Econometría aplicada con SPSS para la investigación

Guía práctica paso a paso para tesis, artículos científicos y análisis de datos

Willian Sebastian Flores Sotelo

Lima, Perú; 2026

Econometría aplicada con SPSS para la investigación
Guía práctica paso a paso para tesis, artículos científicos y análisis de datos
Serie Econometría sin miedo, n.º 3
Autor-editor: © Willian Sebastian Flores Sotelo. Todos los derechos reservados.
Jr. Piura 282, piso 2, Urb. La Libertad, Comas – Lima – Perú
Primera edición digital: octubre de 2026.
Hecho el Depósito Legal en la Biblioteca Nacional del Perú N.° 2026-13312
ISBN: 978-612-03-3394-5
Libro digital disponible en: https://willianfloressotelo.com
Ninguna parte de esta obra puede reproducirse sin autorización escrita del autor, salvo citas breves con fines de reseña o docencia.
Los archivos de sintaxis .sps del libro se ofrecen para uso educativo. Las bases de datos citadas pertenecen a sus instituciones productoras.
Edición digital. Distribuido por Amazon Kindle Direct Publishing, Apple Books y Google Play Books.
IBM y SPSS Statistics son marcas registradas de International Business Machines Corporation (IBM) en muchas jurisdicciones del mundo. Este libro es una obra independiente: no está patrocinado, autorizado ni avalado por IBM. Stata, EViews, R y Excel son marcas de sus respectivos titulares.

Contenido breve

ACERCA DEL AUTOR iv
PREFACIO: CÓMO USAR ESTE LIBRO x
Parte Uno. Antes de analizar 1
Capítulo 1 ¿Por qué tu análisis miente? Correlación y causalidad 2
Capítulo 2 Tus datos listos en SPSS en una tarde: de la encuesta y la ENAHO a una base limpia 17
Capítulo 3 Estadística descriptiva, niveles y baremos: las tablas que pide tu tesis 36
Capítulo 4 Confiabilidad y validez: alfa de Cronbach y análisis factorial 53
Parte Dos. Pruebas de hipótesis 75
Capítulo 5 Prueba de hipótesis y normalidad: Kolmogorov-Smirnov, Shapiro-Wilk y qué prueba elegir 76
Capítulo 6 Comparar grupos: t de Student, ANOVA, Mann-Whitney y Kruskal-Wallis 93
Capítulo 7 Chi cuadrado y tablas de contingencia 110
Capítulo 8 Correlación de Pearson y rho de Spearman: la tesis correlacional paso a paso 124
Parte Tres. La regresión en SPSS, paso a paso 137
Capítulo 9 Regresión lineal simple y múltiple y el sesgo de variable omitida 138
Capítulo 10 Cómo interpretar una regresión sin pánico 150
Capítulo 11 Supuestos de la regresión: heterocedasticidad, multicolinealidad y autocorrelación 164
Capítulo 12 Regresión logística binaria, ordinal y multinomial 176
Parte Cuatro. Efectos, no solo asociaciones 190
Capítulo 13 Datos de panel y medidas repetidas con modelos mixtos 191
Capítulo 14 Evaluación de impacto: diferencias en diferencias, variables instrumentales y propensity score matching 203
Parte Cinco. Reportar 218
Capítulo 15 Cómo redactar los resultados de tu tesis o artículo científico en APA 7 219
APÉNDICE A. EQUIVALENCIAS SPSS · STATA · EVIEWS · R 231
APÉNDICE B. BASES DE DATOS DEL LIBRO 235

Contenido

ACERCA DEL AUTOR iv

PREFACIO: CÓMO USAR ESTE LIBRO x

Parte Uno. Antes de analizar 1

Capítulo 1 ¿Por qué tu análisis miente? Correlación y causalidad 2

Lo que vas a lograr 2

1.1 Un mismo *r* = ,50 puede contar cuatro historias 3

1.2 ¿Qué habría vendido la misma bodega sin el curso? 3

1.3 La cuenta que delata el sesgo 3

1.4 Dibuja las flechas antes de abrir SPSS 4

1.5 Lo que un sorteo hace y ningún programa puede hacer 4

1.6 ¿Y si nadie sorteó nada? 4

1.7 Manos a la obra 5

1.8 El error que se repite en las tesis: «influye significativamente» 11

1.9 El error que cometen los investigadores en la redacción de artículos 11

RESUMEN Y CONCLUSIONES 12

EJERCICIOS RESUELTOS 12

Para tu tesis o artículo 15

Referencias 16

Capítulo 2 Tus datos listos en SPSS en una tarde: de la encuesta y la ENAHO a una base limpia 17

Lo que vas a lograr 17

2.1 Teresa y la encuesta que nadie limpió 18

2.2 Una base de SPSS tiene dos caras 18

2.3 El «No sabe» no es el punto más alto de la escala 18

2.4 Cuando 1 es lo mejor en una pregunta y lo peor en la siguiente 19

2.5 ¿Sumar o promediar? 19

2.6 Las llaves: unir sin perder ni duplicar a nadie 20

2.7 Cada entrevistado vale por miles: el factor de expansión 20

2.8 Sintaxis, no clics 20

2.9 Manos a la obra 20

2.10 El error que se repite en las tesis: el «No sabe» que nadie declaró 29

2.11 El error que cometen los investigadores en la redacción de artículos 30

RESUMEN Y CONCLUSIONES 30

EJERCICIOS RESUELTOS 31

Para tu tesis o artículo 34

Referencias 34

Capítulo 3 Estadística descriptiva, niveles y baremos: las tablas que pide tu tesis 36

Lo que vas a lograr 36

3.1 Gerardo y la pregunta que la tesis repite 37

3.2 Una tabla de frecuencias tiene tres columnas de porcentaje 37

3.3 Media, mediana y moda: tres respuestas a la misma pregunta 37

3.4 El baremo: convertir un puntaje en un nivel 38

3.5 El costo de cortar: lo que se pierde al pasar a niveles 38

3.6 Manos a la obra 39

3.7 El error que se repite en las tesis: el baremo que no se explica 46

3.8 El error que cometen los investigadores en la redacción de artículos 46

RESUMEN Y CONCLUSIONES 47

EJERCICIOS RESUELTOS 47

Para tu tesis o artículo 50

Referencias 51

Capítulo 4 Confiabilidad y validez: alfa de Cronbach y análisis factorial 53
Lo que vas a lograr 53
4.1 Maribel y el instrumento que «ya estaba validado» 54
4.2 El alfa de Cronbach: qué mide y qué no 54
4.3 El análisis factorial exploratorio: buscar las dimensiones 55
4.4 Manos a la obra 55
4.5 El error que se repite en las tesis: tomar el alfa por validez 69
4.6 El error que cometen los investigadores en la redacción de artículos 70
RESUMEN Y CONCLUSIONES 70
EJERCICIOS RESUELTOS 71
Para tu tesis o artículo 73
Referencias 74
Parte Dos. Pruebas de hipótesis 75
Capítulo 5 Prueba de hipótesis y normalidad: Kolmogorov-Smirnov, Shapiro-Wilk y qué prueba elegir 76
Lo que vas a lograr 76
5.1 Celeste y la tabla que «ya sabía» llenar 77
5.2 Hipótesis nula, hipótesis alternativa y valor *p* 77
5.3 Las pruebas de normalidad y su trampa 77
5.4 Manos a la obra 78
5.5 El error que se repite en las tesis: dejar que la prueba de normalidad elija la prueba 88
5.6 El error que cometen los investigadores en la redacción de artículos 88
RESUMEN Y CONCLUSIONES 88
EJERCICIOS RESUELTOS 89
Para tu tesis o artículo 91
Referencias 92
Capítulo 6 Comparar grupos: t de Student, ANOVA, Mann-Whitney y Kruskal-Wallis 93
Lo que vas a lograr 93
6.1 Leandro y las ocho regiones 94
6.2 Dos grupos: la *t* de Student y la *t* de Welch 94
6.3 Tres o más grupos: el ANOVA y las comparaciones múltiples 94
6.4 El tamaño del efecto: la pregunta que el valor *p* no responde 95
6.5 Las alternativas no paramétricas 95
6.6 Manos a la obra 95
6.7 El error que se repite en las tesis: «hay diferencias significativas» 103
6.8 El error que cometen los investigadores en la redacción de artículos 103
RESUMEN Y CONCLUSIONES 103
EJERCICIOS RESUELTOS 104
Para tu tesis o artículo 108
Referencias 108
Capítulo 7 Chi cuadrado y tablas de contingencia 110
Lo que vas a lograr 110
7.1 Nélida y los problemas del país 111
7.2 La tabla de contingencia y las frecuencias esperadas 111
7.3 Después del chi cuadrado: residuos y tamaño del efecto 111
7.4 Manos a la obra 112
7.5 El error que se repite en las tesis: el chi cuadrado sin tamaño del efecto 118
7.6 El error que cometen los investigadores en la redacción de artículos 118
RESUMEN Y CONCLUSIONES 119
EJERCICIOS RESUELTOS 119
Para tu tesis o artículo 122
Referencias 123
Capítulo 8 Correlación de Pearson y rho de Spearman: la tesis correlacional paso a paso 124
Lo que vas a lograr 124
8.1 Ismael y la hipótesis de siempre 125

8.2 La correlación de Pearson: dirección, fuerza y lo que no dice ... 125
8.3 Los baremos de *r*: de dónde salen y por qué desconfiar ... 125
8.4 Spearman, atípicos y restricción de rango ... 126
8.5 Manos a la obra ... 126
8.6 El error que se repite en las tesis: la correlación leída como causa y el baremo sin fuente ... 131
8.7 El error que cometen los investigadores en la redacción de artículos ... 132
RESUMEN Y CONCLUSIONES ... 132
EJERCICIOS RESUELTOS ... 132
Para tu tesis o artículo ... 135
Referencias ... 135
Parte Tres. La regresión en SPSS, paso a paso ... 137
Capítulo 9 Regresión lineal simple y múltiple y el sesgo de variable omitida ... 138
Lo que vas a lograr ... 138
9.1 Rosaura y la brecha rural ... 139
9.2 La regresión lineal simple ... 139
9.3 La regresión múltiple y el sesgo de variable omitida ... 139
9.4 No todo control es bueno ... 139
9.5 Manos a la obra ... 140
9.6 El error que se repite en las tesis: el coeficiente leído como efecto ... 144
9.7 El error que cometen los investigadores en la redacción de artículos ... 145
RESUMEN Y CONCLUSIONES ... 145
EJERCICIOS RESUELTOS ... 145
Para tu tesis o artículo ... 148
Referencias ... 149
Capítulo 10 Cómo interpretar una regresión sin pánico ... 150
Lo que vas a lograr ... 150
10.1 Fabiola y la tabla de coeficientes ... 151
10.2 Cada columna de la tabla de coeficientes 151
10.3 La constante, las unidades y el centrado ... 151
10.4 Variables ficticias e interacciones ... 152
10.5 Manos a la obra ... 152
10.6 El error que se repite en las tesis: leer solo la columna «Sig.» ... 157
10.7 El error que cometen los investigadores en la redacción de artículos ... 157
RESUMEN Y CONCLUSIONES ... 158
EJERCICIOS RESUELTOS ... 158
Para tu tesis o artículo ... 161
Referencias ... 162
Capítulo 11 Supuestos de la regresión: heterocedasticidad, multicolinealidad y autocorrelación ... 164
Lo que vas a lograr ... 164
11.1 Octavio y la sección de supuestos ... 165
11.2 Qué supuestos importan, y para qué ... 165
11.3 Heterocedasticidad y errores robustos ... 165
11.4 Durbin-Watson y los conglomerados ... 166
11.5 Manos a la obra ... 166
11.6 El error que se repite en las tesis: la sección de supuestos como ritual ... 170
11.7 El error que cometen los investigadores en la redacción de artículos ... 171
RESUMEN Y CONCLUSIONES ... 171
EJERCICIOS RESUELTOS ... 171
Para tu tesis o artículo ... 174
Referencias ... 174
Capítulo 12 Regresión logística binaria, ordinal y multinomial ... 176
Lo que vas a lograr ... 176
12.1 Gisela y la corrupción ... 176
12.2 Por qué no basta la regresión lineal ... 177
12.3 El *odds ratio* no es un riesgo relativo ... 177
12.4 Más de dos categorías: ordinal y multinomial ... 177
12.5 Manos a la obra ... 178

12.6 El error que se repite en las tesis: «X veces más probable» ... 184
12.7 El error que cometen los investigadores en la redacción de artículos ... 184
RESUMEN Y CONCLUSIONES ... 184
EJERCICIOS RESUELTOS ... 185
Para tu tesis o artículo ... 187
Referencias ... 188
Parte Cuatro. Efectos, no solo asociaciones ... 190
Capítulo 13 Datos de panel y medidas repetidas con modelos mixtos ... 191
Lo que vas a lograr ... 191
13.1 Esteban y los vecinos que se parecen ... 192
13.2 La correlación intraclase y el efecto de diseño ... 192
13.3 Los modelos mixtos ... 192
13.4 Manos a la obra ... 193
13.5 El error que se repite en las tesis: la pseudorreplicación ... 198
13.6 El error que cometen los investigadores en la redacción de artículos ... 198
RESUMEN Y CONCLUSIONES ... 198
EJERCICIOS RESUELTOS ... 199
Para tu tesis o artículo ... 200
Referencias ... 201
Capítulo 14 Evaluación de impacto: diferencias en diferencias, variables instrumentales y propensity score matching ... 203
Lo que vas a lograr ... 203
14.1 Yesenia y Pensión 65 ... 204
14.2 El contrafactual y el sesgo de selección ... 204
14.3 Manos a la obra ... 204
14.4 El error que se repite en las tesis: llamar «impacto» a una diferencia de medias ... 213
14.5 El error que cometen los investigadores en la redacción de artículos ... 213
RESUMEN Y CONCLUSIONES ... 213
EJERCICIOS RESUELTOS ... 214
Para tu tesis o artículo ... 216
Referencias ... 216
Parte Cinco. Reportar ... 218
Capítulo 15 Cómo redactar los resultados de tu tesis o artículo científico en APA 7 ... 219
Lo que vas a lograr ... 219
15.1 Marisol y el capítulo de resultados ... 220
15.2 Qué reportar: el conjunto mínimo ... 220
15.3 El formato APA 7, pieza por pieza ... 220
15.4 Manos a la obra ... 221
15.5 El error que se repite en las tesis: las tablas de SPSS pegadas y repetidas ... 225
15.6 El error que cometen los investigadores en la redacción de artículos ... 226
RESUMEN Y CONCLUSIONES ... 226
EJERCICIOS RESUELTOS ... 226
Para tu tesis o artículo ... 228
Referencias ... 229
APÉNDICE A. EQUIVALENCIAS SPSS · STATA · EVIEWS · R ... 231
APÉNDICE B. BASES DE DATOS DEL LIBRO 235

Capítulo 1

¿Por qué tu análisis miente? Correlación y causalidad

Entre 2020 y 2021, Noah Haber reunió a 48 revisores, reclutados por invitación personal y por Twitter, para una tarea poco común. No iban a comprobar si los números de otros investigadores eran correctos. Iban a leer cómo esos investigadores **escribían** sus hallazgos.

Eligieron 18 de las revistas más influyentes de medicina, salud pública y epidemiología, entre ellas *The Lancet*, *JAMA* y el *New England Journal of Medicine*. De ellas sacaron 1170 estudios publicados entre 2010 y 2019 en los que nadie había sorteado la exposición. Eran estudios observacionales: se mira lo que la gente hace, no se le asigna qué hacer.

La palabra favorita de los autores fue «asociado». Unía la posible causa con el resultado en 535 resúmenes, el 45,7 %. La palabra «causa» apareció como enlace principal solo en 9. ¿Prudencia ejemplar? Eso parecía.

Luego los revisores calificaron cuánto sugería cada frase una relación de causa y efecto. Solo el 13,8 % de los resúmenes no la insinuaba en absoluto. Más de la mitad de los propios revisores sintió que incluso «asociación» sugiere algo de causa. Y lo más revelador estaba al final. Un tercio de los resúmenes incluía una recomendación de qué hacer, y en el 44,5 % de los artículos esa recomendación implicaba más causalidad que la frase que resumía el hallazgo (Haber et al., 2022).

Los autores evitaban la palabra «causa», pero escribían como si la hubieran demostrado.

Ahora piensa en tu tesis. Abres SPSS, corres una correlación y la tabla te devuelve $r = ,50$ con dos asteriscos. ¿Qué vas a escribir? ¿«Se asocia»? ¿«Influye significativamente»? Y, sobre todo, ¿qué vas a recomendar?

Este capítulo trata de la distancia entre lo que tus datos muestran y lo que te gustaría que mostraran. Es el punto de partida de todo lo que harás con SPSS en el resto del libro.

Lo que vas a lograr

Cuando cierres este capítulo sabrás:

- Reconocer si una cifra de SPSS solo retrata una relación o si de verdad estima un efecto.
- Explicar con una sola cuenta por qué una comparación de promedios puede engañarte.
- Simular en SPSS, paso a paso, un caso en el que la correlación, la prueba *t* y la regresión mienten, y otro en el que dicen la verdad.
- Elegir el verbo correcto para redactar tus resultados en la tesis o en el artículo.

1.1 Un mismo *r* = ,50 puede contar cuatro historias

Imagina a Rocío, economista de la subgerencia de desarrollo económico de una municipalidad de Huancayo. Rocío es un personaje inventado para este capítulo, pero su problema es real y lo vas a reconocer.

La municipalidad ofreció un curso gratuito de gestión para microempresas: costos, inventario, atención al cliente. Rocío quiere hacer su tesis de maestría sobre ese curso. Encuesta a 4000 microempresas del distrito y anota dos cosas: si el negocio asistió al curso y cuánto vende al mes. El 31 % asistió.

Corre una correlación en SPSS y obtiene $r = ,50$, con $p < ,001$. Los negocios que fueron al curso venden más. ¿Ya puede titular su tesis «Influencia de la capacitación municipal en las ventas de las microempresas»?

Todavía no. Una correlación como esa puede nacer de cuatro historias distintas:

1. **El curso sube las ventas.** Es la historia que Rocío quiere contar: causalidad.
2. **Las ventas llevan al curso.** Quizá los negocios que ya venden bien tienen un empleado que atiende mientras el dueño estudia. Es causalidad inversa. Es el riesgo clásico cuando la causa y el resultado se miden a la vez, como en casi toda encuesta transversal (Savitz & Wellenius, 2023).
3. **Un tercer factor mueve a los dos.** Supón que la municipalidad repartió las invitaciones a pie, por las avenidas más transitadas. Esos negocios tienen más clientes en la puerta: venden más con curso o sin él, y además recibieron la invitación. A ese tercer factor se le llama **variable de confusión**.
4. **El azar.** Con 4000 negocios, una correlación de ,50 casi nunca aparece por pura casualidad.

SPSS no sabe cuál de las cuatro historias es la verdadera. Los dos asteriscos de la tabla solo dicen algo sobre la cuarta: que el azar es una explicación muy improbable. Sobre las otras tres, la tabla guarda silencio.

1.2 ¿Qué habría vendido la misma bodega sin el curso?

Piensa en una bodega concreta, la de la señora Carmen, en la Calle Real. Carmen fue al curso. Este mes vendió 18 000 soles.

¿Cuál fue el efecto del curso en su bodega? Para saberlo necesitarías dos cifras: lo que vendió habiendo ido al curso y lo que **habría vendido** si no hubiera ido. El efecto es la resta entre las dos.

Sin embargo, el problema salta a la vista. La segunda cifra no existe. Carmen fue al curso, y ningún programa estadístico puede mostrarte el mes en que no fue. Siempre observas uno solo de los dos resultados posibles de cada negocio.

Como no puedes ver a la misma bodega en dos mundos, haces lo que hace todo investigador: comparas bodegas distintas. Tomas las que fueron al curso y las que no, y restas sus promedios. Esa resta solo sirve si las que no fueron se parecen a lo que habrían sido las que sí fueron. ¿Se parecen? Ahí está todo el problema.

1.3 La cuenta que delata el sesgo

La diferencia de promedios que te da SPSS se puede partir en dos piezas:

```
Diferencia observada = Efecto del curso + Sesgo de selección
```

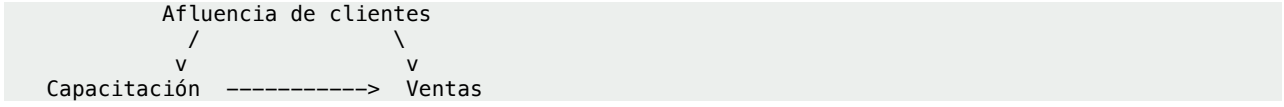
El **sesgo de selección** es la ventaja (o desventaja) con la que un grupo llegó al curso. En palabras: cuánto más habrían vendido, de todos modos, los negocios que asistieron, comparados con los que no asistieron, si ninguno hubiera ido al curso.

Si la municipalidad invitó sobre todo a las bodegas de las avenidas con más clientes, ese sesgo es positivo y grande. Supón que el curso sube las ventas en 1000 soles al mes, y que las bodegas invitadas ya vendían 2600 soles más por estar donde están. La comparación de promedios te dirá 3600 soles. Y te lo dirá con un valor *p* minúsculo.

Por eso conviene fijarse en lo que esto significa. La diferencia observada no es una mala estimación del efecto por culpa de pocos datos o de una fórmula equivocada. Es la respuesta correcta a otra pregunta: ¿cuánto más venden los que fueron al curso? Esa pregunta no es la misma que ¿cuánto más venden **porque** fueron al curso?

1.4 Dibuja las flechas antes de abrir SPSS

Antes de correr cualquier análisis, haz un dibujo. Escribe tus variables y une con una flecha cada causa con su efecto. Para el caso de Rocío queda así:



Este esquema se llama **diagrama causal**. Tiene dos caminos que van de la capacitación a las ventas. El de abajo es el efecto que te interesa. El de arriba pasa por la afluencia de clientes y es el que contamina tu comparación.

Además, el diagrama te dice qué hacer. Si mides la afluencia y la incluyes como control, cierras el camino de arriba. Si no la mides, el camino queda abierto y ninguna prueba estadística lo cerrará por ti.

También te dice qué **no** hacer. Wysocki et al. (2022) muestran que controlar una variable inapropiada, como una consecuencia del tratamiento, puede dejar la estimación más sesgada que no controlar nada. Advierten además que los artículos casi nunca justifican por qué eligieron sus variables de control. La justificación es el dibujo: sin flechas no hay razones.

1.5 Lo que un sorteo hace y ningún programa puede hacer

Imagina ahora que la municipalidad tiene más interesados que vacantes y decide sortearlas. Ganar el sorteo no depende de la avenida, ni del rubro, ni del empuje del dueño. Depende de una moneda.

Así, con 4000 negocios, el sorteo reparte la afluencia de clientes casi por igual entre ganadores y perdedores. Lo mismo hace con todo lo demás, incluso con lo que nadie midió. Los dos grupos son, en promedio, gemelos. El sesgo de selección se vuelve cero, y la diferencia de promedios pasa a medir el efecto del curso.

Por eso el ensayo aleatorizado se considera el patrón de oro para evaluar el efecto de una intervención en salud (Núñez & Lajous, 2025). No es magia estadística. Es diseño: el sorteo resuelve el problema antes de que los datos lleguen a SPSS.

1.6 ¿Y si nadie sorteó nada?

La mayoría de las tesis no tiene sorteo. Tienen una encuesta aplicada una vez, a negocios, estudiantes o pacientes que eligieron por su cuenta. ¿Significa que no puedes decir nada útil?

No. Significa que tienes que trabajar más. Martínez-Abad y León (2023) recuerdan que suponer relaciones de causa y efecto en la investigación *ex post facto* es un problema ampliamente conocido. También muestran que las técnicas de inferencia causal todavía se usan poco en la investigación educativa, y las ilustran con datos de PISA 2018.

Núñez y Lajous (2025) proponen un hábito que puedes aplicar desde hoy: escribir el protocolo del experimento que te gustaría hacer, el **ensayo diana**. Defines quiénes podrían participar, qué tratamientos compararías, cómo los asignarías, cuánto tiempo seguirías a cada caso, qué resultado medirías y cómo lo analizarías. Luego ajustas cada elemento a los datos que sí tienes. El estudio sigue siendo observacional, pero sus sesgos quedan a la vista.

Por otra parte, no todo estudio transversal es inútil para hablar de causas. Savitz y Wellenius (2023) sostienen que su aporte depende de la secuencia real entre la exposición y el resultado y del control de las variables de confusión, no de la etiqueta «transversal». En este libro verás las herramientas para acercarte: la regresión con controles (capítulos 9 a 11), los datos repetidos (capítulo 13) y la evaluación de impacto (capítulo 14).

1.7 Manos a la obra

Ahora vas a hacer lo que ningún investigador puede hacer con una encuesta: **fabricar los datos y saber de antemano la respuesta correcta**. Simularemos las 4000 microempresas de Rocío. Decidiremos nosotros que el curso sube las ventas exactamente en 1 (mil soles al mes). Después le preguntaremos a SPSS, con cada método, cuánto vale ese efecto. Todo lo que sigue usa datos simulados; ninguna cifra viene de negocios reales.

Abre SPSS y crea una ventana de sintaxis con *Archivo › Nuevo › Sintaxis* (*File › New › Syntax*). Copia cada bloque, selecciónalo y pulsa el botón del triángulo, o usa *Ejecutar › Todo* (*Run › All*) para correr todo lo que hay en la ventana (IBM Corp., 2024b). Los resultados aparecen en el Visor. Si prefieres no copiar, abre el archivo `cap01.sps` del repositorio del libro.

Dos reglas de la sintaxis de SPSS te ahorrarán horas. Cada orden termina en punto. Y cada comentario, que empieza con asterisco, también: si un comentario termina en signo de interrogación, SPSS une la línea siguiente al comentario y esa orden desaparece sin dar error.

Paso 1. Crea 4000 microempresas con una afluencia oculta

```
* Bloque 1: 4000 microempresas simuladas con una afluencia de clientes que nadie mide.
* Cierra cualquier otra base abierta: así nadie mezcla datos de dos corridas.
DATASET CLOSE ALL.
SET SEED=2026.
INPUT PROGRAM.
LOOP #i = 1 TO 4000.
COMPUTE afluencia = RV.NORMAL(0, 1).
COMPUTE ruido = RV.NORMAL(0, 2).
END CASE.
END LOOP.
END FILE.
END INPUT PROGRAM.
DATASET NAME microempresas.
* La municipalidad invita sobre todo a los negocios con más clientes.
COMPUTE capacit = (0.8*afluencia + RV.NORMAL(0, 1) > 0.6).
* El efecto verdadero de la capacitación: 1 (mil soles al mes).
COMPUTE ventas = 15 + 1*capacit + 2.5*afluencia + ruido.
FORMATS capacit (F1.0) afluencia ruido ventas (F8.2).
VARIABLE LABELS afluencia 'Afluencia de clientes (oculta)'
  /ruido 'Otros factores (azar)'
  /capacit 'Asistió a la capacitación'
  /ventas 'Ventas mensuales (miles de soles)'.
VALUE LABELS capacit 0 'No asistió' 1 'Asistió'.
DESCRIPTIVES VARIABLES=capacit ventas afluencia /STATISTICS=MEAN STDDEV MIN MAX.
```

`DATASET CLOSE ALL` cierra cualquier otra base abierta, para que nadie mezcle datos de dos sesiones. `SET SEED=2026` fija la semilla: si usas la misma, obtendrás exactamente las mismas cifras que este libro. El bloque `INPUT PROGRAM` … `END INPUT PROGRAM` fabrica 4000 casos sin leer ningún archivo: `LOOP` repite 4000 veces, `END CASE` guarda cada caso y `END FILE` cierra la base (IBM Corp., 2024a).

`RV.NORMAL(0, 1)` sortea un número de una distribución normal con media 0 y desviación 1. Así nace la afluencia de clientes de cada negocio. Nadie la mide en la encuesta; por eso la llamamos oculta. `ruido` reúne todo lo demás que mueve las ventas: suerte, rubro, temporada.

La línea más importante es la de `capacit`. La expresión entre paréntesis vale 1 si es verdadera y 0 si es falsa. Como la afluencia entra en ella, los negocios con más clientes tienen más probabilidad de asistir: es la invitación repartida por las avenidas transitadas. La línea siguiente construye las ventas con un efecto del curso de **exactamente 1** y un efecto de la afluencia de 2,5.

Por último, `FORMATS`, `VARIABLE LABELS` y `VALUE LABELS` dejan la base legible: decimales, etiquetas en español y nombres para el 0 y el 1. `DESCRIPTIVES` equivale a *Analizar › Estadísticos descriptivos › Descriptivos* (*Analyze › Descriptive Statistics › Descriptives*).

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. estándar
Asistió a la capacitación	4000	0	1	.31	.463
Ventas mensuales (miles de soles)	4000	1.86	26.31	15.2085	3.37777
Afluencia de clientes (oculta)	4000	-4.05	3.23	-.0391	.99387
N válido (por lista)	4000				

Como se observa en la tabla de arriba, la media de la capacitación es ,31: el 31 % de los negocios asistió. Las ventas promedian 15,21 miles de soles al mes (*DE* = 3,38), entre 1,86 y 26,31. La afluencia tiene media cercana a cero y desviación cercana a 1, como la pedimos.

Paso 2. Primer intento: lo que haría casi todo tesista

Rocío no conoce la afluencia. Solo tiene la asistencia y las ventas. Hace lo que harías tú: una correlación y una prueba *t*.

```
* Bloque 2: primer intento, lo que haría casi todo tesista.
CORRELATIONS /VARIABLES=capacit ventas /PRINT=TWOTAIL NOSIG.
* La tabla de la prueba t es muy ancha: OUTPUT MODIFY la traspone (filas por columnas).
T-TEST GROUPS=capacit(0 1) /VARIABLES=ventas /ES DISPLAY(FALSE).
OUTPUT MODIFY /SELECT TABLES /IF INSTANCES=LAST /TABLE TRANSPOSE=YES.
```

`CORRELATIONS` es *Analizar › Correlacionar › Bivariadas* (*Analyze › Correlate › Bivariate*). `T-TEST GROUPS=capacit(0 1)` compara las ventas de los dos grupos: es *Analizar › Comparar medias › Prueba T para muestras independientes* (*Analyze › Compare Means › Independent-Samples T Test*). `/ES DISPLAY(FALSE)` apaga la tabla del tamaño del efecto, que veremos en el capítulo 6.

La última línea es un truco de presentación. La tabla de la prueba *t* es tan ancha que no cabe en una página. `OUTPUT MODIFY` toma la última tabla del Visor (`INSTANCES=LAST`) y la traspone: las columnas pasan a ser filas (IBM Corp., 2024a).

Correlaciones

		Asistió a la capacitación	Ventas mensuales (miles de soles)
Asistió a la capacitación	Correlación de Pearson	1	.500**
	Sig. (bilateral)		<.001
	N	4000	4000
Ventas mensuales (miles de soles)	Correlación de Pearson	.500**	1
	Sig. (bilateral)	<.001	
	N	4000	4000

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

Estadísticas de grupo

	Asistió a la capacitación	N	Media	Desv. estándar	Media de error estándar
Ventas mensuales (miles de soles)	No asistió	2755	14.0727	2.97678	.05671
	Asistió	1245	17.7218	2.80732	.07956

Prueba de muestras independientes

			Ventas mensuales (miles de soles)	
			Se asumen varianzas iguales	No se asumen varianzas iguales
Prueba de Levene de igualdad de varianzas	F		4.084	
	Sig.		.043	
prueba t para la igualdad de medias	t		-36.531	-37.347
	gl		3998	2533.844
	Significación	P de un factor	<.001	<.001
		P de dos factores	.000	.000
	Diferencia de medias		-3.64908	-3.64908
	Diferencia de error estándar		.09989	.09771
	95% de intervalo de confianza de la diferencia	Inferior	-3.84492	-3.84068
		Superior	-3.45324	-3.45749

La primera de las tres tablas muestra la correlación que Rocío esperaba: $r = ,50$, con $p < ,001$. Los dos asteriscos que SPSS pone junto a la cifra marcan $p < ,01$ (bilateral). La segunda tabla la traduce a soles. Los negocios que no asistieron venden en promedio 14,07 miles de soles ($DE = 2,98$); los que asistieron, 17,72 ($DE = 2,81$).

La tercera tabla, la de la prueba de muestras independientes, necesita una lectura más lenta. Primero mira la prueba de Levene: $F = 4,08$ con $p = ,043$, menor que ,05. Indica que las varianzas de los grupos difieren, así que se lee la columna «No se asumen varianzas iguales»: $t(2533,84) = -37,35$. La diferencia de medias es −3,65 porque SPSS resta en el orden de los códigos: primero los que no asistieron (0) y después los que sí (1).

Dos rarezas de la salida en español te van a confundir si nadie te avisa. «P de un factor» y «P de dos factores» son una traducción poco feliz de *p* unilateral y bilateral. Y al trasponer la tabla, SPSS muestra la *p* bilateral como ,000. Un valor *p* nunca es cero: lo que SPSS quiere decir es $p < ,001$, y así debes escribirlo.

Conclusión de Rocío: el curso «incrementa significativamente» las ventas en 3,65 miles de soles. Nosotros sabemos que el efecto verdadero es 1. **La comparación más que triplica la verdad.**

Paso 3. Segundo intento: la regresión

Quizá la prueba *t* es demasiado simple. ¿Una regresión lo arregla?

```
* Bloque 3: segundo intento, la regresión simple.
REGRESSION /STATISTICS=COEFF R /DEPENDENT=ventas /METHOD=ENTER capacit.
OUTPUT MODIFY /SELECT TABLES /IF INSTANCES=1 /OBJECTPROPERTIES VISIBLE=NO.
```

`REGRESSION` es *Analizar › Regresión › Lineales* (*Analyze › Regression › Linear*). `/STATISTICS=COEFF R` pide solo los coeficientes y el resumen del modelo. La línea de `OUTPUT MODIFY` oculta la primera tabla del bloque, la de «Variables entradas/eliminadas», que aquí no aporta nada.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados		
		B	Desv. Error	Beta	t	Sig.
1	(Constante)	14.073	.056		252.521	<.001
	Asistió a la capacitación	3.649	.100	.500	36.531	<.001

a. Variable dependiente: Ventas mensuales (miles de soles)

La variable dependiente de la tabla de coeficientes son las ventas mensuales en miles de soles. El coeficiente de la capacitación es $B = 3{,}649$ ($EE = 0{,}100$), con $p < {,}001$. Es la misma diferencia de medias del paso anterior, con tres decimales. No es casualidad: una regresión con una sola variable que vale 0 o 1 **es** una comparación de dos promedios con otro formato. El resumen del modelo da *R*² = ,250, y nada en él avisa del sesgo. Más estrellas, mismo error.

Paso 4. Tercer intento: si pudieras medir la afluencia

Ahora hagamos trampa. Como fabricamos los datos, tenemos la afluencia. Agrégala a la regresión.

```
* Bloque 4: tercer intento, si pudieras medir la afluencia.
REGRESSION /STATISTICS=COEFF R /DEPENDENT=ventas /METHOD=ENTER capacit afluencia.
OUTPUT MODIFY /SELECT TABLES /IF INSTANCES=1 /OBJECTPROPERTIES VISIBLE=NO.
```

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados		
		B	Desv. Error	Beta	t	Sig.
1	(Constante)	14.958	.041		368.521	<.001
	Asistió a la capacitación	1.113	.079	.153	14.158	<.001
	Afluencia de clientes (oculta)	2.440	.037	.718	66.657	<.001

a. Variable dependiente: Ventas mensuales (miles de soles)

En la tabla de coeficientes de este paso, con las mismas ventas como variable dependiente, el coeficiente de la capacitación cae a $B = 1{,}113$ ($EE = 0{,}079$): muy cerca del 1 verdadero. El de la afluencia, $B = 2{,}440$, se acerca al 2,5 que pusimos. El *R*² sube a ,645.

En otras palabras, al comparar negocios con la misma afluencia, el camino de arriba del diagrama queda cerrado. El problema es que Rocío no tiene esta variable. Por eso la llamamos oculta: en la vida real, la variable de confusión más importante casi nunca está en tu base de datos.

Paso 5. Cuarto intento: el sorteo

Volvamos al mundo de Rocío, pero ahora la municipalidad sortea las vacantes.

```
* Bloque 5: cuarto intento, las vacantes se sortean.
COMPUTE sorteo = RV.BERNOULLI(0.5).
COMPUTE ventas_exp = 15 + 1*sorteo + 2.5*afluencia + ruido.
FORMATS sorteo (F1.0) ventas_exp (F8.2).
VARIABLE LABELS sorteo 'Ganó el sorteo de vacantes'
  /ventas_exp 'Ventas mensuales con sorteo (miles de soles)'.
VALUE LABELS sorteo 0 'No ganó' 1 'Ganó'.
T-TEST GROUPS=sorteo(0 1) /VARIABLES=ventas_exp /ES DISPLAY(FALSE).
OUTPUT MODIFY /SELECT TABLES /IF INSTANCES=LAST /TABLE TRANSPOSE=YES.
```

`RV.BERNOULLI(0.5)` es una moneda al aire: vale 1 con probabilidad ,5 y 0 en otro caso. Las ventas del mundo con sorteo usan la misma afluencia y el mismo ruido de antes, y el mismo efecto del curso: 1.

Estadísticas de grupo

	Ganó el sorteo de vacantes	N	Media	Desv. estándar	Media de error estándar
Ventas mensuales con sorteo (miles de soles)	No ganó	1964	14.8627	3.17941	.07174
	Ganó	2036	15.9305	3.16441	.07013

Prueba de muestras independientes

			Ventas mensuales con sorteo (miles de soles)	
			Se asumen varianzas iguales	No se asumen varianzas iguales
Prueba de Levene de igualdad de varianzas	F		.022	
	Sig.		.883	
prueba t para la igualdad de medias	t		-10.644	-10.643
	gl		3998	3991.374
	Significación	P de un factor	<.001	<.001
		P de dos factores	.000	.000
	Diferencia de medias		-1.06775	-1.06775
	Diferencia de error estándar		.10032	.10033
	95% de intervalo de confianza de la diferencia	Inferior	-1.26443	-1.26444
		Superior	-.87107	-.87106

Según la tabla de estadísticas de grupo, los 1964 negocios que perdieron el sorteo venden 14,86 miles de soles y los 2036 que lo ganaron, 15,93. En la tabla de la prueba de muestras independientes, la prueba de Levene no rechaza la igualdad de varianzas (p = ,883), así que se lee la primera columna. La diferencia es de 1,07 miles de soles, con $t(3998) = -10{,}64$ y $p < ,001$.

Con 95 % de confianza, la diferencia está entre 0,87 y 1,26 (SPSS lo imprime con signo negativo por el orden de la resta). Contiene al 1 verdadero. El mismo análisis que fallaba en el paso 2 acierta ahora. No cambió la prueba estadística; cambió quién decidió quién recibía el curso.

Paso 6. La prueba del delito

¿Por qué falló la comparación del paso 2 y acertó la del paso 5? Compara la afluencia entre los grupos.

```
* Bloque 6: la prueba del delito: ¿los grupos eran comparables?.
T-TEST GROUPS=capacit(0 1) /VARIABLES=afluencia /ES DISPLAY(FALSE).
OUTPUT MODIFY /SELECT TABLES /IF INSTANCES=LAST /TABLE TRANSPOSE=YES.
T-TEST GROUPS=sorteo(0 1) /VARIABLES=afluencia /ES DISPLAY(FALSE).
OUTPUT MODIFY /SELECT TABLES /IF INSTANCES=LAST /TABLE TRANSPOSE=YES.
```

Estadísticas de grupo

	Asistió a la capacitación	N	Media	Desv. estándar	Media de error estándar
Afluencia de clientes (oculta)	No asistió	2755	-.3627	.88738	.01691
	Asistió	1245	.6767	.82907	.02350

Estadísticas de grupo

	Ganó el sorteo de vacantes	N	Media	Desv. estándar	Media de error estándar
Afluencia de clientes (oculta)	No ganó	1964	-.0525	.99561	.02247
	Ganó	2036	-.0263	.99225	.02199

La primera tabla, la que compara la afluencia según la asistencia al curso, muestra el delito. Los negocios que no asistieron tienen una afluencia media de −0,36 y los que asistieron, de 0,68: más de una desviación estándar de distancia, con $t(3998) = -35{,}00$ y $p < {,}001$. En la segunda, que la compara según el sorteo, los perdedores y los ganadores del sorteo tienen −0,05 y −0,03. Su diferencia, $t(3998) = -0{,}84$ con $p = {,}404$, es puro azar.

Ahora haz la cuenta de la sección anterior. Los que asistieron tienen 1,04 unidades más de afluencia, y cada unidad sube las ventas en 2,5. Eso da un sesgo de 1,04 × 2,5 = 2,60. Súmale el efecto verdadero y obtienes 3,60, casi exactamente los 3,65 del paso 2. La diferencia que queda es azar.

En la práctica, esta comparación tiene nombre: **prueba de balance**. En una investigación real no puedes hacerla con la variable oculta, pero sí con las que tienes: tamaño del negocio, rubro, años de funcionamiento. Si los grupos ya difieren en lo que mides, sospecha que también difieren en lo que no mides.

Paso 7. Controlar no es magia: el alquiler como proxy

Rocío no tiene la afluencia, pero se le ocurre algo ingenioso. El alquiler del local refleja la ubicación: las avenidas transitadas cobran más. ¿Y si usa el alquiler como sustituto, como **proxy**?

```
* Bloque 7: controlar no es magia, el alquiler como proxy imperfecta.
COMPUTE alquiler = 15 + 4*afluencia + RV.NORMAL(0, 4).
FORMATS alquiler (F8.2).
VARIABLE LABELS alquiler 'Alquiler mensual del local (cientos de soles)'.
REGRESSION /STATISTICS=COEFF R /DEPENDENT=ventas /METHOD=ENTER capacit alquiler.
OUTPUT MODIFY /SELECT TABLES /IF INSTANCES=1 /OBJECTPROPERTIES VISIBLE=NO.
```

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados		
		B	Desv. Error	Beta	t	Sig.
1	(Constante)	10.510	.116		90.878	<.001
	Asistió a la capacitación	2.586	.093	.355	27.704	<.001
	Alquiler mensual del local (cientos de soles)	.261	.008	.435	34.020	<.001

a. Variable dependiente: Ventas mensuales (miles de soles)

La tabla de arriba, otra vez con las ventas mensuales en miles de soles como variable dependiente, da un coeficiente de la capacitación de $B = 2{,}586$ ($EE = 0{,}093$). Mejor que 3,649, pero lejos de 1. El alquiler cerró una parte del camino de arriba, la que comparte con la afluencia, y dejó abierta la otra. El sesgo bajó de 2,65 a 1,59: se fue menos de la mitad.

Ahora mira el *R*²: ,419, mucho mayor que el ,250 de la regresión simple. Un *R*² más alto te dice que el modelo predice mejor las ventas, no que el coeficiente del curso sea más correcto. Controlar con una medida imperfecta del confusor reduce el sesgo, pero no lo elimina.

Recapitulemos. El efecto verdadero es 1. La prueba *t* y la regresión simple dicen 3,65. Con el alquiler como control, 2,59. Con la afluencia, que nadie mide, 1,11. Con sorteo, 1,07. Todas las cifras salieron con $p < ,001$. Solo dos se acercaron a la verdad.

1.8 El error que se repite en las tesis: «influye significativamente»

Busca en el repositorio de tu universidad tesis que empiecen con «Influencia de…» o «Efecto de…». Buena parte de ellas se apoya en un cuestionario que se pasó en un único momento, sin grupo de comparación sorteado y sin una línea sobre las variables de confusión. Y concluyen que una variable «influye significativamente» en otra porque SPSS marcó un asterisco junto a la cifra.

El cálculo suele estar bien. El problema está en el verbo. «Significativo» solo descarta la cuarta historia, la del azar. No dice nada de la causalidad inversa ni de la confusión: lo acabas de ver con el 3,65, que tenía $p < ,001$ y estaba inflado más del triple.

La solución no exige matemáticas nuevas. Exige dos costumbres. La primera: **ajusta el verbo al diseño, no al deseo**. Si nadie sorteó nada, tus verbos son «se asocia con», «varía junto con», «es mayor entre». Guarda «efecto», «impacto» e «influye» para un sorteo o para los métodos de los capítulos 13 y 14.

La segunda: **haz tu lista de confusores antes de abrir SPSS**: para cada uno, anota si lo mediste, cómo lo controlarás y, si no lo tienes, hacia dónde sesgaría tu resultado.

El exceso contrario: meter todas las variables de la base

Hay un tesista que ya aprendió la lección y comete el error opuesto. Pone en la regresión todas las variables del cuestionario, por si acaso. Suena prudente. No lo es.

Algunas variables son consecuencias del tratamiento. Si el curso sube las ventas porque enseña a atraer clientes nuevos, controlar por los clientes nuevos borra justo el camino por el que el curso funciona. Lo verás con cifras en el Ejercicio 1. Wysocki et al. (2022) lo resumen así: el control estadístico solo reduce el sesgo cuando se justifica con una estructura causal explícita.

Resúmelo así: **un control legítimo existía antes del curso; lo que el curso produjo no se controla**. Vuelve a tu diagrama. Si una flecha sale del tratamiento y llega a una variable, esa variable no va como control.

1.9 El error que cometen los investigadores en la redacción de artículos

La frase típica aparece en el resumen: «La capacitación municipal incrementa significativamente las ventas de las microempresas ($r = 0,50$; $p = 0,000$)». En una sola línea hay tres fallas.

Primero, el verbo promete un efecto que el diseño no puede dar. «Incrementa» afirma causa con datos observacionales. Es exactamente lo que encontraron Haber et al. (2022): verbos prudentes en una parte del texto y recomendaciones causales en otra. El revisor lee el resumen, busca el diseño y, si no hay sorteo ni estrategia contra la confusión, devuelve el artículo.

Segundo, las cifras no siguen APA 7. Un valor *p* nunca es 0,000: se escribe $p < ,001$. Y los estadísticos que no pueden pasar de 1, como *r* o *p*, van sin cero inicial.

Tercero, la advertencia aparece demasiado lejos. Una frase de limitaciones en la última página no deshace el verbo causal del resumen. El tipo de diseño tiene que quedar claro desde las primeras líneas, porque muchos lectores no pasarán de ahí.

Cómo se escribe bien: «Las microempresas que asistieron a la capacitación registraron ventas mensuales mayores que las que no asistieron (M = 17,72 frente a M = 14,07 miles de soles; $t(2533{,}84) = -37{,}35$; $p < ,001$; $r = ,50$). Como la asistencia no fue aleatoria y los negocios asistentes tenían más afluencia de clientes, la diferencia se interpreta como una asociación y no como el efecto del curso».

Si las vacantes se hubieran sorteado, podrías escribir: «La asignación por sorteo permite interpretar la diferencia como efecto: 1,07 miles de soles al mes, IC 95 % [0,87; 1,26]; $t(3998) = 10{,}64$; $p < ,001$».

RESUMEN Y CONCLUSIONES

1. Detrás de una correlación puede haber una causa, una causa inversa, una variable de confusión o simple azar. SPSS no distingue entre las cuatro, y el valor p solo habla del azar.
2. Medir un efecto exige comparar a cada negocio consigo mismo con y sin tratamiento, y una de esas dos versiones nunca existe. Por eso comparas grupos distintos.
3. **Diferencia observada = efecto + sesgo de selección.** Si los grupos ya eran distintos, la diferencia te engaña aunque $p < ,001$.
4. Una regresión con una variable 0/1 es una comparación de promedios. Controlar ayuda solo si mides bien el confusor; con un proxy queda sesgo, y controlar una consecuencia del tratamiento lo empeora.
5. Un sorteo deja a los dos grupos iguales en promedio y borra el sesgo de selección. Sin sorteo, escribe el ensayo diana, haz tu diagrama y elige el verbo según tu diseño.

Desde el capítulo 2 trabajarás con datos reales. Vas a dejar lista en SPSS una base de encuesta y otra de la ENAHO: etiquetas, valores perdidos, ítems invertidos y dimensiones sumadas.

EJERCICIOS RESUELTOS

Encontrarás los tres al final de `cap01.sps`, después de la línea `EJERCICIOS RESUELTOS`. Necesitan la base de los pasos 1 a 7 todavía abierta, así que ejecútalos a continuación y uno tras otro. Antes de correr cada uno, intenta predecir el resultado.

Ejercicio 1. Controlar por lo que el curso provocó

Supón que, en el mundo con sorteo, el curso funciona porque enseña a conseguir clientes nuevos: ganar el sorteo suma 8 clientes al mes y cada cliente nuevo suma 0,125 miles de soles. El efecto total sigue siendo 8 × 0,125 = 1. Un tesista decide controlar por los clientes nuevos «para aislar el efecto del curso». ¿Qué obtiene?

```
* Ejercicio 1: controlar por una consecuencia de la capacitación.
COMPUTE clientes = 30 + 8*sorteo + RV.NORMAL(0, 5).
COMPUTE ventas_m = 11.25 + 0.125*clientes + 2.5*afluencia + ruido.
FORMATS clientes ventas_m (F8.2).
VARIABLE LABELS clientes 'Clientes nuevos en el mes'
  /ventas_m 'Ventas mensuales, ejercicio 1 (miles de soles)'.
REGRESSION /STATISTICS=COEFF /DEPENDENT=ventas_m
  /METHOD=ENTER sorteo /METHOD=ENTER clientes.
OUTPUT MODIFY /SELECT TABLES /IF INSTANCES=1 /OBJECTPROPERTIES VISIBLE=NO.
```

Escribir dos subcomandos `/METHOD=ENTER` produce una **regresión jerárquica**: el Modelo 1 incluye solo el sorteo y el Modelo 2 agrega los clientes nuevos.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados		
		B	Desv. Error	Beta	t	Sig.
1	(Constante)	14.836	.073		203.143	<.001
	Ganó el sorteo de vacantes	1.098	.102	.167	10.726	<.001
2	(Constante)	11.025	.305		36.148	<.001
	Ganó el sorteo de vacantes	.043	.130	.007	.336	.737
	Clientes nuevos en el mes	.128	.010	.254	12.856	<.001

a. Variable dependiente: Ventas mensuales, ejercicio 1 (miles de soles)

Solución. La variable dependiente de la tabla de arriba son las ventas del ejercicio 1, en miles de soles. En el Modelo 1, el sorteo tiene $B = 1{,}098$ ($EE = 0{,}102$), con $p < {,}001$: el efecto total, cerca de 1. En el Modelo 2, el coeficiente del sorteo cae a $B = 0{,}043$ ($EE = 0{,}130$), con $p = {,}737$. El curso parece no servir para nada.

Sin embargo, no es que el curso no funcione: el tesista preguntó cuánto suben las ventas por el curso **sin contar** los clientes que el curso trajo. Como ese era su único camino, la respuesta es casi cero. Los clientes nuevos son un mediador, una variable que el tratamiento provoca, y no se controlan cuando quieres el efecto total.

Ejercicio 2. ¿Más datos curan el sesgo?

Un compañero de Rocío le dice que su problema es la muestra: con más casos, el resultado se acercará a la verdad. Comprobémoslo al revés: tomemos al azar solo el 10 % de los negocios y repitamos la prueba *t* del paso 2.

```
* Ejercicio 2: ¿más datos curan el sesgo? Una muestra de 10 %.
COMPUTE muestra = (RV.UNIFORM(0, 1) < 0.10).
FORMATS muestra (F1.0).
FILTER BY muestra.
T-TEST GROUPS=capacit(0 1) /VARIABLES=ventas /ES DISPLAY(FALSE).
OUTPUT MODIFY /SELECT TABLES /IF INSTANCES=LAST /TABLE TRANSPOSE=YES.
FILTER OFF.
```

`RV.UNIFORM(0, 1)` da a cada negocio un número al azar entre 0 y 1; quedarse con los menores que 0.10 equivale a elegir uno de cada diez. `FILTER BY muestra` analiza solo los casos con valor 1, y `FILTER OFF` devuelve la base completa.

Prueba de muestras independientes

			Ventas mensuales (miles de soles)	
			Se asumen varianzas iguales	No se asumen varianzas iguales
Prueba de Levene de igualdad de varianzas	F		.004	
	Sig.		.951	
prueba t para la igualdad de medias	t		-11.818	-11.963
	gl		395	208.356
	Significación	P de un factor	<.001	<.001
		P de dos factores	.000	.000
	Diferencia de medias		-3.85752	-3.85752
	Diferencia de error estándar		.32641	.32246
	95% de intervalo de confianza de la diferencia	Inferior	-4.49924	-4.49321
		Superior	-3.21580	-3.22182

Solución. Con 397 negocios elegidos al azar (285 que no asistieron y 112 que sí), la tabla de arriba da una diferencia de 3,86 miles de soles, con $t(395) = -11{,}82$ y $p < {,}001$. El intervalo de confianza al 95 % es [3,22; 4,50]. Con los 4000 negocios del paso 2, el intervalo iba de 3,45 a 3,84.

En resumen, más datos estrechan el intervalo, pero alrededor del valor equivocado: ninguno de los dos contiene al 1. El tamaño de muestra te da precisión, no te quita sesgo. Una muestra enorme y sesgada te da un error muy preciso.

Ejercicio 3. La correlación parcial, la que tu tesis correlacional no calcula

Muchas tesis correlacionales reportan solo la correlación de Pearson. SPSS tiene, en el mismo menú, la **correlación parcial**: la correlación entre dos variables después de quitarles la parte lineal que comparten con una tercera. Calcula la correlación entre capacitación y ventas controlando primero por el alquiler y luego por la afluencia.

```
* Ejercicio 3: correlación parcial, la que tu tesis correlacional no calcula.
PARTIAL CORR /VARIABLES=capacit ventas BY alquiler /SIGNIFICANCE=TWOTAIL.
PARTIAL CORR /VARIABLES=capacit ventas BY afluencia /SIGNIFICANCE=TWOTAIL.
```

`PARTIAL CORR` equivale a *Analizar › Correlacionar › Parciales* (*Analyze › Correlate › Partial*). Lo que va después de `BY` es la variable de control (IBM Corp., 2024a).

Correlaciones

Variables de control			Asistió a la capacitación	Ventas mensuales (miles de soles)
Alquiler mensual del local (cientos de soles)	Asistió a la capacitación	Correlación	1.000	.401
		Significación (bilateral)	.	<.001
		gl	0	3997
	Ventas mensuales (miles de soles)	Correlación	.401	1.000
		Significación (bilateral)	<.001	.
		gl	3997	0

Correlaciones

Variables de control			Asistió a la capacitación	Ventas mensuales (miles de soles)
Afluencia de clientes (oculta)	Asistió a la capacitación	Correlación	1.000	.219
		Significación (bilateral)	.	<.001
		gl	0	3997
	Ventas mensuales (miles de soles)	Correlación	.219	1.000
		Significación (bilateral)	<.001	.
		gl	3997	0

Solución. La correlación simple era $r = ,50$. Controlando por el alquiler baja a ,401 (primera tabla de arriba), y controlando por la afluencia, a ,219 (segunda tabla); las dos con 3997 grados de libertad y $p < ,001$. La historia es la misma del paso 7, ahora en el lenguaje de tu tesis correlacional: un control imperfecto quita parte de la confusión, el control correcto quita mucho más.

Por cierto, que ,219 no sea cero tampoco es un error. Queda el efecto verdadero del curso. Pero fíjate en que la correlación parcial no te dice cuánto suben las ventas en soles. Para eso necesitas el coeficiente de una regresión, que es el tema de los capítulos 9 y 10.

Para tu tesis o artículo

1. **Escribe las cuatro historias de tu correlación principal.** Para tu variable independiente y tu variable dependiente, redacta en una línea cada explicación: causa, causa inversa, confusión y azar. Al lado, anota qué dato o qué análisis descartaría cada una.

2. **Dibuja tu diagrama causal** con todas las variables de tu instrumento. Marca en rojo las que el tratamiento podría provocar: esas no van como control. Marca en azul los confusores que no mediste: van en las limitaciones, con la dirección del sesgo que producirían.

3. **Haz una prueba de balance.** Compara tus grupos en las variables que tienes, con la prueba *t* del paso 6. Si difieren en lo que mides, dilo, y sospecha que difieren en lo que no mides.

Lo que te va a preguntar el jurado: «Su diseño es correlacional: ¿con qué base afirma que una variable *influye* en la otra?», «¿Qué factor no medido podría estar detrás de ambas?» y «¿Cómo descarta que sea al revés, que la variable dependiente cause la independiente?». Si tienes escritas las tareas 1 y 2, ya tienes las tres respuestas.

Si es un artículo: el revisor busca el diseño antes que los resultados. Declara en el resumen si la asignación fue aleatoria o no, usa el mismo verbo del título a las conclusiones y no recomiendes una política que tu diseño no puede sostener. Si quieres hablar de efectos con datos observacionales, presenta tu ensayo diana y una de las estrategias del capítulo 14.

Referencias

Haber, N. A., Wieten, S. E., Rohrer, J. M., Arah, O. A., Tennant, P. W. G., Stuart, E. A., Murray, E. J., Pilleron, S., Lam, S. T., Riederer, E., Howcutt, S. J., Simmons, A. E., Leyrat, C., Schoenegger, P., Booman, A., Kang Dufour, M.-S., O'Donoghue, A. L., Baglini, R., Do, S., . . . Fox, M. P. (2022). Causal and associational language in observational health research: A systematic evaluation. *American Journal of Epidemiology, 191*(12), 2084–2097. https://doi.org/10.1093/aje/kwac137

IBM Corp. (2024a). *IBM SPSS Statistics 29 command syntax reference*. https://www.ibm.com/docs/SSLVMB_29.0.0/pdf/IBM_SPSS_Statistics_Command_Syntax_Reference.pdf

IBM Corp. (2024b). *IBM SPSS Statistics 29 core system user's guide*. https://www.ibm.com/docs/SSLVMB_29.0.0/pdf/IBM_SPSS_Statistics_Core_System_User_Guide.pdf

Martínez-Abad, F., & León, J. (2023). Inferencia causal en investigación educativa: Análisis de la causalidad en estudios observacionales de carácter transversal. *RELIEVE. Revista Electrónica de Investigación y Evaluación Educativa, 29*(2), Artículo 3. https://doi.org/10.30827/relieve.v29i2.26843

Núñez, I., & Lajous, M. (2025). El ensayo diana para la inferencia causal en estudios observacionales. *Salud Pública de México, 67*(1), 83–90. https://doi.org/10.21149/16056

Savitz, D. A., & Wellenius, G. A. (2023). Can cross-sectional studies contribute to causal inference? It depends. *American Journal of Epidemiology, 192*(4), 514–516. https://doi.org/10.1093/aje/kwac037

Wysocki, A. C., Lawson, K. M., & Rhemtulla, M. (2022). Statistical control requires causal justification. *Advances in Methods and Practices in Psychological Science, 5*(2), 1–19. https://doi.org/10.1177/25152459221095823