estimará el impacto promedio del programa como sigue (Khandker, et al., 2017):

$$DD = E(Y_1^T - Y_0^T / T_1 = 1) - E(Y_1^C - Y_0^C / T_1 = 0) \dots \dots \dots (5.1)$$

En la **ecuación 5.1**, $T_1=1$ denota tratamiento o la presencia del programa en $t = 1$, mientras que $T_1 = 0$ denota áreas no tratadas. La siguiente sección, vuelve a esta ecuación. A diferencia de **PSM**, el estimador **DD** permite una *heterogeneidad no observada* (la diferencia no observada en los resultados contrafactuales medios, entre las unidades *tratadas y no tratadas*) que pueden conducir a un *sesgo de selección*. Por ejemplo, es posible que se requiera tener en cuenta *factores no observados* por el investigador, como las diferencias en la capacidad o liderazgo innato, a través de sujetos *tratados y de control* o los efectos de la colocación de programas no aleatorios a nivel de formulación de políticas de implementación de innovaciones. **DD** supone que esta heterogeneidad no observada es en tiempo invariante, por lo que el *sesgo se cancela mediante la diferenciación*. En otras palabras, el resultado cambia para los *no participantes* revelando los cambios de resultados contrafactuales como se muestra en **ecuación 5.1**.

Método DD

El método **DD** contrasta las diferencias en los resultados a lo largo del tiempo entre una población inscrita en un programa (el grupo de tratamiento) y una población no inscrita (el grupo de comparación). Suponga que se lleva a cabo a nivel de distrito pero que no se puede asignar aleatoriamente entre distritos ni se asigna sobre la base de un índice con un umbral claramente definido, lo cual permitiría un diseño de regresión

discontinua. Las juntas de los distritos pueden decidir inscribirse o no inscribirse en el programa. Uno de los objetivos del programa consiste en mejorar el acceso de la población a los mercados laborales, y uno de los indicadores de resultados es la tasa de empleo. Como se ha visto, el solo hecho de observar el cambio antes y después en las tasas de empleo en los distritos que se inscriben en el programa no capturará el impacto causal del programa porque es probable que muchos otros factores influyan en el empleo a lo largo del tiempo. A la vez, comparar los distritos que se *inscribieron* y los que *no se inscribieron* en el programa de reparación de carreteras será problemático si existen motivos no observables por los que algunos distritos se inscribieron en el programa y otros no lo hicieron (el problema de sesgo de selección en el escenario de inscritos versus no inscritos) **Gertler et al. (2017)**.

El estimador **DD**, se basa en una comparación de *participantes* y *no participantes* antes y después de la intervención. Por ejemplo, después de una encuesta inicial de *línea de base* de referencia de ambos, *no participante* y (posteriores) *participantes*, se puede realizar una encuesta de seguimiento de ambos grupos, después de la intervención. A partir de esta información **(Khanker et al., 2017)**:

- Se calcula la diferencia entre los resultados medios observados para los grupos de *tratamiento* y *control* antes y después de la intervención del programa.
- Cuando los datos de *línea de base* de referencia están disponibles, es posible estimar los impactos asumiendo que la *heterogeneidad* no observada es invariante en el tiempo y uncorrelacionada con el tratamiento sobre el tiempo.
- Esta suposición es más débil que la *exogeneidad condicional* (descrita en capítulos 2 y 3) y representa los cambios de resultado para un grupo comparable de no participantes, es decir, $E(Y_1^C - Y_0^C \mid T_1 = 0)$ como el *contrafactual apropiado*, es decir, igual a $E(Y_1^C - Y_0^C \mid T_1 = 1)$.

- La estimación **DD**, también se puede calcular dentro de un *marco de regresión*; la regresión se puede ponderar para tener en cuenta los posibles sesgos en **DD**. En particular, la ecuación de estimación se especificaría de la siguiente manera (**Khandker et al., 2017**):

$$Y_{it} = \alpha + \beta T_{it} + p T_{it} + \gamma t + \varepsilon_{it} \dots\dots\dots(5.2)$$

- En la **ecuación 5.2**, el coeficiente β en la interacción entre la variable de tratamiento del postprograma (T_{it}) y el tiempo ($t = 1 \dots T$) dan el efecto promedio **DD** del programa. Por lo tanto, usando la notación de la **ecuación 5.1**, $\beta = \text{DD}$. Además de este término de interacción, las variables T_{it} y t se incluyen por separado, para recoger cualquier efecto medio separado del tiempo, así como el efecto de ser el objetivo frente a no serlo.
- De nuevo, mientras los datos de cuatro grupos diferentes estén disponibles para comparar, los datos del panel no son necesarios para implementar el enfoque **DD** (por ejemplo, el subíndice t , normalmente asociado con el tiempo, se puede reinterpretar como un área geográfica particular, $k = 1 \dots K$).
- Para comprender mejor la intuición detrás de la **ecuación 5.2**, se puede escribir en detalle, en forma de expectativas (suprimiendo el subíndice i por el momento, **Khandker, et al., 2017**):

$$E(Y_1^T - Y_0^T / T_1 = 1) = (\alpha + \text{DD} + p + \gamma) - (\alpha + p) \dots\dots\dots(5.3a)$$

$$E(Y_1^C - Y_0^C / T_1 = 0) = (\alpha + \gamma) - \alpha \dots\dots\dots(5.3b)$$

- Siguiendo la **ecuación 5.1**, restando **5.3b** de **5.3a** resulta **DD**. Tenga en cuenta nuevamente que **DD** es *no sesgada*, solo si la fuente potencial de *sesgo de selección* es *aditiva* y en *tiempo invariante*.
- Usando el mismo enfoque, si un simple impacto pre vs un postestimación, en la muestra participante se calcula (*diseño reflexivo*), el

calculo del impacto del programa sería $DD + \gamma$, y el sesgo correspondiente sería γ .

Tenga en cuenta que cuando las medias contrafácticas son invariantes en el tiempo ($E[Y_1^c - Y_0^c | T_1 = 1] = 0$) la estimación **DD** en la ecuación 5.1 se convierte en una *comparación reflexiva*, donde solo los resultados por las unidades de tratamiento son monitoreadas. El capítulo 2, también analiza las *comparaciones reflexivas* con más detalle. Este enfoque, sin embargo, es limitado en la práctica porque es poco probable que los resultados medios para el contrafactual no cambien.

- Como se discutió en capítulo 2, sin un grupo de control, es difícil justificar que otros factores no fueran responsables para afectar los resultados de los participantes. También se podría intentar comparar solo la diferencia del posprograma en los resultados entre las unidades de tratamiento y control; sin embargo, en este caso, el impacto estimado de la política sería $DD + \rho$, y el sesgo sería ρ . Las diferencias sistemáticas, no medidas que podrían correlacionarse con el tratamiento, no se pueden separar fácilmente.
- Recuerde que para que el estimador **DD** anterior se interprete correctamente, lo debe contener (**Khandker, et al., 2017**):
 - a. El modelo en la ecuación (*resultado*) esté correctamente especificado. Por ejemplo, que la estructura impuesta sea correcta.
 - b. El término de error esté uncorrelacionado con las otras variables, en la ecuación:

$$Cov(\epsilon_{it}, T_{it}) = 0$$

$$Cov(\epsilon_{it}, t) = 0$$

$$Cov(\epsilon_{it}, T_{it}t) = 0.$$

- c. Los últimos de estos supuestos se conocen como supuestos de tendencia paralela (*parallel-trend assumption*) que son los más

críticos. Significa que las características no observadas que afectan la participación de un programa, no varían en el tiempo en estatus de tratamiento.

Usos del modelo DD

El modelo de **DD** es el que se utiliza siendo de lo más popular y por lo general, es para los siguientes casos (**Bernal y Peña, 2011**):

1. Para ganar eficiencia en el estimador del efecto del programa. Es decir, la eficiencia es un concepto asociado a la varianza del estimador. Mayor eficiencia significa menor varianza o dispersión del estimador con respecto a un estimador de la misma clase (por ejemplo, sesgado o lineal), lo que indica que el efecto del programa es estimado con mayor precisión (exactitud). Evidentemente, es una propiedad deseada que el estimador del programa sea lo más preciso posible. Si el tratamiento fue aplicado aleatoriamente, entonces el estimador de **DD** puede ser más eficiente que el *estimador del modelo de diferencias*. Este será el caso si algunos de los determinantes no observados de Y son persistentes en el tiempo para un determinado sujeto, por ejemplo, el sexo, el nivel de escolaridad previo a la aplicación del tratamiento, la habilidad innata, etc. Cuál de los dos estimadores (simple diferencia vs **DD**) es más eficiente depende, de qué fracción de la varianza de Y es explicada por estos factores no observados.
2. Para eliminar diferencias preexistentes entre el grupo de tratamiento y el grupo de control. Si, por alguna razón, el tratamiento está correlacionado con el nivel inicial de Y antes de la asignación del tratamiento (antes del experimento o el evento fortuito) pero $E(u_i, \mathbf{I D}_i) = \mathbf{0}$, entonces el estimador de diferencias no es insesgado pero el estimador de **DD** sí lo es, siempre y cuando se

cumpla el supuesto adicional de tendencias paralelas. Como se ha mencionado, en un experimento natural o cuasi experimento no es improbable que esto ocurra, dado que el evento fortuito genera una asignación que parece aleatoria pero que en la mayoría de los casos no corresponde a una aleatorización perfecta.

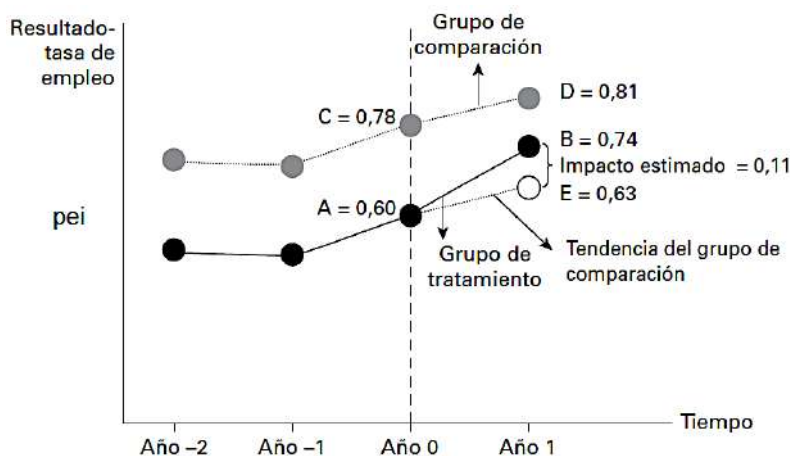
¿Qué sucede en el modelo DD?

Se plantea la cuestión: ¿qué pasaría si se combinan los dos métodos y se comparan los cambios antes-después en los resultados de un grupo que se inscribió en el programa con los cambios *antes-después* de un grupo que no se inscribió en el programa? La diferencia en los resultados *antes-después* para el grupo inscrito, *la primera diferencia*, controla por factores que son constantes a lo largo del tiempo en ese grupo, puesto que se está comparando el propio grupo consigo mismo. Sin embargo, todavía quedan los factores externos que varían con el tiempo (factores variables en el tiempo) en este grupo. Una manera de capturar esos factores que varían en el tiempo es medir el cambio antes-después en los resultados de un grupo que no se inscribió en el programa pero que estuvo expuesto al mismo conjunto de condiciones ambientales (la segunda diferencia). Si se limpia la primera diferencia de otros factores variables en el tiempo que influyen en el resultado de interés sustrayendo la segunda diferencia, se habrá eliminado una fuente de sesgo que resultaba preocupante en las comparaciones sencillas antes-después. El enfoque de **DD** hace lo que su nombre sugiere: combina las dos estimaciones falsas del contrafactual (comparaciones *antes-después* y comparaciones entre quienes se *inscriben* y *quienes deciden no hacerlo*) para producir una mejor estimación del contrafactual. Por ejemplo, de un programa de estímulos a la innovación, el método **DD** podría comparar los cambios en el empleo antes y después de que se ponga en marcha el programa, en los individuos

que viven en distritos que lo introdujeron con los cambios en el empleo en los distritos donde no se implementó el programa (Gertler et al., 2017).

Es importante señalar que el *contrafactual* que se estima en este caso es el cambio en los resultados del grupo de tratamiento. La estimación de este *contrafactual* es el cambio en los resultados del grupo de comparación. Los grupos de *tratamiento* y *comparación* no tienen necesariamente que tener las mismas condiciones antes de la intervención. Sin embargo, para que el método de **DD** sea válido, el grupo de comparación debe mostrar con precisión el cambio en los resultados que habría experimentado el grupo de tratamiento en ausencia de tratamiento. Para aplicar el método **DD**, hay que medir los resultados en el grupo que se beneficia del programa (el grupo de tratamiento) con los resultados del grupo que no se beneficia (el grupo de comparación), tanto antes como después del programa. La **Gráfica 5.1**, se presenta un ejemplo de utilización del método **DD** para entender el impacto de los incentivos de diseño de nuevos productos y servicios, en la implementación hipotética del **pei** en un país y las tasas de deserción de las **pbt**.

Gráfica 5.1. Ejemplo método DD



Fuente: Gertler (2017) con adaptación propia.

El año **0**, es el año de *línea de base*. En el año **1** se inscribe en el **pei** un grupo de distritos de tratamiento, mientras que no lo hace un grupo de distritos de comparación. El nivel de los resultados (la tasa de empleo) en el grupo de tratamiento va de **A**, antes del comienzo del programa, a **B**, después del comienzo del programa, mientras que el resultado para el grupo de comparación va de **C**, antes del comienzo del programa, a **D**, después del comienzo del programa.

Recuerde los dos falsos contrafactuales: *la diferencia de los resultados antes y después de la intervención en el grupo de tratamiento (B-A)* y *la diferencia de los resultados después de la intervención entre los grupos de tratamiento y de comparación (B-D)*. Con **DD**, la estimación del contrafactual se obtiene calculando el cambio en los resultados del grupo de comparación (**D-C**) y luego sustrayéndolo del cambio en los resultados del grupo de tratamiento (**B-A**). Utilizar el cambio en los resultados del grupo de comparación como la *estimación del contrafactual* para el cambio en los resultados del grupo de tratamiento es similar a suponer que si el grupo que se inscribió no hubiera participado en el programa, sus resultados habrían evolucionado a lo largo del tiempo siguiendo la misma tendencia que el grupo que no se inscribió, es decir, la evolución en el resultado del grupo inscrito habría ido de **A** a **E**, como se muestra en la **Gráfica 5.1**.

En resumen, el impacto del programa se computa simplemente como la diferencia entre dos diferencias:

$$\text{Impacto de la DD} = (B - A) - (D - C) = (0.74 - 0.60) - (0.81 - 0.78) = 0.11.$$

Las relaciones que se muestran en la **Gráfica 5.1** también pueden presentarse de forma tabulada como lo muestra la **Tabla 5.2** la cual, describe los componentes de las estimaciones del método **DD**.

Tabla 5.1. Cálculo del método DD de forma tabular

Item	Después	Antes	Diferencia
Tratamiento/ inscritos	B	A	B-A
Comparación/ no inscritos	D	C	D-C
Diferencia	B-D	A-C	DD=(B-D)-(D-C)
	Después	Antes	Diferencia
Tratamiento/ inscritos	0.74	0.60	0.14
Comparación/ no inscritos	0.81	0.78	0.03
Diferencia	-0.07	-0.18	DD=0.14-0.03=0.11

Fuente: Gertler (2017) con adaptación propia.

La primera línea contiene los resultados del *grupo de tratamiento* antes de la intervención (**A**) y después de la intervención (**B**). La comparación *antes-después* del grupo de tratamiento es la primera diferencia (**B-A**). La segunda línea contiene los resultados del grupo de comparación antes de la intervención (**C**) y después de la intervención (**D**), de modo que la segunda diferencia es (**D-C**).

El método **DD** computa la estimación del impacto de la siguiente manera:

- Se calcula la diferencia del resultado (**Y**) entre las situaciones antes y después para el grupo de tratamiento (**B - A**).
- Se calcula la diferencia del resultado (**Y**) entre las situaciones antes y después para el grupo de comparación (**D - C**).
- A continuación, se calcula la diferencia entre la diferencia en los resultados del grupo de tratamiento (**B - A**) y la diferencia del grupo de comparación (**D - C**), o **DD = (B - A) - (D - C)**. Estas diferencias en diferencias constituyen la estimación del impacto.

También se consideran las diferencias en diferencias en la dirección contraria: se calcula primero la diferencia en el resultado entre el grupo de tratamiento y el de comparación en la situación después; luego se calcula la diferencia en el resultado entre el grupo de tratamiento y de comparación en la situación antes, y finalmente se sustrae este último del primero.

$$\text{Impacto DD} = (B - D) - (A - C) = (0.74 - 0.81) - (0.60 - 0.78) = 0.11.$$

Modelo de panel de efectos fijos (*panel fixed-effects*)

El modelo anterior de *dos periodos*, puede generalizarse con *múltiples periodos* de tiempo, que puede llamarse el *modelo de panel de efectos fijos*. Esta posibilidad, es particularmente importante para un modelo que controla no solo la heterogeneidad invariante no observada en el tiempo, sino también para la *heterogeneidad* en las características observadas durante el transcurso de periodos múltiples. Más específicamente, Y_{it} que pueda hacerse regresión en T_{it} , un rango de covariables X_{it} que varían en el tiempo y la heterogeneidad individual no observada invariante en el tiempo η_i que puede correlacionarse tanto con el *tratamiento* y otras características no observadas ε_{it} . Considere la siguiente revisión de la **ecuación 5.2 (Khandker, et al., 2017)**:

$$Y_{it} = \phi T_{it} + \delta X_{it} + \eta_i + \varepsilon_{it} \dots \dots \dots (5.4)$$

Haciendo la diferencia, tanto del lado derecho como el izquierdo de la **ecuación 5.4** sobre tiempo, se obtiene:

$$(Y_{it} - Y_{it-1}) = \phi (T_{it} - T_{it-1}) + \delta (X_{it} - X_{it-1}) + (\eta_i - \eta_i) + (\varepsilon_{it} - \varepsilon_{it-1}) \dots \dots (5.5a)$$

$$\Rightarrow \Delta Y_{it} = \phi \Delta T_{it} + \delta \Delta X_{it} + \Delta \varepsilon_{it} \dots \dots \dots (5.5b)$$

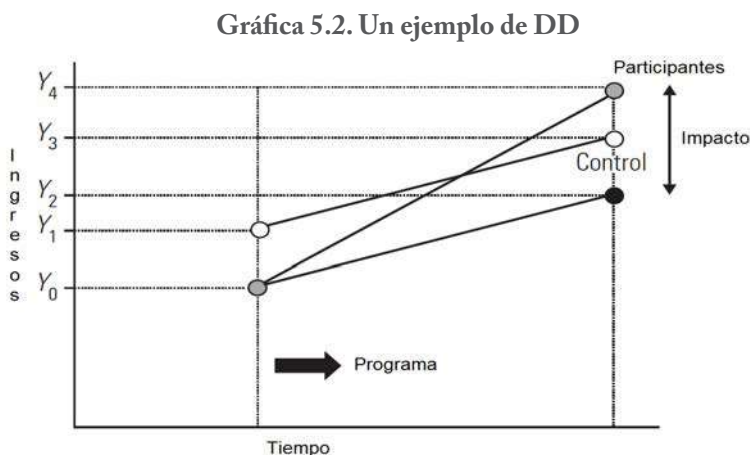
En este caso, dado que la fuente de *endogeneidad* (es decir, las características del sujeto no observado, η_i) se elimina en la diferenciación, los mínimos cuadrados ordinarios (**OLS. Ordinary Least Squares**) pueden ser aplicados a la **ecuación 5.5b** para estimar el efecto imparcial del programa (\emptyset). Con dos periodos de tiempo, \emptyset es equivalente a la estimación **DD** en la **ecuación 5.2**, controlando las mismas covariables X_{it} ; los errores estándar, sin embargo, pueden necesitar corregirse para *correlación serial* (**Bertrand, et al., 2004**). Con más de dos periodos de tiempo, la estimación del impacto del programa diferirá de **DD**.

Implementado DD

Para aplicar un enfoque **DD** utilizando *datos de panel*, los datos de la *línea de base*, deben recopilarse en el programa y las áreas de control, antes de la implementación del programa. Como se describió en el capítulo 2, la información tanto cuantitativa así como cualitativa, sobre estas áreas será útil para determinar quién es probable que participe. Las encuestas de seguimiento, después de la intervención del programa, también deberían realizarse en las mismas unidades. Aunque algunos estudios a gran escala, no son capaces de volver a visitar los mismos sujetos (como las **pbt**) después de un programa de intervención, sí pueden examinar las mismas comunidades, regiones, clústers y, por lo tanto, pueden calcular los impactos del programa **DD** a nivel local, comunitario, regional o clúster. Las encuestas concurrentes de beneficiarios a niveles locales y/o comunitarios, son importantes para mantener esta flexibilidad, ya que las encuestas antes y después de una intervención de programa, pueden abarcar varios años, lo que hace que la recopilación de datos de panel sea más difícil.

Calcular la diferencia promedio en los resultados por separado, para *participantes* y *no participantes*, durante los periodos y luego tomar una

diferencia adicional, entre los cambios promedio en los resultados para estos dos grupos, será calcular el impacto **DD**. Un ejemplo se muestra en la **Gráfica 5.2**, donde $DD = (Y_4 - Y_0) - (Y_3 - Y_1)$ (Khandker, et al., 2017).



Fuente: Khandker et al. (2017) con adaptación propia.

La línea más baja en la **Gráfica 5.2**, también muestra los verdaderos resultados contrafactuales, que nunca se observan (ver capítulo 2). Bajo el enfoque **DD**, las características *no observadas*, que crean una brecha entre los resultados de control medidos y el *verdadero contrafactual*, suponen que los resultados son invariantes en el tiempo, de modo que la brecha entre las dos tendencias es la misma en el período. Este supuesto, implica que: $(Y_3 - Y_2) - (Y_1 - Y_0)$. Utilizando esta igualdad, en la ecuación **DD** precedente, se consigue: $DD = (Y_4 - Y_2)$. Aunque **DD**, normalmente explota una línea de base y datos de panel resultantes, la sección transversal repetida de los datos a lo largo del tiempo también se pueden usar.

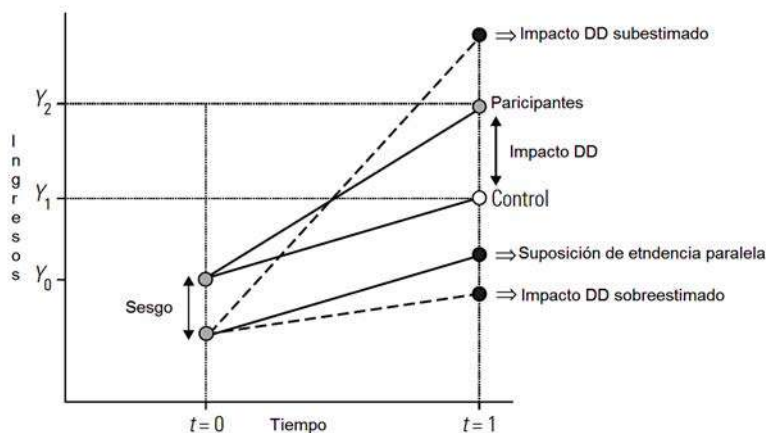
Ventajas y desventajas del uso de la DD

Una ventaja del uso de la técnica **DD**, es que relaja la suposición de *exogeneidad condicional* o de *selección*, solo en las características observadas. También, proporciona una forma manejable e intuitiva, de cuenta para la selección de características *no observadas*. El principal inconveniente, sin embargo, descansa precisamente en esta suposición: la noción de *sesgo de selección* invariante en el tiempo parece *inverosímil* para muchos programas específicos en países en desarrollo. Un argumento similar, contra el método **DD** se aplica en el caso de evaluar un programa usando datos repetidos de encuestas transversales. En la mayoría de los programas a menudo existen enfoque amplios de aplicación como la implementación de innovaciones, por ejemplo, de procesos, para un rango amplio de sectores. Dado que tales programas inician con crecimientos muy bajos, se tiende a esperar que respondan dinámicamente a través de varios periodos o años tanto en las formas *observadas* como *no observadas*; un caso de esto, son los programas de capacitación o entrenamiento. De hecho antes de entrar en vigor cualquier programa a pesar de estar bien planeada su implementación ocurrirá una baja notable a lo esperado en su introducción; este fenómeno se conoce como inmersión de Ashenfelter (*Ashenfelter's Dip*). Por lo tanto, el grupo *tratado* podría haber experimentado un más rápido crecimiento por implementación de una innovación de procesos, incluso, sin participación. En este caso, es probable que un método **DD** tienda a *sobrestimar el efecto de un programa*. Es decir, si los sujetos se *auto-seleccionan* en un programa de acuerdo a alguna regla *desconocida* y datos de sección transversal repetidos, la suposición de heterogeneidad *invariante en el tiempo* puede fallar, si la composición del grupo cambia y la intervención afecta la composición de grupos *tratados* vs. *no tratados*.

La **Gráfica 5.3** refleja este sesgo potencial, cuando la diferencia entre los resultados *no participantes* y *contrafactuales* cambia con el tiempo; el tiempo variante, la *heterogeneidad no observada* podría conducir a

un sesgo hacia arriba o hacia abajo. En la práctica, *ex ante*, la *heterogeneidad no observada* que varía con el tiempo, podría explicarse con un diseño adecuado del programa, lo que incluye garantizar que las áreas de proyecto y control compartan características similares del preprograma.

Gráfica 5.3. Heterogeneidad de tiempo variante no observado



Fuente: Khandker et al. (2017) con adaptación propia.

Si las áreas de comparación, no son similares a los posibles participantes en términos de características *observadas* y *no observadas*, entonces, los cambios en el resultado con el tiempo, pueden ser una función de esta diferencia. Este factor también sesgaría el **DD**. Por ejemplo, en el contexto de un programa de inscripción a una capacitación de innovación, si las *áreas de control* fueran seleccionados para que inicialmente, estuvieran mucho más allá de los centros de capacitación locales que las áreas seleccionadas, **DD sobreestimaría** el impacto del programa en las localidades participantes.

Como se discutió en el capítulo 4, la aplicación de **PSM** podría ayudar a *unir* las unidades de tratamiento con unidades de control *observacionalmente similares*, antes de estimar el impacto **DD**. Específica-

mente, se ejecutaría **PSM** en el año base y luego realizaría un **DD** en las unidades que permanecen en en *soporte común*. Los estudios demuestran que ponderando las observaciones de control, de acuerdo con su *propensión de puntaje*, produce un estimador completamente eficiente (**Hirano et al., 2003**, ver también el capítulo 4 para una discusión). Dado a que el **PSM** efectivo, depende de una rica *línea de base*, sin embargo, durante la recolección inicial de datos, se debe prestar especial atención a características que determinan la participación. Incluso si se pudiera asegurar la *comparabilidad* de las áreas de control y proyecto antes del programa, no obstante, el enfoque **DD** podría fallar si los cambios macroeconómicos durante el programa afectan a los dos grupos de manera diferente.

Suponga que algunas características desconocidas, hacen que los grupos *tratados* y *no tratados* reaccionen de manera diferente a un choque de innovaciones de tecnología disruptivas común. En este caso, un **DD** simple puede *sobrestimar* o *subestimar* los verdaderos efectos de un programa dependiendo, de cómo los grupos *tratados* y *no tratados* reaccionan al choque común. **Bell et al. (1999)** sugieren un **DD** diferencial ajustado a la tendencia temporal, para tal caso. Esta alternativa, se discutirá más adelante en términos del método de *triple diferencia* (*triple-difference method*). Otro enfoque podría ser a través de *variables instrumentales* (*instrumental variables*), que se discuten en Capítulo 6. Si hay suficientes datos disponibles, sobre otros *factores exógenos o independientes* del comportamiento que afecten a los *participantes* y *no participantes* con el tiempo, esos factores pueden ser explotados para identificar impactos, cuando no es constante la *heterogeneidad no observada*. Una variable instrumental (*instrumental variable*) con un enfoque de *panel de efectos fijos* (*panel fixed-effects*), se podría realizar. El capítulo 6 proporciona más detalles.

Utilidad del método DD

Para entender la utilidad de este método, debe tomarse nuestro segundo *contrafactual falso*, visto en capítulos anteriores, que comparaba las unidades inscritas con las no inscritas en un programa. Recuerde que la principal preocupación en este caso era que las dos series de unidades pudieran tener características diferentes y que pueden ser dichas características, y no el programa, las que explican la diferencia en los resultados entre los dos grupos.

Las diferencias no observadas en las características eran especialmente preocupantes: por definición, es imposible incluir las características no observables en el análisis.

El método de **DD** contribuye a resolver este problema en la medida en que se puede razonablemente suponer que muchas características de las unidades o personas son constantes a lo largo del tiempo (o invariables en el tiempo). Suponga, por ejemplo, *características observables*, como el año de nacimiento del CEO, su educación, el tamaño de equipamientos de alta tecnología; el consumo energético de dichos equipos. Es probable que la mayoría de estos tipos de variables, aunque posiblemente relacionadas con los resultados, no cambien en el transcurso de una evaluación. Con el mismo razonamiento, podría llegarse a la conclusión de que muchas características *no observables* de los individuos también son más o menos constantes a lo largo del tiempo. Suponga, por ejemplo, en rasgos de liderazgo o el historial de salud crediticia de la **pbt**. Es posible que estas características intrínsecas de las **pbt** no cambien con el tiempo.

En lugar de contrastar los resultados entre los *grupos de tratamiento* y comparación después de la intervención, los métodos de **DD** estudian las tendencias entre los grupos de *tratamiento y comparación*. La tendencia de un sujeto es la diferencia en los resultados para ese sujeto antes y después del programa. Al sustraer la situación de los resultados antes de la situación después, se anula el efecto de todas las características que son únicas de ese individuo y que no cambian a lo largo del

tiempo. En realidad, se está anulando (o controlando) no solo el efecto de características observables invariables en el tiempo, sino también el efecto de características no observables invariables en el tiempo, como las ya mencionadas (**Gertler et al., 2017**).

Tendencias iguales en el método DD

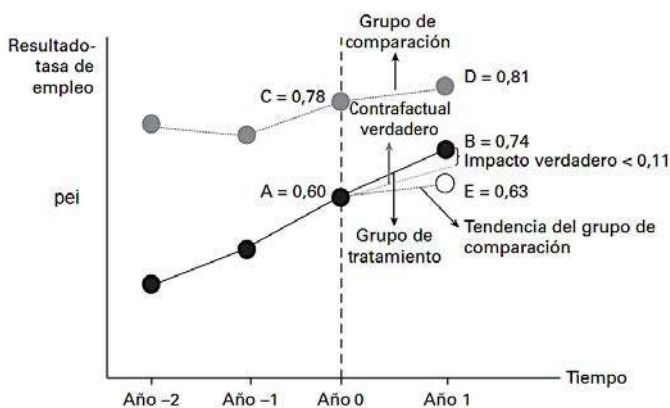
A pesar de que **DD** permite tener en cuenta las diferencias entre los grupos de tratamiento y comparación que son constantes a lo largo del tiempo, no ayudan a eliminar las diferencias entre los grupos de tratamiento y de comparación que cambian con el tiempo. En el ejemplo del programa de reparación de carreteras, si las zonas de tratamiento también se benefician de la construcción de un nuevo puerto marítimo al mismo tiempo que se produce la reparación de las carreteras, el efecto de esta última no podrá separarse de la construcción del puerto marítimo utilizando un enfoque de diferencias en diferencias. Para que el método proporcione una estimación válida del contrafactual, se debe suponer que no existen ese tipo de diferencias que varían en el tiempo entre los grupos de tratamiento y comparación.

Otra manera de pensar en esto es que, en ausencia del programa, las diferencias en los resultados entre los grupos de tratamiento y comparación tendrían que evolucionar de forma paralela. Es decir, sin el tratamiento, los resultados tendrían que aumentar o disminuir en la misma medida en ambos grupos; los resultados tienen que *mostrar tendencias iguales en ausencia de tratamiento*.

Desde luego, no hay manera de demostrar que las diferencias entre los grupos de tratamiento y comparación habrían evolucionado de manera paralela en ausencia del programa. El motivo es que no se puede observar qué habría ocurrido con el grupo de tratamiento en ausencia del tratamiento, es decir, no se puede observar el *contrafactual*.

Por lo tanto, cuando se emplea el método de **DD**, se debe suponer que, en ausencia del programa, los resultados en el grupo de tratamiento habrían evolucionado de forma paralela con los resultados del grupo de comparación. La **Gráfica 5.4**, ilustra un incumplimiento de este supuesto fundamental.

Gráfica 5.4. Método DD cuando las tendencias de los resultados son diferentes



Fuente: Gertler et al. (2017) con adaptación propia.

Si las tendencias de los resultados, son diferentes para los grupos de tratamiento y de comparación, el efecto estimado de tratamiento obtenido mediante métodos de diferencias en diferencias sería inválido o estaría sesgado. Esto se debe a que la tendencia del grupo de comparación no es una estimación válida de la tendencia contrafactual que habría prevalecido en el grupo de tratamiento en ausencia del programa. Como se muestra en la **Gráfica 5.4**, si en realidad los resultados del grupo de comparación aumentan más lentamente que los resultados del grupo de tratamiento en ausencia del programa, utilizar la tendencia del grupo de comparación como estimación del contrafactual de la tendencia del grupo de trata-

miento conduce a una estimación sesgada del impacto del programa. Más concretamente, se estaría sobreestimando el impacto del programa.

Comprobación del supuesto de igualdad de tendencias en el método DD

A pesar de que no se puede demostrar, la validez del supuesto fundamental de igualdad de tendencias, se puede evaluar teniéndose cuatro posibles formas de comprobar el supuesto de las tendencias iguales (**Gertler et al., 2017**):

- Una primera verificación de validez consiste en contrastar los cambios en los resultados en los grupos de *tratamiento* y *comparación* en repetidas ocasiones antes de la implementación del programa. En el programa de estímulos a la innovación (**pei**), esto significa que se mediría el cambio en la tasa de empleo entre los grupos de tratamiento y comparación antes del comienzo del programa, es decir, entre el segundo y el primer año, y entre el primer año y el año cero. Si se ve que los resultados evolucionaban de forma paralela antes del comienzo del programa, es razonable suponer que habrían seguido evolucionando de la misma manera después de la intervención. Para verificar la igualdad de las tendencias antes de la intervención, se requieren al menos, dos rondas de observaciones, en los grupos de tratamiento y comparación, antes del comienzo del programa. Esto significa que la evaluación requeriría *tres rondas de observaciones: dos observaciones previas a la intervención para valorar las tendencias anteriores al programa, y al menos una observación posterior a la intervención para evaluar el impacto mediante el método de DD*.
- Una segunda manera de comprobar el supuesto de las tendencias iguales sería llevar a cabo lo que se denomina *prueba de placebo*. Para esta prueba, se realiza una segunda estimación de **DD** utili-

zando un grupo de tratamiento *falso*, es decir, un grupo que, según lo que el evaluador sabe, no ha sido afectado por el programa. Por ejemplo, se quiere estimar cómo el soporte técnico para las **pbt** del séptimo ciclo, influyen en su probabilidad de solicitarlo, y entonces se eligen **pbt** del octavo ciclo como grupo de comparación. Para comprobar si los **pbt** de séptimo y octavo ciclo, tienen las mismas tendencias en términos requerimiento del soporte técnico, podría verificarse que los **pbt** de octavo y de sexto ciclo tengan las mismas tendencias. El evaluador sabe que los **pbt** de sexto ciclo no se verán afectados por el programa, de modo que si realiza una estimación de **DD** utilizando a los **pbt** de octavo ciclo como *grupo de comparación* y a los de sexto como el grupo de *tratamiento falso*, tiene que obtener un *impacto de cero*. De lo contrario, el impacto que encuentre, se deberá necesariamente a alguna diferencia fundamental en las tendencias entre los **pbt** de sexto y octavo ciclo. Esto, a su vez, *arroja dudas sobre si es válido el supuesto de que los alumnos de séptimo y octavo grado tienen tendencias iguales en ausencia del programa*.

- Una tercera manera de probar el supuesto de tendencias iguales, sería llevar a cabo la *prueba de placebo* no solo con un *grupo de tratamiento falso*, sino también con un *resultado falso*. En el ejemplo de las clases de apoyo, conviene comprobar la validez de utilizar a los **pbt** de octavo ciclo como grupo de *comparación* estimando el impacto de las clases de apoyo en un resultado que, según lo que se sabe, no se ve afectado por dichas clases, como, por ejemplo, el número de empleados que los **pbt** tienen. Si la estimación de **DD** encuentra un impacto de las clases de apoyo en el número de empleados de las **pbt**, entonces ya se sabe que el grupo de comparación debe tener alguna falla.
- Una cuarta manera de comprobar el supuesto de las tendencias iguales consistiría en aplicar el método de **DD** utilizando diferentes grupos de comparación. En el ejemplo de las clases de

apoyo, primero se llevaría a cabo la estimación con las **pbt** de octavo ciclo como grupo de *comparación*, y luego se realizaría una segunda estimación tomando a los alumnos de sexto ciclo como grupo de *comparación*. Si los dos grupos son válidos, se observará que el impacto estimado es aproximadamente el mismo en ambos cálculos.

Modelos DD alternativos

El enfoque de *doble diferencia* (**DD**) descrito, produce estimaciones consistentes de los impactos del proyecto si el grupo *no observado* y la *heterogeneidad individual* son invariantes en el tiempo. Sin embargo, se conciben varios casos donde las características *no observadas* de una población, pueden de hecho cambiar con el tiempo, derivado, por ejemplo, de cambios en preferencias o normas sobre una serie de tiempo más larga. Algunas variantes del método **DD**, tpor lo tanto, se ha propuesto controlar los factores que afectan estos cambios en los casos no observables (**Khan-dker et al., 2017**).

Un caso en el que la *heterogeneidad no observada*, puede no permanecer constante en el tiempo, es donde las inversiones públicas dependen de las condiciones iniciales (preprograma) del área local. No controlar las condiciones iniciales del área, al evaluar el impacto de, por ejemplo, una política de estímulos a la innovación el programa puede conducir a un significativo *sesgo de variable omitida*, si las condiciones locales, también fueran responsables de la mejora de los resultados de la **pbt** o la focalización del programa fue correlacionada con tales características del área.

Combinando PSM con DD

Como se mencionó anteriormente, *siempre que existan datos completos sobre las áreas de control y tratamiento, el PSM se puede combinar con métodos DD* para hacer una mejor coincidencia de las unidades de control y proyecto sobre las características del preprograma. Específicamente, recordando la discusión en el capítulo 4, se observa que la *propensión del puntaje* puede usarse para hacer coincidentes o emparejables a las unidades participante y de control en el año base (*preprograma*), y el impacto del tratamiento, se calcula a través de las unidades del participante y de control coincidentes o emparejadas, dentro de la zona de soporte común. Para dos períodos $t = \{1, 2\}$, la estimación **DD** para cada área de tratamiento i se calculará como (Khandker et al., 2017):

$$DD_i = (Y_{i2}^T - Y_{i1}^T) - \sum_{j \in C} \omega(i, j) (Y_{j2}^C - Y_{j1}^C)$$

Donde :

$\omega(i, j)$ es el peso (usando el enfoque **PSM**) dado al área de control j -ésima coincidente o emparejado al *área de tratamiento* i . Son aplicables los diferentes enfoques de coincidencia o emparejamiento discutidos en el capítulo 4.

En términos de un marco de regresión (también discutido en el capítulo 4), Hirano, et al. (2003) muestran que una *regresión ponderada* de mínimos cuadrados, al *ponderar* el control de observaciones, de acuerdo con su propensión de puntaje, produce un estimador completamente eficiente (Khandker et al., 2017):

$$\Delta Y_{it} = \alpha + \beta T_i + \gamma \Delta + \epsilon_{it}, \beta = DD \dots \dots \dots (5.6)$$

Los pesos en la regresión de la **ecuación 5.6**, son igual 1 para las unidades tratadas y $P'(X) / (1 - P'(X))$ para las unidades de comparación.

Método de la triple diferencia (*triple-difference method*)

¿Qué sucede si los datos de referencia, no están disponibles? Tal podría ser el caso durante una crisis económica, por ejemplo, donde un programa o red de seguridad tiene que configurarse rápidamente. En este contexto, Se puede utilizar un método de triple diferencia. Además de un *primer experimento* que compara a los grupos de proyecto y de control, este método explota el uso de un experimento completamente separado, después de un programa de intervención (**Khandker et al., 2017**). Es decir, este grupo de control separado, refleja un conjunto de *no participantes* en áreas *tratadas* y *no tratadas* que no forman parte del primer grupo de control. Estas nuevas unidades de control, pueden ser diferentes del primer grupo de control en diversas características por ejemplo, tecnológicas, o socioeconómicas, si los evaluadores desean examinar el impacto del proyecto en *participantes* en relación con otro grupo socioeconómico. Otra diferencia, del primer experimento se tomaría del cambio en la muestra adicional de control para examinar el impacto del proyecto, teniendo en cuenta otros factores que cambian con el tiempo (**Gruber, 1994**). Por lo tanto, este método requeriría datos de varios años después de la intervención del programa, aún y falten datos de referencia.

Ajuste de las tendencias de diferencial de tiempo

Como se mencionó anteriormente, supongamos que se requiere evaluar un programa como el de capacitación en diseño e innovación de productos y servicios, introducida durante una crisis macroeconómica. Con datos disponibles para los grupos de *tratamiento* y *no tratados* antes y después del programa, se podría usar un enfoque **DD** para estimar el efecto del programa en las ganancias, por ejemplo. Sin embargo, tales eventos son probables que creen condiciones donde los *grupos tratados* y *no tratados*

respondan diferente ante la crisis. **Bell et al. (1999)** han construido un método **DD** que explica estos efectos diferenciales de tendencia temporal. Aparte de los datos en grupos tratados y no tratados antes y después del tratamiento, otro intervalo de tiempo es necesario ($t - 1$ a t) para examinar los mismos grupos tratados y no tratados. El ciclo pasado reciente, es probablemente el intervalo de tiempo más apropiado para tal comparación. Más formalmente, el **DD** ajustado de tendencia temporal se define como (**Khandker et al., 2017**):

$$DD = [E (Y_1^T - Y_0^T \mid T_1 = 1) - E (Y_1^C - Y_0^C \mid T_1 = 0)] - [E (Y_t^T - Y_{t-1}^T \mid T_1 = 1) - E (Y_t^C - Y_{t-1}^C \mid T_1 = 0)] \dots\dots\dots (5.7)$$

Limitaciones del método de DD

Aun cuando las tendencias sean iguales antes del comienzo de la intervención, el sesgo en la estimación de diferencias en diferencias puede producirse y pasar inadvertido. Esto se debe a que el método **DD** atribuye a la intervención cualquier diferencia de las tendencias entre los grupos de tratamiento y de comparación que se producen desde el momento en que la intervención comienza. Si hay otros factores presentes que influyen en la diferencia en las tendencias entre los dos grupos, y la regresión multivariante no rinde cuenta de ellos, la estimación será inválida o sesgada (**Gertler et al., 2017**).

Suponga que se intenta estimar el impacto del uso de una innovación tecnológica agrícola (un nuevo sistema de riego con monitoreo electrónico) con la subvención de sus componentes (software y hardware) y que esto se lleva a cabo midiendo la producción de un cultivo en particular de los agricultores subvencionados (*tratamiento*) y de los agricultores no subvencionados (*comparación*) antes y después de la distribución de las subvenciones. Si en el año 1, tiene lugar una *sequía* que afecta solamente

a los agricultores subvencionados, la estimación de **DD** producirá una estimación inválida del impacto de subvencionar la innovación tecnológica agrícola. En general, cualquier factor que afecte a uno de los dos grupos de forma *desproporcionada*, y lo hace al mismo tiempo en que el grupo de tratamiento recibe el tratamiento, sin que esto se tome en cuenta en la regresión, puede *potencialmente invalidar o sesgar la estimación del impacto del programa*. El método **DD** supone que no hay factores de este tipo presentes.

Verificación DD

El método **DD** supone que las tendencias de los resultados son similares en los grupos de comparación y tratamiento, antes de la intervención y que los únicos factores que explican las diferencias en los resultados entre ambos grupos, aparte del propio programa, *son constantes a lo largo del tiempo*. Por lo tanto (**Gertler et al., 2017**):

1. Cabe preguntarse por lo tanto, los resultados ¿habrían evolucionado de forma paralela en los grupos de tratamiento y comparación en ausencia del programa? Esto se puede evaluar utilizando diversas *pruebas de falsificación*, como las siguientes:
 - Los resultados en los grupos de tratamiento y comparación ¿evolucionaban de modo paralelo antes de la intervención? Si hay dos rondas de datos disponibles antes del comienzo del programa, se debe probar si existen diferencias en las tendencias que aparecen entre ambos grupos.
 - ¿Qué sucede con los resultados falsos que no deberían verse afectados por el programa? ¿Evolucionan de forma paralela antes y después del inicio de la intervención en los grupos de tratamiento y comparación?

2. Realizar el análisis de **DD** utilizando varios grupos plausibles de comparación. Deberían obtenerse estimaciones similares del impacto del programa.
3. Efectuar el análisis de diferencias en diferencias usando los grupos de tratamiento y comparación elegidos, y un resultado falso que no debería verse afectado por el programa. Debería encontrarse un impacto nulo del programa en ese resultado.
4. Llevar adelante el análisis de **DD** utilizando la variable de resultados elegida con dos grupos que, según lo que se sabe, no se vieron afectados por el programa. Debería observarse un impacto cero del programa.

Ejemplo utilizando STATA

Los métodos de coincidencia o emparejamiento, discutidos en ejercicios anteriores, están destinados a reducir el sesgo por elección de los grupos de *tratamiento* y *comparación*, en función de las características observables. Por lo general, son implementados después de que el programa ha estado funcionando, durante algún tiempo y se han recopilado datos de encuestas. Otra forma poderosa de medir el impacto de un programa, es mediante el uso de *datos de panel*, recopilados de una encuesta de referencia antes que el programa sea implementado y después de que el programa ha estado funcionando por algún tiempo. Estas dos encuestas deben ser comparables en las preguntas y métodos de encuesta utilizados y deben ser administrados a *participantes* y *no participantes*. Usar los *datos del panel* permite la eliminación del sesgo variable no observado, siempre que no cambie con el tiempo. Los *datos de panel* no son estrictamente necesarios para la estimación de DD. Cómo puede ser esta técnica, aplicado a *datos transversales*, se muestra más adelante. El método de la **DD**, (también conocido comúnmente como *diferencia en diferencia*) ha sido popular en evaluaciones no experimentales. Como

hemos visto, el método **DD** estima la diferencia en el resultado durante el período posterior a la intervención entre un grupo de *tratamiento* y un grupo de *comparación*, en relación con los resultados observados durante una encuesta de referencia previa a la intervención.

Comparación simple (uso de test)

La forma más sencilla de calcular el estimador **DD**, es tomar la *diferencia manualmente* en los resultados entre las encuestas de los grupos de *tratamiento* y *control*. Los datos del panel, se ubican en el archivo **pei_0108.dta** que se utilizan para este propósito, por lo que los siguientes comandos abren el archivo de datos y crearán una nueva variable de resultado a nivel del periodo **01** (*gasto per cápita de la pbt*) para que esté disponible en observaciones de ambos años. Luego, solo se mantienen las observaciones del período **08**, y se crea un registro como gasto per cápita por lo que se crea, una diferencia de gasto per cápita **pbt** entre el periodo **08** y **01**. Por lo tanto, se proponen realizar los siguientes pasos para su resolución:

- Teclear los comandos (**ver Tabla 5.2**):

```
clear
use "/Users/DCA/Desktop/STATA log files/pei_0108.dta"
gen exptot0=exptot if year==0
egen exptot01=max(exptot0), by(id)
keep if year==1
gen lexptot01=ln(1+exptot01)
gen lexptot08=ln(1+exptot)
gen lexptot0801=lexptot08-lexptot01
```

Tabla 5.2. Comandos iniciales y estructura de archivo modificada para preparación prueba simple comparación

```

. clear

. use "/Users/DCA/Desktop/STATA log files/ps1_0100.dta"

. gen exptot0=exptot if year==0
(826 missing values generated)

. egen exptot01=max(exptot0), by(id)

. keep if year==1
(826 observations deleted)

. gen lexptot01=ln(1+exptot01)

. gen lexptot00=ln(1+exptot)

. gen lexptot0001=lexptot00-lexptot01

```

Command

Fuente: STATA con datos propios.

Variable	
name	Label
id	ID
year	year of observation: 0=01, 1=08
region	id region
cluster	id cluster
agecat	age of doc: years
gender	gender of doc: 1=M, 0=F
educat	education of doc: years
innovat	innv are based on personnel
innfac	innv assets based on facilities
innmtr	innv increase
expcost	total expenditure
expcost	innv innovation expenditure: T&E...
exptot	total expenditure
innv01	innovation scale matched01 per...
innv01	innovation female matched01 p...
weight	innovation sampling weight
innv0001	innv broadband accessible all s...
pcbr	proportion of regions with intern...
raw1	raw1
raw2	raw2
raw3	raw3
raw4	raw4
raw5	raw5
raw6	raw6
exptot0	
exptot01	
lexptot00	
lexptot0001	

Fuente: STATA con datos propios.

- El comando *ttest* toma la variable de diferencia de los resultados creados como *lexptot0801*) y lo compara con participantes y no participantes de microcréditos. En esencia, crea una segunda diferencia con *lexptot 0801* para aquellos con *dfmfd* = 1, o sea, CEOs femeninas con gasto per cápita **pbt** que acceden al **pei** y excluyendo a los de género masculino con *dfmfd* == 0. Así, se debe teclear (ver **Tabla 5.3**):

```
ttest lexptot0801, by(dfmfd)
```

**Tabla 5.3. Comandos iniciales tabla de resultados
prueba simple comparación ttest**

```
. ttest lexptot0801, by(dfmfd)
```

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
0	391	.1473188	.0269923	.5337372	.0942502	.2003873
1	435	.2586952	.024194	.5046057	.2111432	.3062472
combined	826	.2059734	.018137	.5212616	.1703733	.2415735
diff		-.1113764	.03614		-.1823136	-.0404392

```
diff = mean(0) - mean(1)                                t = -3.0818
Ho: diff = 0                                             degrees of freedom = 824
```

```
Ha: diff < 0                Ha: diff != 0                Ha: diff > 0
Pr(T < t) = 0.0011          Pr(|T| > |t|) = 0.0021          Pr(T > t) = 0.9989
```

*

Command

Fuente: STATA con datos propios.

El resultado muestra, que la participación de CEOs femeninas de las **pbt** en el **pei** con acceso a microcrédito aumenta por consumo capita en un **11.1%** y que este impacto es significativo en al menos de **1 %** de nivel.

El signo negativo significa que el resultado de los *participantes* ($dfmfd = 1$) es mayor que el de *no participantes* ($dfmfd = 0$), lo que implica que el impacto de la participación es de hecho positivo.

Implementación de regresión

En lugar de tomar manualmente la diferencia de los resultados, se puede implementar **DD** usando una regresión. Sobre la base de la discusión en **Ravallion (2008)**, la estimación de **DD** se puede calcular a partir de la regresión:

$$Y_{it} = a + DD \cdot T_i t + \beta T_i + \delta t_i + \varepsilon_{it}$$

Donde:

T es la variable del grupo de tratamiento

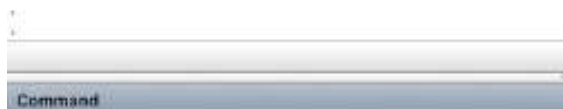
t es el tiempo de relleno (*dummy*) y el coeficiente de interacción de T y t (**DD**), que produce la estimación del impacto del tratamiento, en el resultado **Y**. Por lo tanto, se deberá teclear:

- Los siguientes comandos abren el archivo de datos del panel, crean el registro de la variable de resultado, y crear una variable de participación a nivel del periodo **08** disponible para ambos años, es decir, aquellos quienes participan en programas de microcrédito en el período **08** son el grupo de tratamiento asumido (ver **Tabla 5.4**).

```
use "/Users/DCA/Desktop/STATA log files/pei_0108.dta", clear
gen lexptot=ln(1+exptot)
gen dfmfd1=dfmfd==1 & year==1
egen dfmfd08=max(dfmfd1), by(id)
```

Tabla 5.4. Comandos iniciales y estructura de archivo modificada para preparación prueba por regresión

```
. use "/Users/DCA/Desktop/STATA log files/pei_0100.dta", clear
. gen lexptot=ln(1+exptot)
. gen dfmfd1=dfmfd=1 & year=1
. egen dfmfd00=max(dfmfd1), by(id)
```



Variables	
Name	Label
id	ID
year	year of observation: 0=0...
region	id region
cluster	id cluster
ageceo	age of ceo: years
genceo	gender of ceo: 1=M, 0=F
educceo	education of ceo: years
sme size	sme size based on persa...
sme fac	sme assets based on fac...
incinnov	sme incinnov
expopn	total expenditure
expinnov	sme innovation expendit...
exptot	total expenditure
dmmfd	innovation male microcre...
dfmfd	innovation female microc...
weight	innovation sampling weight
bdb...ess	sme broadband accessib...
pcirr	proportion of region with...
raw1	raw1
raw2	raw2
raw3	raw3
raw4	raw4
raw5	raw5
raw6	raw6
lexptot	
dfmfd1	
dfmfd08	

Fuente: STATA con datos propios.

- El siguiente commando crea la interacción entre la variable tratamiento y el tiempo de relleno (*time dummy*), año en este caso, el cual es **0** para **01** and **1** for **08**):

```
gen dfmfdyr=dfmfd08*year
```

- El siguiente comando ejecuta la *regresión* que implementa el método **DD** (ver **Tabla 5.5**):

```
reg lexptot year dfmfd08 dfmfdyr
```

- Los resultados muestran el mismo impacto de la participación de CEOs de género femenino en los **pei** con acceso a microcréditos sobre el gasto total anual per cápita de las **pbt**, tal como se obtuvo en el periodo anterior ejercicio:

Tabla 5.5. Comandos y tabla de resultados de la regresión para cálculo DD

```

. gen dfmfdyr=dfmfd08*year
. reg lexptot year dfmfd08 dfmfdyr

```

Source	SS	df	MS	Number of obs	=
1,652				F(3, 1648)	=
32.18				Prob > F	=
Model	20.2263902	3	6.74213005	R-squared	=
0.0000				Adj R-squared	=
Residual	345.321048	1,648	.209539471	Root MSE	=
0.0553					
0.0536					
Total	365.547438	1,651	.221409714		
.45775					

	Coef.	Std. Err.	t	P> t	[95% Conf. Int.	
lexptot						
errall						
year	.1473188	.0327386	4.50	0.000	.0831052	.2
113323						
dfmfd08	-.1145671	.0318999	-3.59	0.000	-.1771358	-.0
519984						
dfmfdyr	.1113764	.0451133	2.47	0.014	.0228909	.1
998619						
_cons	8.310481	.0231497	358.00	0.000	8.265875	8.
355887						

Command

Fuente: STATA con datos propios.

- Una suposición básica detrás de la implementación simple de **DD**, es que otras covariables no cambien a través del tiempo (años). Pero si esas variables varían, deberán ser controladas en la regresión, para obtener el *efecto neto* de la participación del programa en los resultados. Entonces, *el modelo de regresión se extiende*, al incluir otras covariables que pueden afectar los resultados de interés. Así, se debe teclear (**ver Tabla 5.6**):

gen lnsmfac=ln(1+smefac/100)

reg lexptot year dfmfd08 dfmfdyr genceo ageceo educeo lnsmfac
bdbaccess pcirr raw1 raw2 raw3 raw5 raw6 [pw=weight]

Tabla 5.6. Comandos, tabla de resultados con la regresión incluyendo todas las covariables involucradas y estructura del archivo

```
. gen lnsmfac=ln(1+smefac/100)

. reg lexptot year dfmfd08 dfmfdyr genceo ageceo educeo lnsmfac bdbaccess pcirr raw1 raw2 raw3 raw5 raw6 [pw=weight]
(sum of wgt is 1.6445e+03)
```

Linear regression

Number of obs	=	1,652
F(14, 1637)	=	24.90
Prob > F	=	0.0000
R-squared	=	0.2826
Root MSE	=	.42765

	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
year	.2768099	.0679939	4.07	0.000	.1434456	.4101741
dfmfd08	.0012122	.0326585	0.04	0.970	-.0628446	.0652691
dfmfdyr	.0514655	.0530814	0.97	0.332	-.0526491	.1555802
genceo	-.0455035	.053903	-0.84	0.399	-.1512296	.0602227
ageceo	.0017445	.0011041	1.58	0.114	-.0004212	.0039102
educeo	.0385333	.0049841	7.73	0.000	.0287575	.0483092
lnsmfac	.226467	.0309236	7.32	0.000	.165813	.2871209
bdbaccess	-.011292	.0498495	-0.23	0.821	-.1090674	.0864835
pcirr	.0628715	.0453625	1.39	0.166	-.0261031	.1518461
raw1	-.0023961	.0109958	-0.22	0.828	-.0239634	.0191712
raw2	.0071376	.0120905	0.59	0.555	-.0165769	.0308521
raw3	.0158481	.005106	3.10	0.002	.0058332	.025863
raw5	.1458875	.0475718	3.07	0.002	.0525794	.2391956
raw6	.0011434	.0031013	0.37	0.712	-.0049395	.0072263
_cons	7.399387	.2715525	27.25	0.000	6.86676	7.932014

Command

Fuente: STATA con datos propios.

Variables	
Name	Label
id	ID
year	year of observation: 0=01, 1=08
region	id region
cluster	id cluster
ageceo	age of ceo: years
genceo	gender of ceo: 1=M, 0=F
educceo	education of ceo: years
sme size	sme size based on personnel
sme fac	sme assets based on facilities
incinnov	sme incinnov
expopn	total expenditure
expinnov	sme innovation expenditure: Tk/year
exptot	total expenditure
dmfmd	innovation male microcredit participant: 1=Y, 0=N
dfmfd	innovation female microcredit participant: 1=Y, 0=N
weight	innovation sampling weight
bdbaccess	sme broadband accessible all year: 1=Y, 0=N
pcirr	proportion of region with internet broadband
raw1	raw1
raw2	raw2
raw3	raw3
raw4	raw4
raw5	raw5
raw6	raw6
lexptot	
dfmfd1	
dfmfd08	
dfmfdyr	
insmefac	

Fuente: STATA con datos propios.

Al mantener constantes otros factores, se aprecia que el impacto del programa, han cambiado de significate a no significativo ($t = 0.97$).

Checando robustez de DD con regresión de efectos fijos

Otra forma de medir la estimación de **DD**, es usar una *regresión de efectos fijos en lugar de mínimos cuadrados ordinarios (OLS. Ordinary Least Squares)*. Lo anterior por:

- La *regresión de efectos fijos* hace un mejor control de características no observadas en las **pbt** y de las características invariantes en el tiempo, que pueden influir en la variable de resultado.
- El comando **STATA** *xtreg* es el que se utiliza para ejecutar la regresión de efectos fijos. En particular, se complementa con la opción *fe*, que se ajusta a los modelos de efectos fijos. Así, se deberá teclear (ver **Tabla 5.7**):

```
xtreg lexptot year dfmfd08 dfmfdyr, fe i(id)
```

Los resultados mostraron nuevamente un impacto positivo significativo de la participación de las CEOs de género femenino.

Tabla 5.7. Comandos y tabla de resultados de la regresión de efectos fijos

```
. xtreg lexptot year dfmfd08 dfmfdyr, fe i(id)
note: dfmfd08 omitted because of collinearity
```

Fixed-effects (within) regression	Number of obs	=	1,652
Group variable: id	Number of groups	=	826
R-sq:	Obs per group:		
within = 0.1450	min =		2
between = 0.0061	avg =		2.0
overall = 0.0415	max =		2
corr(u_i, Xb) = -0.0379	F(2,824)	=	69.90
	Prob > F	=	0.0000

lexptot	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
year	.1473188	.0262266	5.62	0.000	.0958399	.1987976
dfmfd08	0 (omitted)					
dfmfdyr	.1113764	.03614	3.08	0.002	.0404392	.1823136
_cons	8.250146	.0127593	646.60	0.000	8.225101	8.27519
sigma_u	.38132289					
sigma_e	.36670395					
rho	.51953588	(fraction of variance due to u_i)				

F test that all u_i=0: F(825, 824) = 2.14 Prob > F = 0.0000

.

Command

Fuente: STATA con datos propios.

- Al incluir el resto de las covariables en la regresión, *regresión de efectos fijos* se extienden, de la siguiente forma (**ver Tabla 5.8**):

```
xtreg lexptot year dfmfd08 dfmfdyr genceo ageceo educeo  
lnsmefac bdbaccess pcirr raw1 raw2 raw3 raw5 raw6, fe i(id)
```

Tabla 5.8. Comandos y tabla de resultados de la regresión de efectos fijos incluyendo a todas las covariables involucradas

. xtreg lexptot year dfmfd08 dfmfdyr genceo ageceo educeo lnsmefac bdbaccess pcirr raw1 raw2 raw3 raw5 raw6, fe i(id) note: dfmfd08 omitted because of collinearity					
Fixed-effects (within) regression Group variable: id			Number of obs = 1,652 Number of groups = 826		
R-sq:			Obs per group:		
within = 0.1715			min = 2		
between = 0.1914			avg = 2.0		
overall = 0.1737			max = 2		
corr(u_i, Xb) = 0.1222			F(13,813) = 12.95 Prob > F = 0.0000		
lexptot	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
year	.2211178	.063087	3.50	0.000	.0972851 .3449504
dfmfd08	0 (omitted)				
dfmfdyr	.0906308	.0367358	2.47	0.014	.0185226 .1627391
genceo	-.0577238	.0722968	-0.80	0.425	-.1996342 .0841866
ageceo	-.0003766	.0016985	-0.22	0.825	-.0037106 .0029574
educeo	.0137419	.0082935	1.66	0.098	-.0025373 .030021
lnsmefac	.1381659	.0619682	2.23	0.026	.0165293 .2598025
bdbaccess	-.0932955	.053396	-1.75	0.081	-.1981057 .0115147
pcirr	.0823594	.0642728	1.28	0.200	-.0438009 .2085196
raw1	.0107911	.010209	1.06	0.291	-.0092481 .0308303
raw2	-.0227681	.0123379	-1.85	0.065	-.046986 .0014498
raw3	-.0014743	.0064578	-0.23	0.819	-.0141503 .0112016
raw5	.1439482	.047915	3.00	0.003	.0498965 .238
raw6	.0038546	.0031366	1.23	0.219	-.0023022 .0100113
_cons	7.853111	.2482708	31.63	0.000	7.365784 8.340439
sigma_u	.34608097				
sigma_e	.3634207				
rho	.47557527	(fraction of variance due to u_i)			
F test that all u_i=0: F(825, 813) = 1.59			Prob > F = 0.0000		

Fuente: STATA con datos propios.

Los resultados muestran que, después de controlar los efectos de los factores de tiempo invariante no observado, la participación de CEOs femenina con acceso al microcrédito, tiene un impacto positivo del **9.1%** en el consumo per cápita de las **pbt**, y el impacto es muy significativo.

DD en datos de sección transversal

El método **DD**, también se puede aplicar a *datos de sección transversal*, no solo a *datos de panel*. La idea es muy similar a la utilizada en los datos del panel. *En lugar de una comparación entre años, son comparadas regiones programadas y no programados, en lugar de una comparación entre participantes y no participantes, se comparan grupos objetivo y no objetivo.* De esta forma:

- Se utilizan los datos de **pei_01.dta** para crear una variable ficticia (*dummy*) llamada *target*, para aquellas **pbt** que sean elegibles para participar en programas de microcrédito (es decir, aquellos que tienen menos de 50 decimales de activos sme). Luego, se crea otra variable ficticia *dummy* del programa a nivel cluster (*progcluster*) para los que están identificados como tal. Por lo tanto, se tiene (ver **Tabla 5.9**):

```
use "/Users/DCA/Desktop/STATA log files/pei_01.dta", clear
gen lexptot=ln(1+exptot)
gen lnsmejac=ln(1+smejac/100)
gen target=smejac<50
gen progcluster=cluster<25
```

Tabla 5.9. Comandos para preparar estructura de archivo en análisis de datos de sección transversal

```
. use "/Users/DCA/Desktop/STATA log files/pei_01.dta", clear
(Mexico's small and medium enterprises based on technology)

.. use "/Users/DCA/Desktop/STATA log files/pei_01.dta", clear
(Mexico's small and medium enterprises based on technology)

. gen lexptot=ln(1+exptot)

. gen lnsmefac=ln(1+smefac/100)

. gen target=smefac<50

. gen progcluster=cluster<25
```



Fuente: STATA con datos propios.

Variables	
Name	Label
region	id region
cluster	id cluster
ageceo	age of ceo
gender	gender of ceo: 1=M, 0=F
educceo	education of ceo: years
sme size	sme size based on personal
smefac	sme assets based on facilities
incinnov	sme incinnov
expopn	sme opn expenditure
expinnov	sme innovation expenditure
exp tot	total expenditure
dmfmd	innovation male microcredit participant: 1=Y, 0=N
dfmfd	innovation female microcredit participant: 1=Y, 0=N
weight	innovation sampling weight
bdbaccess	sme broad band accessible all year: 1=Y, 0=N
pickr	Proportion of region with internet broadband
raw1	raw1
raw2	raw2
raw3	raw3
raw4	raw4
raw5	raw5
raw6	raw6
oldceo	CEO is over 50: 1=Y, 0=N
avgage	
avgagemf	
lexptot	
lnsmefac	
target	
progcluster	

Fuente: STATA con datos propios.

- Posteriormente, genere una variable que interactúe con la *prog-cluster* y el *target* del programa:

```
gen progtarget=progcluster*target
```

- Así, calcule la estimación de **DD** con regresión del registro del gasto total per cápita de la **pbt** contra las variables cluster, target y su interacción (ver **Tabla 5.10**):

```
reg lexptot progcluster target progtarget
```

Los resultados muestran, que el impacto de la colocación del **pei** con acceso a microcrédito en el *grupo objetivo* (*target*) no es significativo ($t = -0.61$).

Tabla 5.10. Comandos y tabulación de la regresión

```
. gen progtarget=progcluster*target
```

```
. reg lexptot progcluster target progtarget
```

Source	SS	df	MS	Number of obs	=	826
Model	10.9420259	3	3.64734195	F(3, 822)	=	27.38
Residual	109.485295	822	.133193789	Prob > F	=	0.0000
				R-squared	=	0.0909
				Adj R-squared	=	0.0875
Total	120.427321	825	.14597251	Root MSE	=	.36496

lexptot	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
progcluster	-.0646577	.0770632	-0.84	0.402	-.2159215 .086606
target	-.2996852	.0815261	-3.68	0.000	-.459709 -.1396614
progtarget	.0529438	.0867976	0.61	0.542	-.1174272 .2233147
_cons	8.485567	.0729914	116.25	0.000	8.342296 8.628039

Command

Fuente: STATA con datos propio.

- El coeficiente de la variable de impacto (*progtarget*), que es **0.053**, no da el impacto real de los programas de microcrédito; tiene que ajustarse dividiendo entre una proporción de los **pbt** objetivo en los cluster del programa. El siguiente comando sirve para encontrar la proporción (**ver Tabla 5.11**):

```
sum target if progcluster==1
```

Tabla 5.11. Comandos y tabla de resultados tras aplicar proporción

```
. sum target if progcluster==1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
target	700	.6885714	.4634087	0	1

```

+-----+
| Command |
+-----+

```

Fuente: STATA con datos propios.

- De los **pbt** en los cluster del programa, el **68.9%** pertenece al grupo objetivo. Por lo tanto, el coeficiente de regresión de *progtarget* se divide por este valor, (**0.529438/0.6885714**) resultando **0.077**, que es el verdadero impacto de los programas de microcrédito en la población objetivo, incluso, aunque no sea significativo.
- Como antes, el modelo de regresión puede especificarse ajustando las covariables que afectan los resultados de interés (**ver Tabla 5.12**):

```
reg lexptot progcluster target progtarget genceo ageceo educeo  
lnsmefac bdbaccess pcirr raw1 raw2 raw3 raw5 raw6 [pw=weight]
```

Manteniendo constantes otros factores, se encuentran cambios en el nivel de significación de impactos del microcrédito en el gasto total anual per cápita de las **pbt**.

Tabla 5.12. Comandos y tabulación de la regresión incluyendo todas las covariables para análisis de datos de sección transversal

```
. xtreg lexptot progcluster target progtarget genceo ageceo educeo lnsmefac, fe i(cluster)
note: progcluster omitted because of collinearity
```

```
Fixed-effects (within) regression              Number of obs   =      826
Group variable: cluster                      Number of groups =      29

R-sq:                                         Obs per group:
    within = 0.2213                          min           =      21
    between = 0.0050                          avg           =     28.5
    overall  = 0.1901                          max           =     38

                                           F(6,791)        =     37.47
corr(u_i, Xb) = -0.0782                      Prob > F        =     0.0000
```

lexptot	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
progcluster	0	(omitted)				
target	.001118	.0805374	0.01	0.989	-.1569742	.1592102
progtarget	.0040862	.0790204	0.05	0.959	-.1510282	.1592007
genceo	-.0236468	.056175	-0.42	0.674	-.1339164	.0866228
ageceo	-.0001933	.0010037	-0.19	0.847	-.0021636	.001777
educeo	.0267043	.0037686	7.09	0.000	.0193066	.034102
lnsmefac	.2466233	.0378532	6.52	0.000	.1723188	.3209278
_cons	8.120513	.0799974	101.51	0.000	7.96348	8.277545
sigma_u	.13064017					
sigma_e	.32648911					
rho	.13801212	(fraction of variance due to u_i)				

F test that all u_i=0: F(28, 791) = 4.41 Prob > F = 0.0000

Command

Fuente: STATA con datos propios.

- Nuevamente, la *regresión de efectos fijos* se puede usar en lugar de **OLS** (*Ordinary Least Squares*), para verificar la robustez de los resultados. Sin embargo, con *datos transversales*, los efectos fijos a nivel del **pbt** es posible que no se ejecuten, porque cada **pbt** o sujeto, aparece solo una vez en los datos. Por lo tanto, a nivel de cluster sí se ejecuta la *regresión de efectos fijos*, con la siguiente programación (**ver Tabla 5.13**):

```
xtreg lexptot progcluster target progtarget, fe i(cluster)
```

Esta vez, hay un impacto negativo (insignificante) de los programas de microcrédito en el gasto per cápita de los **pbt**:

Tabla 5.13. Comandos y tabulación de regresión de efectos fijos para verificar robustez de resultados a nivel clúster

```
. xtreg lexptot progcluster target progtarget, fe i(cluster)
note: progcluster omitted because of collinearity
```

```
Fixed-effects (within) regression               Number of obs   =       826
Group variable: cluster                       Number of groups =       29

R-sq:                                         Obs per group:
    within = 0.1037                          min =          21
    between = 0.0035                         avg =         28.5
    overall = 0.0900                         max =          38

corr(u_i, Xb) = -0.0455                      F(2,795)        =       45.98
                                           Prob > F         =       0.0000
```

lexptot	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
progcluster	0 (omitted)					
target	-.2457652	.0791694	-3.10	0.002	-.4011708	-.0903595
progtarget	-.0212434	.0844614	-0.25	0.801	-.187037	.1445503
_cons	8.436006	.0231597	364.25	0.000	8.390545	8.481467
sigma_u	.1268703					
sigma_e	.34940166					
rho	.1164883	(fraction of variance due to u_i)				

F test that all u_i=0: F(28, 795) = 3.66 Prob > F = 0.0000

```
.
.
```

Command

Fuente: STATA con datos propios.

- La misma *regresión de efectos fija* se vuelve a correr incluyendo todas las covariables (ver **Tabla 5.14**):

```
xtreg lexptot progcluster target progtarget genceo ageceo educeo
lnsmefac, fe i(cluster)
```

No se observan cambios significativos de nivel si se ubicaran clústers de 29-87.

Tabla 5.14. Comandos y tabulación de regresión de efectos fijos para verificar robustez de resultados a nivel clúster con todas las covariables

```
. xtreg lexptot progcluster target progtarget genceo ageceo educeo lnsmejac, fe i(cluster)
note: progcluster omitted because of collinearity
```

```
Fixed-effects (within) regression      Number of obs   =      826
Group variable: cluster               Number of groups =      29

R-sq:                                Obs per group:
    within = 0.2213                    min       =      21
    between = 0.0050                   avg       =     28.5
    overall  = 0.1901                   max       =      38

corr(u_i, Xb) = -0.0782                F(6,791)         =     37.47
                                         Prob > F         =     0.0000
```

	lexptot	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
progcluster		0 (omitted)				
target		.001118	.0805374	0.01	0.989	-.1569742 .1592102
progtarget		.0040862	.0790204	0.05	0.959	-.1510282 .1592007
genceo		-.0236468	.056175	-0.42	0.674	-.1339164 .0866228
ageceo		-.0001933	.0010037	-0.19	0.847	-.0021636 .001777
educeo		.0267043	.0037686	7.09	0.000	.0193066 .034102
lnsmejac		.2466233	.0378532	6.52	0.000	.1723188 .3209278
_cons		8.120513	.0799974	101.51	0.000	7.96348 8.277545
sigma_u		.13064017				
sigma_e		.32648911				
rho		.13801212				(fraction of variance due to u_i)

```
F test that all u_i=0: F(28, 791) = 4.41                Prob > F = 0.0000
```

```
Command
```

Fuente: STATA con datos propios.

Tomando en cuenta las condiciones iniciales

Aunque la implementación **DD** a través de controles de regresión (OLS o regresión efectos fijos) para covariables a nivel de **pbt** y clúster, las condiciones iniciales durante las encuestas en *línea de base*, pueden tener una influencia separada sobre los cambios posteriores en el resultado o asignación al tratamiento. Por lo tanto, ignorar el efecto separado de las condiciones iniciales, puede sesgar las estimaciones de **DD**.

Sin embargo, incluir las *condiciones iniciales* en la regresión, es complicado. Porque las observaciones de referencia (línea base) en la muestra del panel, ya contienen características iniciales. Las variables extra para las condiciones iniciales, no se pueden agregar directamente. Una forma de agregar condiciones iniciales, es tener en cuenta una imple-

mentación alternativa de la regresión de efectos fijos. En esta implementación, se crean variables de diferencia para todas las variables (resultado y covariables) entre los años, y luego estas variables de diferencia se usan en regresiones, en lugar de las variables originales. En este conjunto de datos modificado, las variables de condición inicial, se pueden agregar como regresores adicionales sin problemas de colinealidad, por lo que:

- Los siguientes comandos crean las variables de diferencia del panel **pei_0108.dta**. Se debe asegurar que en los archivos **pei_01.dat** y **pei08.dat**, se encuentren en la estructura:

```
gen lexptot=ln(1+exptot)
gen lnsmefac=ln(1+smefac/100)
gen dfmfd1=dfmfd==1 & year==1
egen dfmfd08=max(dfmfd1), by(id)
gen dfmfdyr=dfmfd08*year
```

- Con lo anterior, proceder a teclear:

```
clear
use "/Users/DCA/Desktop/STATA log files/pei_0108.dta"
sort id year
by id: gen dlexptot=lexptot[2]-lexptot[1]
by id: gen dlsmefac=lnsmefac[2]-lnsmefac[1]
by id: gen ddfmfd08=dfmfd08[2]-dfmfd08[1]
by id: gen ddfmfdyr=dfmfdyr[2]-dfmfdyr[1]
by id: gen ddmfd=dmmfd[2]-dmmfd[1]
by id: gen ddfmfd=dfmfd[2]-dfmfd[1]
by id: gen dgenceo=genceo[2]-genceo[1]
by id: gen dageceo=ageceo[2]-ageceo[1]
```



```

by id: gen deduceo=educeo[2]-educeo[1]
by id: gen dbdbaccess=bdbaccess[2]-bdbaccess[1]
by id: gen dpcirr=pcirr[2]-pcirr[1]
by id: gen draw1=row1[2]-row1[1]
by id: gen draw2=row2[2]-row2[1]
by id: gen draw3=row3[2]-row3[1]
by id: gen draw5=row5[2]-row5[1]
by id: gen draw6=row6[2]-row6[1]

```

- **STATA** crea estas variables de diferencia para ambos años. Entonces una regresión **OLS** se ejecuta, con las variables de diferencia más las covariables originales como regresores adicionales, restringiendo la muestra al año de *línea de base* o referencia (año = 0). Esto se debe realizar porque la *línea de base* contiene tanto las variables de diferencia como las variables de condición inicial. Por lo tanto, se teclea (**ver tabla 5.15**):

```

reg dlexptot ddfmfd08 ddfmfdyr dgenceo dageceo deduce dlns-
mefac dbdbaccess dpcirr draw1 draw2 draw3 draw5 draw6 genceo
ageceo educeo lnsmeffac bdbaccess pcirr row1 row2 row3 row5
row6 if year==0 [pw=weight]

```

Tabla 5.15. Tabla de resultados de la regresión
incluyendo todas las covariables

dlexptot	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
ddfmfd08	(dropped)					
ddfmfdyr	.0619405	.0435103	1.42	0.155	-.0234671	.1473481
dgenceo	-.0615416	.0871488	-0.71	0.480	-.2326083	.1095251
dageceo	.0013583	.0023165	0.59	0.558	-.0031889	.0059055
deduceo	.0153497	.0117889	1.30	0.193	-.0077909	.0384904
dlnsmefac	.1260302	.0701158	1.80	0.073	-.011602	.2636624
bdbaccess	-.1365889	.0702504	-1.94	0.052	-.2744853	.0013075
dpcirr	.1042085	.1124156	0.93	0.354	-.1164551	.3248721
draw1	.0065267	.0147616	0.44	0.659	-.0224493	.0355027
draw2	-.04828	.0261598	-1.85	0.065	-.0996297	.0030697
draw3	-.0071707	.0143637	-0.50	0.618	-.0353656	.0210241
draw5	.0137635	.0062199	2.21	0.027	.0015542	.0259727
draw6	.1991899	.101613	1.96	0.050	-.0002689	.3986486
genceo	-.1157563	.0844686	-1.37	0.171	-.281562	.0500494
ageceo	.0054212	.002046	2.65	0.008	.001405	.0094375
educeo	.0230352	.008891	2.59	0.010	.0055828	.0404876
lnsmetac	-.0690961	.0545822	-1.27	0.206	-.1762369	.0380448
bdbaccess	-.1142214	.1065896	-1.07	0.284	-.323449	.0950062
pcirr	.1471455	.109057	1.35	0.178	-.0669254	.3612164
raw1	-.0047485	.0317983	-0.15	0.881	-.0671661	.0576691
raw2	-.0337045	.0306002	-1.10	0.271	-.0937705	.0263614
raw3	-.0047502	.0129723	-0.37	0.714	-.0302138	.0207134
raw5	.0205757	.0083353	2.47	0.014	.0042142	.0369373
raw6	.1015795	.1273284	0.80	0.425	-.1483568	.3515158
_cons	-.704969	.5861648	-1.20	0.229	-1.855567	.4456292

Fuente: STATA con datos propios.

Los resultados muestran que, después de controlar las condiciones iniciales, el impacto de la participación del microcrédito desaparece (t = en función de la variable de relación que escoja a interactuar).

DD combinada con PSM

El método **DD** se puede refinar de varias maneras:

1. Una es usando la propensión de puntaje de coincidencia o emparejamiento (**PSM**) con los datos de referencia o *línea de base*, para asegurarse de que el *grupo de comparación* sea similar al *grupo de tratamiento* y luego, aplicando **DD** a la muestra e coincidencia o emparejada. De esta manera, se puede tratar la *heterogeneidad observable* en las condiciones iniciales. Usando el comando `pscore`, la variable de participación en 08/09 (aquí como **dfmfd08** para ambos años) se hace su regresión con variables exógenas de **01/02** para obtener puntajes de propensión a partir de los datos de referencia. Estos comandos son los siguientes (**ver Tabla 5.16**):

```
clear
use "/Users/DCA/Desktop/STATA log files/pei_0108.dta"
gen lnsmejac=ln(1+smefac/100)
gen dfmfd1=dfmfd==1 & year==1
egen dfmfd08=max(dfmfd1), by(id)
keep if year==0
pscore dfmfd08 genceo ageceo educeo lnsmejac bdbaccess pcirr
raw1 raw2 raw3 raw5 raw6 [pw=weight], pscore(ps98) blockid(-
blockf1) comsup level(0.001)
```

La propiedad de balanceo del **PSM** debe ser satisfecha, lo cual significa que las **pbt** con la misma propensión de puntaje tienen las mismas distribuciones de todas las covariables en todos los cinco bloques. La *región de soporte común* es **[0.606030439, 0.78893426]**, con 26 regiones.

Tabla 5.16. Comandos y tabulación de Resultados utilizando PSM

```
. pscore dfmfd08 genceo ageceo educeo lnsmfefac bdbaccess pcirr raw1 raw2 raw3 raw5 raw6 [pweight], pscore(pscore) blo
> ckid(blockf1) consup level(0.001)
(0 real changes made)
```

```
*****
Algorithm to estimate the propensity score
*****
```

The treatment is dfmfd08

dfmfd08	Freq.	Percent	Cum.
0	391	47.34	47.34
1	435	52.66	100.00
Total	826	100.00	

Estimation of the propensity score

```
(sum of wgt is 8.2233e+02)
Iteration 0: log pseudolikelihood = -554.25786
Iteration 1: log pseudolikelihood = -480.05123
Iteration 2: log pseudolikelihood = -475.25432
Iteration 3: log pseudolikelihood = -475.17443
Iteration 4: log pseudolikelihood = -475.1744
```

```
Probit regression                                Number of obs   =      826
                                                Wald chi2(11)   =      78.73
                                                Prob > chi2     =      0.0000
Log pseudolikelihood = -475.1744                Pseudo R2      =      0.1427
```

dfmfd08	Robust					
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
genceo	-.1512794	.2698723	-0.56	0.575	-.6802194	.3776605
ageceo	-.0073102	.0046942	-1.56	0.119	-.0165106	.0018903
educeo	-.0261142	.018235	-1.43	0.152	-.0618542	.0096257
lnsmfefac	-.9010234	.137662	-6.55	0.000	-1.170836	-.6312109
bdbaccess	.2894359	.2626682	1.10	0.271	-.2253843	.804256
pcirr	.0367083	.1999013	0.18	0.854	-.3550911	.4285077
raw1	.1682276	.0606261	2.77	0.006	.0494028	.2870525
raw2	.0603593	.0500646	1.21	0.228	-.0377655	.1584841
raw3	-.0472819	.0205877	-2.30	0.022	-.087633	-.0069309
raw5	-.2991866	.184372	-1.62	0.105	-.660549	.0621759
raw6	.009133	.0141985	0.64	0.520	-.0186954	.0369615
_cons	-1.002465	1.241022	-0.81	0.419	-3.434823	1.429894

Note: the common support option has been selected
The region of common support is [.06030439, .78893426]

**Description of the estimated propensity score
in region of common support**

Estimated propensity score				
	Percentiles	Smallest		
1%	.0800224	.0603044		
5%	.1415098	.061277		
10%	.2124288	.0622054	Obs	800
25%	.3583033	.0647113	Sum of Wgt.	800
50%	.481352		Mean	.4579494
		Largest	Std. Dev.	.1612539
75%	.570064	.7616697		
90%	.6600336	.7650957	Variance	.0260028
95%	.688278	.7716357	Skewness	-.4881678
99%	.7515092	.7889343	Kurtosis	2.637857

Step 1: Identification of the optimal number of blocks
 Use option detail if you want more detailed output

The final number of blocks is 4

This number of blocks ensures that the mean propensity score
is not different for treated and controls in each blocks

Step 2: Test of balancing property of the propensity score
 Use option detail if you want more detailed output

The balancing property is satisfied

This table shows the inferior bound, the number of treated
and the number of controls for each block

Inferior of block of pscore	dfmfd08		Total
	0	1	
.0603044	53	16	69
.2	110	70	180
.4	151	250	401
.6	51	99	150
Total	365	435	800

```
Note: the common support option has been selected
```

```
*****
End of the algorithm to estimate the pscore
*****
```

```
.
```

```
Command
```

Fuente: STATA con datos propios.

Los siguientes comandos, mantienen la coincidencia o emparejamiento de las **pbt**, en el año de referencia o *línea de base* y los combina con los datos del panel para mantener solo los **pbt** coincidentes en la muestra del panel (ver **Tabla 5.17**):

```
keep if blockf1!=.
keep id
sort id
merge id using "/Users/DCA/Desktop/STATA log files/pei_0108.dta"
keep if _merge==3
```

Tabla 5.17. Comandos y estructura de archivo fusionado

```
. keep if blockf1!=.
{26 observations deleted}

. keep id

. sort id

. merge id using "/Users/DCA/Desktop/STATA log files/pei_0108.dta"
{note: you are using old merge syntax; see [0] merge for new syntax}
variable id does not uniquely identify observations in /Users/DCA/Desktop/STATA log files/pei_0108.dta

. keep if _merge==3
{32 observations deleted}

.
```

Fuente: STATA con datos propios.

Variables	
Name	Label
id	ID
year	year of observation: 0=01, 1=08
region	id region
cluster	id cluster
ageceo	age of ceo: years
gencceo	gender of ceo: 1=M, 0=F
educceo	education of ceo: years
sme size	sme size based on personnel
sme fac	sme assets based on facilities
incinnov	sme incinnov
expopn	total expenditure
expinnov	sme innovation expenditure: Tk/year
exptot	total expenditure
dmmfd	innovation male microcredit participant: 1=Y, 0=N
dfmfd	innovation female microcredit participant: 1=Y, 0=N
weight	innovation sampling weight
bdbaccess	sme broadband accessible all year: 1=Y, 0=N
pcirr	proportion of region with internet broadband
raw1	raw1
raw2	raw2
raw3	raw3
raw4	raw4
raw5	raw5
raw6	raw6
_merge	

Fuente: STATA con datos propios.

2. El siguiente paso, es implementar el método **DD**. Para este ejercicio solo la implementación de los efectos fijos es mostrada: (**ver Tabla 5.18**):

```

gen lexptot=ln(1+exptot)
gen lnsme fac=ln(1+sme fac/100)
gen dfmfd1=dfmfd==1 & year==1
egen dfmfd08=max(dfmfd1), by(id)
gen dfmfdyr=dfmfd08*year
xtreg lexptot year dfmfd08 dfmfdyr gencceo ageceo educceo
lnsme fac bdbaccess pcirr raw1 raw2 raw3 raw5 raw6, fe i(id)

```

Los resultados muestran, que la aplicación de **PSM a DD** conserva el *impacto positivo* original de participación de CEOs de género femenino en el pei de acceso a microcrédito para los gastos per cápita de las **pbt**.

Tabla 5.18. Comandos y tabulación de resultados

```

. gen lexptot=ln(1+expot)
. gen lnsmefac=ln(1+smefac/100)
. gen dfmfd08=dfmfd0 & year=01
. egen dbdbaccess=dfmfd1, bgl(4)
. gen dfmfdyr=ln(1+dbdbaccess)

. xtreg lexptot year dfmfd08 dfmfdyr genceo ageceo educeo lnsmefac dbdbaccess pcirr raw1 raw2 raw3 raw5 raw6, fe i(id)
      (note: dfmfd08 omitted because of collinearity)

Fixed-effects (within) regression              Number of obs   =    2,400
Group variable: id                          Number of groups =     200

Time:
      within = 0.1291                        obs =           2
      between = 0.1257                       sep =          2.0
      overall = 0.1400                       res =           2

      F(13,197) = 59.25
      Prob > F   = 0.0000

varrta_3, X(1) = 0.0634

```

lexptot	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
year	.222509	.0639108	3.48	0.001	.0970532	.3479647
dfmfd08	0 (omitted)					
dfmfdyr	.0925741	.0371517	2.49	0.013	.019646	.1655023
genceo	-.084584	.0739679	-1.14	0.253	-.2297818	.0606138
ageceo	-.0003225	.001732	-0.19	0.852	-.0037223	.0030773
educeo	.0132322	.0084471	1.57	0.118	-.0033494	.0298138
lnsmefac	.2003341	.0778701	2.57	0.010	.0474766	.3531917
dbdbaccess	-.0857169	.0542065	-1.58	0.114	-.1921234	.0206896
pcirr	.083983	.0644159	1.30	0.193	-.0424644	.2104303
raw1	.0131877	.0102657	1.28	0.199	-.0069638	.0333392
raw2	-.0272757	.0123259	-2.21	0.027	-.0514712	-.0030802
raw3	-.0015386	.0064937	-0.24	0.813	-.0142857	.0112084
raw5	.1400882	.0485296	2.89	0.004	.0448254	.2353509
raw6	.0047885	.0031592	1.52	0.130	-.001413	.0109899
_cons	7.815588	.2504303	31.21	0.000	7.323998	8.307179
sigma_u	.33642591					
sigma_e	.36009944					
rho	.46605118	(fraction of variance due to u_i)				

F test that all u_i=0: F(799, 787) = 1.58

Prob > F = 0.0000

•
•

Command

Fuente: STATA con datos propios.

ANEXO.

Introducción a STATA

Esta sección está diseñada para conocer los recursos de **STATA** <https://www.stata.com/>. Ver **imagen 1**.

Imagen 1. Sitio web de STATA



Fuente: sitio web STATA.

Con el fin de manipular los archivos del caso de estímulos a la innovación, adjuntos a la obra, utilizando **STATA 14.1** (en adelante **STATA**). Es un caso hipotético de información recolectada de pymes de base tecnológica (**pbt**) correspondiente al sector de desarrollo de software de la industria electrónica, en varias regiones de un país, con observaciones de año 1 a año 8.

Estructura de archivos (*File Structure*)

Estos ejercicios usan y generan muchos archivos. Existen principalmente tres tipos de archivos **STATA**:

- Los que contienen el conjuntos de datos, identificados por el sufijo **.dta**
- Los que contienen los programas **STATA**, identificados por el sufijo **.do**, y
- Los que contienen el registro y salida del trabajo hecho en **STATA** identificados por el sufijo **.log**.

Para mantener estos archivos organizados, se sugiere la siguiente estructura de directorios, ubicados en un disco c, de un equipo Macintosh:

“/Users/DCA/Desktop/eval/data”

“/Users/DCA/Desktop/eval/do”

“/Users/DCA/Desktop/eval/log”

Descripción de archivos y ejemplos en STATA

Los archivos de datos, son localizados en “/Users/DCA/Desktop/eval/data”, del disco c de un equipo Macintosh, con tres archivos:

- **pei_01.dta**. Este archivo comprende los datos de pymes de base tecnológica (**pbt**) del año **1** que con **826** registros con **24** variables de información.
 - 1.ID
 2. year. Year of observation
 3. region. Region

4. cluster. Cluster
 5. ageceo. Age of CEO
 6. genceo. Gender of CEO
 7. educeo. Education of CEO
 8. smesize. Sme size based on quantity of personnel
 9. smefac. Facilities, resources for innovations
 10. incinnov. Incentives for innovations per year
 11. expopn. Sme operation expenditure per year
 12. expinnov. Sme innovation expenditure per year
 13. exptot. Sme total expenditure per year
 14. dmmfd. Innovation male microcredit participant: 1= Y; 0=N
 15. dfmfd. Innovation female microcredit participant: 1= Y; 0=N
 16. weight. Innovation sampling weight
 17. bdbaccess. Broadband access
 18. pcirr. Proportion of region with internet broadband
 19. raw1. Raw 1
 20. raw2. Raw 2
 21. raw3. Raw 3
 22. raw4. Raw 4
 23. raw5. Raw 5
 24. raw6. Raw 6
- **pei_08.dta.** Este archivo es la versión levantada al **año 8** de la versión **pei_01.dta**. Incluye **303** nuevos pymes de base tecnológica (**pbt**), haciendo el número total **1,129** pymes.
 - **pei_0108.dta.** Este es un conjunto de datos de panel restringido a **826** pymes de base tecnológica (**pbt**) entrevistados en los dos años. en formato de serie temporal.

El subdirectorio **.do** tiene los *archivos del programa (.do)* específicos para las diferentes técnicas de evaluación de impact. Estos archivos contienen todo el código **STATA** necesario para implementar los ejemplos de los capítulos aquí contenidos.

El subdirectorio **.log** contiene todas las salidas generadas al ejecutar los archivos **.do**.

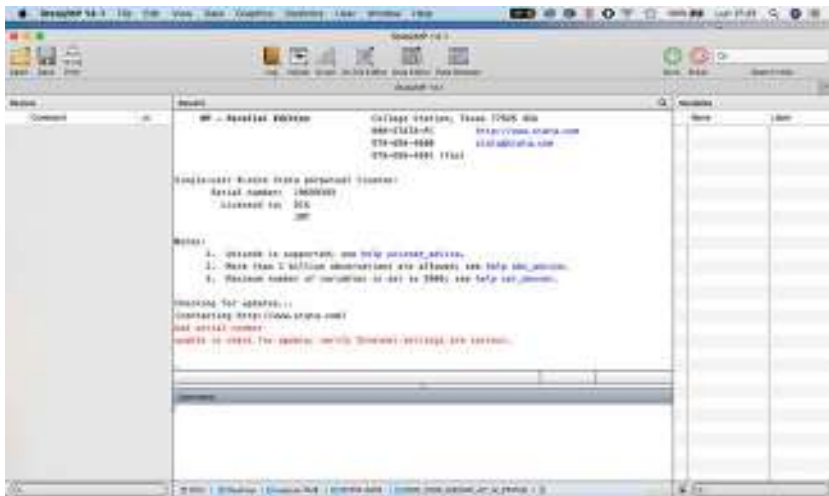
STATA es un paquete de software estadístico que ofrece una gran cantidad de funciones estadísticas y procedimientos de estimación econométrica. Iniciar una sesión de **STATA** se realiza oprimiendo doble en el icono instalado en su escritorio. Ver **imagen 2**.

**Imagen 2. Icono de STATA
instalado en el escritorio de un equipo Macintosh**



El **STATA** El entorno informático consta de cuatro ventanas principales. El tamaño y la forma de estos las ventanas pueden cambiarse y moverse en la pantalla. En su versión **STATA 14**. Ver **Imagen A.2**.

Imagen 2. Vista de STATA 14

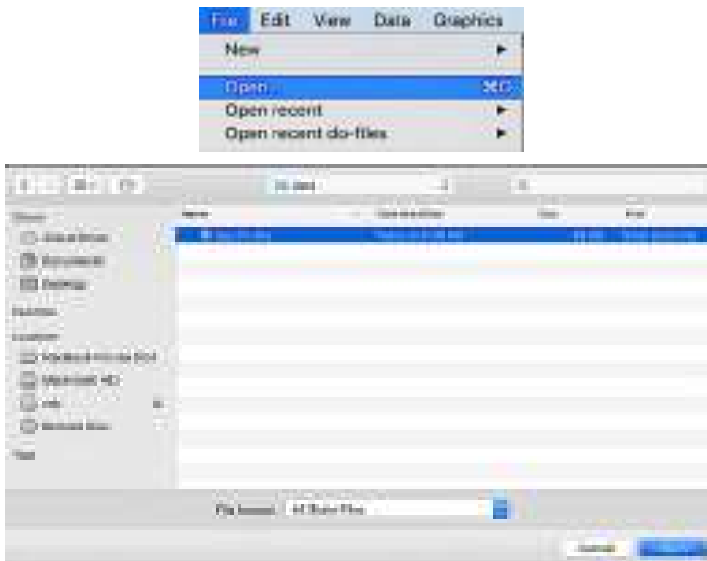


Fuente: STATA 14.1.

Además de estas ventanas, el entorno **STATA** tiene un *menú* y una *barra de herramientas* en la parte superior (para realizar operaciones **STATA**) y una *barra de estado del directorio* en la parte inferior (que muestra el directorio actual). Es posible, usar el *menú* y la *barra de herramientas* para emitir diferentes comandos **STATA** (como abrir y guardar archivos de datos), aunque la mayoría de las veces se usa la *ventana de comandos STATA* para ejecutar programas de aplicación. Si está creando un registro archivo (discutido con más detalle más adelante), el contenido se puede mostrar en la pantalla, lo cual es útil si desea regresar y ver resultados anteriores de la sesión actual. Aunque no es el propósito de dar una capacitación en **STATA**, sí es la intención de dar a conocer al lector de sus principales características y forma de usarse a través de una serie de ejercicios prácticos, para su mayor claridad.

Ejemplo 1. Abriendo el conjunto de datos (*Open data set*)

Se hace al ingresar el siguiente comando en la *ventana de comandos STATA 14.1*. File→Open→pei_01.dta





La línea que se observa, es producto de la instrucción de apertura de archivo y con la ausencia de un mensaje de error en una segunda línea, implica que el comando se ha ejecutado con éxito. De ahora en adelante, solo se mostrará la *ventana de resultados STATA*. El seguimiento los puntos deben tenerse en cuenta:

- **STATA** asume que el archivo está en su formato con una extensión **.dta**. Por lo tanto, escribiendo **pei_08** es lo mismo que escribir **pei_08.dta**.
- Solo se puede abrir un conjunto de datos a la vez en **STATA**. En este caso, por ejemplo, si abre otro conjunto de datos **pei_01.dta**, reemplazará **pei_08.dta** con **pei_01.dta**.
- Suponiendo que el archivo **pei_08.dta** no está en el actual directorio. Para hacer que cambie el uso de su directorio de:

use “/Users/DCA/Desktop/eval/data/pei_01.dta”

ahora a **Desktop**, teclee en *ventana de comando STATA*:

use “/Users/DCA/Desktop/pei_01.dta”

. use “/Users/DCA/Desktop/eval/data/pei_01.dta”



Con vista de la ventana de variables de **STATA**.

Variables	
Name	Label
id	ID
year	year of obser...
region	village ID
cluster	idcluster
ageceo	age of ceo
genceo	gender of ce...
educeo	education of...
smesize	sme size bas...
smefac	sme assets b...
incinnov	sme incinnov
expopn	sme opn exp...
expinnov	sme innovatio...
exptot	total expendit...
dmmfd	innovation ma...
dfmfd	innovation fe...
weight	innovation sa...
bdbaccess	sme broad ba...
pcirr	Proportion of...
raw1	raw1
raw2	raw2
raw3	raw3
raw4	raw4
raw5	raw5
raw6	raw6

Ejemplo 2. Guardando el conjunto de datos (*Saving data set*)

Se cuenta con las siguientes alternativas:

- Si realiza cambios en un archivo de datos **STATA** abierto y desea guardar esos cambios, puede hacerlo, usando el comando *save* de **STATA**. Por ejemplo, el siguiente comando guarda el archivo **pei_01.dta**:

```
. save pei_01, replace
```

```
. save pei_01, replace
file pei_01.dta saved
```



- La opción *replace* le indica a **STATA** que sobrescriba la versión original con la nueva versión. Si no quieres perder la versión original, tienes para especificar un nombre de archivo diferente en el comando *save*.

Ejemplo 3. Saliendo de STATA (*exiting STATA*)

Una manera fácil de salir de **STATA** es emitir el comando *exit*. Sin embargo, si tiene un conjunto de datos sin guardar abierto, **STATA** emitirá el siguiente mensaje de error:

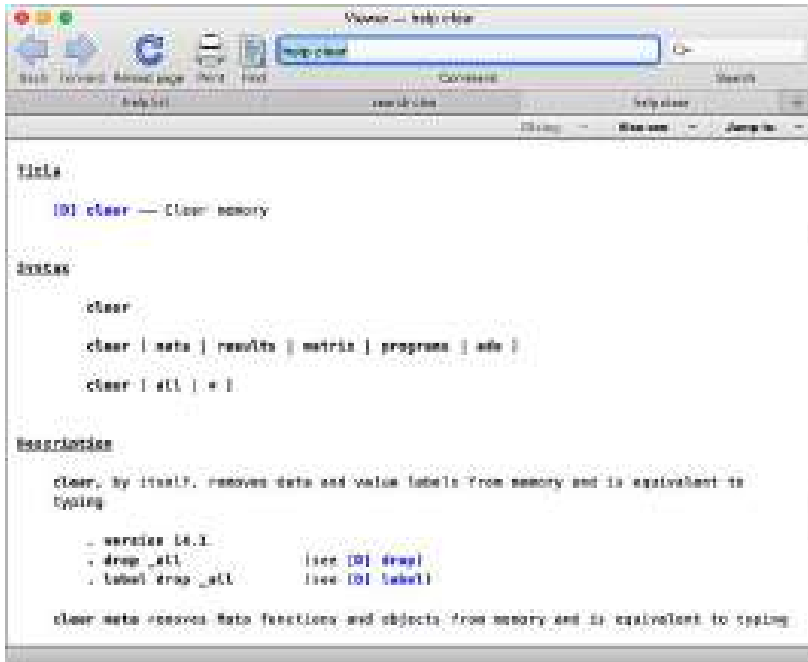
```
. exit
```



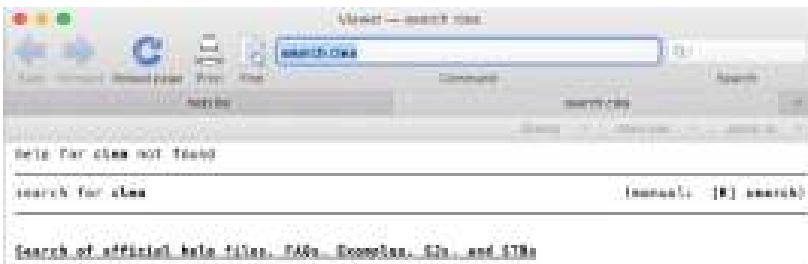
Ejemplo 4. Apoyo de STATA (*Help*)

STATA viene con un excelente, enorme y útil conjunto de manuales multivolumen. Con acceso a la Web, incluso dispone de un conjunto más grande de macros y otra información útil. Desde **STATA**, puede escribir el comando *help* y la palabra clave para solicitar información de ayuda sobre. Este comando, solo funciona si escribe el comando el nombre completo nombre o palabra clave sin abreviaturas. Por ejemplo:

. help clear



. help clea



En **STATA**, por cierto, un valor faltante se representa con un punto (.). Se considera un valor faltante más grande que cualquier número. El comando *summarize* ignora las observaciones con valores faltantes, y el comando *tabulate* hace lo mismo, a menos que se vea obligado a incluir los valores faltantes.

Ejemplo 5. Apoyo de STATA (*Lookup/Search*)

Si no puede recordar el nombre completo del comando o la palabra clave, o si no está seguro acerca de qué comando desea, puede usar el comando *lookup* o *search* seguido el nombre del comando o palabra clave, escribiéndose así:

```
. search mem
```



Este comando, enumerará todos los asociados con esta palabra clave y mostrará una breve descripción de cada uno de ellos. Entonces puede elegir el comando que cree que es relevante y utilizar la ayuda para obtener la referencia específica. El sitio web <https://www.stata.com/> tiene excelentes facilidades de ayuda, como un tutorial en línea y preguntas frecuentes (**FAQ**).

Ejemplo 6. Notas sobre los comandos STATA

Aquí hay algunos comentarios generales sobre los comandos **STATA**:

- Los comandos **STATA** se escriben en *minúsculas*.
- Todos los nombres, incluidos los comandos o nombres de variables, *se pueden abreviar siempre que no existe ambigüedad*. Por ejemplo: *describe*, *des* y simplemente *hacen el mismo trabajo porque no existe confusión*.
- Además de escribir, se pueden usar algunas teclas para representar algunos comandos **STATA** o secuencias. Los más importantes de ellos son las teclas *Page-Up* y *Page-Down*. Para mostrar el comando anterior en la *ventana de comando STATA* puede presionar la tecla *Page-Up*. Puedes seguir haciéndolo, hasta que el primer comando de la sesión aparece. Del mismo modo, la tecla *Page-Down* muestra el comando que sigue al que actualmente usa, en la misma *ventana de comandos STATA*.
- Al oprimir una vez en un comando, en la ventana de revisión (*review window*), lo colocará también, en el *ventana de comando STATA*. Al hacer oprimir doble, le indicará a **STATA** que ejecute el comando. Esto puede ser útil, cuando los comandos deben repetirse o editarse ligeramente en la *ventana de comando STATA*.

Ejemplo 7. Listando las variables (*Listing variables*)

Para ver todas las variables en el conjunto de datos, utilice el comando *describe* (completo o abreviado). Este comando proporciona información sobre el conjunto de datos y *enumera* todas las variables. Para ver solo una variable o lista de variables, utilice el comando de *describe* seguido del nombre o los nombres de las variables:

```
. desc id region
```

```
. desc id year region
```

variable name	storage type	display format	value label	variable label
id	long	%12.0g		HH ID
year	byte	%3.0g		Year of observation
region	byte	%8.0g		Village ID

Command

Como puede ver, el comando *describe* también muestra el *tipo* y la *longitud* de la variable, como una breve descripción de la variable (si está disponible). Deben tenerse en cuenta los siguientes puntos:

- *Puede abreviar una lista de variables* escribiendo el nombre tan solo de la primera y la última variable, separados por un guión (-); la ventana de variables (*variable window*) muestra el orden en que las variables se almacenan. Por ejemplo, para ver todas las variables desde *id* hasta *raw6* se escribe:

```
. desc id-raw6
```

variable name	storage type	display format	value label	variable label
id	long	%12.0g		HH ID
year	byte	%3.0g		Year of observation
region	byte	%8.0g		Village ID
cluster	byte	%8.0g		idcluster
agecoo	byte	%8.0g		Age of HH head: years
gender	byte	%8.0g		Gender of HH head: 1=H, 2=F
educat	byte	%8.0g		Education of HH head: years
agehh	byte	%8.0g		HH age
landarea	float	%9.0g		HH land area: decimale
lnclmuv	double	%11.0g		HH total asset: Tk.
exposho	float	%9.0g		HH per capita food expenditure: Tk/year
exphmuv	float	%9.0g		HH per capita nonfood expenditure: Tk/year
exptot	float	%9.0g		HH per capita total expenditure: Tk/year
maleid	byte	%8.0g		HH has male microcredit participant: 1=H, 2=F
female	byte	%8.0g		HH has female microcredit participant: 1=H, 2=F
weight	float	%9.0g		HH sampling weight

Command

- El símbolo de comodín (*) es útil para guardar algo de escritura. Por ejemplo, para ver todas las variables que comienzan con *sme*, se escribe:

. desc exp *

```
. desc smex
```

variable name	storage type	display format	value label	variable label
smesize	byte	%8.0g		MM size
smelend	float	%9.0g		MM lend asset: decimals

Command

- Puede abreviar una variable o lista de variables de esta manera, en cualquier comando (donde tenga sentido), no solo con *describe*.

Ejemplo 8. Listando los datos (*Listing data*)

Para ver los datos al momento, almacenados en las variables, utilice el comando *list* (abreviado como *l*). Si escribe el comando *list* por sí mismo, **STATA** mostrará valores para todas las variables y todas las observaciones, que pueden no ser deseables para ningún propósito práctico (así, se usa la combinación *Ctrl-Break*, para evitar que los datos se desplacen sin fin por la pantalla). Por lo general, se desea ver los datos de ciertas variables y de ciertas observaciones. Esto se logra escribiendo un comando *list* con una lista de variables y con condiciones. El siguiente comando enumera todas las variables de las primeras tres observaciones:

. list in 1/3

```
. list in 1/3
```

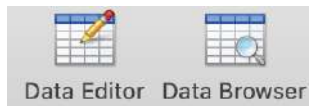
1.	id 11054	year 0	region 1	cluster 1	ageceo 72	genceo 1	educeo 0	smesize 3	smeland 36
	incinnov 33295	expopn 3055.856	expinnov 902.2549	exptot 3958.112	dmmfd 0	dfmfd 0	weight 1.042735	vaccess 1	pcirr .45
	raw1 11	raw2 8.5	raw3 12	raw4 8	raw5 2.5	raw6 55			

2.	id 11061	year 0	region 1	cluster 1	ageceo 35	genceo 1	educeo 5	smesize 10	smeland 116
	incinnov 180325	expopn 3031.017	expinnov 421.2493	exptot 3452.266	dmmfd 1	dfmfd 0	weight 1.011396	vaccess 1	pcirr .45
	raw1	raw2	raw3	raw4	raw5	raw6			

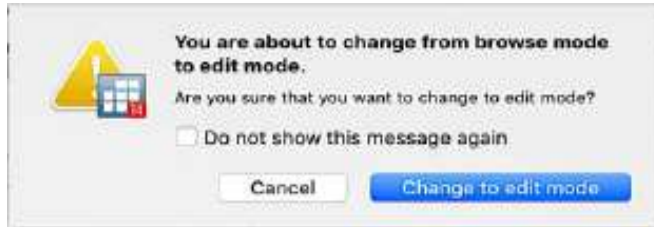
Command									
---------	--	--	--	--	--	--	--	--	--

Aquí, **STATA** muestra todos los registros que comienzan con la observación 1 y terminan con observación 3.

STATA, también puede mostrar los datos como una *hoja de cálculo*. Para hacerlo, use los dos íconos en la *barra de herramientas* llamada *Data Editor* y *Data Browser*). Oprimiendo una vez, hará que aparezca una nueva ventana donde los datos se mostrarán como una tabla, con los registros como filas y las variables como columnas. *Data Browser* solo mostrará los datos, mientras que para editarlos necesitará de *Data Editor*. Los comandos *Edit* y *Browser*, también abrirán la ventana de la *hoja de cálculo*.



	id	year	region	cluster	ageceo	genceo	educeo	smesize	smeland	incinnov	expopn	expinnov	exptot	dmmfd	dfmfd	weight	vaccess	pcirr
1	11054	0	1	1	72	1	0	3	36	33295	3055.856	902.2549	3958.112	0	0	1.042735	1	.45
2	11061	0	1	1	35	1	5	10	116	180325	3031.017	421.2493	3452.266	1	0	1.011396	1	.45
3	11062	0	1	1	54	1	6	50	50	98725	3347.578	343.3877	343.3877	1	0	1.011396	1	.45
4	11063	0	1	1	40	1	7	8	8	10750	3186.833	265.8044	265.8044	1	0	1.011396	1	.45
5	12012	0	2	1	38	1	8	10	10	10750	3405.124	308.987	308.987	1	0	1.011396	1	.45



Ejemplo 9. Listando datos específicos

El siguiente comando enumera el tamaño de una pyme de base tecnológica y la educación del ceo de género femenino menor de 45 años:

```
. list smesize educeo if (gencoe==0 & ageceo<45)
```

La declaración anterior usa dos *operadores relacionales* (`==` y `<`) y un operador lógico (`&`). Los *operadores relacionales* imponen una condición en una variable, mientras que los *operadores lógicos* combinar dos o más operadores relacionales. Abajo se muestra la relación y operadores lógicos utilizados en **STATA**.

Relational operators	Logical operators
> (greater than)	! (not)
< (less than)	(or)
== (equal)	& (and)
>= (greater than or equal)	
<= (less than or equal)	
= or != (not equal)	

Puede usar operadores relacionales y lógicos en cualquier comando Stata (donde haga sentido), no solo en el comando *list*.

```
. list smesize educeo if (gencoe==0 & ageceo<45)
```

	smesize	educeo
35.	4	0
44.	7	2
54.	5	0
68.	2	0
99.	6	0
108.	2	0
135.	4	0
170.	2	0
180.	6	0
228.	3	0
229.	3	0
234.	5	0
274.	5	8
389.	3	0

Command

Ejemplo 10. Sumando datos (*Summarizing data*)

El comando *summarize* (el cual puede abreviarse como *sum*) calcula y muestra algunas sumas de estadísticas, que incluyen *medias* y *desviaciones estándar*. Sino se especifica alguna variable, este comando aplicará para todas las variables incluidas en el conjunto de datos.

Como ejemplo, se muestra la suma de las variables tamaño de la pyme de base tecnológica y la educación del ceo, como sigue:

```
. sum smesize educeo
```

```
. sum smesize educeo [aw=weight]
```

Variable	Obs	Weight	Mean	Std. Dev.	Min	Max
smesize	826	822.32837	5.728505	2.567759	1	19
educeo	826	822.32837	2.657877	3.615297	0	16

Command

STATA excluye cualquier observación que tenga un valor faltante para las variables que se *suman* a partir de este cálculo (los valores faltantes se analizan más adelante). Si se requiere saber la mediana y los percentiles de una variable, agregue la opción *detail* (abreviado *d*), por ejemplo:

```
. sum smesize educeo, d
```

```
. sum smesize educeo, d
```

HH size				
Percentiles		Smallest		
1%	2	1		
5%	2	1		
10%	3	1	Obs	826
25%	4	1	Sum of Wgt.	826
50%	5		Mean	5.470944
		Largest	Std. Dev.	2.378643
75%	7	15		
90%	8	17	Variance	5.657943
95%	10	17	Skewness	1.308173
99%	14	19	Kurtosis	6.355481
Education of HH head: years				
—more—				
Command				

Ejemplo 11. Ponderaciones (*Weight*)

Una gran fortaleza de **STATA** es que permite el uso de ponderaciones (*weights*). La opción de *weight* es útil, si la probabilidad de muestreo de una observación es diferente de la de otra. Por ejemplo, en nuestro caso es de suponer que en las pymes de base tecnológica (**pbt**), el marco de muestreo estaría estratificado, donde el primer muestreo primario las

unidades (a menudo regiones) se muestrean y dependen de la selección de las unidades primarias, de las que se extraen unidades de muestreo secundarias (a menudo **pbt** de cierto tamaño). Las encuestas, generalmente deben aplicar ponderaciones para corregir las diferencias de diseño de muestreo y, a veces problemas de recolección de datos. La implementación en **STATA** es sencilla:

```
. sum smesize educeo [aw=weight]
```

. sum smesize educeo					
Variable	Obs	Mean	Std. Dev.	Min	Max
smesize	826	5.478944	2.378643	1	19
educeo	826	2.405569	3.333782	0	16

Aquí, la variable *weight* tiene la información sobre el peso que se debe dar a cada observación y *aw* es una opción de **STATA** para incorporar el peso en el cálculo. El uso de ponderaciones se analiza más adelante en ejercicios de capítulos posteriores. Para las variables que son cadenas (*strings*), el comando *summarize* no aplica para producir datos de estadística descriptiva, excepto que el número de observaciones es cero. También para *variables categóricas* (por ejemplo, educación secundaria = 1, educación superior = 2, etc.), interpretar el resultado del comando *summarize* puede ser difícil. En ambos casos, una tabulación completa puede ser más significativa, lo cual se discute enseguida.

En la mayoría del levantamiento de encuestas, las observaciones se seleccionan mediante un proceso aleatorio y puede tener diferentes probabilidades de selección. Por lo tanto, se deben usar *ponderaciones* que son igual a la inversa de la probabilidad de ser muestreada. Una ponderación de *wn* para el *enésima* observación significa, en términos generales, que

la observación n representa elementos w_j en la población de la cual se extrajo la muestra. Omitir las ponderaciones de muestreo en el análisis, generalmente proporciona estimaciones sesgadas, que pueden estar lejos de los valores verdaderos.

Por lo general, son necesarios varios ajustes a las ponderaciones, son necesarios posteriores al muestreo. Por ejemplo, la ponderación de muestreo de las pymes de base tecnológica (**pbt**) de la **pei_01.dta** es la ponderación correcta a usar, cuando se utiliza para sumarizar los datos relacionados.

STATA tiene cuatro tipos de ponderaciones:

- *Ponderación por frecuencia* (frequency weights. *fweight*), que indica cuántas observaciones en la población están representados por cada una de la muestra y deben tomar valores enteros.
- *Ponderación analítica* (analytic weights. *aweight*) son especialmente apropiados, cuando se trabaja con datos que contienen promedios (por ejemplo, ingreso promedio per cápita en una pyme de base tecnológica). La variable de ponderación, es proporcional al número de pymes sobre las cuales se calculó el promedio (por ejemplo, número de miembros de una pyme). Técnicamente, *la ponderación analítica están en proporción inversa a la varianza de una observación* (es decir, una mayor ponderación significa que la observación se basó en más información y, por lo tanto, es más confiable en el sentido de tener menos varianza).
- Las *ponderaciones de muestreo* (sampling weights. *pweight*) son lo inverso de la probabilidad de selección, debido al diseño de la muestra.
- Las *ponderaciones de importancia* (importance weights. *iweight*) que indican la importancia relativa de la observación.

Los más comúnmente usados son `pweight` y `aweight`. Más información sobre las ponderaciones, se pueden obtener escribiendo:

```
. help weight
```



Los siguientes comandos muestran más sobre la aplicación de ponderaciones:

```
. tabstat smesize [aweight=weight], statistics(mean sd) by(dfmfd)
. table dfmfd [aweight=weight], contents(mean smesize sd smesize)
```

```
. tabstat smesize [aweight=weight], statistics(mean sd) by(dfmfd)
```

Summary for variables: smesize
by categories of: dfmfd (innovation female microcredit participant: 1=Y, 0=N)

dfmfd	mean	sd
0	5.816998	2.672921
1	5.358539	2.036723
Total	5.728505	2.567758

```
. table dfmfd [aweight=weight], contents(mean smesize sd smesize)
```

innovation female microcredit participant: 1=Y, 0=N	mean(smesize)	sd(smesize)
0	5.817	2.672921
1	5.35854	2.036723

Command

Ejemplo 12. Ordenamiento con ponderación (*Sort / weight*)

A menudo, se requieren ver estadísticas sumariadas por grupos de ciertas variables, no solo para todo el conjunto de datos. Supongamos que quiere ver el tamaño de la pyme de base tecnológica y la educación del CEO para *participantes* y *no participantes*. Así:

- Primero, ordene los datos por variable de grupo (en este caso, *dfmfd*).
- Puede verificar el ordenamiento con el comando *describe*, después de abrir cada archivo.
- El comando *describe*, después de enumerar todas las variables, indica si el conjunto de datos está ordenado por alguna variable.

- Si no hay información del ordenamiento, o el conjunto de datos se ordena por una variable que es diferente de la que desea, puede utilizar el comando *sort* y posteriormente, guardar el conjunto de datos en este formulario.
- Los siguientes comandos ordenan el conjunto de datos por la variable *dfmfd* y muestran la sumarización estadística del tamaño de la yme con base tecnológica y la educación CEO para *participantes y no participantes*:

```
. sort dfmfd
```

```
. by dfmfd: sum smesize educeo [aw=weight]
```

```
. sort dfmfd
```

```
. by dfmfd: sum smesize educeo [aw=weight]
```

```
=> dfmfd = 0
```

Variable	Obs	Weight	Mean	Std. Dev.	Min	Max
smesize	539	663.598263	5.816999	2.672921	1	15
educeo	539	663.598263	2.85233	3.771715	0	16

```
=> dfmfd = 1
```

Variable	Obs	Weight	Mean	Std. Dev.	Min	Max
smesize	287	158.730107	5.358539	2.836723	1	14
educeo	287	158.730107	1.84493	2.734086	0	12

Command

Ejemplo 13. Uso del comando *tabstat*

Una alternativa útil al comando *summary* es el comando *tabstat*, que le permite especificar la lista de estadísticas que desea mostrar en una sola tabla. Puede estar condicionado por otra variable. El siguiente comando muestra la media y desviación estándar del tamaño de pyme de base tecnológica y educación del CEO por la variable *dfmfd*:

```
. tabstat smesize educeo, statistics(mean sd) by(dfmfd)
```

```
. tabstat smesize educeo, statistics(mean sd) by(dfmfd)
```

```
Summary statistics: mean, sd  
by categories of: dfmfd (HH has female microcredit participant: 1=Y, 0=N)
```

dfmfd	smesize	educeo
0	3.560297 2.582033	2.669759 3.559692
1	3.303136 1.932992	1.909406 2.800877
Total	3.470944 2.378643	2.405569 3.333762

Command

Ejemplo 14. Distribuciones de frecuencia (*Tabulations*)

Las distribuciones de frecuencia y las tabulaciones cruzadas (*cross-tabulations*) a menudo son necesarias. El comando *tabulate* (abreviado *tab*) se realiza, como sigue:

```
. tab dfmfd
```

```
. tab dfmfd
```

innovation female microcredit participant : 1=Y, 0=N	Freq.	Percent	Cum.
0	539	65.25	65.25
1	287	34.75	100.00
Total	826	100.00	

Command

El siguiente comando, proporciona la distribución de género de los CEO de las pymes de base tecnológica (**pbt**) participantes:

```
. tab genceo if dfmfd==1
```

```
. tab genceo if dfmfd==1
```

gender of ceo: 1=M, 0=F	Freq.	Percent	Cum.
0	23	8.01	8.01
1	264	91.99	100.00
Total	287	100.00	

Command

Tenga en cuenta aquí, el uso del signo = = . Esto indica que si la variable es igual a uno, entonces se realiza la tabulación. El comando *tabulate* también puede ser utilizado para mostrar una distribución bidireccional (*two-way distribution*). Por ejemplo, si desea verificar si existe algún *sesgo* de género en la educación de los CEO de las pymes de base tecnológica (**pbt**), tiene el comando:

```
. tab educeo genceo
```

```
. tab educeo genceo
```

education of ceo: years	gender of ceo: 1=M, 0=F		Total
	0	1	
0	32	414	446
1	0	27	27
2	1	56	57
3	2	39	41
4	1	43	44
5	1	77	78
6	0	23	23
7	0	21	21
8	1	26	27
9	0	15	15
10	0	26	26
11	0	2	2
12	0	14	14

Command

Para ver porcentajes por fila o columnas, agregue opciones al comando *tabulate*:

. tab dfmfd genceo, col row

. tab dfmfd genceo, col row

Key
<i>frequency</i>
<i>row percentage</i>
<i>column percentage</i>

innovation female microcredi t participan t: 1=Y, 0=N	gender of ceo: 1=M, 0=F		Total
	0	1	
0	15 2.78 39.47	524 97.22 66.50	539 100.00 65.25
1	23 8.01 60.53	264 91.99 33.50	287 100.00 34.75
Total	38 4.60 100.00	788 95.40 100.00	826 100.00 100.00

.
Command

Ejemplo 15. Estadística descriptiva (Table Command)

Otro comando muy conveniente es *table*, que combina características de los comandos *sum* y *tab*. Además, muestra los resultados en una forma más presentable. El siguiente comando *table*, muestra la *media* del tamaño de la pyme de base tecnológica y de la educación del CEO, por su participación en programa de microfinanzas:

```
.table dfmfd, c (mean smesize mean educeo)
```

```
. table dfmfd, c (mean smesize mean educeo)
```

innovatio n female microcred it participa nt: 1=Y, 0=N	mean(smesize)	mean(educ
0	5.5603	3
1	5.30314	2

```
.
```

Command

Ejemplo 16. Despliegue de punto decimal

Los resultados son los esperados. Pero, ¿por qué la media de educeo se muestra como entero y no una fracción? Esto ocurre porque la variable educeo se almacena como un número entero, y **STATA** simplemente trunca los números después del decimal. Observe la descripción de la variable:

```
.d educeo
```

```
. d educeo
```

variable name	storage type	display format	value label	variable label
educ	byte	%2.0f		education of ceo: years

```
.
```

Command

Tenga en cuenta que `educeo`, es una variable con punto flotante: su formato (`% 2.0f`) muestra que sus dígitos ocupan dos lugares y no tiene ningún dígito después del decimal. Es posible forzar cambiar dicho formato; suponga que desea que muestre dos lugares después del decimal para una visualización de tres dígitos, el cual se realiza como sigue:

```
. format educeo %3.2f
. table dfmfd, c (mean smesize mean educeo)
```

innovatio n female microcred it participa nt: 1=Y, 0=N	mean(smesize) mean(educeo)	
0	5.5603	2.67
1	5.30314	1.91

```
.
.
```

Command

El formato solo cambia la visualización de la variable, no su representación interna de la variable en la memoria. El comando `table`, puede mostrar hasta cinco estadísticas de las variables como de la media, la suma, el mínimo o máximo. Se pueden mostrar tablas bidimensionales, tridimensionales o incluso de dimensiones superiores. Por ejemplo, mostramos una tabla bidimensional que desglosa la educación del CEO, no solo por región, sino también género del mismo.

```
. table dfmfd genceo, c (mean smesize mean educeo)
```

```
. table dfmfd genceo, c (mean smesize mean educeo)
```

innovation female microcredit participation: 1=Y, 0=N	gender of ceo:	
	1=M, 0=F	
	0	1
0	3.13333 0.53	5.62977 2.73
1	4.6087 0.74	5.36364 2.01

```
.
```

```
Command
```

Ejemplo 17. Comando contar (*Count*)

El comando *count*, se utiliza para contar el número de observaciones en el conjunto de datos:

```
. count
```

```
. count
826
```

```
.
```

```
Command
```

El comando *count*, se puede usar con condiciones. Por ejemplo, el siguiente comando proporciona el número de pymes de base tecnológica (*pbt*) cuyos CEO tienen una edad de más de más de 50 años:

```
. count if ageceo>50
```

```
. count if ageceo>50
181
```

```
.
```

```
Command
```

Hasta ahora, la discusión se ha limitado a los comandos **STATA** que muestran información en el datos de diferentes maneras *sin cambiar los datos*. En realidad, las sesiones más frecuentes de su uso, implica hacer cambios en los datos (por ejemplo, crear nuevas variables o cambiar valores de variables existentes). Los siguientes ejercicios demuestran cómo esos cambios se puede incorporar en **STATA**.

Ejemplo 18. Generando nuevas variables

En **STATA**, el comando *genérate* (abreviado *gen*) crea nuevas variables, mientras que el comando *replace* cambia los valores de una variable existente. Los siguientes comandos crean una nueva variable llamada *oldceo* y posteriormente, establecen su valor a una de las pymes con CEO de edad de más 50 años y a cero de lo contrario:

```
. gen oldceo=1 if ageceo>50
```

```
. gen oldceo=1 if ageceo>50
(645 missing values generated)
```

Command	
oldceo	1
	.
	1
	.
	.
	1

```
. replace oldceo=0 if ageceo<=50
```

```
. replace oldceo=0 if ageceo<=50
(645 real changes made)
```

Command

oldceo
1
0
1
0
0
0
1

Así, para cada observación, el comando *gen* verifica la condición (por ejemplo, si el CEO de la pyme de base tecnológica, tiene más de 50 años) y establece el valor de la variable *oldceo*, uno para esa observación, si la condición es verdadera o a valor perdido, de lo contrario. El comando *replace* funciona de manera similar. Después del comando *generate*, **STATA** indica que 645 observaciones *no cumplieron con la condición*, y después del comando *replace* **STATA** indica que esas 645 observaciones tienen nuevos valores (cero en este caso). Vale la pena señalar los siguientes puntos:

- Si se emite un comando *gen* o *replace* sin ninguna condición, ese comando se aplica a todo el archivo de datos.
- Al usar el comando *gen*, se debe tener cuidado de manejar los valores faltantes correctamente.
- El lado derecho del signo = en los comandos *gen* o *replace* puede ser cualquier expresión que involucre nombres de variables, no solo un valor. Así, por ejemplo, el comando *gen young = (ageceo <= 32)* crearía una variable llamada *young* que tomaría el valor de


```
. egen avgagemf=mean (ageceo), by (genceo)
```

```
. egen avgagemf=mean (ageceo), by (genceo)
```

Command		
oldceo	avgage	avgagemf
1	40.76634	40.70304
0	40.76634	40.70304
1	40.76634	40.70304
0	40.76634	40.70304
0	40.76634	40.70304
0	40.76634	40.70304
1	40.76634	40.70304

Ejemplo 19. Etiquetando variables (*Labeling variables*)

Puede etiquetar a las variables para darles una descripción. Por ejemplo, la variable *oldceo* no tiene ninguna etiqueta ahora. Puede adjuntar una etiqueta a esta variable escribiendo:

```
. label var oldceo "CEO is over 50: 1=Y, 0=N"
```

Command		row5	row5
label var oldceo "CEO is over 50: 1=Y, 0=N"		oldceo	row5
		avgage	
		avgagemf	
Command		row1	row3
		row1	row1
		row2	row2
		row3	row3
		oldceo	CEO is over 50: 1=Y, 0=N
		avgage	
		avgagemf	

Ahora para ver lo nueva etiqueta, escriba lo siguiente:

```
.des oldceo
```



```
. des oldcon
```

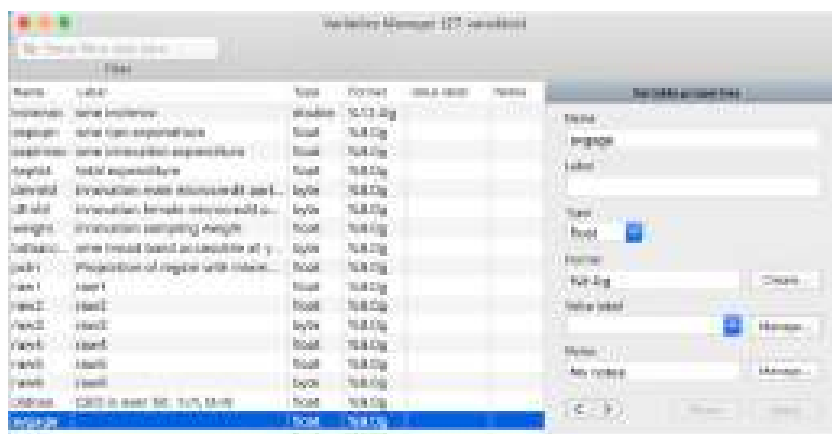
variable name	storage type	display format	value label	variable label
oldcon	float	%9.0g		CEO is over 50: 1=Y, 0=N

Command

Ejemplo 20. Etiquetando variables (*Labeling variables*) desde barra de herramientas

También, es posible corregir desde la barra de herramientas oprimiendo:

Data→Variables Manager:

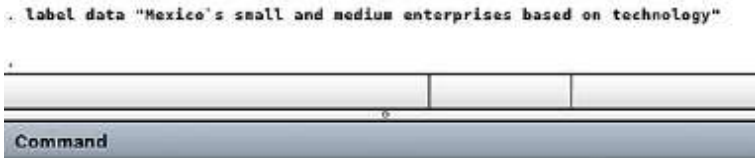


Hacer las modificaciones necesarias y oprimir el botón Apply para finalizar.

Ejemplo 21. Etiquetando datos (*Labeling data*)

Se pueden crear otros tipos de etiquetas. Para adjuntar una etiqueta a todo el conjunto de datos, teclee:

```
. label data "Mexico`s small and medium enterprises based on  
technology"
```



Para ver la etiqueta, teclee:

. des

. des

Contains data from /Users/DCA/Desktop/eval/data/pei_01.dta				
obs:	826	Mexico's small and medium enterprises based on technology		
vars:	27	14 Jun 2020 19:00		
size:	62,776			
variable name	storage type	display format	value label	variable label
id	long	%12.0g		ID
year	byte	%8.0g		year of observation
region	byte	%8.0g		village ID
cluster	byte	%8.0g		idcluster
ageceo	byte	%8.0g		age of ceo
genceo	byte	%8.0g		gender of ceo: 1=M, 0=F
educeo	byte	%8.0g		education of ceo: years
sme size	byte	%8.0g		sme size based on personel
sme fac	float	%9.0g		sme assets based on facilities
incinnov	double	%12.0g		sme incinnov
expopn	float	%9.0g		sme opn expenditure
expinnov	float	%9.0g		sme innovation expenditure
exptot	float	%9.0g		total expenditure
dmmfd	byte	%8.0g		innovation male microcredit participant: 1=Y, 0=N
dfmfd	byte	%8.0g		innovation female microcredit participant: 1=Y, 0=N
weight	float	%9.0g		innovation sampling weight
—more—				
n				
Command				

Observe que el conjunto de datos de **pei_01.dta** tiene el nombre de “Mexico’s small and medium enterprises based on technology”, el cual permite asignar nombres al conjunto de datos.

Ejemplo 22. Etiquetando valores de variables (*Labeling values of variables*)

Las variables categóricas, como *genceo* (1 = hombre, 0 = mujer), pueden tener etiquetas que ayudan a recordar cuáles son las categorías. Por ejemplo, usando **pei_08.dta**, tabular la variable *genceo* muestra solo valores 0 y 1:

.tab genceo

```
. tab genceo
```

gender of ceo: 1=M, 0=F	Freq.	Percent	Cum.
0	104	9.21	9.21
1	1,025	90.79	100.00
Total	1,129	100.00	

Command	
---------	--

Para adjuntar etiquetas a los valores de una variable, y volviendo a **pei_01.dat**, se deben hacer dos cosas:

- Primero, definir una etiqueta de valor.
- Posteriormente, asigne esta etiqueta a la variable. Usando las nuevas categorías para *genceo*, teclee:

. label define genlabel 0 "Female" 1 "Male"

```
. label define genlabel 0 "Female" 1 "Male"
```

Command	
---------	--

. label values genceo genlabel

```
. label values genceo genlabel
```

Command	
---------	--

Ahora, para ver las etiquetas, escriba:

```
. tab genceo
```

```
. tab genceo
```

gender of ceo: 1=M, 0=F	Freq.	Percent	Cum.
"Female"	38	4.60	4.60
"Male"	788	95.40	100.00
Total	826	100.00	

Command	
---------	--

	id	year	region	cluster	ageceo	genceo
1	11054	0	1	1	72	"Male"
2	11061	0	1	1	35	"Male"
3	11081	0	1	1	54	"Male"
4	11101	0	1	1	44	"Male"
5	12021	0	2	1	28	"Male"
6	12035	0	2	1	25	"Male"

Si desea ver los valores reales de la variable genceo, que siguen siendo ceros y unos, puede agregar una opción para no mostrar las etiquetas asignadas a los valores de variable. Por ejemplo, teclee:

```
. tab genceo, nolabel
```

```
. tab genceo, nolabel
```

gender of ceo: 1=M, 0=F	Freq.	Percent	Cum.
0	38	4.60	4.60
1	788	95.40	100.00
Total	826	100.00	

Command	
---------	--

Ejemplo 23. Soltar variables así como observaciones

Las variables y observaciones de un conjunto de datos, se pueden seleccionar mediante el uso de los comandos *keep* o *drop*. Suponga que tiene de la base de datos .dta seis variables: var1, var2,..., var6 y desea mantener un archivo con solo tres de ellos (por ejemplo, var1, var2 y var3). Así, se es posible usar cualquiera de los siguientes dos comandos:

- “keep var1 var2 var3” (or “keep var1-var3” si es este el orden deseado).
- “drop var4 var5 var6” (or “drop var4-var6” si es este el orden deseado).

Tenga en cuenta el uso de un guión (-) en ambos comandos. Es una buena práctica, usar el comando que implica el menor tecleo posible de las variables (y, por lo tanto, tener menos riesgo de error). También es posible usar operadores relacionales o lógicos. Por ejemplo, el siguiente comando muestra todos los CEOs de pymes de base tecnológica (**pbt**) con edad de 80 años o más:

```
drop if ageceo>=80
```

```
. drop if ageceo>=80
(0 observations deleted)
```



Y este comando, por ejemplo, mantiene esas observaciones donde el tamaño de la pyme de base tecnológica, es de seis o menos miembros:

```
keep if smesize <=6
```

```
. keep if smesize <=6
(221 observations deleted)
```

```
.
```

Command

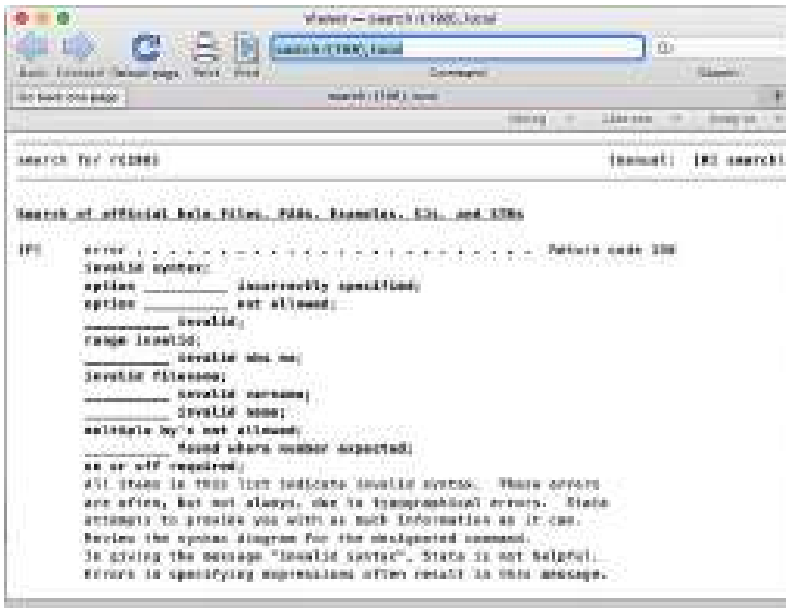
Los dos comandos anteriores eliminan o mantienen todas las variables según las condiciones. Así, no es posible incluir una lista de variables en un comando *drop* o *keep* que también utiliza condiciones. Por ejemplo, el siguiente comando provocará errores:

keep id smesize if smesize<=6

```
. keep id smesize if smesize<=6
invalid syntax
r(198);
```

```
.
```

Command



Para lograrlo, se utilizan dos comandos:

keep if smesize<=6

```
. keep if smesize<=6
(0 observations deleted)
```

.

Command

keep id smesize

```
. keep id smesize
```

.

Command

Variables	
Name	Label
id	ID
smesize	sme size based on personel

También, es posible usar la palabra clave del comando *drop* o *keep*. Por ejemplo, para soltar (*drop*) las primeras 20 observaciones:

drop in 1/20

```
. drop in 1/20
(20 observations deleted)
```

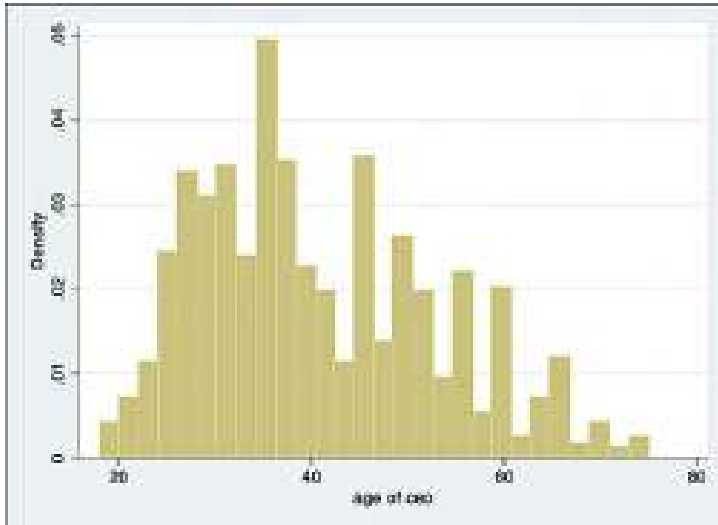
.

Command

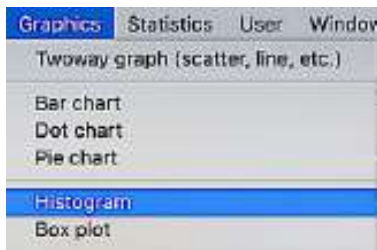
Ejemplo 24. Produciendo gráficos

STATA es bastante eficiente en la producción de gráficos básicos, aunque requerirá de considerable experimentación, para producirlos más atractivos. El siguiente comando muestra la distribución de la edad del ceo de las pymes de base tecnológica (**pbt**) en un gráfico de barras (histograma):

histogram ageceo



En muchos casos, la forma más fácil de producir gráficos es usando los menús; en esto caso, oprima Gráficos y posteriormente, en Histograma y siga las indicaciones.



histogram - Histograms for continuous and categorical variables

Menu: Main | Bin | Weights | Density plots | Add plots | Y axis | X axis | Titles | Legend | Overall | By

Data:

Variable: ☒ Data are continuous ☐ Data are discrete

Bin:

☐ 10 Number of bins

☐ Width of bins

☐ Lower limit of first bin

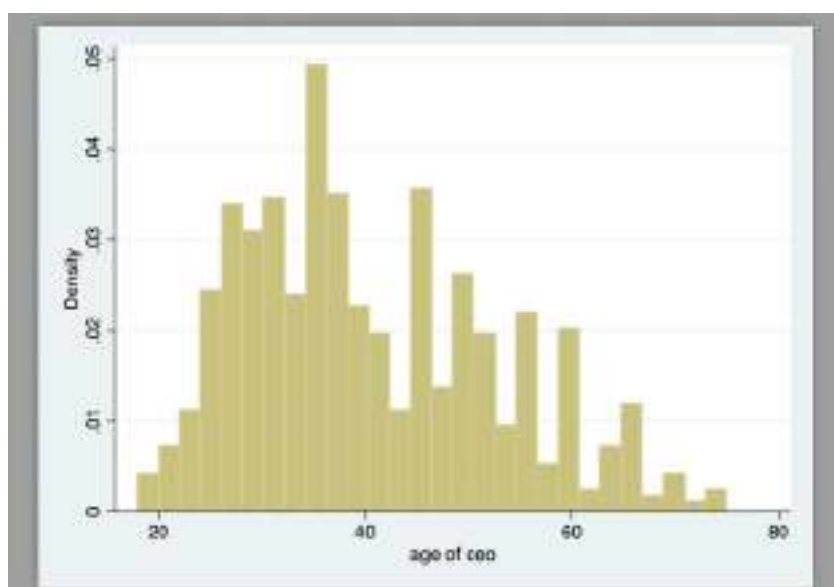
Y axis:

☒ Density ☐ Fraction ☐ Frequency ☐ Percent

☐ Add height labels to bars

☐ Recalculate bin sizes when by() is specified

Buttons: ? | [] | [] | Submit | Cancel | OK

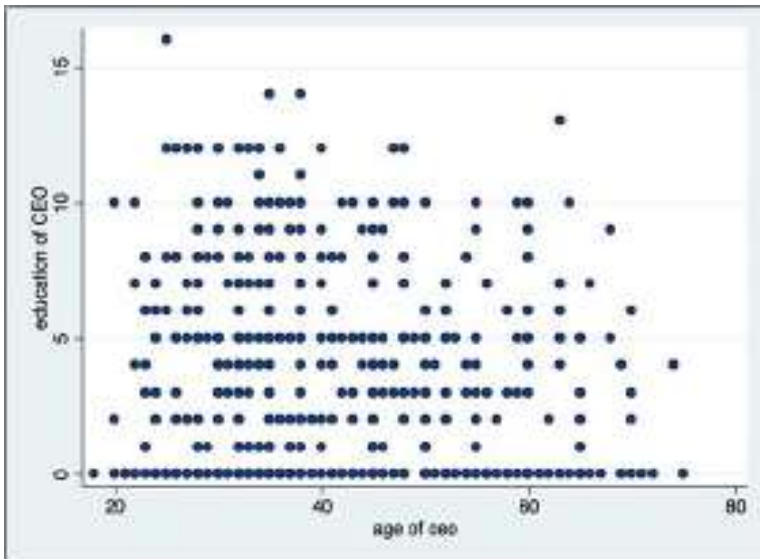


Una manera fácil de guardar un gráfico es hacer clic derecho sobre él y copiarlo para pegarlo en un Microsoft Word o Documento de Excel.

Aquí hay un comando para un diagrama de dispersión de dos variables:

```
twoway (scatter educeo ageceo), ytitle(education of CEO)
xtitle(age of CEO) title(education by age)
```

Command
<pre>twoway (scatter educeo ageceo), ytitle(education of CEO) xtitle(age of CEO) title(education by age)</pre>



Ejemplo 25. Combinando archivos de datos

La combinación de archivos de datos, se realiza cuando se necesita usar variables que se distribuyen en dos o más archivos. Como ejemplo de combinación, se tiene el archivo el **pei_08.dta**, el cual, se dividirá en dos conjuntos de datos, de tal manera que uno contiene una o más variables que la otra no y entonces y entonces los conjuntos de datos se combinarán para recuperar el **pei_08.dta** original. Así, se requiere abrir el archivo **pei_08.dta**, descartar las variables de participación del programa y guarde el archivo de datos como **pei_98_1.dta**, con los siguientes comandos:

```
use "/Users/DCA/Desktop/eval/data/pei_08.dta", clear
drop dmmfd dfmfd
save pei_98_1.dta, replace
```

```
. use "/Users/DCA/Desktop/eval/data/pei_08.dta"

. drop dmmfd dfmfd

. save pei_98_1.dta, replace
(note: file pei_98_1.dta not found)
file pei_98_1.dta saved
```

Command

StataMP 14.1 — pei_98_1.dta

Variables		
	Name	Label
	id	ID
	year	year of observation
See all data	region	id region
	cluster	id cluster
N	ageceo	age of ceo: years
	gender	gender of ceo: 1=M, 0=F
N of obs	educceo	education of ceo: years
	smesize	sme size based on personnel
	smeasoc	sme assets based on facilities
	smcinnov	sme innovation
	expopen	sme open expenditure
	expinnov	sme innovation expenditure
	exptot	total expenditure
	weight	innovation sampling weight
	broadband	sme broadband accessible all year: 1=Y, 0=N
	prbr	proportion of region with internet broadband
	rsw1	rsw1
	rsw2	rsw2
	rsw3	rsw3
	rsw4	rsw4
	rsw5	rsw5
	rsw6	rsw6

Si desea asignarle un nuevo nombre a este archivo (**pei_98_1.dta**) porque no desea cambiar el **pei_08.dta** original de forma permanente. Ahora abra el **pei_08.dta** nuevamente. Esta vez, mantenga solo las variables de participación. Guarde este archivo como **pei_98_2.dta**.

```
use "/Users/DCA/Desktop/eval/data/pei_08.dta", clear
keep id dmmfd dfmfd
save pei_98_2.dta, replace
```

```
. use "/Users/DCA/Desktop/eval/data/pei_08.dta", clear

. keep id dmmfd dfmfd

. save pei_98_2.dta, replace
file pei_98_2.dta saved
```

.

Command	

Name	Label
id	id
clearid	innovation male microcredit participant: 1=V, 0=N
clearid	innovation female microcredit participant: 1=V, 0=N

Como se observa, se mantiene la *id* de cada registro pyme de base tecnológica, además del resto de los campos. Esto es necesario, porque la combinación requiere al menos una variable de identificación común entre los dos archivos que se van a fusionar. Aquí *id* es esa variable común, entre los dos archivos. Ahora se tienen dos archivos de datos: uno contiene las variables de participación (**pei_98_2.dta**), y el otro no las tiene (**pei_98_1.dta**). Si necesita usar las variables de ambos archivos, tendrá que fusionar los dos archivos. Sin embargo, antes de fusionarlos, necesita asegurarse de que ambos archivos estén ordenados por la variable de identificación. Esto puede hacerse de la siguiente forma:

```
use "/Users/DCA/Desktop/eval/data/pei_98_1.dta", clear
sort id
save,replace
use pei_98_2, clear
sort id
save,replace
```

```
. use "/Users/DCA/Desktop/eval/data/pei_98_1.dta", clear
. sort id
. save,replace
file /Users/DCA/Desktop/eval/data/pei_98_1.dta saved
. use pei_98_2, clear
. sort id
. save,replace
file pei_98_2.dta saved
.
```

Command	



Name	Label
id	ID
gender	interview male microcovid participant: 1=Y, 0=N
gender	interview female microcovid participant: 1=Y, 0=N

Ahora, se está listo para fusionar los dos archivos. Uno de los archivos, tiene que estar abierto (no importa cual archivo). Abra el archivo **pei_98_1.dta** y posteriormente, combine el archivo **pei_98_2.dta** con él, de la siguiente manera:

```
use pei_98_1, clear
merge 1:1 id using pei_98_2
```

```
. use pei_98_1, clear
. merge 1:1 id using pei_98_2
```

Result	# of obs.
not matched	0
matched	1,129 (_merge==3)

4

Command

En este contexto, **pei_98_1.dta** se denomina *archivo maestro (master file)*, es el archivo que permanece en la memoria, antes de la fusión) y **pei_98_2.dta** se llama el archivo de uso (*using file*). Para ver cómo se hizo la operación de fusión, escriba el siguiente comando:

342


```
tab _merge
```

_merge	Freq.	Percent	Cum.
matched (3)	1,129	100.00	100.00
Total	1,129	100.00	

Command		
---------	--	--

STATA crea una nueva variable `_merge` durante la operación de fusión. Una operación de tabulación a esta variable, muestra los diferentes valores de `_merge` y, por lo tanto, el estado de la operación de fusión. Aunque en este caso, `_merge` solo tiene un valor (3), puede tener hasta tres valores posibles, dependiendo de la naturaleza de la operación de fusión:

- Un valor de 1 muestra el número de observaciones que provienen únicamente del archivo maestro.
- Un valor de 2 muestra el número de observaciones que provienen únicamente del archivo que usa.
- Un valor de 3 muestra el número de observaciones comunes en ambos archivos.

El número total de observaciones en el archivo resultante, es la suma de estos tres frecuencias `_merge`. En este ejemplo, sin embargo, cada observación pyme con base tecnológica, en el archivo **pei_98_1.dta** tiene una coincidencia exacta en el archivo **pei_98_2.dta**, por lo que `_merge` = 3 y no 1s o 2s (obviamente, porque los dos archivos se crean a partir de el mismo archivo). Pero en los ejemplos de la vida real, 1s y 2s pueden permanecer después de la fusión. Más a menudo, se requiere trabajar con las observaciones que son comunes en ambos archivos (`_merge` = 3), de la siguiente forma:

```
keep if _merge==3
```

```
. keep if _merge==3
(0 observations deleted)
```



Además, es una buena práctica eliminar la variable `_merge` del archivo de datos después de la fusión, con:

```
drop _merge
```

Ahora, se tiene un archivo de datos que es idéntico a **pei_98.dta** en el contenido.

Ejemplo 26. Haciendo apéndices en archivos (*Appending data sets*)

Hace agregados a los archivos de datos, es necesario cuando se requiere combinar dos archivos que tienen las mismas (o casi las mismas) variables, pero sus unidades de observación (por ejemplo, las **pbt**) son mutuamente excluyentes. Para demostrar la operación *append* (agregar), nuevamente dividirá el archivo **pei_98.dta**. Esta vez, sin embargo, en lugar de descartar variables, retire algunos registros. Así, abra el archivo **pei_98.dta**, borre de 1 a 700 registros y guarde este archivo como **pei_98_1.dta**:

```
use "/Users/DCA/Desktop/eval/data/pei_98.dta"
drop in 1/700
save pei_98_1.dta, replace
sum
```

```

. use "/Users/DCA/Desktop/eval/data/pei_98.dta", clear

. drop in 1/700
(700 observations deleted)

. save pei_98_1.dta,replace
file pei_98_1.dta saved

. sum

```

Variable	Obs	Mean	Std. Dev.	Min	Max
id	429	259977.1	35668.11	203163	323103
year	429	1	0	1	1
region	429	2.11655	.8836158	1	4
cluster	429	25.77855	3.562513	20	32
ageceo	429	45.82051	12.33387	20	95
genceo	429	.9067599	.2911081	0	1
educeo	429	2.034965	3.304088	0	14
smesize	429	5.407925	2.322992	1	18
smefac	429	83.28788	290.4088	1	4208
incinnov	429	235740.7	1355599	1	2.42e+07
expopn	429	3769.862	1734.835	1149.456	15270.67

Command

A continuación, vuelva a abrir **pei_98.dta** pero mantenga las observaciones 1 a 700 y guarde este archivo como **pei_98_2.dta**:

```

use "/Users/DCA/Desktop/eval/data/pei_98.dta", clear
keep in 1/700
save pei_98_2.dta,replace
sum

```

```

. use "/Users/DCA/Desktop/eval/data/pei_98.dta", clear

. keep in 1/700
(429 observations deleted)

. save pei_98_2.dta, replace
file pei_98_2.dta saved

. sum

```

Variable	Obs	Mean	Std. Dev.	Min	Max
id	700	114809.2	59655.12	11054	203133
year	700	1	0	1	1
region	700	2.217143	.9826785	1	4
cluster	700	11.25286	5.962717	1	20
ageceo	700	46.13	12.89271	18	88
genceo	700	.9085714	.2884237	0	1
educeo	700	2.49	3.568847	0	16
smesize	700	5.234286	2.128752	1	18
smefac	700	72.87588	124.4461	.2	1090
incinnov	700	106447.2	183288.6	1	1733076

Ahora, tiene dos archivos de datos; ambos tienen variables idénticas pero diferentes pymes de base tecnológica (**pbt**). En esta situación, debe agregar (*append*) dos archivos. De nuevo, un archivo tiene que estar en la memoria (cuál, no importa). Abra **pei_98_1.dta** y posteriormente, agregue **pei_98_2.dta**.

```

. use "/Users/DCA/Desktop/eval/data/pei_98_1.dta", clear

. append using pei_98_2

```

Tenga en cuenta que no es necesario ordenar los archivos individuales para la operación *append* (agregar), y **STATA** no crea ninguna variable nueva como *_merge* después de la operación de agregar. Se puede verificar que la operación de anexar se haya ejecutado correctamente emitiendo el conteo de **STATA** comando, que muestra el número de observaciones en el conjunto de datos resultante, que debe ser la suma de las observaciones en los dos archivos individuales (es decir, 1,129).

La siguiente sección, discute el uso de dos tipos de archivos que son extremadamente eficientes en las aplicaciones de **STATA**. Uno alma-

cena los comandos y resultados de para su posterior revisión (*archivos .log*), y el otro, almacena comandos para ejecuciones repetidas más tarde (*archivos do*). Los dos tipos de archivos funcionan de forma interactivamente, lo cual es muy útil para depurar comandos y obtener mayor información de los datos.

Ejemplo 27. Archivos *.log* (log files)

A menudo, se desea guardar los resultados de los comandos **STATA** y también imprimirlos, lo cual es posible realizar del comando *.log*. Dicho archivo, se crea mediante la emisión de un comando *log using* (inicio de sesión) y cerrado por un comando *log close*; todos los comandos emitidos en el transcurso, así como los correspondientes a los resultados (excepto los gráficos), se guardan en el archivo *.log*. Por ejemplo, al utilizar **pei_08.dta** suponga que desea guardar solo la ejecución de la resumen educativo de los jefes por género familiar aquí son los comandos:

```
use "/Users/DCA/Desktop/STATA logfiles/pei_08.dta
log using educeo.log
by genceo, sort: sum educeo
log close
```

```

- use "/Users/DCA/Desktop/STATA log Files/pei_08.dta"

- log using educeo.log

name: <unnamed>
log: /Users/DCA/Library/Mobile Documents/com-apple-CloudDocs/.Trash/969_2000_8210WP_v01_8_STATA8/educ
> ex.log
log type: text
opened on: 18 Jun 2020, 14:58:24

- by geneco, sortisom educeo

-# geneco = 0

+-----+-----+
| Variable | Obs | Mean | Std. Dev. | Min | Max |
+-----+-----+
| educeo | 184 | .8461538 | 2.534932 | 0 | 12 |
+-----+-----+

-# geneco = 1

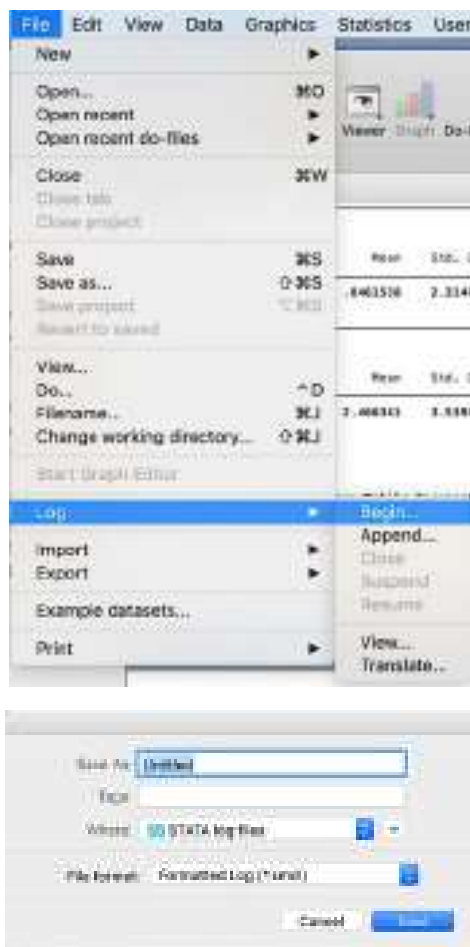
+-----+-----+
| Variable | Obs | Mean | Std. Dev. | Min | Max |
+-----+-----+
| educeo | 1,829 | 2.466341 | 3.539825 | 0 | 16 |
+-----+-----+

- log close
name: <unnamed>
log: /Users/DCA/Library/Mobile Documents/com-apple-CloudDocs/.Trash/969_2000_8210WP_v01_8_STATA8/educ
> ex.log
log type: text
closed on: 18 Jun 2020, 14:58:24

+-----+-----+
| Command |
+-----+-----+

```

STATA crea un archivo de texto llamado *educeo.log* en el subdirectorio y guardando los resultados en ese archivo. Si desea que el archivo *.log* se guarde en una carpeta que no sea la carpeta actual, puede especificar la ruta completa, dentro del archivo *.log*. También puede usar la opción **File**→**Log**→**Begin**.



Así, se sugiere tener en cuenta las siguientes opciones:

- Si ya existe un archivo *.log*, puede reemplazarlo utilizando el comando: *log using educeo.log* o también utilizando los comandos: *replace* o *append* a una nueva salida con ingresar *log using educeo.log, append*.

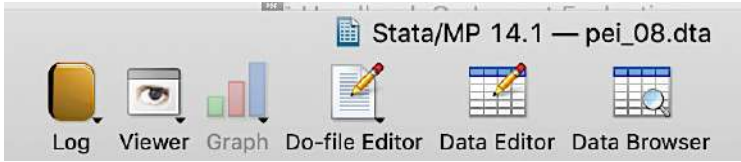
- Si se desea mantener el archivo *.log* existente sin cambios, entonces puede cambiar el nombre de este archivo o el archivo en el comando de creación *.log*.
- Si desea suprimir una parte de un archivo *.log*, puede emitir un comando *log off* antes de esa parte, seguido de un comando *log on* para la sección que desea guardar. Debe cerrar el archivo *.log* antes de abrir uno nuevo; de lo contrario, recibirá un mensaje de error.

Ejemplo 28. Archivos *.do* (*.do files*)

Hasta ahora, se ha visto el uso interactivo de los comandos **STATA**, que es útil para la depuración comandos y obtener más información de los datos. Al escribir una línea de comando cada vez, y **STATA** procesa ese comando, muestra el resultado (si lo hay) y espera el siguiente mando. Aunque este enfoque tiene sus propios beneficios, el uso más avanzado de **STATA** implica ejecutar comandos en un *lote*, es decir, los comandos se agrupan y se envían juntos en lugar de uno a la vez. *Si se encuentra usando el mismo conjunto de comandos repetidamente, puede guardar los comandos en un archivo y ejecutarlos juntos siempre que los necesite.* Estos archivos de comando se les conoce como *archivos .do*; son el equivalente **STATA** de las macros. Puede crear archivos *.do* al menos en tres formas:

- La primera, es escribir simplemente los comandos en un archivo de texto, etiquetado como en nuestro ejemplo: *educeo.do* (el sufijo *.do* es importante), y ejecute el archivo usando *do educeo* en la ventana de comando **STATA** (**STATA** command window).
- La segunda, oprima clic derecho en cualquier lugar de la ventana revisar **STATA** (**STATA** review window) para guardar todos los comandos que fueron utilizados de forma interactiva. El archivo en el que se guardaron, se puede editar y etiquetar y usar como un archivo *.do*.

- Tercero, utilice el *Do-file editor* incorporado en la barra de menú de **STATA**.



Los comandos se pueden escribir en el editor.

Ejecute estos comandos resaltándolos y utilizando el icono apropiado. Con tiempo y práctica, este procedimiento se convierte en una forma muy rápida y conveniente de trabajar con **STATA**.

Aquí hay un ejemplo de un archivo *.do*:

```
log using educeo.log
use pei_98
sort id
save, replace
sort genceo
by genceo: sum educeo
log close
```

Las principales ventajas de usar archivos *.do* en lugar de escribir comandos línea por línea son *replicabilidad* y *repetibilidad*. Con un archivo *.do*, uno puede replicar resultados que fueron trabajados semanas o meses antes. Además, los archivos *.do* son especialmente útiles cuando los conjuntos de comandos deben repetirse, por ejemplo, con diferentes archivos de datos.

Por ejemplo:

```

*This is a Stata comment that is not executed
/*****This is a do file that shows some very useful
commands used in do files. In addition, it creates a
log file and uses some basic Stata commands ***/
#delimit ;
set more 1;
drop _all;
cap log close;
log using c:\eval\log\try.log, replace;
use c:\eval\data\pei_98.dta ;
describe ;
list in 1/3 ;
list id smesize educeo if genceo==0 & ageceo<45;
summarize smesize;
summarize smesize, detail;
sum smesize educeo [aw=weight], d;
tab genceo;
tab educeo genceo, col row;
tab educeo, summarize(ageceo);
label define genlabel 1 "MALE" 0 "FEMALE";
label values genceo genlabel;
tabulate genceo;
label variable sexhead "Gender of Head: 1=M, 0=F";
save c:\eval\data\temp.dta, replace;
#delimit cr
use c:\eval\data\pei_01.dta
append using temp
tab year
_l_o_g_c_l_o_s_e_____

```

La primera línea del archivo, es un comentario. **STATA** trata cualquier línea que comience con un *asterisco* (*) como comentario y lo ignora. Puede escribir un comentario multilínea utilizando (/ *) como inicio del comentario, y finalizarlo con (* /). Los comentarios son muy útiles para fines de documentación y deben incluir al menos: *el propósito general del archivo .do y la fecha y hora de la última modificación*. Usted puede incluir comentarios en cualquier parte del archivo *.do*, no solo al principio. Así, de lo contenido, se recomienda tomar en cuenta :

#delimit. Por defecto, **STATA** asume que cada comando termina con el comando retorno de carro (es decir, presionando la tecla *Enter*). Si acaso, un comando, es demasiado largo para quedar en una línea, puede extenderlo en más de una línea con este comando. El comando en el ejemplo, dice que un punto y coma (;) finaliza un comando. Cada comando siguiente el comando *delimit* tiene que terminar con un punto y coma. Aunque para este archivo *.do* particular, el comando *#delimit* no es necesario (todos los comandos son lo suficientemente cortos), se hace para explicar el comando con mayor claridad.

set more 1. **STATA**, generalmente muestra los resultados por pantalla a la vez y espera para que el usuario presione cualquier tecla. Pero este proceso, pronto puede convertirse en una molestia si, después de dejar correr un archivo *.do*, tiene que presionar una tecla para cada pantalla hasta que finalice el programa. Este comando, muestra la salida completa saltando página tras página automáticamente.

drop all. Este comando borra la memoria.

cap log close. Este comando cierra cualquier archivo abierto *.log file*. Si no hay alguno abierto, **STATA** lo ignora.

Ejemplo 29. Archivos .ado (.ado files)

Los archivos *.ado* son programas **STATA** destinados a realizar tareas específicas. Muchos comandos **STATA** se implementan como archivos *.ado* (por ejemplo, el comando *summarize*). Para ejecutar dicho programa, simplemente escriba su nombre en la línea de comando. Los usuarios pueden escribir sus propios programas *.ado* para cumplir requisitos especiales. De hecho, los usuarios de **STATA** y los desarrolladores escriben continuamente dichos programas, que a menudo están disponibles a la gran comunidad de usuarios de **STATA** en Internet. Así, se usarán tales comandos a lo largo de los ejemplos sobre diferentes técnicas de evaluación de impacto. Por ejemplo, la técnica de propensión de coincidencia de puntaje (**PSM**. *Propensity Score Matching*) se implementa mediante un archivo *.ado* llamado *pscore.ado*.

```
findit pscore
```

STATA responde con una lista de implementaciones de programa *.ado*. Al hacer clic en uno de ellos dará sus detalles y presentará la opción de instalarlo. Cuando Stata instala un *.ado*, también instala los archivos de ayuda asociados.

GLOSARIO

A continuación se ofrece una parte del vocabulario usado en el campo de la evaluación de impacto social en la introducción de innovaciones. Se intentó mantener el énfasis en términos específicos, pero dado que los evaluadores operan en múltiples campos, el lenguaje empleado comprende no solo términos específicos sino también otros que provienen de la evaluación de impacto ambiental y de otros tipos de evaluaciones de impacto, así como también de otros sectores en los cuales esos especialistas se desempeñan. Las descripciones, sin ser definiciones, se incluyen para ayudar a la comprensión de los conceptos. Pese a que muchas descripciones son originales (inspiradas a veces de múltiples fuentes), otras son de uso estándar y de empleo común en varios campos.

Actividad. Medidas adoptadas o trabajo realizado a través del cual los insumos, como los fondos, la asistencia técnica y otro tipo de recursos que se movilizan para generar productos específicos, como el dinero gastado, los libros de texto distribuidos o el número de participantes en un programa de empleo.

Acuerdo de desarrollo. Concepto muy similar al del Acuerdo de Impactos y Beneficios (AIB) pero, en vez de comunitario constituir un acuerdo bilateral entre una empresa y una comunidad, que es lo que el AIB tiende a ser, podría ser iniciado por un gobierno.

Acuerdo de participación. Véase Acuerdo de impactos y beneficios.

Acuerdos comunitarios. Véase Acuerdo de impactos y beneficios (*Community agreements*).

Acuerdos de impactos y beneficios (IAB). Acuerdos vinculantes entre una empresa y una comunidad (a veces también interviene el gobierno)

que esboza los impactos negativos probables que un proyecto creará, los esfuerzos de mitigación que una empresa llevará a cabo, y el alcance de las contribuciones que la empresa proporcionará a la comunidad en forma de empleos y otros beneficios, como en inversión social.

Adquisiciones locales. Se refiere a las políticas y estrategias deliberadas de las empresas que les permiten aprovisionarse de bienes y servicios a partir de proveedores locales, de modo de intensificar los beneficios del proyecto para la comunidad local.

Afrenta cultural. Acto deliberado que insulta u ofende profundamente, como la violación o profanación de sitios sagrados, o la rotura deliberada de tabúes u otras costumbres culturales significativas.

Agravio (*Grievance*). Toda inquietud percibida que suscite en un individuo o grupo el sentido de tener derechos o de haber sido agraviado, sobre la base de la ley, de un contrato, de promesas explícitas o implícitas, de prácticas habituales o de conceptos generales de equidad.

AIPP (IAP2) Asociación Internacional para la Participación Pública www.iap2.org.

Alienación. Concepto de las ciencias sociales que refiere a los procesos sociales que alienan a los individuos (los convierte en extranjeros o extraños). El resultado de la alienación es una falta de pertenencia y se experimenta una desconexión, falta de sentido, impotencia y falta de agencia. Así, es un proceso que repercute severamente en el bienestar mental (y a la larga, en la salud física) de los individuos afectados. Se manifiesta en aislamiento social, desesperación, depresión médica y una variedad de otros comportamientos relacionados con la salud.

Análisis de costo-beneficio. Estima los beneficios totales previstos de un programa, en comparación con sus costos totales previstos. Su fin es cuantificar todos los costos y beneficios de un programa en términos monetarios y evaluar si los beneficios superan a los costos. Método económico que evalúa alternativas para un negocio usualmente mediante la determinación del ratio de beneficios con respecto a los costos. En décadas pasadas, se empleaba extensamente en las evaluaciones de impacto, pero

no tanto en las de impacto social porque intenta representar todos los impactos únicamente en términos monetarios.

Análisis de costo-efectividad. Compara el costo relativo de dos o más programas o alternativas de programa en términos de alcanzar un resultado común, como la producción agrícola o las calificaciones de los alumnos en los exámenes.

Análisis de género. Un proceso utilizado para examinar y comprender las implicancias de género de una intervención planeada dirigida a las mujeres como asimismo a los hombres, en el contexto cultural de las comunidades afectadas. Un análisis de género debería observar el sexo y las diferencias de género.

Análisis FODA. Considera las fortalezas, oportunidades, debilidades y amenazas de una organización o comunidad.

Análisis de regresión. Método estadístico para analizar las relaciones entre una variable dependiente (la variable que se debe explicar) y variables explicativas. El análisis de regresión normalmente no es suficiente para capturar los efectos causales. En la evaluación de impacto, el análisis de regresión es una manera de representar la relación entre el valor de un indicador de resultado Y (variable dependiente) y una variable independiente que captura la asignación al grupo de tratamiento o grupo de comparación, mientras se mantienen constantes otras características. Tanto la asignación al grupo de tratamiento y de comparación como las demás características son variables explicativas. El análisis de regresión puede ser univariante (si hay solo una variable explicativa; en el caso de la evaluación de impacto, la única variable explicativa es la asignación al grupo de tratamiento o de comparación) o multivariante (si hay varias variables explicativas).

Análisis de sensibilidad. Trata de la sensibilidad del análisis ante los cambios en los supuestos. En el contexto de los cálculos de potencia, contribuye a comprender cuánto tendrá que aumentar el tamaño requerido de la muestra bajo supuestos más conservadores (como un menor

impacto esperado, una mayor variación en el indicador de resultado o un nivel más alto de potencia).

Antiguos terrenos. Hace referencia al proyecto que tiene lugar en un sitio en el que hubo otros proyectos industriales anteriormente y donde es necesario abordar problemas heredados.

Arraigo a un lugar. Hasta qué punto un individuo tiene sentimientos positivos con respecto a su entorno o comunidad local.

Área de influencia. En una evaluación de impacto ambiental (EIA), es el área física (y componentes tales como aire, agua, suelo) sobre la cual un proyecto produce impactos (incluso abiótico, biótico y socioeconómico) causados por un proyecto (y sus actividades asociadas). Por tanto, comprende no solo el área de la superficie de tierra pero también el funcionamiento de todo ecosistema marino y terrestre, cuencas atmosféricas y cuencas hidrográficas o hidrológicas, y todos los agrupamientos sociales de individuos, comunidades, empresas (especialmente pequeñas y medianas empresas), organizaciones y entidades de gobierno. (Véase Área de influencia social)

Área de influencia social. Término cuyo significado es prácticamente el mismo que el de *área de influencia* pero que enfatiza los impactos sociales del proyecto. Dada la movilidad de las personas y el alcance de los impactos sociales, el área de influencia social probablemente sea mucho más amplia en área física que el área física de influencia.

Arraigo. Es estar fuertemente integrado en la comunidad local, enraizado, como un árbol que tiene raíces. El arraigo, un componente del apego al lugar junto con el sentido de pertenencia, se refiere a los vínculos sociales de una persona (capital social) con otros miembros de la comunidad, por ejemplo, el tener parientes en ella, tener amigos de larga data, tener amigos íntimos, etc.

Artesanal. Trabajar con las manos. Un artesano es una persona con habilidad para hacer artesanías. A menudo se aplica a mineros de pequeña escala o a pescadores de subsistencia.

Asalariados ingreso fijo. Se refiere a aquellas personas cuyos ingresos (provenientes de trabajos o pensiones) son, es decir que no son objeto de ajuste por cuestiones de inflación o costo de vida. Las ciudades en auge habitualmente tienen una inflación localizada. Aunque las personas vinculadas con el proyecto reciben en pago una cantidad de dinero suficiente como para hacer frente a los aumentos del costo de vida (que son por lo general la causa de la inflación), mucha gente percibe ingresos que no se ajustan a los niveles locales de inflación. Esta gente incluye no sólo jubilados, sino también personas con salarios que son fijados a nivel nacional, como enfermeros, docentes, policías y otras personas que trabajan en el gobierno.

Asentamientos informales. Asentamientos que se desarrollan informalmente, esto es, sin planificación adecuada y por lo general en violación de los códigos de construcción y planeamiento, usualmente en tierras para las cuales los residentes no tienen títulos de propiedad. A menudo se emplean materiales desechados para construir viviendas temporarias.

Asignación aleatoria o ensayo controlado aleatorio. Método de evaluación de impacto por el cual cada unidad elegible (por ejemplo, un individuo, un hogar, una empresa, una escuela, un hospital o una comunidad) tiene la misma probabilidad de ser seleccionada para ser tratada en un programa. Con un número suficientemente grande de unidades, el proceso de asignación aleatoria garantiza la equivalencia tanto en las características observables como no observables entre el grupo de tratamiento y el grupo de comparación, y así se descarta cualquier *sesgo de selección*. La asignación aleatoria se considera el método más robusto para estimar los contrafactuales y se le suele considerar como la regla de oro de la evaluación de impacto.

Autodeterminación. Uno de los derechos humanos designados que pertenece a grupos (y no a individuos) y que establece que todos los pueblos deben poder determinar libremente su condición política y perseguir libremente su desarrollo económico, social y cultura.

Autoridad competente. Toda persona u organización que tiene la autoridad, capacidad o facultad, asignada por ley o investida, de llevar a cabo una función determinada. En términos de una evaluación de impacto social para introducción de innovaciones de lo ambiental en general refiere a la autoridad que otorga la licencia ambiental. (Véase también Organismo regulador).

Beneficiarios. El estado social, económico, psicológico, espiritual o de salud de un individuo o grupo.

Bienestar subjetivo. Experiencia personal de la vida de uno, nivel de satisfacción con la vida de uno, cuán feliz la gente se siente en general sobre su vida en conjunto.

Buena gobernanza. Una interpretación normativa de cómo debería producirse la gobernanza (de cualquier institución), lo cual incluye un compromiso con la responsabilidad, la transparencia, el imperio de la ley, la creación de capacidad, un proceso inclusivo y participativo.

Buena práctica. Lo que se considera habitualmente apropiado y esperado (es decir, convencional más que vanguardista) en un área. En cambio, mejores prácticas quiere decir vanguardista o líder, y por ende adecuado para propugnar, pero no es de esperarse en todas las circunstancias.

Cadena de resultados. Establece la lógica causal del programa explicando cómo se logra el objetivo de desarrollo. Articula la secuencia de insumos, actividades y productos que se espera que mejoren los resultados.

Cálculos de potencia. Método para determinar cuál es el tamaño de la muestra requerida para que una evaluación de impacto estime con precisión el impacto de un programa, es decir: la muestra más pequeña que permitirá detectar el efecto mínimo detectable. Los cálculos de potencia dependen de parámetros como la potencia (o la probabilidad de un error de tipo II), el nivel de significancia, la media, la varianza y la correlación intra-clústers de los resultados de interés.

Calidad de vida. El bienestar general de un individuo, que difiere del estándar de vida en cuanto a que comprende todas las dimensiones no económicas, subjetivas, de la vida.

Capacidad de carga social. Se refiere a la cantidad de personas que un lugar en particular puede sostener. Mientras que la capacidad de carga (ecológica) es un término adecuadamente establecido en ecología, la capacidad de carga social es un término raramente empleado por los profesionales de las ciencias sociales. Los encargados de los parques hablan de capacidad de carga social para referirse a la cantidad de turistas que ellos creen que sus parques pueden recibir, y en el ámbito del turismo existe un concepto de capacidad de carga social como la cantidad de turistas que una atracción turística específica (sitio, localidad, región, cultura) puede manejar con comodidad.

Capital humano. La educación, destrezas, conocimiento, habilidad para trabajar y salud que, en conjunto, permiten que las personas persigan sus medios de vida objetivos. (Véase también Capitales)

Capital social. Los recursos sociales de los cuales se sirven las personas en pos de alcanzar sus objetivos de subsistencia, lo cual incluye sus redes y la amplitud de sus conexiones y sus relaciones de confianza, reciprocidad e intercambio que facilitan la cooperación, reducen los costos de transacción y conforman la base de las redes de seguridad informales. Incluye las instituciones, relaciones, actitudes, valores y valores compartidos y reglas de conducta social que rigen las interacciones entre las personas y contribuyen al desarrollo económico y social.

Capitales. El marco de referencia para considerar la sostenibilidad y el logro de resultados de desarrollo en términos de activos (o capitales), tal como el capital natural, capital humano, capital social, capital financiero, capital manufacturero, y a veces capital político e institucional, y capital cultural y espiritual. Hay varios enfoques que usan los capitales como elemento central, entre los cuales se encuentra el enfoque de los medios de subsistencia sostenibles.

Captura por la élite. Una situación en la cual los recursos que estaban previstos en beneficio de la mayoría de la población son usurpados (capturados) por un pequeño grupo de ricos y poderosos que son parte de la sociedad, una elite económica, política, educativa o étnica.

Carácter irremediable. Daño irremparable, impactos negativos que no pueden ser mitigados ni remediados.

Carácter irreparable. Que no puede ser reparado.

Censo. Empadronamiento total de una población. Los datos censales abarcan todas las unidades de la población. Compárese con muestra.

Cerco perimetral. La delimitación que se extiende alrededor del emplazamiento de un proyecto. Se la utiliza con el propósito de diferenciar entre cuestiones internas y cuestiones externas. No obstante, aunque esto pueda resultar obvio en el sentido técnico, no es un término claro, y en términos de cuestiones de derechos sociales y humanos es irrelevante.

CFI. Corporación Financiera Internacional (*International Finance Corporation*), el brazo del Grupo del Banco Mundial para el financiamiento del sector privado. Es de particular importancia ya que sus estándares de desempeño se convirtieron en referencia internacional y son la base de los Principios de Ecuador.

Ciencia abierta. Movimiento cuyo fin es elaborar métodos de investigación más

transparentes, mediante el registro de los ensayos, la utilización de planes de preanálisis, documentación de datos y registros.

Ciudad de crecimiento rápido. Una comunidad, pueblo o ciudad que experimenta un crecimiento excesivamente rápido.

Ciudad industrial Un asentamiento en el cual la gran mayoría de las personas trabaja para la misma empresa o por lo menos en una de las muchas empresas que operan en el mismo proyecto.

CLPI (FPIC). Consentimiento libre, previo e informado (véase el desarrollo del tema al respecto en el presente documento).

Clúster. También llamado conglomerado. Grupo de unidades que pueden compartir características similares. Por ejemplo, los niños que asisten a

la misma escuela pertenecen a un mismo clúster porque comparten el mismo centro escolar, los mismos profesores y viven en el mismo barrio.

Cohesión comunitaria. Se refiere al sentido de armonía en un lugar (área rural, pueblo o ciudad), que se puede establecer por los niveles de: aceptación y valoración de la diversidad social, sentido compartido de pertenencia en todos los grupos, una visión e imagen ampliamente aceptada del lugar, oportunidades en la vida y acceso a servicios razonablemente similares, y relaciones sociales positivas entre las personas de diferentes contextos.

Coincidencia (Matching). Método no experimental de evaluación de impacto que utiliza grandes bases de datos y técnicas estadísticas para construir el mejor grupo de comparación posible para un determinado grupo de tratamiento sobre la base de características observables.

Communitas. Término en latín de las ciencias sociales que significa un fuerte sentido de comunidad, especialmente el que se construye a través de la participación en ritos o celebraciones comunitarias.

Comoditización. Procesos por los cuales la cultura local y los objetos de la cultura local se convierten en mercancías (*commodities*) y por lo tanto tradiciones religiosas, costumbres y festivales del lugar se reducen a conformar las expectativas de aquellos que los adquieren. Es un concepto que proviene de los impactos sociales del turismo, pero puede ocurrir en cualquier situación donde una cultura local entra en contacto con un grupo acaudalado de extranjeros.

Comparación antes y después. También conocida como *comparación previa posterior* o *comparación reflexiva*. Se trata de un seguimiento de los cambios en los resultados para los beneficiarios del programa a lo largo del tiempo, utilizando mediciones antes y después de la implementación del programa o la política, sin utilizar un grupo de comparación.

Comparaciones de inscritos y no inscritos. También conocidas como *comparaciones autoseleccionadas*. Esta estrategia compara los resultados de las unidades que decidieron inscribirse y las unidades que decidieron no inscribirse en un programa.

Complicidad. Palabra empleada en el ámbito de los derechos humanos. Las empresas no deben ser cómplices de violaciones de derechos humanos por parte de terceros. La complicidad comprende cualquiera de las siguientes situaciones: causa o contribuye a la violación de derechos humanos a través del permiso, agravamiento o facilitamiento de aquel; saber o deber haber sabido que las violaciones de derechos humanos probablemente ocurrieran a partir de su conducta; y estar próximo a la violación de derechos humanos ya sea geográficamente o a través de la fuerza, duración o tono de sus relaciones.

Compromiso comunitario. Término que abarca las muy diversas maneras en que las personas interactúan con los procesos de toma de decisiones y participan en estos. Es similar a la “participación pública” y a “intervención pública”, por lo que a menudo estos términos se emplean indistintamente. No obstante, el compromiso comunitario es el que se prefiere actualmente ya que pone énfasis en un grado mayor de participación o compromiso en el proceso de toma de decisiones y mayor respeto por las personas. También se vincula con un discurso diferente y representa el desarrollo de la comprensión en la transición de la consulta al empoderamiento.

Comunidad. Concepto de uso común pero disputado que puede aplicarse a diferentes niveles, si bien en general refiere a la agrupación de personas en un lugar determinado y que se presume comparten un sentido de identidad y algunas interacciones de la vida diaria, y tienen algunas instituciones sociales y políticas comunes. Si bien los individuos experimentan impactos sociales a un nivel personal, para la evaluación de impacto social para la introducción de innovaciones las personas viven, trabajan y juegan en agrupamientos sociales llamados comunidades, lo que por lo tanto los hace el foco principal de las dichas evaluaciones.

Comunidades a ambos lados del cerco perimetral. Las comunidades que se encuentran en la vecindad inmediata del proyecto y que son particularmente vulnerables a los impactos directos del proyecto, tal como ruido, polvo y vibraciones.

Comunidades afectadas. Aquellas comunidades (y comunidades anfitrionas) que reciben el impacto de un proyecto.

Comunidades anfitrionas. Aquellas comunidades que se encuentran cerca de los emplazamientos de los proyectos, que albergan el proyecto y sus trabajadores, dicho de otro modo, las comunidades que reciben el impacto. En la planificación de reasentamientos, se refiere a las comunidades existentes que absorberán las personas que están siendo reubicadas.

Condicionalidades. En asistencia al desarrollo, se refiere a las condiciones que acompañan a un préstamo, alivio de deuda o ayuda bilateral; en licenciamiento ambiental, se refiere a las condiciones de la licencia.

Conflicto latente. Conflicto que está latente, oculto o encubierto en una comunidad. El personal del proyecto podría no notar la presencia de tensión en una comunidad.

Conocimiento indígena. Véase Conocimiento local.

Conocimiento local. El conocimiento que la gente de una localidad o comunidad dada ha desarrollado con el tiempo y que seguirá desarrollando. Hace referencia a la recopilación de hechos y sistemas de conceptos, creencias y percepciones que las personas tienen con respecto al mundo que los rodea. Incluye también la manera en que las personas observan y miden su entorno, cómo resuelven problemas y convalidan la información. Se emplean una amplia variedad de términos: conocimiento indígena, conocimiento técnico indígena, conocimiento tradicional, conocimiento ecológico tradicional o conocimiento medioambiental tradicional y conocimiento de los aborígenes.

Conocimiento tradicional. Véase Conocimiento local.

Consentimiento. Otra palabra para expresar que se está de acuerdo con algo. También puede significar tener la facultad de poder dar o retirar la aprobación para un proyecto. Así, un organismo regulador debe determinar las condiciones del consentimiento. En algunas circunstancias, las personas del lugar también podrían tener la habilidad de dar o retirar su consentimiento para un proyecto (véase **CLPI**). El concepto de consentimiento está altamente asociado con la confianza. Un evaluador de impacto

social a menudo puede crear consenso para un proceso de evaluación de impacto social para introducción de innovaciones (o para el proyecto), para lo cual muestra que existe un problema que debe abordarse, que una gran variedad de grupos están abordándolo y que el proceso para resolver el problema es justo.

Consentimiento informado. El principio ético más básico para llevar a cabo investigación con humanos; en esencia, que el participante en la investigación tiene derecho a decidir si quiere participar o no, y que para tomar esa decisión debe basarse en información completa sobre la investigación y los posibles riesgos que conlleva. El consentimiento informado a menudo se lo registra mediante la firma del participante en un formulario a tal efecto. Los especialistas en evaluación de impacto social para introducción de innovaciones deben recibir el consentimiento informado de los participantes en la recolección de datos para la evaluación de impacto social para introducción. Uno de los fundamentos de la protección de los derechos de los sujetos humanos. En el caso de las evaluaciones de impacto, requiere que los encuestados comprendan claramente los fines, procedimientos, riesgos y beneficios de la recopilación de datos en la que se les pide participar.

Consulta pública, intervención pública, participación pública y compromiso comunitario. Si bien a menudo se emplean indistintamente, existen distinciones importantes. Se puede decir que *intervención pública* es un concepto genérico relacionado a los procesos o que implica al público en los procesos de toma de decisiones. *Consulta* implica pedir las opiniones de la comunidad, mientras que *participación* implica realmente que el público colabora en el proceso de toma de decisiones.

Contenido local. Se refiere a la exigencia, expectativa o compromiso de una empresa para asegurar que el valor quede retenido en el ámbito local a través del empleo o del aprovisionamiento.

Contrafactual. Valor que habría tenido el resultado (Y) para los participantes del programa sino hubieran participado en el programa (P). Por definición, el contrafactual no se puede observar. Por lo tanto, debe esti-

marse utilizando un grupo de comparación. En psicología, refiere a una representación mental o imagen de una trayectoria alternativa, pasada o futura, como manera de concebir otras posibilidades para lo realmente ocurrido. Así, se permite a los individuos procesar sus sentimientos sobre eventos pasados (como en relación a la culpa, remordimiento, arrepentimiento, y preocupaciones de tipo *¿por qué a mí?*, etc.) y también como forma de aprender a partir de las experiencias. Este aprendizaje se puede formalizar en un análisis de escenarios. En el ámbito de una evaluación, sin embargo, tiene un significado diferente: se refiere a la comparación entre lo realmente ocurrido y lo que podría haber ocurrido en ausencia de intervención.

Correlación. Medida estadística que indica hasta qué punto dos o más variables fluctúan juntas.

Correlación intra-clústers. También conocida como “correlación intra-clase”. Se trata del nivel de similitud en los resultados o características entre las unidades de los grupos previamente existentes o clusters en relación con unidades de otros clústers. Por ejemplo, los niños que asisten a la misma escuela normalmente serían más similares o estarían más correlacionados en términos de sus zonas de residencia o antecedentes socioeconómicos, en comparación con niños que no asisten a esa escuela.

Cosmología. La creencia sobre el origen, la historia, la evolución y las leyes culturales relacionados con el cosmos o universo en una cultura particular o sistema mitológico.

Cosmovisión. La perspectiva a través de la cual una persona (y a veces una sociedad o cultura) concibe su mundo; en otras palabras, un marco de referencia cognitivo. Es un término frecuentemente empleado por especialistas en ciencias sociales y especialistas en evaluación de impacto social para la introducción de innovaciones.

Costo de remplazo. Un concepto de la economía y los seguros que se refiere al costo total de reemplazar un activo. La valorización de los activos destruidos por un proyecto con fines de indemnización puede resultar controvertida. Los peritos tasadores de las compañías de seguros

hacen uso frecuentemente del valor depreciado de un activo. En el reasentamiento y el desplazamiento inducido por un proyecto, debería cubrirse el costo total de reemplazo de modo de asegurar que no se empeora la situación de las personas.

Cribado. Un proceso de selección para determinar si es necesaria una evaluación de impacto y, si así fuera, qué tipo de evaluación debería realizarse. En un entorno regulatorio, esto se especifica en las regulaciones aplicables. Aunque no se emplea siempre de esta manera, el término podría aplicarse de manera más general a la determinación de los requisitos que deben cumplirse, o a los procedimientos que deben seguirse, que están implícitos en los procedimientos de las empresas, la legislación nacional e internacional, y/o las exigencias de los socios financieros, especialmente cuando tienen participación el Banco Mundial, la CFI y los Principios de Ecuador.

Cuenta de depósito en garantía. Se refiere a sumas de dinero retenidas en concepto de garantía por parte de un tercero, de conformidad con los términos de un acuerdo, y que son liberadas una vez que se hayan cumplido las condiciones del acuerdo y/o con el consentimiento de las partes contratantes y/o mediante una orden judicial u otra acción legal.

Cuestiones heredadas. El significado general de herencia es algo que se deja, como cuando una persona muere (lo que deja a otras personas en su testamento), o cuando se jubila de su trabajo (cuando han creado valor o han hecho un desastre); pero también puede referirse a lo que una empresa o proyecto deja. Aunque la palabra pueda ser neutra/de doble interpretación y pueda significar una situación positiva y la contribución a los resultados del desarrollo, en la evaluación de impacto social (EIS), por lo general, hace referencia al desorden (problema) dejado por proyectos anteriores. En el sector extractivo, se refiere mayormente a la polución y la contaminación que queda, tal como el drenaje ácido de las minas. Es importante observar que las cuestiones heredadas de otros proyectos afectan la confianza de la comunidad para aceptar un nuevo proyecto.

Cultura. Aspectos materiales y no materiales de una forma de vida que se moldean y transmiten entre los miembros de una comunidad o de una sociedad mayor. A veces se la denomina creencias, valores, normas, comportamientos, lengua y objetos materiales compartidos que pasan de una generación a la siguiente.

Cumplimiento. Fenómeno que se produce cuando las unidades adhieren a su asignación como parte del grupo de tratamiento o del grupo de comparación. Significa cumplir con la ley y todo reglamento que regule la actividad. En el contexto de una evaluación de impacto, se refiere al grado de cumplimiento de las condicionalidades para el licenciamiento del proyecto. En general, se espera que exista una auditoría periódica o un monitoreo para garantizar el cumplimiento.

Cumplimiento imperfecto. Discrepancia entre el estatus de tratamiento asignado y la condición de tratamiento real. Se produce cuando algunas unidades asignadas al grupo de comparación participan en el programa, o cuando algunas unidades asignadas al grupo de tratamiento no participan.

Datos administrativos. Datos recopilados asiduamente por organismos públicos o privados como parte de la administración de un programa, normalmente con frecuencia periódica y a menudo en el lugar de la prestación de servicios, e incluyen los servicios prestados, los costos y la participación en el programa. Los datos de monitoreo constituyen un tipo de datos administrativos.

Datos de encuesta. Datos que cubren una muestra de la población de interés. Compárese con censo.

Datos de monitoreo. Datos provenientes del monitoreo del programa que proporcionan información esencial a propósito de la prestación de una intervención, e incluyen quiénes son los beneficiarios y qué beneficios o productos del programa pueden haber recibido. Los datos de monitoreo constituyen un tipo de datos administrativos.

Deber de diligencia. Una obligación de adoptar un cuidado razonable para evitar causar un perjuicio previsible a otra persona o a sus bienes. Un especialista en la evaluación de impacto social para la introducción de innovaciones tiene el deber de diligencia profesional con respecto al cliente y una responsabilidad ética hacia la comunidad, a fin de asegurar que se hayan planteado todas las cuestiones apropiadas.

Debida diligencia. En términos generales, es la investigación que una persona o empresa realizaría antes de celebrar un contrato o antes de realizar una adquisición, especialmente en situaciones en las que pueden existir riesgos. En el ámbito de la evaluación de impacto social para la introducción de innovaciones), refiere prácticamente al mismo concepto excepto que con mayor referencia al Principio 17 de los Principios rectores de las Naciones Unidas sobre las empresas y los derechos humanos que afirma que *“con el fin de identificar, prevenir, mitigar y responder de las consecuencias negativas de sus actividades sobre los derechos humanos, las empresas deben proceder con la debida diligencia en materia de derechos humanos. Este proceso debe incluir una evaluación del impacto real y potencial de las actividades sobre los derechos humanos, la integración de las conclusiones, y la actuación al respecto; el monitoreo de las respuestas y la comunicación de la forma en que se hace frente a las consecuencias negativas”*.

Debida diligencia en materia de derechos humanos. Se entiende el objetivo de los Principios Rectores de las Naciones Unidas sobre las Empresas y los Derechos Humanos de que las empresas deben adoptar un proceso de debida diligencia para garantizar que una acción, transacción o adquisición comercial propuesta no oculte riesgos en materia de derechos humanos (en otras palabras, riesgos para las personas y comunidades, no únicamente riesgos para la empresa). Dado que muchos impactos sociales también son impactos de derechos humanos, los actores afectados son titulares de derechos jurídicos.

Declaración de impacto ambiental (DIA). Documento formal elaborado por una evaluación de impacto ambiental que se presenta ante la autoridad competente.

Declaración de impacto social. El equivalente a una Declaración de Impacto Ambiental en el marco de una evaluación de impacto social para la introducción de innovaciones, un documento formal que se presenta ante una entidad reguladora.

Déficit de gobernanza. Existen muchos tipos de déficits de gobernanza, pero en general se refieren a la diferencia entre la práctica ideal (o al menos lo que debería esperarse normalmente como correcto) y la práctica real de la gobernanza. En la evaluación de impacto y los discursos de empresas y de la sociedad, el déficit tiende a referirse a la falta de observación de las actividades de las empresas multinacionales en los países en desarrollo.

Deliberación. (Y estado deliberativo) Un concepto multidimensional que puede definirse como un diálogo destinado a inducir una reflexión profunda (es decir, un análisis serio) acerca de las opciones y posibilidades, de manera abierta e inclusiva (o sea, sin la intrusión del poder o de la política), y que contemple las inquietudes de todas las partes interesadas.

Denuncia y descrédito. Se refiere a una estrategia que intenta lograr un compromiso o el cumplimiento de las normas o buenas prácticas previstas haciendo públicos los nombres de los infractores o delincuentes. A menudo se la contrasta con *saber* y *hacer saber*.

Dependencia a un lugar. Similar al apego a un lugar pero se emplea para destacar el alcance de la dependencia o atadura a un lugar determinado, que lo imposibilita de moverse y por lo tanto lo hace vulnerable a cambios.

Dependencia económica. Una situación en la cual una comunidad local o una región dependen fuertemente de una empresa o industria. Dicho de otro modo, cuando una elevada proporción de la población de la región trabaja para esa empresa o industria.

Derecho consuetudinario. Prácticas y creencias culturales que son parte vital e intrínseca del sistema social y económico de una cultura particular que son tratadas como si fueran leyes y están (semi)legitimadas; esto es, costumbres que se aceptan como requisitos jurídicos o reglas obligatorias de conducta.

Derechos consuetudinarios. Derechos que rigen por costumbre o cultura.

Derechos humanos. Garantías jurídicas universales que protegen a individuos y grupos de personas contra actos y omisiones que afectan las libertades fundamentales, los derechos y la dignidad humana. Las normas de derechos humanos obligan a los gobiernos (principalmente) y a otros órganos de protección a hacer determinadas cosas y les impiden hacer otras.

Desahogo. El proceso de “dejar salir el vapor”. En el marco de la EIS, se refiere a situaciones en las que las personas que están enojadas o emotivas pueden expresar sus sentimientos. Las oportunidades de desahogarse son una parte necesaria de un buen proceso de compromiso comunitario.

Desarraigo. Puede referirse a lugares que carecen de un “espíritu de lugar”, no son auténticos o están desconectados de su entorno ambiente, o puede referirse a la desconexión que sienten los individuos como consecuencia de su reasentamiento o en razón del grado de cambio ocurrido en su comunidad.

Desarrollo comunitario. Proceso a largo plazo en el cual personas marginadas o que viven en la pobreza aún con esfuerzos para identificar sus necesidades, crear el cambio, ejercer más influencia en las decisiones que afectan sus vidas y trabajan para mejorar la calidad de sus vidas, las comunidades en las que viven y la sociedad que integran.

Desarrollo sostenible. El Informe Brundtland (Nuestro futuro común), define al desarrollo sostenible como el desarrollo que satisface las necesidades del presente sin comprometer la habilidad de las generaciones futuras de satisfacer sus propias necesidades.

Desarticulación social. Proceso mediante el cual se perturban las redes sociales y mecanismos de apoyo dentro de un grupo social. A menudo resulta de la fragmentación que ocurre con un reasentamiento.

Desempeño social. La interfaz entre un proyecto y la sociedad; la configuración de los principios de responsabilidad social de una empresa, sus procesos de sensibilidad social y las políticas, programas y resultados perceptibles en la medida en que se vinculan con las relaciones de la empresa con la sociedad.

Desgaste. El desgaste se produce cuando algunas unidades abandonan la muestra entre una ronda de datos y la siguiente. Por ejemplo, cuando las personas cambian su residencia y no se les puede localizar. El desgaste de la muestra es un caso de falta de respuesta de la unidad. Puede introducir un sesgo en la estimación de impacto.

Desplazamiento. Mientras el reasentamiento es el proceso activo de reubicar personas en razón de la ejecución de un proyecto, el desplazamiento es la experiencia personal y social de pasar por el trastorno de la reubicación, el proceso de pérdida del sentido de pertenencia de cada uno. En los procesos de reasentamiento, el desplazamiento físico refiere a la pérdida de la vivienda como consecuencia de la adquisición de tierras en relación con el proyecto o de restricciones en el uso de las tierras que requieren que las personas afectadas se trasladen a otro lugar. El desplazamiento económico se refiere a situaciones en las que las casas de la gente no se ven afectadas pero pierden otros bienes o el acceso a los bienes (por ejemplo, tierras agrícolas) que generará un perjuicio a los medios de subsistencia y la correspondiente pérdida de ingresos.

Desplazamiento económico. Se refiere al trastorno y los impactos sociales (es decir, desplazamiento) debidos, no ya a la reubicación del lugar de residencia, sino a la pérdida de medios de subsistencia económica, tal como cuando los agricultores pierden sus tierras de cultivo, o cuando la contaminación del agua causa la destrucción de los medios de vida de los pescadores.

Desplazamiento físico. Véase Desplazamiento.

Determinación del alcance. El proceso de establecer las principales cuestiones de interés y los actores interesados afectados para determinada intervención planeada.

Determinantes sociales de la salud. Las condiciones económicas y sociales, y su distribución en una población que influyen sobre las diferencias entre individuos y grupos con respecto al estado de salud. Se refiere a los factores de riesgo social en la vida y las condiciones laborales de una persona, más que a los factores individuales (tal como los factores de riesgo del comportamiento o la genética) que influyen en el riesgo de enfermedad o de vulnerabilidad ante una enfermedad o una lesión. Las distribuciones de los determinantes sociales están delineadas por la salud pública y otras políticas, una gobernanza deficiente y ordenamientos económicos injustos donde los que ya son pudientes y sanos pasan a estar en mejores condiciones mientras los pobres que ya tienen más probabilidad de enfermarse se vuelven aún más pobres.

Diferencias en diferencias (DD). También conocido como *doble diferencia o diferencia en diferencia*. La **DD** comparan los cambios en los resultados a lo largo del tiempo entre el grupo de tratamiento y el grupo de comparación. Esto elimina cualquier diferencia entre estos grupos que sea constante a lo largo del tiempo.

Discurso. Implica todo lo que pueda tener que ver con el idioma y la conversación, inclusive todo lo que el uso del idioma conlleva, tal como la construcción activa de pensamientos, identidades y acciones. Es una construcción social que ofrece un conjunto de posibles expresiones acerca de un área determinada y organiza y brinda un marco a la manera como va a hablarse de un tema, un objeto o un proceso en particular. El discurso proporciona descripciones, reglas, autorizaciones y prohibiciones de la acción social e individual.

Diseño cruzado. También denominado diseño transversal. Se produce cuando hay una asignación aleatoria con dos o más intervenciones, lo que permite estimar el impacto de las intervenciones individuales y combinadas.

Diseño de regresión discontinua (DRD). Método de evaluación de impacto cuasi experimental que se puede utilizar en programas que dependen de un índice límite que determina si los participantes potenciales son elegibles para recibir el programa o no. El umbral de elegibilidad del programa proporciona un punto divisorio entre el grupo de tratamiento y el grupo de comparación. Los resultados para los participantes en un lado del umbral se comparan con los resultados de los no participantes al otro lado del umbral. Cuando todas las unidades cumplen con la asignación que corresponde sobre la base de su índice de elegibilidad, se dice que el DRD es *nítido*. Si hay incumplimiento en el otro lado del umbral, se dice que el DRD es *difuso* o *borroso*.

Diversidad social. Se refiere a la mezcla localizada de grupos sociales e individuos basada en características tales como género, etnia, edad, cultura y contexto económico.

Divulgación. Con frecuencia denominado *divulgación total y franca*, una política de *divulgación abierta* o un *deber de divulgación*, se trata de un término con connotaciones jurídicas y cuasijurídicas que se refiere a la obligación de las partes de una negociación de revelar todo lo que sea considerado pertinente para el tema en estudio (es decir, importancia relativa).

Dominio eminente. Se refiere al poder del Estado para adquirir propiedad privada de manera forzosa. Esto podría hacerse para recuperar (expropiar) tierra para autopistas, aeropuertos, etc. Ocasionalmente, el Estado extiende su poder para habilitar la ejecución de proyectos del sector privado cuando se considera que son de interés nacional.

Economía informal. Se refiere a esa parte de la economía que evita la regulación, imposición o control del gobierno, ya sea porque funciona de manera ilegal o porque funciona en pequeña escala, a base de dinero en efectivo.

Efecto causal. Véase impacto.

Efecto de derrame (*Spillover*). También denominado efecto de contagio. Ocurre cuando el grupo de tratamiento influye directa o indirectamente en los resultados del grupo de comparación (o a la inversa).

Efecto mínimo detectable. El efecto mínimo detectable es un insumo en los cálculos de potencia, es decir, proporciona el tamaño del efecto que una evaluación de impacto está diseñada para estimar con un determinado nivel de significancia y potencia. Las muestras de la evaluación tienen que ser lo bastante grandes para distinguir al menos el efecto mínimo detectable. Este efecto se determina teniendo en cuenta el cambio en los resultados que justificaría la inversión que se ha hecho en una intervención.

Efecto Hawthorne. Se produce cuando, por el simple hecho de ser observadas, las unidades se comportan de manera diferente.

Efecto John Henry. Acontece cuando las unidades de la comparación se esfuerzan más para compensar que no se les haya ofrecido el tratamiento. Cuando se comparan las unidades tratadas con las unidades de la comparación que se esfuerzan más, la estimación del impacto del programa tiene un sesgo; es decir, se estima un impacto menor del programa en comparación con el impacto real que se obtendría si las unidades de la comparación no realizaran un esfuerzo adicional.

Efecto local promedio del tratamiento (LATE. *Local Average Treatment Effect*). Impacto de un programa estimado para un subconjunto específico de la población, como las unidades que cumplen con su asignación al grupo de tratamiento o de comparación en presencia de un cumplimiento imperfecto, o en torno al umbral de elegibilidad cuando se aplica un diseño de regresión discontinua. Por lo tanto, el LATE proporciona solo una estimación local del impacto del programa y no debería generalizarse al conjunto de la población.

Efecto de demostración. Las consecuencias que trae a ciertos individuos el observar a otras personas. En la esto puede incluir situaciones en las que los miembros de una comunidad receptora tratan de imitar los estilos de vida, comportamientos, actitudes o idioma de los recién llegados tal

como trabajadores o turistas extranjeros. Esto puede generar muchos impactos sociales negativos, entre los que se encuentran el aumento del costo de vida, la frustración y un problemático cambio cultural. También podría eventualmente causar impactos sociales positivos en forma de transferencia de conocimientos entre los contratistas internacionales y las pequeñas y medianas empresas.

Efectos de equilibrio de contexto. Efectos de derrame que se producen cuando una intervención influye en las normas de conducta o sociales en un contexto determinado, como una localidad tratada.

Efectos de equilibrio general. Estos efectos de derrame se producen cuando las intervenciones afectan la oferta y demanda de bienes y servicios y, por lo tanto, cambian el precio de mercado de esos servicios.

Efecto multiplicador. Véase Multiplicador regional.

Efecto promedio del tratamiento (ATE, por sus siglas en inglés). Impacto del programa bajo el supuesto de cumplimiento total; es decir, todas las unidades que hayan sido asignadas a un programa se inscriben realmente en él, y ninguna de las unidades de comparación recibe el programa.

Efecto tarro de miel. Foco de atracción para migrantes laborales o migración inducida por un proyecto; las personas podrían desplazarse hacia el emplazamiento del proyecto en una tentativa de ser consideradas personas afectadas y por consiguiente con derecho a indemnización, o en busca de trabajo u oportunidades económicas que surjan del proyecto.

Elaboración de perfiles. Proceso de reunir información básica sobre las características de una comunidad y su contexto local en el estado previo al proyecto.

Elefante blanco. Expresión de uso corriente que consiste en una instalación (edificio, etc.) de alto costo y utilidad limitada. Históricamente, la expresión tenía un significado más preciso relacionado con un enorme costo de mantenimiento (un activo que es un pasivo) pero también a la incapacidad de disponer de aquel. Salvo que se tomen las precauciones

necesarias en la selección del gasto para la inversión social, muchos proyectos pueden ser elefantes blancos.

Empobrecimiento. El proceso de empobrecerse (que lleva a la pobreza). La pérdida de medios de subsistencia a partir del desplazamiento causado por los proyectos puede conducir al empobrecimiento.

Empoderamiento. Aumentar los recursos y capacidades de diversos individuos y grupos con el fin de comprometer e influir en las instituciones económicas y sociales, e incrementar la responsabilidad de las instituciones públicas. Un proceso participativo que transfiere o coloca la responsabilidad de la toma de decisiones y los recursos en manos de aquellos que se beneficiarán con ellos. Esto puede incluir (i) la creación de capacidad para las organizaciones interesadas; (ii) el fortalecimiento de la capacidad jurídica de las organizaciones interesadas, (iii) la autoridad de las partes interesadas para administrar fondos, contratar y despedir trabajadores, supervisar el trabajo y adquirir materiales; (iv) la autoridad de las partes interesadas para certificar la realización satisfactoria del proyecto y establecer indicadores de monitoreo y evaluación, y (v) el apoyo a iniciativas nuevas y espontáneas de las partes interesadas.

Encuesta de seguimiento. También conocida como encuesta *posterior a la intervención*. Se trata de una encuesta realizada después de que el programa ha comenzado, una vez que los beneficiarios se han beneficiado de él durante algún tiempo. Una evaluación de impacto puede incluir varias encuestas de seguimiento, a veces denominadas encuestas *intermedias* y *finales*.

Enfoque (o marco) de medios de subsistencia sostenibles. Forma de analizar el efecto de los proyectos con respecto a los medios de subsistencia de las personas y comunidades. Emplea los capitales (activos de medios de subsistencia) como base del análisis.

Enfoque basado en derechos. Véase Enfoque basado en los derechos humanos.

Enfoque basado en los derechos humanos (lente de los derechos humanos) Marco conceptual y operativo orientado hacia asegurar la

promoción y protección de los derechos humanos. Es un enfoque (para la salud, la cooperación en desarrollo, etc.) que procura: (1) ubicar a los derechos humanos y sus principios como el componente central de las acciones; (2) exigir rendición de cuentas y transparencia de parte de los garantes del respeto de los derechos; (3) fomentar el empoderamiento de los titulares de derechos y fortalecer su capacidad para, entre otras cosas, hacer responsables a los garantes referidos; (4) garantizar que la participación significativa de los titulares de derechos en procesos de desarrollo e intervenciones planeadas se reconozca como un derecho intrínseco, no simplemente como mejor práctica; y (5) garantizar el compromiso no discriminatorio de los titulares de derechos y la priorización de individuos o grupos especialmente vulnerables.

Equidad del impacto. La noción de que los impactos que se producen en una sociedad o generados por un proyecto deben ser compartidos de manera equitativa, de que al menos debería prestarse atención a la justa distribución de los impactos positivos y negativos. Por ejemplo, las trayectorias de vuelo para un aeropuerto deberían ajustarse de modo que la carga de ruido sea compartida, en lugar de que la misma gente reciba todo el ruido.

Equipo de evaluación. Equipo que lleva a cabo la evaluación. Se trata en esencia de una asociación entre dos grupos: un equipo de responsables de las políticas públicas (equipo de políticas) y un equipo de investigadores (equipo de investigación).

EMARF (en inglés, SMART). Específico, medible, atribuible, realista y focalizado. Los buenos indicadores tienen estas características.

Error de tipo I. También conocido como falso positivo. Este error se comete cuando se rechaza una hipótesis nula que, en realidad, es válida. En el contexto de una evaluación de impacto, se comete un error de tipo I cuando una evaluación llega a la conclusión de que un programa ha tenido un impacto (es decir, se rechaza la hipótesis nula de falta de impacto), aunque en realidad el programa no ha tenido impacto (es decir,

la hipótesis nula se mantiene). El nivel de significancia es la probabilidad de cometer un error de tipo I.

Error de tipo II. También conocido como falso negativo. Este error se comete cuando se acepta (no se rechaza) la hipótesis nula, pese a que esta última, de hecho, no es válida. En el contexto de una evaluación de impacto, se comete un error de tipo II cuando se llega a la conclusión de que un programa no ha tenido impacto (es decir, no se rechaza la hipótesis nula de falta de impacto), aunque el programa de hecho tuvo impacto (es decir, la hipótesis nula no es válida). La probabilidad de cometer un error de tipo II es 1 menos el nivel de potencia.

ESHIA. Evaluación de Impacto Ambiental, Social y de Salud (“Environmental, Social and Health Impact Assessment”).

Espacio deliberativo. Un entorno físico que conduce al estado deliberativo.

Espíritu de lugar. Hace referencia a los aspectos únicos, distintivos y preciados de un lugar. Mientras el *sentido de pertenencia* implica los sentimientos personales que un individuo tiene acerca de un lugar, el espíritu de lugar se refiere a las características inherentes al lugar.

Estándar de vida. Indicadores físicos, objetivos del bienestar de un individuo o grupo.

Estimación de tratamiento en los tratados. La estimación del impacto del tratamiento en aquellas unidades que en la práctica se han beneficiado del tratamiento. Compárese con intención de tratar.

Estimador. En Estadística, es una regla utilizada para calcular una característica desconocida de una población a partir de los datos (técnicamente conocido como *parámetro*); una estimación es el resultado de la aplicación real de una regla a una muestra concreta de datos.

Estrategia de salida. En el marco de una EIS o de un proyecto, refiere a la atención que una empresa debe prestar a la manera en que se apartará de las obligaciones a largo plazo que no forman parte de su actividad principal.

Estudio de efectividad. Analiza si un programa funciona en condiciones normales al aumentar la escala. Cuando están adecuadamente diseñados e implementados, los resultados de estos estudios pueden ser más generalizables que en el caso de los estudios de eficacia.

Estudio de eficacia. Analiza si un programa puede funcionar en condiciones ideales. Estos estudios se llevan a cabo en circunstancias muy específicas, a menudo con una fuerte participación técnica de los investigadores durante la implementación del programa. Suelen emprenderse para probar la viabilidad de un programa nuevo. Sus resultados no son generalizables más allá del alcance de la evaluación.

Evaluación. Valoración periódica y objetiva de un proyecto, un programa o una política planificados, en curso o finalizados. Las evaluaciones se utilizan para responder preguntas específicas, a menudo relacionadas con el diseño, la implementación o los resultados.

Evaluación de impacto. Evaluación que intenta establecer un vínculo causal entre un programa o intervención y un conjunto de resultados. Una evaluación de impacto procura responder a la pregunta: ¿cuál es el impacto (o efecto causal) de un programa en un resultado de interés? Proceso de identificar las consecuencias futuras de una acción actual o propuesta.

Evaluación de impacto ambiental (EIA). Proceso formal empleado para predecir las probables consecuencias ambientales (positivas o negativas) de un plan, política, programa o proyecto previo a su implementación, a menudo como parte del procedimiento normativo (licenciamiento ambiental).

Evaluación de impacto cultural. Forma de evaluación de impacto que considera los impactos de un proyecto específicamente sobre la cultura de un grupo social en particular (como un grupo étnico o indígena específico). Consideraría, entre otras cosas, los valores, sistemas de creencias, leyes no escritas, lengua(s), costumbres, economía, relaciones con el medio ambiente local y especies particulares, organización social y tradiciones de la comunidad afectada. Dado que los impactos culturales debe-

rían formar parte de una evaluación de impacto social para la introducción de innovación, la evaluación de impacto cultural es un subcomponente de ésta, pero está estrechamente vinculada a otros impactos sociales con lo cual esa distinción no tendría sentido, excepto en la medida que indique la perspectiva y el propósito de la evaluación de impacto; esto es, la evaluación de impacto social par ala introducción de innovaciones que apunta a comprender los impactos sociales de un proyecto sobre la cultura de la comunidad local.

Evaluación de impacto de igualdad (EqIA). Evalúa aspectos de equidad y discriminación en todas las políticas y estrategias, especialmente con respecto a los grupos vulnerables.

Evaluación de impacto sobre el patrimonio cultural. Proceso de evaluar los impactos posibles de un proyecto propuesto sobre las manifestaciones materiales del patrimonio cultural de una comunidad que incluye sitios, estructuras y restos de valor o importancia arqueológica, arquitectónica, histórica, religiosa, espiritual, cultural, ecológica o estética. Los impactos sobre el patrimonio cultural intangible se evaluarían en un estudio de impacto cultural.

Evaluación de impacto social. Comprende los procesos de analizar, monitorear y gestionar las consecuencias sociales deseadas y no deseadas, tanto positivas como negativas, de las intervenciones planeadas (políticas, programas, planes, proyectos) y todo proceso de cambio social invocado por tales intervenciones. Su principal objetivo es lograr un ambiente biofísico y humano más sostenible y equitativo. Evaluación de impacto social (forma abreviada) analizar, monitorear y gestionar las consecuencias sociales del desarrollo.

Evaluación de necesidades. Procedimiento sistemático para determinar problemas del cliente o comunitarios y clasificar su importancia como componente del programa de desarrollo. La evaluación de necesidades es el predecesor de los programas de intervención pública actuales.

Evaluación ex-ante. Evaluacion anticipada. La mayoría de las evaluaciones de impacto se realizan exante, constituyen un pronóstico sobre

los probables impactos de una intervención planeada; en otras palabras, acerca de algo que todavía no ha sucedido.

Evaluación ex-post. Evaluación posterior al proyecto. Las evaluaciones ex-post son en realidad evaluaciones de los impactos de un proyecto o política en particular.

Evaluación de proceso. Evaluación que se centra en cómo se implementa y funciona un programa, y que evalúa si se corresponde con su diseño original y documenta su desarrollo y funcionamiento. Compárese con evaluación de impacto.

Evaluación prospectiva. Evaluaciones diseñadas y aplicadas antes de que se implemente un programa. Las evaluaciones prospectivas están incorporadas en los planes de implementación del programa. Compárese con evaluación retrospectiva.

Evaluación retrospectiva. Evaluación diseñada después de que se ha implementado un programa (ex post). Compárese con evaluación prospectiva.

Exclusión social. Procesos que retrasan el logro de la inclusión social y de la integración social, o que conducen a la marginalización de grupos vulnerables.

Experimento de mecanismo. Evaluación de impacto que prueba un mecanismo causal particular dentro de la teoría del cambio de un programa, en lugar de probar el efecto causal (impacto) del programa como un todo.

Expropiar/expropiación. Se refiere a la capacidad de un organismo del sector público, o de una empresa autorizada por el gobierno, para adquirir tierras y otros recursos, sin el consentimiento de los usuarios existentes o residentes. (Véase también Dominio eminente).

Externalidad. Un término económico que se refiere a los costos (y por ende a los impactos) que se consideran externos o ajenos a los análisis que se están haciendo, por consiguiente se trata por lo general de los impactos ambiental y social. Un principio de la sostenibilidad es internalizar las externalidades.

Facilitación. El proceso de posibilitar que grupos y organizaciones alcancen sus metas, ayudándolos en los procesos que utilizan para colaborar entre sí. Un facilitador es habitualmente independiente, está capacitado y posee experiencia en tareas de facilitación, y cuenta con un repertorio de técnicas para emplear, seleccionadas conforme al objetivo y a los intereses del grupo al que está dirigida esa facilitación. En una evaluación de impacto social para introducción de innovaciones y/o proceso de compromiso comunitario, es un facilitador el que normalmente maneja los procesos inherentes al compromiso comunitario. Las principales aptitudes son la imparcialidad, una capacidad para hacer que las personas se sientan cómodas, una buena comprensión del proceso social y un buen conocimiento de una amplia variedad de técnicas y de cuándo utilizar cada una.

Facipulación. Una palabra inventada que combina las palabras facilitación y manipulación. Se refiere al sentimiento que tienen las personas cuando han intervenido en un proceso de participación comunitaria pero se quedaron con la sensación de haber sido manipuladas.

Factor invariante en el tiempo. Factor que no varía a lo largo del tiempo; es constante.

Factor variante en el tiempo. Factor que varía a lo largo del tiempo.

Falta de rango común. Cuando se utiliza el método de coincidencia (*matching*), la falta de rango común es una falta de superposición entre los puntajes de propensión del grupo de tratamiento, o inscrito, y los del grupo de no inscritos.

Falta de respuesta. Se produce cuando faltan datos o los datos son incompletos para algunas unidades de la muestra. La falta de respuesta de la unidad surge cuando no hay información disponible para algunas unidades de la muestra, es decir, cuando la muestra real es diferente de la muestra planificada. Una forma de falta de respuesta a nivel de la unidad es el desgaste. La falta de respuesta de una entrada se produce cuando los datos son incompletos para algunas unidades de la muestra en un deter-

minado momento del tiempo. La falta de respuesta puede generar sesgos en los resultados de una evaluación si está asociada con la condición de tratamiento.

Falta de respuesta de la unidad. Surge cuando no hay información disponible para un subconjunto de unidades; es decir, cuando la muestra real es diferente de la muestra planificada.

Falta de respuesta de una entrada. Ocurre cuando los datos son incompletos para algunas unidades de la muestra.

Fecha límite. En los procesos de reasentamiento indica la fecha luego de la cual la población no será incluida en la lista de personas identificadas como afectadas por un proyecto, y por consiguiente no tendrán derecho a asistencia e indemnización por reasentamiento.

FIFO. Una sigla referida al uso de trabajadores que normalmente no residen en la vecindad del proyecto y que por lo general son introducidos como trabajadores, habitualmente por determinados períodos de tiempo. Se origina en la expresión en inglés *fly in, fly out* (llevar y traer en avión).

Fondo social. La provisión de financiamiento (quizás, pero no necesariamente en carácter de indemnización) proveniente de un proyecto para ser dirigida a las poblaciones afectadas. Habitualmente los administra la comunidad y se destinan a la promoción de proyectos sociales en beneficio de la comunidad toda. Con frecuencia se calculan como un porcentaje de algo, tal como dólares por megavatio hora de electricidad producida, o millar de toneladas de mineral extraído de una mina. Un fondo social no es un pago en concepto de regalías, aunque los pagos de regalías también podrían ingresar a los fondos sociales.

Fortalecimiento de la capacidad. Proceso coordinado de intervenciones, tales como programas de capacitación usualmente centrados en fortalecer capital humano y en mejorar prácticas institucionales y mecanismos de gobernanza.

Garante. En el enfoque basado en los derechos humanos, los derechos humanos conllevan tanto derechos (y por ende titulares de los derechos) como obligaciones (y en consecuencia, garantes de esos derechos). En la legislación de derechos humanos, los garantes tienden a ser principalmente Estados, pero también podrían incluir a todos los individuos, en particular las empresas y sus proveedores y contratistas. En el marco del derecho internacional, los Estados asumen obligaciones y deberes de respetar, proteger y cumplir con los derechos humanos. La obligación de respetar significa que los Estados deben abstenerse de restringir o interferir con el ejercicio de los derechos humanos. La obligación de proteger requiere que los Estados protejan a los individuos y grupos contra las violaciones de los derechos humanos. La obligación de cumplir significa que los Estados deben adoptar medidas firmes para facilitar el ejercicio de los derechos humanos básicos. A nivel individual, aunque estemos facultados para ejercer nuestros derechos humanos, también debemos respetar los derechos humanos de los demás.

Garantía ambiental. Un instrumento financiero (a menudo en forma de cuenta de depósito de garantía) que proporciona una garantía que asegure que un proyecto cumplirá con los requisitos de rehabilitación ambiental.

Gastos iniciales. Un término propio de la gestión de proyectos que implica gastar más dinero al comienzo de modo de contar con un mejor diseño y por consiguiente ahorrar dinero en una etapa posterior.

Generabilidad. La medida en que los resultados de una evaluación en un ámbito local serán válidos en otros contextos y en otros grupos de población.

Género. Se refiere a los roles socialmente construidos y asignados a hombres y mujeres. Estos roles se aprenden, cambian con el tiempo y varían ampliamente dentro de una misma cultura y entre culturas diferentes.

Gentrificación. El proceso gradual por el cual un lugar (por lo general un barrio céntrico pobre o una pequeña localidad ubicada en un lugar estético) se transforma, dejando de pertenecer a la clase trabajadora y

pasando a ser de clase media o clase media alta en cuanto a sus características y composición. Una consecuencia de este proceso es un aumento del valor de las propiedades y del costo del alquiler de las viviendas, lo cual provoca el desplazamiento de los residentes anteriores. Cuando la gentrificación conduce a la turistificación, también puede llevar a un conflicto entre los residentes establecidos (antiguos) y los turistas y/o los nuevos residentes (recién llegados).

Gestión de flujos migratorios. El proceso de manejo de grandes cantidades de personas que llegan a los emplazamientos de los proyectos en búsqueda de oportunidades económicas. (Véase *Efecto tarro de miel*)

Gobernanza. Hace referencia al modo de gobernar de organizaciones, instituciones, empresas y gobiernos. Es el acto de gobernar, el que implica la aplicación de leyes y reglamentos, pero también costumbres, estándares éticos y normas. (Véase también *Buena gobernanza*)

Grupo de comparación. También conocido como grupo de control. Un grupo de comparación válido tendrá las mismas características, en promedio, que el grupo de beneficiarios del programa (grupo de tratamiento), con la única diferencia de que las unidades del grupo de comparación no se benefician del programa que se evalúa. Los grupos de comparación se utilizan para estimar el contrafactual.

Grupo de control. También conocido como grupo de comparación (véase la definición).

Grupo de tratamiento. También conocido como grupo tratado o grupo de intervención. El grupo de tratamiento es el grupo de unidades que es objeto de una intervención versus el grupo de comparación, que no es objeto de ella.

Grupos marginados. Véanse *Grupos minoritarios* y *Grupos vulnerables*.

Grupos minoritarios. Un término de las ciencias sociales empleado para referirse a agrupaciones sociales que se diferencian o pueden diferenciarse de la mayoría de la sociedad. Por lo general sufren discriminación y subordinación, tienen rasgos físicos y/o culturales que los apartan, y en razón de los cuales son marginados por el grupo dominante, un sentido

compartido de identidad colectiva y preocupaciones en común; reglas sociales compartidas acerca de quién pertenece y quién no determina la condición de minoría; y la tendencia a contraer matrimonio dentro del grupo.

Grupos vulnerables. Grupos caracterizados por alguna vulnerabilidad. Si bien esta depende del contexto y puede incluir una amplia gama de grupos, en general el concepto abarca: pueblos indígenas, minorías étnicas, migrantes, personas discapacitadas, personas sin hogar, pobres, quienes luchan contra el abuso de sustancias y personas mayores aisladas.

Guardián. Una persona o institución que controla el acceso a algo. Los guardianes pueden ejercer roles formales o informales. En el contexto de una evaluación de impacto social para la introducción de innovaciones, el concepto indica con frecuencia aquellos individuos que tienen poder para facilitar o impedir el acceso de los consultores a una comunidad específica. Dicho de otro modo, ocupan una posición influyente clave.

Habitabilidad. Los aspectos de un lugar que hace que sus habitantes estén felices de vivir y trabajar allí, y que ofrecen una alta calidad de vida para todos sus habitantes.

Hipótesis. Explicación propuesta de un fenómeno observable. Véase también hipótesis nula e hipótesis alternativa.

Hipótesis alternativa. Suposición de que la hipótesis nula es falsa. En una evaluación de impacto, la hipótesis alternativa suele ser la hipótesis de que la intervención tiene un impacto en los resultados.

Hipótesis nula. Hipótesis que puede ser falsificada sobre la base de los datos observados. Normalmente, la hipótesis nula propone una posición general o por defecto. En la evaluación de impacto, la hipótesis nula suele ser que el programa no tiene impacto, es decir: la diferencia entre el resultado del grupo de tratamiento y el grupo de comparación es cero.

Historial de impacto. Se refiere a la experiencia previa que una comunidad ha tenido con otros proyectos. Afecta la manera en que se relacionan con nuevos proyectos y el nivel de confianza que podrían tener.

Significa también que podrían existir cuestiones heredadas de las cuales un operador nuevo debe encargarse.

Huella social. Un concepto que intenta ser el equivalente social de la *huella ecológica*, es decir una metáfora con referencia al grado de daño social generado por un proyecto o producto. Es un concepto que no se ve propiciado por los profesionales de las ciencias sociales y por ende no se aplica en la evaluación de impacto social para la introducción de innovaciones, pero algunos profesionales de las ciencias físicas promueven este concepto junto con el de huella ecológica.

IAIA. Asociación Internacional para la Evaluación de Impactos (International Association for Impact Assessment) www.iaia.org.

ICMM. Consejo Internacional de Minería y Metales (International Council on Mining and Metals) www.icmm.com.

Impacto. También conocido como efecto causal. En el contexto de las evaluaciones de impacto, un impacto es un cambio directamente atribuible a un programa, a una modalidad del programa o a innovaciones de diseño. Consecuencia económica, social, ambiental u otra que se puede predecir razonablemente y medir previamente en caso de implementarse una acción propuesta.

Impacto (efecto) indirecto. Ocurre como resultado de otro cambio causado por una intervención planeada. En el ámbito de la evaluación de impacto social para la introducción de innovaciones, un efecto indirecto podrá ser causado por un cambio físico en el medio ambiente. Por ejemplo, la mina podrá aumentar la turbiedad del río, lo que podrá reducir el abastecimiento de peces, lo que puede llegar a disminuir los medios de vida económicos de los pescadores de ese pueblo. También pueden existir efectos secundarios, efectos de segundo o mayor grado.

Impacto (o efecto) directo. Ocurre como resultado directo de una intervención planeada. También denominado impacto principal o de primer orden. En la evaluación de impacto social para introducción de innovaciones refiere a los cambios sociales e impactos sociales directamente

causados por el proyecto en sí, como las molestias que ocasiona a las personas el ruido de maquinaria del proyecto.

Impacto de orden superior. Los impactos sociales indirectos que se producen luego de los impactos de primer orden inmediatamente anteriores en la cadena de impactos que surgen de un proyecto.

Impacto de primer orden. Los impactos que constituyen las consecuencias directas e inmediatas de un proyecto.

Impacto percibido. Algo que se cree que es un posible impacto en vez de algo que se haya establecido como un impacto real. Nótese que los impactos percibidos afectan lo que las personas sienten con respecto al proyecto y cómo se sienten y se comportan en general, por lo que la percepción para ellos es real.

Impacto potencial. Impacto que se ha predicho, y no un impacto real que ya haya ocurrido.

Impacto real. Los impactos sociales que las comunidades sufren efectivamente y no los que se pronostica que ocurrirán.

Impacto residual. Los impactos adversos predichos y que permanecen incluso luego de que se han implementado medidas de mitigación.

Impacto social. Algo que se experimenta o se siente, en el sentido perceptual o corporal a nivel de individuo, unidad social (familia/hogar/colectividad) o comunidad/sociedad. (Véase también *Proceso de cambio social*)

Impactos acumulativos. Los impactos sucesivos, incrementales y combinados de uno o más proyectos (existentes, actuales y futuros proyectos predecibles) sobre la sociedad, la economía o el medio ambiente. Pueden resultar de una agregación o interacción de impactos dentro de un sistema social o ambiental y se definen desde la perspectiva de las personas o del medio ambiente que los experimenta.

Impactos transfronterizos. Se refiere a los impactos, ambientales y sociales, que se trasladan más allá de las fronteras, en general fronteras nacionales, pero se puede decir también cualquier frontera jurisdiccional.

Importancia (evaluación determinación). El acto de asignar alguna forma de prioridad a los problemas que deben tratarse, para su posterior

análisis y mitigación. Luego del proceso de determinación del alcance, se procede a evaluar los impactos según su importancia, de conformidad con algunos criterios predeterminados o a través de un proceso deliberativo con una comisión de enlace con la comunidad.

Importancia relativa. Concepto jurídico que hace referencia a si algo es relevante con respecto al asunto por tratar. Por ejemplo, puede referirse a lo que debería divulgarse en una negociación amistosa. También puede referir a lo que podría comprender un informe sobre sostenibilidad.

Inclusión social. Concepto de la justicia social que refiere a un compromiso de políticas y a las estrategias activas del gobierno, en todos sus niveles, y de la sociedad civil para mejorar el acceso de individuos y grupos diversos a oportunidades de desarrollo y a una participación íntegra en la sociedad mediante la eliminación de limitaciones institucionales y la provisión de incentivos y mecanismos.

Indemnización. En los lugares en los que los impactos no se pueden evitar, la indemnización implica restituir a las personas, ya sea individual o colectivamente. Puede ser en forma de pagos en efectivo o de suministro de otras actividades de desarrollo, como la de un hospital, una escuela o una biblioteca pública. Si bien la indemnización puede ser la respuesta que surge de los derechos de propiedad de la comunidad afectada, también la puede realizar el proponente como gesto de buena voluntad o como resultado de una negociación.

Indicador. Variable que mide un fenómeno de interés para el evaluador. El fenómeno puede ser un insumo, un producto, un resultado, una característica o un atributo. Véase también EMARF. Véase también *indicador social*.

Índice de elegibilidad. También conocido como variable forzada. Se trata de una variable que permite clasificar a la población de interés a lo largo de una línea continua y tiene un umbral o una puntuación límite que determina quién es elegible y quién no lo es.

Indicador social. Medida estadística (variable) empleada para monitorear el cambio en un fenómeno social. La evaluación de impacto social

para la introducción de innovaciones identifica indicadores sociales para todos los problemas sociales como temas importantes que necesitan ser monitoreados.

Indicadores clave desempeño. Se establecen para medir el desempeño de una empresa o persona. Los indicadores desempeño de monitoreo por lo general no se consideran indicadores clave de, pero la idoneidad de ese monitoreo debería constituir uno para el directivo responsable de este; y la ausencia de incumplimientos debería ser otro indicador clave de desempeño para una empresa.

Indignación moral. Indignación que provoca la percepción o creencia que se ha violado alguna norma moral, como puede ser la norma de imparcialidad o justicia.

Infraestructura compartida. Se refiere a la infraestructura construida para el proyecto, pero que también se pone al servicio de las necesidades de las comunidades locales. Puede hacer referencia a la generación de electricidad, aprovisionamiento de agua, tratamiento de aguas servidas, como asimismo para puentes, carreteras, líneas ferroviarias, puertos y aeropuertos.

Infraestructura de la comunidad. Servicios e instalaciones públicas y privadas que contribuyen a la calidad de vida general (esto es, salud, transporte, energía, educación, agua, calidad del agua y servicios sanitarios).

Inmigración inducida por un proyecto. Véase *Efecto tarro de miel*.

Innovación social. Un discurso sobre las nuevas maneras de satisfacer las necesidades sociales de las comunidades o de entregar a estas beneficios sociales, a través de un rediseño o la creación de nuevos productos, servicios, estructuras orgánicas, estructuras de gobernanza, políticas, procedimientos y actividades que sean más eficaces que los métodos tradicionales ya existentes del sector público, filantrópicos y dependientes del mercado que se aplican para responder a la exclusión social.

Insumos. Los recursos financieros, humanos y materiales utilizados en la intervención. Intención de tratar (ITT, por sus siglas en inglés). Las estimaciones de ITT miden la diferencia en los resultados entre las unidades

asignadas al grupo de tratamiento y las asignadas al grupo de comparación, independientemente de si las unidades de cada grupo recibieron en realidad el tratamiento. **Intervención.** En el contexto de la evaluación de impacto, se trata del proyecto, del programa o de la política que se evalúa. También conocida como tratamiento.

Integración social. La capacidad de los diferentes grupos de la sociedad para vivir juntos en armonía productiva y cooperativa, y para zanjar diferencias en un marco de interés común en beneficio de todos. La integración social implica justicia para cada individuo y armonía entre los diferentes grupos sociales y países. Significa la integración de grupos desfavorecidos y vulnerables, haciendo que todas las instituciones de la sociedad sean más accesibles para ellos.

Intervención planeada. Proyecto, plan, política o programa. Fundamentalmente, cualquier medida estimada que busca lograr una meta o un resultado definido.

Inversión extranjera directa. Un término comercial referido a las inversiones relacionadas con la obtención de una participación mayoritaria (que por lo general se considera el 10% o más del paquete accionario) en una empresa comercial en un país, por parte de una persona jurídica con sede en otro país.

Inversión social. En el marco de una evaluación de impacto social (EIS), un proyecto debería aportar al desarrollo local facilitando financiamiento para proyectos que contribuyan con los resultados en materia de desarrollo local. La inversión social estratégica se emplea cuando para la empresa que facilita los fondos existe una clara viabilidad comercial.

Inversión social estratégica. Específicamente diseñada para lograr tanto los propósitos estratégicos del proyecto o empresa como las necesidades y aspiraciones de las comunidades locales.

IPIECA. La asociación mundial del sector del petróleo y del gas especializada en cuestiones medioambientales y sociales <http://www.ipieca.org/>.

Jerarquía de la mitigación. Marco de referencia para planificar medidas de mitigación: evitar, reducir, restablecer e indemnizar. En definitiva, evitar en la fuente, minimizar en la fuente, reducir en el sitio, reducir en el receptor, reparar, indemnizar en especie, indemnizar por otros medios.

Junta de revisión institucional (JRI). Comité nombrado para examinar, aprobar y monitorear la investigación con sujetos humanos. También conocido como Comité de ética independiente o Junta de revisión ética.

Jurado ciudadano. Técnica deliberativa en la que la decisión (quizás sobre seleccionar la mejor alternativa) la toma un panel de (alrededor de 12) ciudadanos legos seleccionados entre el público a quienes se les ha confiado que deliberen sobre temas relevantes en nombre de la comunidad. Si bien la intención es que el panel sea lego y no experto, se espera que aprendan sobre los temas relevantes y consulten a expertos y pidan información según sea necesario. Las decisiones tomadas por un jurado ciudadano tienen mayor legitimidad entre los miembros de la comunidad local que aquellas tomadas en procesos a cargo de expertos.

Justicia social. Se refiere a nociones de justicia y equidad en una sociedad. Se trata de una filosofía acerca del respeto por los derechos humanos, el concepto de que cada uno debería tener la oportunidad de mejorar, y de que deberían tener la posibilidad de participar en decisiones que afectan su propia vida.

Legitimidad. Un concepto que significa que las acciones de una parte son consideradas por un individuo o grupo como convenientes, correctas, apropiadas o al menos aceptables desde el punto de vista normativo de la otra parte. La legitimidad puede interpretarse de varias maneras, entre ellas la legitimidad jurídica, legitimidad política, legitimidad moral y legitimidad social. Por sí sola, al menos en el contexto de una evaluación de impacto social para la introducción de innovaciones, en general por *legitimidad* se entiende *legitimidad social*; la medida en que una acción es aceptable socialmente.

Ley común. La legislación y la interpretación de las leyes que existirían en una comunidad según se manifiestan a través de las decisiones judiciales (precedentes) de los jueces. Se puede contrastar con el derecho legislado, y está legitimado en países cuyos sistemas jurídicos derivan del sistema inglés. Los sistemas continentales europeos (esto es, derecho romano y francés o código napoleónico) no reconocen la Ley Común (*Common law*). De alguna manera, es como un derecho consuetudinario.

Ley blanda. Se refiere a los procesos informales de control de las actividades de una empresa, tal como las normas y directivas de la industria. Hace referencia también a acuerdos, convenios, etc. a nivel mundial que influyen en las prácticas de las empresas y a la manera en que son juzgadas o percibidas.

LGBT/LGBTIQ. La comunidad de lesbianas, homosexuales, bisexuales, transexuales, intersexuales y queer. Algunas variantes de esta sigla en inglés incluyen también *en duda* (*questioning*). LGBT, LGBT+ y LGBTIQ son las siglas inclusivas habituales que se emplean en las discusiones sobre la orientación sexual y la identidad de género (SOGI, por sus siglas en inglés). Quienes pertenecen a este colectivo se ven con frecuencia discriminados, y por consiguiente se encuentran entre los grupos marginados y vulnerables de la sociedad.

Licencia ambiental. Permiso administrativo expedido por una autoridad competente mediante el cual se autoriza a un operador de una actividad productiva u obra de infraestructura a llevar a cabo acciones para las cuales se solicita la licencia, si bien quizás dependiendo de si se cumplan ciertas condiciones operativas, se respeten ciertos límites de uso y se implementen ciertas medidas para contener, minimizar y evitar cualquier impacto social o ambiental que la actividad o la obra puedan causar.

Licencia social para operar. Expresión popular que implica que la aceptación por parte de la comunidad también es necesaria para el éxito de un proyecto.

Licencia social para operar y crecer. Variante de *licencia social para operar* que pone énfasis en que la aceptación de todos los actores también es necesaria para que el negocio se amplíe.

Línea de base. Situación previa a una intervención, con respecto a la cual se puede valorar el progreso o se pueden hacer comparaciones. La línea de base se recopila en forma previa a la implementación de un programa o política para observar la situación antes. La disponibilidad de datos de línea de base es fundamental para documentar el equilibrio en las características anteriores al programa entre los grupos de tratamiento y de comparación. Los datos de línea de base son necesarios para algunos diseños cuasi experimentales. Los datos de una serie de indicadores seleccionados que se miden cerca del inicio del proyecto y que permiten hacer un monitoreo de un cambio en el tiempo. Se convierten en un punto de referencia, junto a otros valores de referencia, frente a los cuales se pueden comparar situaciones futuras. Si bien la línea base original refiere a un punto específico en el tiempo, el perfil comunitario debería destacar tendencias en el zona del proyecto para que se pueda comparar entre lo que probablemente hubiera ocurrido con y sin ese proyecto.

Línea base social. Véase *Línea de base*.

Lugar. Espacio geográfico que tiene un significado (sentimiento negativo tanto como positivo) para individuos o grupos por experiencias personales y/o por las relaciones que estos tienen con aquel.

Mapa mental. Un diagrama que es una manera de presentar ideas y pensamientos; la creación de mapas mentales es un proceso de organización de información e ideas.

Marco de políticas de reasentamiento. Una política o guía operativa para el proyecto acerca de la manera en que deberán abordarse las cuestiones de la adquisición de tierras, reasentamiento, indemnización y recuperación de los medios de vida durante todo el transcurso del proyecto.

Marco muestral. Lista exhaustiva de las unidades de la población de interés. Se requiere un marco muestral adecuado para asegurar que las

conclusiones a las que se arribe a partir del análisis de una muestra se puedan generalizar a toda la población. Las diferencias entre el marco muestral y la población de interés crea un sesgo de cobertura. Ante la presencia de dicho sesgo, los resultados de la muestra no tienen validez externa para toda la población de interés.

Marginación. Procesos sociales y económicos que empeoran la situación de los grupos minoritarios o vulnerables.

Mecanismo comunitario de reclamos (Community grievance mechanism). Específicamente diseñado para que accedan a este los miembros de las comunidades afectadas por un proyecto.

Mecanismo de reclamos. Un proceso formal, judicial o extrajudicial para el tratamiento de reclamos, al que pueden acceder individuos, trabajadores, comunidades y entidades de la sociedad civil que han sido perjudicados, o sienten que lo han sido, por las actividades de un proyecto o empresa.

Medio ambiente. Un concepto muy vago que se define de diferentes maneras en diferentes circunstancias. En algunas jurisdicciones, incluye los ecosistemas y las partes que los constituyen, incluidas las personas y las comunidades; recursos naturales y físicos; las calidades y características de las ubicaciones, lugares y áreas; y los aspectos sociales, económicos, culturales, estéticos y patrimoniales de todos estos elementos. En otras jurisdicciones, el medio ambiente se refiere solamente a los elementos biofísicos tal como el agua, el aire, el suelo, la flora y la fauna.

Medio de subsistencia. Un medio de subsistencia comprende las capacidades, bienes (véase *Capitales*) y sostenibles actividades requeridas para un medio de subsistencia. Un medio de vida es sostenible cuando puede hacer frente y recuperarse de tensiones y conmociones y mantener o mejorar sus capacidades y recursos tanto ahora como en el futuro, sin socavar la base de los recursos naturales.

Medios de subsistencia. Hace referencia al modo de vida de una persona o de un hogar y a la manera en que se ganan la vida, en particular el modo en que cubren las necesidades básicas, por ejemplo, sus alimentos, agua,

abrigo y vestimenta, y viven en comunidad. (Véase *Medios de subsistencia sostenibles*).

Medios de subsistencia alternativos. También llamados oportunidades económicas alternativas, se refieren al proceso de identificar, seleccionar y desarrollar una variedad de actividades generadoras de ingresos para reemplazar o aumentar las actividades corrientes que proporcionan medios de subsistencia a las personas afectadas por un proyecto. Esto tiene particular importancia en el caso de un desplazamiento económico o físico, pero también pueden integrar los programas de creación de beneficios o de inversión social del proyecto.

Mejoramiento. Medidas tomadas en el diseño y las fases subsiguientes de los proyectos de modo de garantizar la obtención de una amplia variedad de resultados directos e indirectos en materia de desarrollo.

Mejores prácticas (*Best practice*). El conjunto de lineamientos, ética, ideas, procedimientos y métodos que representan la mejor (más adecuada) forma de actuar en un sector determinado. Si bien la mejor práctica sería la indicada por una asociación profesional, en general el concepto es vago y refiere a un punto de referencia que se aspira alcanzar.

Mentalidad asocial. Actitud según la cual las personas no cuentan, o que los problemas sociales no son importantes y no es necesario considerarlos.

Mercado emergente. Otra forma de referirse a un país en desarrollo.

Método cuasi experimental. Métodos de evaluación de impacto que no dependen de la asignación aleatoria del tratamiento. Las diferencias en diferencias, el diseño de regresión discontinua y el de coincidencia (*matching*) son ejemplos de métodos cuasi experimentales.

Método de control sintético. Un método de coincidencia (*matching*) específico que permite estimar el impacto en contextos donde una única unidad (como un país, una empresa o un hospital) es objeto de una intervención o es expuesto a un suceso. En lugar de comparar esta unidad tratada con un grupo de unidades no tratadas, el método utiliza información sobre las características de la unidad tratada y las unidades no

tratadas para construir una unidad de comparación sintética o artificial, ponderando cada unidad no tratada de tal manera que la unidad de comparación sintética se parezca todo lo posible a la unidad tratada. Esto requiere una larga serie de observaciones a lo largo del tiempo, tanto de las características de la unidad tratada como de las unidades no tratadas. Esta combinación de unidades de comparación en una unidad sintética proporciona una mejor comparación para la unidad tratada que cualquier unidad no tratada individualmente.

Métodos mixtos. Enfoque analítico que combina datos cuantitativos y cualitativos.

Microfinanzas. Variedad de servicios bancarios (en general, la provisión de pequeños préstamos) para asistir a personas de bajos ingresos (en particular mujeres) para establecer pequeñas empresas. Los contratos de microfinanzas son muy importantes porque proveen fondos a personas que de otra manera no acceden al capital debido al pequeño valor del préstamo requerido, sus bajos ingresos, la falta de garantía y antecedentes de préstamos muy pobres o no existentes.

Miedo al delito. Se refiere a un miedo anormal que la gente siente ante la posibilidad de ser víctima de un delito. Un cambio rápido en una comunidad puede incrementar el miedo de las personas a sufrir un delito, en modo tal que su miedo sea desproporcionado con respecto a la posibilidad real del delito. Es un estado muy debilitante, porque modifica el comportamiento de las personas, afecta su sentimiento con relación a su comunidad y afecta su bienestar general.

Minería de datos. Práctica de manipular los datos en busca de resultados concretos.

Mitigabilidad. Capaz de ser mitigado.

Mitigación. Proceso de concebir e implementar procesos, procedimientos y/o cambios a una intervención planeada con el fin de evitar, reducir o minimizar, o para compensar, impactos que probablemente ocurran.

Monitoreo. Proceso continuo de recopilar y analizar información para evaluar el desempeño de un proyecto, un programa o una política. El

monitoreo suele hacer un seguimiento de los insumos, actividades y productos, aunque ocasionalmente también incluye los resultados. Se utiliza para fundamentar la gestión y las decisiones diarias. También se puede emplear para hacer un seguimiento del desempeño en relación con los resultados previstos, establecer comparaciones entre programas y analizar las tendencias a lo largo del tiempo. Puede referir a un proceso de control de cumplimiento de las condiciones necesarias para obtener el consentimiento para llevar a cabo una intervención planeada, pero en general hace referencia al proceso de prueba continuo que determina que no haya impactos no anticipados.

Monitoreo participativo. La participación de actores en actividades de monitoreo y la verificación de información para garantizar la legitimidad del proceso de monitoreo y del proyecto en su conjunto.

Muestra aleatoria. Muestra extraída a partir de un muestreo probabilístico, por lo cual cada unidad en el marco muestral tiene una probabilidad conocida de ser extraída. Seleccionar una muestra aleatoria es la mejor manera de evitar una muestra no representativa. El muestreo aleatorio no debería confundirse con la asignación aleatoria.

Muestra conglomerada. Una muestra compuesta de clústers.

Muestra estratificada. Se obtiene dividiendo la población de interés (marco muestral) en grupos (por ejemplo, hombres y mujeres) y posteriormente, definiendo una muestra aleatoria en cada grupo. Una muestra estratificada es una muestra probabilística: todas las unidades de cada grupo (o estrato) tienen la misma probabilidad de ser asignadas. Siempre que todos los grupos sean lo bastante grandes, el muestreo estratificado permite elaborar inferencias a propósito de los resultados no solo a nivel de la población sino también dentro de cada grupo.

Muestra. En Estadística, una muestra es un subconjunto de una población de interés. Normalmente, la población es muy grande, lo cual hace impracticable o imposible realizar un censo o un registro completo de todos sus valores. En cambio, los investigadores pueden seleccionar un subconjunto representativo de la población (utilizando un marco mues-

tral) y recopilar estadísticas sobre la muestra. Estas se pueden utilizar para hacer inferencias o para extrapolar a la población. Este proceso se conoce como muestreo. Compárese con censo.

Muestreo. Proceso por el cual las unidades se extraen del marco muestral creado a partir de la población de interés. Se pueden utilizar diversas alternativas de procedimientos de muestreo. Los métodos de muestreo probabilístico son los más rigurosos, ya que asignan una probabilidad bien definida a cada unidad que será extraída. El muestreo aleatorio, el muestreo aleatorio estratificado y el muestreo conglomerado son métodos de muestreo probabilístico. El muestreo no probabilístico (por ejemplo, el muestreo intencional o por conveniencia) puede generar errores de muestreo.

Muestreo probabilístico. Proceso de muestreo que asigna una probabilidad bien definida a cada unidad que será extraída de un marco muestral. Incluye el muestreo aleatorio, el muestreo aleatorio estratificado y el muestreo de clusters.

Multiplicador regional. Término económico que refiere al ratio de la magnitud del efecto total sobre la economía regional de un estímulo inicial específico (como un proyecto) relativo a la magnitud del impacto directo; de otro modo, el alcance de la amplificación de una inversión local en una economía local.

Mundo de la vida. Un concepto (en alemán, *Lebenswelt*) de las ciencias sociales que se refiere a las experiencias vividas por las personas y a sus vidas cotidianas. Implica tener en cuenta su perspectiva en el análisis o narrativa que se presente.

NIMBY. Acrónimo que significa *no en mi patio trasero*” también se lo conoce en español como **SPAN** o *sí, pero aquí no*. Califica a la reacción que se recibe a menudo a la hora de decidir la localización de usos de suelos localmente no deseados (**LULU**, por su acrónimo en inglés), como aeropuertos, vertederos, centrales eólicas, etc.

Norma. Estándar, convención o resolución que se emplea como referencia u objetivo.

Norma de desempeño. Concepto genérico que articula el estándar previsto de una práctica o un logro. En el contexto de una evaluación de impacto social para introducción de innovaciones, se refiere normalmente a las Normas de Desempeño sobre Sostenibilidad Ambiental y Social de la Corporación Financiera Internacional (**CFI**).

Normalización. Intento por evitar que se desarrolle una ciudad industrial o de que se cree un sistema de valores de una ciudad de crecimiento rápido, o por darle a una ciudad industrial una vida comunitaria más normal.

Normativo. En una perspectiva o juicio normativo se afirma como tienen que ser las cosas, o como deberían ser, como las valoramos, qué es bueno o malo y qué medidas están bien o mal; a menudo en referencia a un principio ético, o se puede decir en referencia a un código o estándar internacional que va más allá de los requisitos jurídicos (si fuera requerido jurídicamente, no sería un requisito normativo).

Oficina del Asesor en Cumplimiento/Defensoria. Oficina independiente (depende directamente del Presidente de la Corporación Financiera Internacional (**CFI**)/Grupo del Banco Mundial) que responde a reclamaciones de aquellos afectados por proyectos financiados por la **CFI**. Únicamente considera si la **CFI** ha seguido los procedimientos apropiados.

ONG de control. Una organización no gubernamental (**ONG**) cuya misión u objetivos en particular incluyen una función de monitoreo para supervisar las actividades de empresas u otras organizaciones. Estas **ONG** de control tienden a emplear la estrategia de denuncia y descrédito para fomentar el uso de mejores prácticas industriales.

Organismo regulador. (Autoridad reguladora, entidad reguladora, autoridad competente) Es la autoridad pública o entidad gubernamental encargada de la toma de decisiones y supervisión, con función fiscalizadora o de control, sobre un área de actividad humana. En el marco de una

evaluación de impacto social para la introducción de innovaciones, puede ser la Dirección de Medio Ambiente o la Oficina de Planificación. El organismo es responsable por determinar la aceptabilidad de una Declaración de Impacto Ambiental o de un informe de Evaluación de Impacto Social y por expedir la licencia para llevar adelante el proyecto.

Paradigma. Una serie de prácticas, métodos, teorías y entendimientos respaldados por una disciplina científica (que la define) en un determinado momento de la historia. En otras palabras, la visión del mundo dentro de un área de investigación; esto es, lo que el conjunto de académicos y profesionales que ejercen en un área particular de investigación consideran que es buena práctica corriente.

Participación. Proceso a través del cual los actores influyen y controlan las iniciativas de desarrollo y las decisiones y recursos que las afectan. Proceso que puede mejorar la calidad, efectividad y sostenibilidad de proyectos y fortalecer la propiedad y compromiso del gobierno y los diferentes grupos interesados.

Patrimonio cultural. El legado de objetos materiales y las cualidades intangibles de un grupo o sociedad que provienen de generaciones pasadas, mantenidas en el presente y concedidas para el beneficio de futuras generaciones. El patrimonio cultural comprende la cultura tangible (edificios, monumentos, libros, artesanías y objetos), cultura intangible (folclor, tradiciones, lengua y saber tradicional) y patrimonio natural (paisajes significativos desde el punto de vista cultural, hábitats importantes para la vida silvestre y biodiversidad).

Patrimonio cultural intangible. Véase *Patrimonio cultural*.

Patrimonio vernáculo. La palabra *vernáculo* remite por lo general al idioma cotidiano de la gente común en un lugar determinado. De manera similar, *patrimonio vernáculo* hace referencia al patrimonio cultural que atañe a la vida de la gente común o que es específico de pequeños grupos que comparten un idioma o una serie de experiencias comunes.

Perfil comunitario. Descripción de las comunidades que probablemente se vean afectadas por una intervención planeada.

Perfil social. Véase *Perfil comunitario*.

Periodo de recepción de comentarios públicos. En procesos regulatorios de evaluación de impactos en general se exige que la Declaración de Impacto completa esté disponible para permitir que el público comente al respecto durante un periodo de tiempo designado (esto es, 30, 45, 60 o 90 días).

Personas afectadas. Quienes residen en zonas aledañas y escucharán, verán, sentirán u olerán el proyecto propuesto. Se las reubica ya sea voluntaria o involuntariamente. Tienen un interés en el proyecto o en cambios de política (ya sea que residan o no en zonas de influencia primaria o secundaria) y también en los recursos potencialmente afectados. Es probable que usen regularmente la tierra afectada o que las afecte el influjo de residentes zafrales, temporarios o permanentes asociados con el proyecto.

Personas afectadas por el proyecto (PAP). Término de la CFI/Banco Mundial que puede a veces significar cualquier persona adversamente afectada por un proyecto, pero a veces se refiere principalmente a personas que deben reasentarse en otro lugar o que de otro modo han sido desplazadas como resultado de un proyecto.

Personas interesadas y afectadas. Grupos, organizaciones o individuos que creen que una acción podría afectarlos o que se encuentran implicados en el proceso de decisión (también denominados partes *interesadas*. Véase también *Personas afectadas*).

Pertenencia. El sentido de pertenecer a un grupo social, que es una necesidad emocional humana importante. Muchas veces, como consecuencia de los proyectos, el sentido de pertenencia se reduce, ya sea por los cambios físicos y sociales que ocurren o la presencia de personas nuevas, pero también por los procesos que ocurren y que inducen a la alienación.

PGIS. Plan de Gestión del Impacto Social.

PIMBY. Acrónimo que en inglés significa “por favor, en mi patio trasero”, corresponde a la actitud contraria a **NIMBY**, es decir, obtener

una licencia social para operar de forma que la gente desee que haya proyectos en las inmediaciones.

Plan. Estrategia para alcanzar objetivos identificados o un programa de implementación.

Plan de acción de reasentamiento. Elaborado en línea con el marco de políticas de reasentamiento, consiste en una estrategia detallada en cuanto a la manera en que se llevará a cabo un proceso de reasentamiento específico. Detalla los procesos de relevar las condiciones iniciales, consultar a las personas afectadas, y brinda una estrategia detallada para: (i) minimizar o evitar el reasentamiento; (ii) indemnizar por las pérdidas; (iii) reubicar y reconstruir según sea necesario; (iv) asegurar que se brinde a las personas afectadas la oportunidad de mejorar sus ingresos, las actividades generadoras de ingresos y los niveles de vida que tenían antes de que el proyecto las afectara.

Plan de Gestión de Impacto Social (PGIS). Un documento formal y un sistema de gestión asociado que delinea las estrategias a adoptar durante las diversas fases de un emprendimiento (incluido el cierre) a fin de monitorear, rendir informes, evaluar, examinar y responder de manera proactiva al cambio. De algún modo similar a la Declaración de Impacto Social, la idea de base de un **PGIS** consiste en concentrarse en las estrategias de gestión para abordar los impactos, en lugar de ser solo un listado de impactos potenciales.

Plan de recuperación y mejora de medios de subsistencia. Un plan creado como parte de un proceso de reasentamiento para restablecer y mejorar los medios de subsistencia de las personas luego de haber sido afectadas por un reasentamiento o desplazamiento económico.

Planificación del cierre. El proceso de planificación y gestión del emplazamiento de un proyecto con miras a la situación posterior al cierre; en otras palabras, luego de que se cierra la mina o fábrica. La evaluación de impacto social para la introducción de innovaciones realizada de conformidad con las buenas prácticas permite que se incorpore la planificación posterior al cierre al proceso de planificación y sea prevista en una fase

inicial de la construcción. Esto es muy importante, especialmente en un proyecto de minería donde la volatilidad de los precios de los recursos afecta la viabilidad de la mina.

Planificación posterior. Véase *Planificación del cierre*.

Población de interés. Grupo exhaustivo de todas las unidades (como individuos, hogares, empresas, centros) elegibles para recibir una intervención o un tratamiento, y para los cuales una evaluación de impacto se propone estimar los impactos del programa.

Política. Documento elaborado por una organización y que es una declaración de principios, o una declaración general de metas o de fases procedimentales acerca de algún aspecto de importancia organizacional.

Políticas de salvaguardia. Serie de políticas del Banco Mundial que estipulan el desempeño ambiental y social.

Políticas de salvaguardia del Banco Mundial. Serie de 10 políticas de salvaguardia identificadas por el Banco Mundial como importantes a la hora de garantizar que sus operaciones no perjudiquen a las personas o al medio ambiente: Evaluación ambiental (OP 4.01); Zonas en disputa (OP 7.60); Bosques (OP 4.36); Pueblos indígenas (OP 4.10); Cursos de agua internacionales (OP 7.50); Reasentamiento involuntario (OP 4.12); Hábitats naturales (OP 4.04); Control de plagas (OP 4.09); Recursos culturales físicos (OP 4.11); y Seguridad de las presas (OP 4.37).

Políticas de salvaguardia social del Banco Mundial. Un subgrupo de políticas de salvaguardia (véase arriba) específicas sobre cuestiones sociales (o que las gestiona el Departamento de Desarrollo Social del Banco) y que comprende: Pueblos indígenas (OP 4.10) y Reasentamiento involuntario (OP 4.12).

Potencia (o potencia estadística). Probabilidad de que una evaluación de impacto detecte un impacto (es decir, una diferencia entre el grupo de tratamiento y el grupo de comparación) cuando, de hecho, hay un impacto. La potencia es igual a 1 menos la probabilidad de un error de tipo II, que oscila entre 0 y 1. Los niveles habituales de potencia son 0,8 y 0,9. Los niveles altos de potencia son más conservadores, lo que signi-

fica que hay una baja probabilidad de no detectar los impactos reales del programa.

Potencia estadística. La potencia de una prueba estadística es la probabilidad de que la prueba rechace la hipótesis nula cuando la hipótesis alternativa es verdadera (es decir, que no se cometerá un error de tipo II). A medida que la potencia aumenta, la probabilidad de un error de tipo II disminuye. La probabilidad de un error de tipo II se denomina tasa negativa falsa (β). Por lo tanto, la potencia es igual a $1 - \beta$.

Práctica avanzada. En realidad es lo mismo que las mejores prácticas, pero quizás con la idea de que lo que es avanzado siempre cambia, y por consiguiente es un concepto relativo. Tiende a hacer referencia a una actitud de una empresa (que sea una organización que emplee prácticas avanzadas) más que a una propiedad de la práctica.

Predicción. Tarea de identificar posibles impactos futuros para una intervención planeada.

Principios del Ecuador. Marco de principios en materia de sostenibilidad y responsabilidad social corporativa en el sector financiero global. Más específicamente, es un marco de gestión de riesgo que entidades financieras (esto es, bancos) han adoptado para determinar, evaluar y gestionar el riesgo social y ambiental en proyectos en cualquier parte del mundo y para todos los sectores industriales. Su principal propósito es brindar un estándar mínimo para que se realicen procedimientos de debida diligencia en la toma de decisiones responsable sobre riesgos. Los bancos que han adherido a estos principios se comprometen a implementarlos en sus normas ambientales y sociales internas, procedimientos y estándares para financiar proyectos y acuerdan *“no ofrecer financiación de proyectos ni préstamos corporativos vinculados a proyectos; cuando el cliente no cumpla, o no tenga la capacidad de cumplir los Principios de Ecuador”*. En esencia, los Principios de Ecuador constituyen una serie de principios de alto nivel; para obtener los lineamientos operativos, los Principios de Ecuador requieren que se cumplan las Normas de Desempeño de la CFI.

Principios Rectores. Tiende a hacer referencia a los Principios Rectores de las Naciones Unidas sobre las Empresas y los Derechos Humanos.

Principios voluntarios. En el marco de una evaluación de impacto social para la introducción de innovaciones, generalmente refieren a los Principios voluntarios de seguridad y derechos humanos.

Proceso de cambio social. Un proceso de cambio identificable en comunidades afectadas por proyectos, que es creado, iniciado, habilitado, facilitado y/o exacerbado por medio de una intervención planeada. El proceso de cambio social no es en sí mismo un impacto social, pero podría o no conducir a una situación de impactos sociales, lo cual dependerá del contexto local. Por ejemplo, la inmigración y el reasentamiento son procesos de cambio social que pueden o no provocar impactos sociales.

Proceso de concesión. El proceso mediante el cual un operador obtiene una licencia o permiso en materia de licencias medioambiental, y el proceso por el cual el organismo regulador determina si debe otorgarse.

Proceso de obtención de permiso. Proceso regulatorio que consiste en evaluar y aprobar un proyecto. (Véase también *Licencia ambiental*)

Proceso de resolución de controversias. Un concepto relacionado con acuerdos y contratos (y que con frecuencia constituye una cláusula de ellos), ligado a lo que sucederá si se produce un desacuerdo acerca de la interpretación del contrato/acuerdo o un desacuerdo entre las partes del acuerdo. Es diferente de un mecanismo de reclamos.

Producto. Productos, bienes y servicios tangibles producidos (suministrados) directamente por las actividades de un programa. La generación de productos está directamente bajo el control del organismo ejecutor del programa. El uso de los productos por parte de los beneficiarios contribuye a cambios en los resultados. En evaluación, lógica de programas y gestión de proyectos, el programa entrega productos. El concepto que importa es que los productos rara vez son los resultados previstos, únicamente constituyen un paso hacia alcanzarlos.

Profanización. Proceso por el cual objetos sagrados pasan a ser ordinarios (profanos). La exposición a otras culturas, y especialmente la venta de objetos culturales, los hace perder su valor sagrado.

Programa. Plan coherente y organizado o conjunto ordenado de compromisos, propuestas, instrumentos y/o actividades que describe e implementa una política y que, a la larga, enmarcará a varios proyectos.

Promoción aleatoria. Método de variables instrumentales para estimar los impactos de un programa. El método asigna de forma aleatoria a un subgrupo de unidades una promoción, o incentivo, para participar en el programa. La promoción aleatoria busca aumentar la participación voluntaria en un programa en una submuestra de la población seleccionada aleatoriamente. La promoción puede adoptar la forma de un incentivo, estímulo o información adicional que motiva a las unidades a inscribirse en el programa, sin influir directamente en el resultado de interés. De esta manera, el programa puede quedar abierto a todas las unidades elegibles.

Propensión por coincidencia de puntajes (PSM. *Propensity Score Matching*). Método de coincidencia (*matching*) o pareamiento que depende de los puntajes de propensión para encontrar el mejor grupo de comparación posible para un determinado grupo de tratamiento.

Propensión de puntaje. En el contexto de la evaluación de impacto, la propensión de puntaje es la probabilidad de que una unidad participe en el programa sobre la base de las características observables. Esta puntuación es un número real entre 0 y 1 que resume la influencia de todas las características observables en la probabilidad de inscribirse en el programa.

Propietarios tradicionales. Un término que se utiliza en Australia para designar los pueblos indígenas que tienen el reclamo válido de ser considerados dueños de la tierra en virtud de un título autóctono.

Proponente. En el marco de una evaluación de impacto social (EIS), las proyecciones refieren a las estimaciones, extrapolaciones, predicciones o pronósticos sobre el estado futuro de una o más variables del impacto social en consideración.

Proyección. En el marco de una evaluación de impacto social (EIS), las proyecciones refieren a las estimaciones, extrapolaciones, predicciones o pronósticos sobre el estado futuro de una o más variables del impacto social en consideración.

Proyecto (Project). Un emprendimiento de inversión propuesto que por lo general conlleva la planificación, el diseño y la implementación de actividades específicas.

Proyecto totalmente nuevo (*“Terreno verde”*). Proyecto nuevo, que tiene lugar donde no hubo desarrollo previo; dicho de otro modo, que no implica cuestiones heredadas.

Proyectos lineales. Proyectos como oleoductos o gasoductos, autopistas, ferrocarriles, líneas de transmisión y sistemas de riego que afectan una franja angosta de terreno pero se extienden por muchos kilómetros.

Prueba de placebo. Prueba falsificada que se utiliza para evaluar si los supuestos de un método se mantienen. Por ejemplo, cuando se aplica el método de diferencias en diferencias, se puede implementar una prueba de placebo utilizando un grupo de tratamiento falso o un resultado falso, es decir: un grupo o resultado que se sabe que no se ve afectado por el programa. Las pruebas de placebo no pueden confirmar que los supuestos sean válidos, pero pueden poner de manifiesto los casos en que los supuestos no se sostienen.

Prueba de significancia. Prueba de si la hipótesis alternativa alcanza el nivel predeterminado de significancia con el fin de que esta se acepte de preferencia a la hipótesis nula. Si una prueba de significancia da un valor p menor que el nivel de significancia estadística (α), la hipótesis nula es rechazada.

Pueblo indígena. A grandes rasgos se lo define como un grupo social y cultural distinto y que manifiesta en alguna medida las siguientes características: autoidentificación como miembro de un grupo cultural distinto y reconocimiento de esta identidad por otros; vínculo colectivo con un hábitat determinado geográficamente o territorio ancestral y a los recursos naturales que allí se encuentran; instituciones culturales, econó-

micas social y/o políticas habituales diferentes de las de la sociedad o cultura dominante; una lengua que a menudo difiere de la lengua oficial del país o de la región.

Punto de referencia. Norma de comparación. Para cada indicador social seleccionado para hacer monitoreo, se debería identificar algún estándar o valor que pueda emplearse como referencia. Por ejemplo, una referencia podrá ser los niveles aceptables para la **OMS** en materia de contaminación del aire o exposición al ruido, o la cantidad esperada de médicos por mil habitantes.

PYMES (SME). Pequeñas a medianas empresas, su definición varía de un país a otro según los criterios sobre la cantidad de trabajadores o los ingresos anuales. Cuando se trata exclusivamente de empresas muy pequeñas o unipersonales, *microempresa* es el término que se emplea. Las PYMES son importantes en una evaluación de impacto social para la introducción de innovaciones, ya que a menudo están dispersas, desorganizadas y con frecuencia se ven afectadas por los proyectos. Con la debida consideración, también pueden tener una gran participación activa a través de un compromiso a realizar las adquisiciones en el ámbito local.

Reasentamiento. El proceso planificado de reubicar personas y comunidades de un lugar a otro como parte de la adquisición de tierras, inducida por el proyecto, necesaria para que se pueda ejecutar un proyecto. El reasentamiento se considera involuntario cuando el lugar del proyecto es fijo y las comunidades locales realmente no tienen otra opción que ser reubicados; mientras que el reasentamiento se considera voluntario cuando el Estado no ejerce ni amenaza con ejercer el derecho de expropiación, ni se percibe una tal amenaza, y los individuos afectados tienen una oportunidad real con respecto a si aceptan o no su reasentamiento.

Reasentamiento involuntario. El reasentamiento de personas en razón de un proyecto, en situaciones en las que el Estado ha ejercido o amenazado con ejercer su facultad de dominio eminente. Véase también *Reasentamiento*)

Reasentamiento voluntario. Véase *Reasentamiento*.

Recién llegados. Residentes que llegan a instalarse en un determinado lugar.

Regalía. En términos generales, una regalía es lo que paga una parte (el licenciatario) a otra parte (el licenciante) por el derecho de uso continuado de un bien. En el marco de una evaluación de impacto social para la introducción de innovaciones las regalías son pagos que las empresas dedicadas a la extracción de recursos realizan a los gobiernos y/o a los propietarios Tradicionales de las tierras para acceder a los recursos que se extraen.

Registro de compromisos. Un documento público formal que registra toda declaración o promesa de la empresa a la comunidad especialmente en relación con cualquier beneficio o mitigación prometido.

Rehabilitación. En la evaluación de impactos, hace referencia en general a retrotraer el paisaje al estado en que se encontraba antes del proyecto (*restauración*), o cuando esto no es posible, al menos hacer que el paisaje resulte aceptable para la gente (*recuperación*). En minería, por ejemplo, se refiere al reemplazo de las capas del suelo y a la revegetación de la tierra.

Residentes de larga data. Hace referencia a los que residen desde hace tiempo en un lugar, se emplea en contraste a los recién llegados. Los intereses de los veteranos a menudo difieren de los de los nuevos.

Resiliencia. Capacidad de una comunidad para recuperarse de los impactos que la amenazan.

Resistencia. Capacidad de una comunidad para resistir el cambio, ser capaz de resistirse frente a un proyecto de desarrollo injustificado.

Resolución/gestión de conflictos. Las intervenciones a menudo modifican las relaciones de poder entre los diferentes grupos de la sociedad. Algunos grupos pueden perder mientras otros ganan, y como consecuencia pueden surgir conflictos. Los conflictos son un componente normal de la interacción social, pero cuando se vuelven problemáticos ejercen un impacto negativo sobre todas las partes intervinientes. Por consiguiente,

se hace necesario contar con mecanismos y técnicas eficaces para la prevención, manejo y resolución de los conflictos, de modo de solucionarlos o de mantenerlos dentro de límites aceptables. La transparencia y el intercambio de información pueden contribuir a eliminar conflictos provocados por un conocimiento incompleto o distorsionado. La aceptación y la implementación de un amplio espacio para la expresión de diferentes puntos de vista pueden ayudar a prevenir el desarrollo de formas más destructivas de conflicto.

Restauración del paisaje. *Véase Rehabilitación*

Resultado. Resultado de interés que se mide a nivel de los beneficiarios del programa. Resultados que deben alcanzarse una vez que la población beneficiaria utilice los productos del proyecto. Los resultados no están directamente bajo el control de un organismo ejecutor del programa. En ellos influye tanto la implementación de un programa (las actividades y productos que genera) como las respuestas de las conductas de los beneficiarios expuestos a ese programa (el uso que los beneficiarios hacen de los beneficios a los que están expuestos). Un resultado puede ser intermedio o final (de largo plazo). Los resultados finales son resultados más distantes. La distancia se puede interpretar en términos de tiempo (se tarda más en conseguir el resultado) o en términos de causalidad (se requieren numerosos vínculos causales para alcanzar el resultado y en ello influyen múltiples factores). Lo que se espera que logre una intervención específica en el ámbito del desarrollo.

Retorno social sobre la inversión. Metodología para calcular el beneficio adicional más allá del retorno de inversión financiero inmediato en proyectos o actividades, por ejemplo, de la financiación en inversión social. Generalmente se expresa como proporción relativa a los recursos invertidos.

Riesgo. Puede significar la probabilidad de que se produzca algún acontecimiento, pero el término puede emplearse con un significado levemente diferente para hacer referencia a un suceso incierto (o probabilidad desconocida) que, si ocurre, afectará la concreción de uno o más objetivos. A

menudo se crean subcategorías de riesgos. Por ejemplo, los riesgos no técnicos (o riesgos sociales) se relacionan con cuestiones administrativas, jurídicas, sociales y políticas de un proyecto; mientras que los riesgos técnicos son los aspectos físicos, estructurales, ingenieriles y medioambientales.

Riesgo reputacional. El riesgo potencial para la reputación de una organización, al vincularse o ser vinculada con una práctica en particular.

Riesgo social. El Banco Mundial define el riesgo social como “*la posibilidad de que la intervención cree, refuerce o profundice la desigualdad o el conflicto social o ambos, o de que las actitudes y acciones de los principales interesados puedan subvertir el logro del objetivo de desarrollo o de que los principales interesados no asuman la responsabilidad por el logro del objetivo de desarrollo ni cuenten con los medios para lograrlo. Tales riesgos pueden surgir del contexto sociocultural, político, operativo o institucional*”.

Riesgos no técnicos. Están relacionados con la gestión y los aspectos jurídicos, sociales y políticos de un proyecto. En el sector industrial, a veces se los denomina riesgos *externos* ya que ocurren como resultado de circunstancias ajenas al control de los directores de proyectos.

Rito/ritual. Acto ceremonial que es expresión de cultura y comunidad.

Saber y hacer saber. Una frase popularizada por John Ruggie. En lugar de ser vulnerables y quedar expuestas a las campañas de denuncia y descrédito, las empresas deberían demostrar que han internalizado su respeto por los derechos humanos, a través de su compromiso con los procesos de debida diligencia.

Sabotaje. Un acto deliberado destinado a perjudicar a una empresa por medio de obstrucción, perturbación o destrucción. Las partes interesadas damnificadas podrían ejercer una acción directa contra un proyecto bloqueándolo o destruyendo equipos ya sea en un intento por atraer publicidad, retrasar el proyecto o simplemente por venganza.

Sagrado. Que tiene significancia espiritual o religiosa.

Salud mental. En pocas palabras, es el nivel de bienestar psicológico y la ausencia de cualquier trastorno mental. Comprende la habilidad de un individuo de disfrutar de la vida, así como su bienestar subjetivo, auto-eficacia percibida, autonomía, competencia y desarrollo personal de su potencial intelectual y emocional.

Seguidores. Este término en general refiere a los civiles que siguen ejércitos, que usualmente satisfacen las necesidades de los soldados con bienes y servicios no provistos por las autoridades militares, esto es, determinados alimentos, lavado de ropa, alcohol y drogas, cuidado de enfermos y servicios sexuales. En su empleo actual más generalizado describe a aquellos empresarios que acosan de servicios a los obreros en campamentos de construcción y terrenos de proyectos. (Véase *Efecto tarro de miel*)

Seguridad socialmente responsable. Forma de brindar servicios de seguridad en torno a un proyecto atendiendo a los derechos humanos y otras cuestiones sociales, y que probablemente esté en consonancia con los Principios voluntarios de seguridad y derechos humanos.

Selección. Se produce cuando la participación en el programa se basa en las preferencias, decisiones o características no observables de los participantes o de los administradores del programa.

Sensibilidad cultural. Cualidad individual de tener conciencia de diferencias culturales y de saber cómo manejarse en situaciones multiculturales. Muchos impactos sociales surgen por falta de sensibilidad cultural de gran parte del personal del proyecto.

Sentido comunitario. El grado de noción que los individuos de un determinado lugar tienen de formar parte de una comunidad, y de ayudar a través de la participación en actividades comunitarias y de ser buen vecino (esto es, ejercer la buena vecindad).

Sentido de comunidad. Véase *Sentido comunitario*.

Sentido de pertenencia. Una relación personal de un individuo con su entorno local, tanto social como natural, que mantiene en su vida cotidiana.

Servicios del ecosistema. La noción de que el medio ambiente (un ecosistema) provee una variedad de servicios (y productos) de los cuales dependen los seres humanos. Por lo general, estos servicios se identifican como: aprovisionamiento (por ejemplo, la producción de alimentos y agua); regulación (por ejemplo, el control del clima y de las enfermedades); secundarios (por ejemplo, los ciclos de los nutrientes y la polinización de los cultivos); y culturales (por ejemplo, beneficios espirituales y recreativos). Para ayudar a informar a los responsables de la toma de decisiones, frecuentemente se asigna a los servicios del ecosistema un valor económico.

Sesgo. En la evaluación de impacto, el sesgo es la diferencia entre el impacto que se calcula y el verdadero impacto del programa.

Sesgo de cobertura. Se produce cuando un marco muestral no coincide exactamente con la población de interés.

Sesgo de selección. El impacto estimado sufre un sesgo de selección cuando se desvía del impacto verdadero en presencia de la selección. Esto suele ocurrir cuando se correlacionan motivos no observados para participar en el programa con los resultados. Este sesgo normalmente acontece cuando el grupo de comparación es no elegible o se autoexcluye del tratamiento.

Sesgo de sustitución. Efecto no intencionado de la conducta que afecta al grupo de comparación. Las unidades que no fueron seleccionadas para recibir el programa pueden encontrar buenos sustitutos para el tratamiento a través de su propia iniciativa.

Significancia. La significancia estadística señala la probabilidad de cometer un error de tipo I; es decir, la probabilidad de detectar un impacto que en realidad no existe. El nivel de significancia suele señalarse con el símbolo griego α (alfa). Los niveles más habituales de significancia son del 10%, 5% y 1%. Cuanto menor sea el nivel de significancia, mayor será la confianza de que el impacto estimado es real. Por ejemplo, si el nivel de significancia se fija en 5%, se puede tener un 95% de confianza

al concluir que el programa ha tenido impacto, si de hecho se observa un impacto significativo.

Simulaciones ex ante. Evaluaciones que utilizan datos disponibles para simular los efectos previstos de un programa o de la reforma de una política en los resultados de interés.

Sistema de gestión ambiental. Serie de actividades continuas y planeadas, basadas en el concepto de mejora continua, que una empresa lleva a cabo para gestionar mejor sus impactos ambientales. Está estipulada por la norma ISO 14001.

Sistema de gestión social. Sistema de gestión que específicamente aborda los problemas sociales en una empresa o proyecto.

Sistema de tenencia de tierras. Las disposiciones legales por las cuales se establecen formalmente la propiedad de la tierra y todo proceso para la transferencia intergeneracional o venta de la propiedad de la tierra.

Sitio sagrado. Un sitio (lugar en un paisaje) con un significado espiritual especial para la población local. Aunque se lo asocia habitualmente con los pueblos indígenas, podría aplicarse en un sentido más amplio para referirse a otros lugares y santuarios espirituales y religiosos con una gran significación de patrimonio cultural.

Situación de calle. La situación personal de no tener una vivienda, de no tener un lugar permanente donde estar. En un nivel, puede estar referido a personas sin domicilio fijo que dependen de los servicios de vivienda de emergencia y que de lo contrario dormirían en las calles (es decir, *dormir a la intemperie*); en otro nivel, puede significar un sentimiento de alienación tal que aunque pueda haber un lugar de residencia físicamente adecuado, otros elementos emocionales ligados a él significan que no se lo percibe como un hogar.

Sociedad civil. La red de individuos y grupos (tanto formal como informal)—y sus conexiones, prácticas y normas sociales—que comprenden las actividades de una sociedad independientemente de las instituciones estatales y del mercado. Incluye organizaciones religiosas, grupos comunitarios, fundaciones, gremios, asociaciones profesionales, sindicatos,

instituciones académicas, medios de comunicación, grupos de presión o activistas, partidos políticos, etc.

Subsistencia. Este concepto se encuentra en términos como economía de subsistencia, medios de vida para subsistencia, agricultura de subsistencia, pesca de subsistencia y minería de subsistencia (esto es, minería artesanal). Se refiere a la economía informal o no mercantil (en vez de la economía monetaria) en la cual las personas producen sus propios bienes y servicios o los intercambian en trueques y no por efectivo

Supuesto de estabilidad del valor de la unidad de tratamiento (SUTVA). Requisito básico de que el resultado de una unidad no debería verse afectado por la asignación del tratamiento a otras unidades. Esto es necesario para asegurar que la asignación aleatoria produzca estimaciones de impacto no sesgadas.

Tabú. Algo que está culturalmente prohibido. Los proyectos a menudo transgreden tabúes por desconocimiento, generando mucho agravio.

Tamaño del efecto. Magnitud del cambio en un resultado, que es causado por una intervención.

Tecnocrático. Un término peyorativo que hace referencia a personas e instituciones que poseen excelentes conocimientos técnicos pero carecen de conciencia social y comprensión social; y en particular que toman decisiones sobre la base de su conocimiento técnico, sin prestar la debida atención al contexto social y político.

Teoría del cambio. Explica los canales a través de los cuales los programas pueden influir en los resultados finales. Describe la lógica causal de cómo y por qué un programa, una modalidad de programa o una innovación de diseño en particular logrará sus resultados deseados. Una teoría del cambio es una pieza clave en cualquier evaluación de impacto, dada la focalización de causa y efecto de la investigación.

Titular de derechos. Individuos y grupos cuyos derechos han sufrido un impacto. Podría incluir a todas las partes interesadas. El término es en

realidad similar a *parte interesada* pero el empleo de *titular de derechos* implica una conexión con el enfoque basado en los derechos humanos y una conciencia precisa de que estas personas podrían tener capacidad jurídica.

Tratamiento. Véase intervención.

Tratamiento en los tratados (TOT, por sus siglas en inglés). Las estimaciones **TOT** miden la diferencia en los resultados entre las unidades que en efecto reciben el tratamiento y el grupo de comparación.

Unidad. Persona, hogar, comunidad, empresa, escuela, hospital u otra unidad de observación que pueda ser objeto de un programa o verse afectada por él.

Validez externa. Una evaluación es externamente válida si la muestra de la evaluación representa con precisión a la población de unidades elegibles. Los resultados de la evaluación, posteriormente, se pueden generalizar a la población de unidades elegibles. Estadísticamente, para que una evaluación de impacto sea externamente válida, la muestra de la evaluación debe ser representativa de la población de interés. Véase también validez interna.

Validez interna. Una evaluación es internamente válida si proporciona una estimación precisa del contrafactual mediante un grupo de comparación válido.

Valor compartido. Una manera de pensar acerca del rol de una empresa, que reconoce que son las necesidades de la sociedad, y no solo las necesidades económicas convencionales, las que definen los mercados, y que el objetivo de la empresa debe definirse como la creación de valor compartido y no solo utilidades, de modo que la sociedad se beneficie tanto como la empresa. Esta visión también contempla que los perjuicios sociales generan frecuentemente costos para las empresas, en forma de riesgos sociales, y por consiguiente deben manejarse cuidadosamente.

Valor intrínseco. Concepto filosófico según el cual un objeto o entidad (como la naturaleza, un lugar determinado, una especie rara) tiene un valor o calidad inherente más allá de su valor de uso o instrumental para los humanos.

Valores. Suposiciones abstractas y a menudo subconscientes que los individuos hacen acerca de lo que está bien o tiene importancia en sus vidas. Por lo general, se organizan en un sistema de valores. Los valores y los sistemas de valores pueden variar sustancialmente de un grupo cultural a otro.

Variable. En la terminología estadística, se trata de un símbolo que representa un valor que puede variar.

Variable dependiente. Normalmente, es la variable de resultado. Se trata de la variable que hay que explicar, por oposición a las variables explicativas.

Variable explicativa. También conocida como variable “independiente”. Se trata de una variable utilizada en el lado derecho de una regresión para ayudar a explicar la variable dependiente en el lado izquierdo de la regresión.

Variable instrumental (VI). También conocida como instrumento. Se basa en el uso de una fuente externa de variación para determinar la probabilidad de participación en el programa cuando la participación en el mismo está relacionada con los resultados potenciales. El instrumento se encuentra fuera del control de los participantes y no tiene relación con las características de los mismos.

Variables no observadas. Se trata de características no observables. Pueden incluir particularidades como la motivación, las preferencias u otros rasgos de la personalidad que son difíciles de medir.

Vías causales. Un concepto vinculado con el de pensamiento sistémico y el de evaluación. Refiere a las relaciones causales (al menos a las correlaciones) entre diversos elementos dentro de un sistema. En una evaluación de impacto social para introducción de innovaciones, refieren a las

secuencias de las experiencias de impactos primarios y secundarios y entre los procesos de cambio social y los impactos sociales.

Vías o cadenas de impacto. Se refiere a los vínculos entre los impactos primarios (impactos de primer orden) y los impactos secundarios; así como a los vínculos entre los procesos de cambio social tal como la inmigración.

Visión comunitaria. Proceso de crear consenso sobre el futuro que la comunidad quiere, y luego decidir sobre lo que sea necesario para alcanzarlo. Es tanto el proceso de creación de una visión como el producto de esa visión.

Viviendas de fin de semana. Moradas de personas quienes durante los fines de semana y vacaciones se trasladan a sus segundos hogares. A veces el término también se usa para referirse a las casas. A menudo son un elemento importante en las evaluaciones de impacto social de introducción de innovaciones, pero acceder a ellos es difícil. Algunas veces los proyectos los afectan adversamente pero sus problemas son diferentes de los del resto de la comunidad afectada.

Vulnerabilidad. Una situación o estado caracterizado por una baja resistencia o un mayor riesgo y la capacidad reducida de un individuo, grupo o comunidad de hacer frente a conmociones o impactos adversos. La vulnerabilidad está asociada con una situación socioeconómica baja, incapacidad, etnicidad o uno o más de los muchos factores que influyen en la capacidad de las personas para acceder a los recursos y a las oportunidades de desarrollo.

Zona de influencia o del proyecto. Comprende: el(los) emplazamiento(s) principal(es) del proyecto e instalaciones relacionadas que un proponente (o sus contratistas) desarrolla o controla; instalaciones asociadas construidas como resultado del proyecto (incluso si no fueran financiadas directamente por el proyecto sino por un cliente o un tercero, que puede ser el gobierno), y cuya viabilidad y existencia depende exclusivamente del proyecto y cuyos bienes y servicios son esenciales para la opera-

ción exitosa del proyecto; áreas potencialmente afectadas por impactos acumulativos de otro emprendimiento planeado del proyecto; y áreas potencialmente afectadas por impactos que surgen de acontecimientos no planeados pero predecibles causados por el proyecto y que pueden ocurrir posteriormente o en otro emplazamiento.

Zona de influencia primaria. Se refiere a los impactos sociales que ocurren en una zona de influencia primaria por la medida propuesta y que tienen lugar en el mismo momento y lugar.

Bibliografía

- Abadie, A. J.; Imbens, G. (2008). On the failure of the Bootstrap for Matching Estimators. *Econometrica* 7 (6): 1537-1557
- Abadie, A. J.; Angrist, D.; Imbens, G. W. (2002). “Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings. *Econometrica* 70 (1): 91–117.
- AIEI (2015). *Evaluación de Impacto Social: Lineamientos para la evaluación y gestión de impactos sociales de proyectos*. Asociación Internacional para la Evaluación de Impactos. Banco Interamericano de desarrollo, en: <https://www.iaia.org/pdf/Evaluacion-Impacto-Social-Lineamientos.pdf>
- Bamberger, M.; Rao, V. y Woolcock, M. (2010). *Using Mixed Methods in Monitoring and Evaluation: Experiences from International Development*. Documento de trabajo de investigación de políticas Núm. 5245. Washington, D.C.: BancoMundial.
- Becker, S. O. e Ichino (A). (2002a). *Stata programs for ATT estimation based on propensity score matching*, en: <http://sobecker.userweb.mwn.de/pscore.html>
- Becker, S. O. e Ichino, A. (2002b). Estimation of Average Treatment Effects Based on Propensity Scores. *Stata Journal* 2 (4): 358–77.
- Bell, B.; Blundell, R. y Van Reenen, J.(1999). Getting the Unemployed Back to Work: An Evaluation of the New Deal Proposals. *International Tax and Public Finance* 6 (3): 339–60.
- Bernal, R. y Peña X. (2011). *Guía práctica para la evaluación de impacto*. Perú: Universidad de los Andes: Facultad de Economía.
- Bertrand, M.; Dufl E. y Mullainathan, S. (2004). How Much Should We Trust Differences-in-Differences Estimates? *Quarterly Journal of Economics* 119 (1): 249–75.

- Björklund, A. y Moffitt, R. (1987). The Estimation of Wage Gains and Welfare Gains in Self-Selection Models. *Review of Economics and Statistics* 69 (1): 42–49.
- Bourguignon, F. y Ferreira, F. H. G. (2003). *Ex Ante Evaluation of Policy Reforms Using Behavioral Models*. In *The Impact of Economic Policies on Poverty and Income Distribution: Evaluation Techniques and Tools*, ed. François Bourguignon and Luiz A. Pereira da Silva, 123–41. Washington, DC: World Bank and Oxford University Press.
- Bryson, A.; Dorsett, R. y Purdon, S. (2002). *The use of Propensity Score Matching in the Evaluation of Labour Market Policies*. Working Paper No. 4, Department of Work and Pensions, UK.
- Caliendo, M. y Kopeinig, S. (2008). Some Practical Guidance for the Implementation of Propensity Score Matching. *Journal of Economic Surveys* 22 (1): 31–72.
- Cochran, W. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*: 295–314
- Creswell, J. W. (2014). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Thousand Oaks, CA: Sage Publications.
- Dehejia, R. (2005). Practical Propensity Score Matching: A Reply to Smith and Todd. *Journal of Econometrics* 125 (1–2): 355–64.
- Duflo, E.; Glennerster, R. y Kremer, M. (2008). *Using Randomization in Development Economics Research: A Toolkit*. In *Handbook of Development Economics*, vol. 4, ed. T. Paul Schultz and John Strauss, 3895–962. Amsterdam: North-Holland.
- Fan, J. (1992). Design-Adaptive Nonparametric Regression. *Journal of the American Statistical Association* 87 (420): 998–1004.
- Fan, J. (1993). Local Linear Regression Smoothers and Their Minimax Efficiencies. *Annals of Statistics* 21 (1): 196–216.
- Gertler, P. J.; Martínez, S.; Premand, P.; Rwlings, L. B. y Vermeersch, Ch. M. J. (2017). *La evaluación de impacto en la práctica 2a. ed.*

- Washington, D. C.: Banco Interamericano de Desarrollo y Banco Mundial.
- GII (2019). *Global Innovation Index*. INSEAD-WIPO: Soumitra Dutta, Bruno Lanvin, and Sacha Wunsch-Vincent, en: https://www.wipo.int/edocs/pubdocs/en/wipo_pub_gii_2019.pdf
- Gruber, J. (1994). The Incidence of Mandated Maternity Benefits. *American Economic Review* 84 (3): 622–41.
- Hahn, J.; Hirano, K. y Karlan, D. (2008). *Adaptive Experimental Design Using the Propensity Score*. Working Paper 969, Economic Growth Center. New Haven, CT: Yale University.
- Harberger A. C. (1972). On Measuring the Social Opportunity Cost of Public Funds. In: Project Evaluation. Palgrave Macmillan, London. https://doi.org/10.1007/978-1-349-01653-2_4
- Heckman, J. J.; Ichimura, H. y Todd, P. (1997). Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *Review of Economic Studies* 64 (4): 605–54.
- Heckman, J. J.; Ichimura, H. y Todd, P. (1998). Matching as an Econometric Evaluation Estimator. *Review of Economic Studies* 65 (2): 261–94.
- Heckman, J. J.; LaLonde, R. y Smith, J. (1999). *The Economics and Econometrics of Active Labor Market Programs*. In *Handbook of Labor Economics*, vol. 3, ed. Orley Ashenfelter and David Card, 1865–2097. Amsterdam: North-Holland.
- Heckman, J. J. y Vytlačil, E.J. (2000). *Local Instrumental Variables*. NBER Technical Working Paper 252. National Bureau of Economic Research, Cambridge, MA.
- Heckman, J. J. y Vytlačil, E. (2005). Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica* 73 (3): 669–738.
- Hirano, K.; Imbens, G. W. y Ridder, G. (2003). “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica* 71 (4): 1161–89.

- Imas, L. G. M. y Rist R. C. (2009). *The Road to Results: Designing and Conducting Effective Development Evaluations*. Washington, D.C.: Banco Mundial.
- Imbens, G. y Angrist, J. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica* 62 (2): 467–76.
- Imbens, G. (2004). Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review. *Review of Economics and Statistics* 86 (1): 4–29.
- Imbens, G. y Rubin, D. (2010). Causal Inference in Statistics and Social sciences, manuscrito Harvard University.
- Jalan, J. y Ravallion, M. (2003). “Estimating the Benefit Incidence of an Antipoverty Program by Propensity-Score Matching”. *Journal of Business & Economic Statistics* 21 (1): 19–30.
- Khandker, S. R.; Koolwal, G. B. y Samad, H. A. (2017). *The Handbook on Impact Evaluation. Quantitative Methods and Practices*. Washington, D. C.: The Worldbank.
- Kish, L. (1987). *Statistical Design for Research*. New York: Wiley.
- Kusek, J. Z. y Rist.R.C. (2004). *A Handbook for Development Practitioners: Ten Steps to a Results-Based Monitoring and Evaluation System*. Washington, DC: World Bank.
- Lechner, M. (1999). Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany after Unification. *Journal of Business Economic Statistics* 17 (1): 74–90.
- Mejía-Trejo, J.; Sánchez-Gutiérrez, J.; y Vázquez-Ávila, G. (2015). Empirical Model for Mobile Learning and their Factors. Case Study: Universities Located in the Urban City of Guadalajara, México. *Revista Apertura*, 7 (2):1-13, en:<http://www.udgvirtual.udg.mx/apertura/index.php/apertura/article/view/725/498>
- Mejía-Trejo, J. (2017). Mercadotecnia Digital. Una descripción de las herramientas que apoyan la planeación estratégica de toda innovación de campaña web. México: Editorial Patria, en: <https://play>.

- google.com/books/reader?id=AUbJDgAAQBAJ&printsec=front-cover&pg=GBS.PP1
- Mejía-Trejo, J. y Bravo-Rodríguez, A. (2019). *Values-Based Innovation. Designing a model to be Applied in Management Sciences*. CUCEA-Universidad de Guadalajara-Universidad Autónoma de Barcelona, en: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3512169
- Miguel, E. y Kremer, M. (2004). “Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica* 72 (1): 159–217.
- Moffitt, R. (2003). *The Role of Randomized Field Trials in Social Science Research: A Perspective from Evaluations of Reforms from Social Welfare Programs*. NBER Technical Working Paper 295, National Bureau of Economic Research, Cambridge, MA.
- OIT (2020). *Covid-19 y el mundo del trabajo: repercusiones y respuestas*. Organización Internacional del Trabajo, en: https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/documents/briefing-note/wcms_739158.pdf.
- ONU CEPAL (2015). *La Agenda 2030 y los Objetivos de Desarrollo Sostenible Una oportunidad para América Latina y el Caribe*. Santiago, Chile: ONU, Comisión Económica para América Latina y el Caribe, en: <https://www.un.org/sustainabledevelopment/es/>
- PNUD (2009). *Handbook on Planning, Monitoring and Evaluating for Development Results*. Nueva York: PNUD (Programa de las Naciones Unidas para el Desarrollo).
- Quandt, R. (1972). Methods for Estimating Switching Regressions. *Journal of the American Statistical Association* 67 (338): 306–10.
- Rao, V. y Woolcock, M. (2003). *Integrating Qualitative and Quantitative Approaches in Program Evaluation* en : F. J. Bourguignon y L. Pereira da Silva, *The Impact of Economic Policies on Poverty and Income Distribution: Evaluation Techniques and Tools*: 165–90. Nueva York: Oxford University Press.

- Ravallion, M. (2003). *Assessing the Poverty Impact of an Assigned Program*. In *The Impact of Economic Policies on Poverty and Income Distribution: Evaluation Techniques and Tools*, ed. François Bourguignon and Luiz A. Pereira da Silva, 103–22. Washington, DC: World Bank and Oxford University Press.
- Ravallion, M. (2008). Evaluating Anti-Poverty Programs. In *Handbook of Development Economics*, vol. 4, ed. T. Paul Schultz and John Strauss, 3787–846. Amsterdam: North-Holland.
- Rosenbaum, P. (2002). *Observational Studies* (2da. edición), Springer Series in Statistics. Nueva York: Springer-Verlag.
- Rosenbaum, P. R. y Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70 (1): 41–55.
- Roy, A. D. (1951). *Some Thoughts on the Distribution of Earnings*. Oxford Economic Papers 3 (2): 135–46.
- Smith, J. y Todd, P.; (2005). Does Matching Overcome LaLonde's Critique of Non-experimental Estimators?. *Journal of Econometrics*, 125 81-2), 305-353. *The Economist* (2020), "New trade barriers could hamper the supply of masks and medicines" 11 de marzo [en línea] <https://www.economist.com/finance-and-economics/2020/03/11/new-trade-barriers-could-hamper-the-supply-of-masks-and-medicines>
- Todd, P. y Wolpin, K. (2006). *Ex Ante Evaluation of Social Programs*. PIER Working Paper 06-122, Penn Institute for Economic Research, University of Pennsylvania, Philadelphia.
- Todd, P. (2007). *Evaluating Social Programs with Endogenous Program Placement and Selection of the Treated*. In *Handbook of Development Economics*, vol. 4, ed. T. Paul Schultz and John Strauss, 3847–94. Amsterdam: North-Holland.



**Evaluación de impacto social en proyectos
de innovación vía STATA
Métodos: Inferencia Causal, Aleatorización, Propensión de
Coincidencia de Puntaje y Doble Diferencia
Tomo I**

Registrado y producido como libro electrónico en mayo de 2021
en los talleres gráficos de TRAUCO Editorial
Camino Real a Colima 285 int. 56
Teléfono: (33) 32.71.33.33
Tlaquepaque, Jalisco.

En el mundo empresarial e industrial, el diseño e implementación de proyectos normalmente toma en cuenta el impacto económico, financiero y hasta el político o ambiental en la introducción de innovaciones. Sin embargo, en los primeros veinte años del siglo XXI, han sucedido diversos acontecimientos que han demostrado que la evaluación de impacto social en proyectos de innovación es de vital importancia. De hecho, en los tiempos de la nueva normalidad que se avizora como la era PosCOVID-19, esto toma particular relevancia dado que todas las políticas y acciones que emitan empresas y gobiernos, deberán contar con el aval necesario de una evaluación de impacto social a la introducción de innovaciones.

Es por esta razón, que la presente obra está orientada a describir tanto a propios como ajenos al tema, lo qué es el impacto social, sus características, condiciones e implicaciones, los principales métodos utilizados para calcularla, que demanda que los recursos y acciones de innovación a diseñar e implementar, reflejen altos estándares de impacto social que fomenten el bienestar, particularmente en los países emergentes.

Para lograrlo, esta obra está dividida en una colección de dos tomos, correspondiendo al Tomo I:

Conceptos básicos así como la preparación, inferencia causal y contrafactuales que dan pauta a la descripción y aplicación de los métodos por aleatorización; propensión de coincidencia de puntaje (Propensity Score Matching) y doble diferencia describiéndose una variedad de ejemplos de proyectos de innovación. Se anexa al final, un breve manual de la configuración y manejo del software STATA.



CUCEA
El mejor lugar para el talento

