14.310x – Data Analysis for Social Scientists

MicroMasters en Statistics and Data Science (SDS) – MITx

BREIT (Brescia Institute of Technology) · Aporta — Grupo Breca

Clase 8

Causalidad, Experimentos Aleatorizados y Regresión No Paramétrica

Soluciones

Módulos oficiales del MicroMaster:
M7 – Causality, Analyzing Randomized Experiments and Nonparametric Regression

B.Sc. Cristian Lazo Quispe
clazoq@uni.pe
linkedin.com/in/cristian-lazo-quispe · github.com/CristianLazoQuispe

Material del *Advanced Program in Data Science & Global Skills*, diseñado y desarrollado por **BREIT (Brescia Institute of Technology)**, iniciativa de **Aporta** —plataforma de innovación e impacto social del **Grupo Breca**— en alianza con el **MIT IDSS** (Institute for Data, Systems, and Society).

11 de agosto de 2026

Niveles de dificultad *(difficulty levels)*

Cada problema está rotulado por color según su nivel, de lo más accesible (Nivel 1) a lo más difícil (Nivel 5). Empieza por donde te sientas cómodo y avanza a tu ritmo; los niveles 4–5 son extensiones para profundizar, no un requisito.

- **Nivel 1 – Fundamentos**: calentamiento muy guiado; para quien parte de casi cero.
- **Nivel 2 – Núcleo**: el nivel objetivo del curso/examen.
- **Nivel 3 – Aplicación**: combina varias ideas en contexto realista.
- **Nivel 4 – Reto**: difícil, integra varios temas.
- **Nivel 5 – Reto+**: muy difícil / fuera del scope típico.

1. Problemas y soluciones

Nivel 1 – Fundamentos (warm-up)

Para quien parte de casi cero: cálculos directos y guiados.

Ejercicio 1.1 (Correlación no es causalidad: encuentra el confusor)**.** Para cada afirmación observacional, identifica una *tercera variable* (confusor) que pueda explicar la asociación sin que exista una relación causal directa entre las dos primeras.

(a) "En los distritos donde hay más heladerasías vendiendo helado, hay más ahogamientos."

(b) "Los niños con pies más grandes leen mejor."

(c) "Los hogares que tienen más libros en casa tienen hijos con mejores notas."

Solución

(a) El **calor / la temporada de verano**: en verano se vende más helado y más gente va a nadar (más ahogamientos). El helado no causa ahogamientos.
(b) La **edad**: niños mayores tienen pies más grandes y leen mejor.
(c) El **nivel socioeconómico / educativo de los padres**: hogares con más recursos compran más libros y apoyan más el estudio. En los tres casos la correlación es real, pero la causa es una variable de fondo (confusor) que mueve a ambas a la vez.

Ejercicio 1.2 (El confusor en un diagrama: nombrar la flecha que falta)**.** Un titular dice: "Los estudiantes que duermen más horas sacan mejores notas." Un amigo concluye que dormir *causa* buenas notas. Antes de creerle, dibujemos las relaciones. Sea X = horas de sueño, Y = nota, y Z un posible confusor.

(a) Propon un confusor Z concreto (una variable que afecte *tanto* a las horas de sueño *como* a la nota).

(b) Para que Z sea de verdad un confusor, debe haber una flecha $Z \to X$ y otra $Z \to Y$. Describe en palabras esas dos flechas para tu Z.

(c) Si Z explica toda la asociación, ¿qué pasaría con la relación X–Y *entre estudiantes con el mismo valor de* Z?

Solución

(a) Un confusor típico es la **carga de trabajo / ansiedad** (o, equivalentemente, "buena organización del tiempo"). También sirve la **salud** o el **nivel socioeconómico**.
(b) Con Z = organización del tiempo: $Z \to X$ (quien se organiza bien *duerme más*, no trasnocha) y $Z \to Y$ (quien se organiza bien *estudia mejor y saca mejores notas*). Las dos flechas salen de Z hacia X y hacia Y.
(c) Si Z lo explica todo, entonces *a igual organización* (mismo Z) dormir más ya **no** se asociaría con mejores notas: la relación X–Y se desvanecería. Eso es exactamente lo que significa "controlar por el confusor".

Ejercicio 1.3 (Resultados potenciales: el efecto individual)**.** Para una persona i definimos dos *resultados potenciales*: $Y_i(1)$ es su resultado si recibe el tratamiento y $Y_i(0)$ si no lo recibe. El efecto causal individual es $\tau_i = Y_i(1) - Y_i(0)$. Considera un programa de tutorías y el puntaje final (sobre 100):

Persona i	$Y_i(1)$ (con tutoría)	$Y_i(0)$ (sin tutoría)	τ_i
Ana	78	70	?
Beto	65	65	?
Carla	90	84	?

(a) Completa la columna τ_i para cada persona.

(b) ¿Para quién *no* tuvo efecto el programa?

Solución

(a) $\tau_{\text{Ana}} = 78 - 70 = 8$; $\tau_{\text{Beto}} = 65 - 65 = 0$; $\tau_{\text{Carla}} = 90 - 84 = 6$.
(b) Para **Beto**: su efecto individual es 0, el puntaje es el mismo con o sin tutoría.

Ejercicio 1.4 (Leer una tabla de $Y(1)/Y(0)$ y calcular efectos individuales)**.** Cuatro personas participan (hipotéticamente) en un taller de finanzas personales. La tabla da, para cada una, sus dos resultados potenciales en ahorro mensual (en soles): $Y_i(1)$ si toma el taller y $Y_i(0)$ si no lo toma.

Persona i	$Y_i(1)$	$Y_i(0)$	efecto $\tau_i = Y_i(1) - Y_i(0)$
Rosa	120	100	?
Luis	90	90	?
Marta	150	130	?
Pedro	80	95	?

(a) Completa la columna de efectos individuales τ_i.

(b) ¿A quién *perjudicó* el taller (efecto negativo)? ¿A quién no le afectó?

(c) ¿Qué persona obtuvo el mayor beneficio?

Solución

(a) $\tau_{\text{Rosa}} = 120 - 100 = 20$; $\tau_{\text{Luis}} = 90 - 90 = 0$; $\tau_{\text{Marta}} = 150 - 130 = 20$; $\tau_{\text{Pedro}} = 80 - 95 = -15$.
(b) El taller **perjudicó** a **Pedro** ($\tau = -15$, ahorra 15 soles menos); a **Luis** *no* le afectó ($\tau = 0$).
(c) El mayor beneficio es de 20 soles, empatan **Rosa** y **Marta**.

Ejercicio 1.5 (ATE como diferencia de medias en un RCT (guiado paso a paso))**.** En un pequeño experimento aleatorizado se mide el rendimiento (sobre 20) de 6 alumnos. Tres fueron asignados *al azar* al **tratamiento** (una nueva guía de estudio) y tres al **control**. Resultados observados:

Tratamiento ($W = 1$)	16	14	18
Control ($W = 0$)	12	11	13

(a) Calcula el promedio del grupo tratado $\bar{Y}_t$ y del control $\bar{Y}_c$.

(b) El estimador del ATE en un RCT es la *diferencia de medias* $\hat{\tau} = \bar{Y}_t - \bar{Y}_c$. Calcúlalo.

(c) Interpreta el resultado en una frase.

Solución

(a) $\bar{Y}_t = \dfrac{16 + 14 + 18}{3} = \dfrac{48}{3} = 16$; $\bar{Y}_c = \dfrac{12 + 11 + 13}{3} = \dfrac{36}{3} = 12$.
(b) $\hat{\tau} = \bar{Y}_t - \bar{Y}_c = 16 - 12 = 4$ puntos.
(c) En promedio, usar la nueva guía se asocia a un rendimiento **mayor** en 4 puntos. Como la asignación fue al azar, $\hat{\tau}$ estima el efecto causal promedio (ATE).

Ejercicio 1.6 (Tratamiento y control en un RCT)**.** Una ONG evalúa un programa de microcréditos. De 500 solicitantes elegibles, asigna **al azar** 250 a recibir el crédito y 250 a una lista de espera.

(a) ¿Quiénes forman el grupo de **tratamiento** y quiénes el de **control**?

(b) ¿Qué variable indicadora W_i vale 1 y cuál 0? Da su valor para alguien en la lista de espera.

Solución

(a) **Tratamiento**: los 250 que reciben el microcrédito; **control**: los 250 en lista de espera (no reciben el crédito durante el estudio).
(b) $W_i = 1$ si la persona i fue asignada al tratamiento y $W_i = 0$ si fue asignada al control. Para alguien en lista de espera, $W_i = 0$.

Ejercicio 1.7 (El problema fundamental de la inferencia causal)**.** En la tabla de tutorías pudimos ver $Y_i(1)$ *y* $Y_i(0)$ para cada persona, pero en la realidad eso es imposible: a cada persona la observamos *o* tratada *o* de control, nunca ambas.

(a) Si Ana recibió tutoría ($W_{\text{Ana}} = 1$), ¿cuál de sus dos resultados potenciales *observamos* y cuál es el "dato faltante" (*counterfactual*)?

(b) Explica en una frase por qué esto obliga a comparar *distintas* personas (grupos) en lugar de a la misma persona consigo misma.

Solución

(a) Observamos $Y_{\text{Ana}}(1) = 78$ (el resultado bajo el tratamiento que recibió); el faltante (*counterfactual*) es $Y_{\text{Ana}}(0)$, lo que habría pasado sin tutoría.
(b) Como nunca vemos los dos potenciales de la misma persona, no podemos calcular τ_i directamente; estimamos el efecto *promedio* comparando un grupo tratado con un grupo de control (el "problema fundamental de la inferencia causal", Holland 1986).

Ejercicio 1.8 (¿Por qué ayuda la aleatorización?)**.** Un colega propone evaluar un curso de capacitación comparando a quienes *decidieron* inscribirse con quienes *no* se inscribieron.

(a) Da una razón por la que esa comparación puede no reflejar el efecto causal del curso (piensa en quiénes se inscriben).

(b) Explica en una frase cómo la asignación *al azar* resuelve ese problema.

Solución

(a) Quienes se inscriben suelen ser más motivados, más jóvenes o con más tiempo; esas características también mejoran sus resultados aunque el curso no sirviera. La diferencia mezcla el efecto del curso con esas diferencias previas (*sesgo de selección*).
(b) Al asignar al azar, los grupos de tratamiento y control quedan *en promedio* iguales en todo (motivación, edad, etc.), así que cualquier diferencia posterior en resultados puede atribuirse al tratamiento.

Nivel 2 – Núcleo (core)

El nivel objetivo del curso/examen; todos deberían llegar aquí.

Ejercicio 1.9 (¿Causal u observacional?)**.** Clasifica cada diseño como **experimental (causal por construcción)** u **observacional (requiere supuestos extra)**.

(a) Se sortea quién recibe un bono educativo y se compara con quien no.

(b) Se comparan los salarios de quienes ya tenían y no tenían título universitario.

(c) Una app asigna al azar a la mitad de sus usuarios una nueva pantalla y mide clics.

Solución

(a) **Experimental**: el sorteo es una asignación aleatoria (RCT).
(b) **Observacional**: la gente eligió estudiar o no; hay confusores (habilidad, familia) ⇒ requiere supuestos.
(c) **Experimental**: es un A/B test, asignación al azar de la pantalla.

Ejercicio 1.10 (Calcular el ATE con resultados potenciales completos)**.** El efecto promedio de tratamiento (*Average Treatment Effect*) es $\text{ATE} = \mathbb{E}[Y(1) - Y(0)] = \frac{1}{N}\sum_{i=1}^{N}\big(Y_i(1) - Y_i(0)\big)$. Para una población hipotética de $N = 4$ personas conocemos ambos potenciales (ingreso mensual, en cientos de soles):

i	$Y_i(1)$	$Y_i(0)$
1	14	12
2	11	10
3	18	15
4	13	13

(a) Calcula los efectos individuales τ_i.

(b) Calcula el ATE.

Solución

(a) $\tau_1 = 14 - 12 = 2$; $\tau_2 = 11 - 10 = 1$; $\tau_3 = 18 - 15 = 3$; $\tau_4 = 13 - 13 = 0$.
(b) $\text{ATE} = \dfrac{2+1+3+0}{4} = \dfrac{6}{4} = 1.5$ (es decir, 150 soles en promedio). Equivale a $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \dfrac{14+11+18+13}{4} - \dfrac{12+10+15+13}{4} = 14 - 12.5 = 1.5$.

Ejercicio 1.11 (Estimar el efecto como diferencia de medias)**.** En el RCT de microcréditos se mide el ingreso mensual (en soles) un año después. El promedio observado es $\bar{Y}_t = 1\,420$ en tratamiento y $\bar{Y}_c = 1\,300$ en control.

(a) El estimador del efecto en un RCT es $\hat{\tau} = \bar{Y}_t - \bar{Y}_c$. Calcúlalo.

(b) Interpreta el signo y la magnitud en una frase.

Solución

(a) $\hat{\tau} = 1\,420 - 1\,300 = 120$ soles.
(b) En promedio, recibir el microcrédito se asocia a un ingreso mensual **mayor** en 120 soles. Como el tratamiento se asignó al azar, $\hat{\tau}$ es un estimador insesgado del efecto causal promedio (ATE).

Ejercicio 1.12 (Y^{obs}, Y^{miss} y notación W_i)**.** Con la notación de resultados potenciales, lo que observamos es $Y_i^{\text{obs}} = Y_i(W_i)$ y el faltante es $Y_i^{\text{miss}} = Y_i(1 - W_i)$.

(a) Para una persona con $W_i = 1$ y $Y_i(1) = 9$, $Y_i(0) = 6$, ¿cuánto vale Y_i^{obs} y cuánto Y_i^{miss}?

(b) Para otra con $W_i = 0$ y $Y_i(1) = 8$, $Y_i(0) = 5$, repite.

Solución

(a) $W_i = 1 \Rightarrow Y_i^{\text{obs}} = Y_i(1) = 9$ y $Y_i^{\text{miss}} = Y_i(0) = 6$.
(b) $W_i = 0 \Rightarrow Y_i^{\text{obs}} = Y_i(0) = 5$ y $Y_i^{\text{miss}} = Y_i(1) = 8$.
En cada fila solo uno de los dos potenciales es "dato"; el otro es contrafactual.

Ejercicio 1.13 (Leer una tabla de RCT (estilo Oregon))**.** Inspirado en el *Oregon Health Insurance Experiment*, una tabla reporta, para el indicador "se reporta en buena salud" (proporción), la media de control y el *efecto* estimado (coeficiente sobre el indicador de tratamiento):

Variable	Media control	Efecto (E.E.)
Buena salud (proporción)	0.548	0.039 (0.008)

(a) ¿Cuál es la proporción estimada en el grupo de **tratamiento**?

(b) ¿Cuántos puntos porcentuales subió "buena salud" por el tratamiento?

Solución

(a) Tratamiento = control + efecto $= 0.548 + 0.039 = 0.587$.
(b) El efecto es 0.039, es decir ≈ 3.9 puntos porcentuales más de personas que se reportan en buena salud.

Ejercicio 1.14 (Error estándar e IC del efecto en un RCT)**.** En un RCT, el estimador del efecto es $\hat{\tau} = \bar{Y}_t - \bar{Y}_c$ y su varianza estimada es $\widehat{\text{Var}}(\hat{\tau}) = \dfrac{s_t^2}{N_t} + \dfrac{s_c^2}{N_c}$ (suma de varianzas de dos grupos independientes). Un programa de nutrición da, sobre la talla (cm): $\bar{Y}_t = 102$, $s_t^2 = 36$, $N_t = 100$ y $\bar{Y}_c = 100$, $s_c^2 = 64$, $N_c = 100$.

(a) Calcula $\hat{\tau}$, $\widehat{\text{Var}}(\hat{\tau})$ y el error estándar $\text{ee} = \sqrt{\widehat{\text{Var}}(\hat{\tau})}$.

(b) Da el IC del 95 %, $\hat{\tau} \pm 1.96\,\text{ee}$ (muestra grande $\Rightarrow$ normal).

Solución

(a) $\hat{\tau} = 102 - 100 = 2$. $\widehat{\mathrm{Var}}(\hat{\tau}) = \dfrac{36}{100} + \dfrac{64}{100} = 0.36 + 0.64 = 1$. ee $= \sqrt{1} = 1$.
(b) IC: $2 \pm 1.96(1) = [0.04,\ 3.96]$. Con 95 % de confianza el efecto está entre 0.04 y 3.96 cm.

Ejercicio 1.15 (Prueba de hipótesis del efecto (conexión con la Clase 6))**.** Con los datos del problema anterior ($\hat{\tau} = 2$, ee $= 1$) se prueba si el efecto es nulo: $H_0 : \tau = 0$ vs. $H_1 : \tau \neq 0$. El estadístico es $t = \dfrac{\hat{\tau}}{\text{ee}}$ (con N grande, $\approx \mathcal{N}(0,1)$).

(a) Calcula el estadístico y compáralo con $z_{0.975} = 1.96$.

(b) Conecta con el IC: ¿es coherente la decisión con que el IC del 95 % excluya el 0?

Solución

(a) $t = \dfrac{2}{1} = 2 > 1.96$: **se rechaza** H_0 al 5 %. Hay evidencia de un efecto positivo del programa sobre la talla.
(b) Sí: el IC del 95 % fue [0.04, 3.96], que *no* contiene a 0; por la dualidad IC↔prueba de dos colas, rechazar H_0 y excluir el 0 del IC son la misma decisión.

Ejercicio 1.16 (A/B testing: comparar dos proporciones)**.** Una plataforma educativa prueba dos versiones de un correo de recordatorio. La versión A (control) se envía a $N_c = 1\,000$ usuarios y 200 vuelven a la app; la versión B (tratamiento) a $N_t = 1\,000$ y vuelven 250. Sea $\hat{p}_t - \hat{p}_c$ el efecto, con ee $= \sqrt{\dfrac{\hat{p}_t(1-\hat{p}_t)}{N_t} + \dfrac{\hat{p}_c(1-\hat{p}_c)}{N_c}}$.

(a) Calcula $\hat{p}_c$, $\hat{p}_t$ y la diferencia $\hat{p}_t - \hat{p}_c$.

(b) Calcula el error estándar y el estadístico $z = (\hat{p}_t - \hat{p}_c)/\text{ee}$; ¿es significativo al 5 %?

Solución

(a) $\hat{p}_c = \dfrac{200}{1000} = 0.20$, $\hat{p}_t = \dfrac{250}{1000} = 0.25$; diferencia $= 0.05$.
(b) ee $= \sqrt{\dfrac{0.25 \cdot 0.75}{1000} + \dfrac{0.20 \cdot 0.80}{1000}} = \sqrt{0.0001875 + 0.00016} = \sqrt{0.0003475} \approx 0.01864$.
$z = \dfrac{0.05}{0.01864} \approx 2.68 > 1.96$: **sí** es significativo al 5 %. La versión B aumenta el retorno.

Ejercicio 1.17 (Diferencia observada ≠ ATE: sesgo de selección numérico)**.** En una población hipotética de $N = 6$ conocemos $Y(1)$, $Y(0)$ y quién se trató (W). Quienes más se benefician tienden a tratarse (auto-selección):

i	$Y_i(1)$	$Y_i(0)$	W_i
1	10	6	1
2	9	5	1
3	11	7	1
4	6	5	0
5	5	4	0
6	7	6	0

(a) Calcula el verdadero ATE $= \frac{1}{6}\sum_i (Y_i(1) - Y_i(0))$.

(b) Calcula la diferencia de medias *observada* $\bar{Y}_t - \bar{Y}_c$ (usando $Y_i(1)$ para los tratados y $Y_i(0)$ para los controles). Compárala con el ATE.

Solución

(a) Efectos: $4, 4, 4, 1, 1, 1$. ATE $= \frac{4+4+4+1+1+1}{6} = \frac{15}{6} = 2.5$.

(b) Tratados (filas 1–3) observan $Y(1)$: $\bar{Y}_t = \frac{10+9+11}{3} = 10$. Controles (filas 4–6) observan $Y(0)$: $\bar{Y}_c = \frac{5+4+6}{3} = 5$. Diferencia observada $= 10 - 5 = 5 \neq 2.5$. La auto-selección *infla* el efecto: los tratados tienen mejores resultados *aun sin tratarse*.

Ejercicio 1.18 (Confusión: la diferencia ingenua puede tener el signo equivocado)**.** En un hospital, la tasa de mortalidad entre quienes recibieron cirugía es *mayor* que entre quienes recibieron tratamiento conservador.

(a) Propon un confusor que explique por qué la comparación ingenua puede sugerir que "la cirugía mata" aunque salve vidas.

(b) ¿Qué diseño eliminaría ese confusor?

Solución

(a) La **gravedad inicial**: a los pacientes más graves se les opera; su mayor mortalidad se debe a su estado, no a la cirugía. El confusor (gravedad) afecta tanto a recibir cirugía como al resultado.
(b) Un **RCT**: asignar la cirugía al azar (cuando sea ético) equilibra la gravedad entre grupos y aisla el efecto causal. Sin RCT, habría que *controlar* por gravedad (estratificar/regresión).

Ejercicio 1.19 (p-valor de un efecto de RCT)**.** Un RCT educativo estima un efecto $\hat{\tau} = 6$ puntos con error estándar ee $= 3$, muestra grande.

(a) Calcula $z = \hat{\tau}/\text{ee}$ y el p-valor de dos colas $p = 2\,\mathbb{P}(Z > z)$. *(Usa* $\mathbb{P}(Z > 2) \approx 0.0228$*.)*

(b) Decide al 5 % y al 1 %.

Solución

(a) $z = \dfrac{6}{3} = 2$. $p = 2\,\mathbb{P}(Z > 2) = 2(0.0228) = 0.0456$.
(b) Al 5 %: $0.0456 < 0.05$, **se rechaza** $H_0 : \tau = 0$. Al 1 %: $0.0456 > 0.01$, **no** se rechaza. Conviene reportar el p-valor para que el lector juzgue.

Nivel 3 – Aplicación

Combina dos o más ideas en contexto realista; consolida lo aprendido.

Ejercicio 1.20 (IC del efecto en la tabla de Oregon)**.** De la tabla estilo Oregon, el efecto sobre "buena salud" es 0.039 con error estándar 0.008.

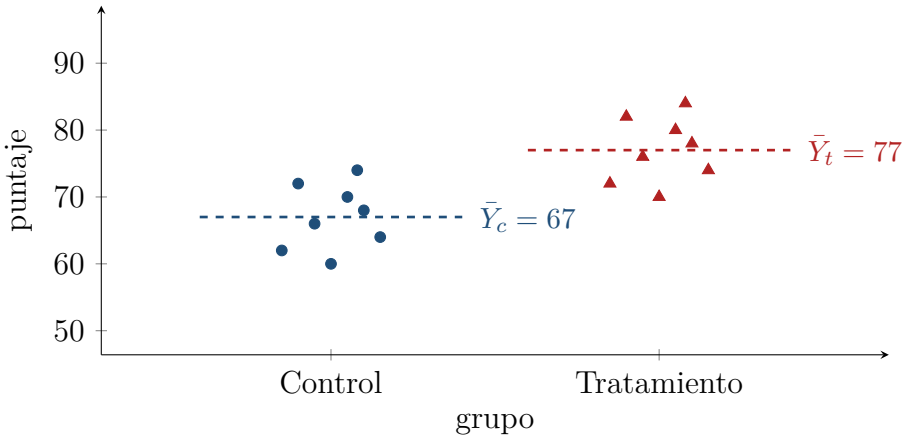
(1) Construye el IC del 95 %, $0.039 \pm 1.96(0.008)$.

(2) ¿El efecto es significativamente distinto de cero al 5 %? Justifica con el IC.

Solución

(1) Margen $= 1.96 \times 0.008 = 0.01568$. IC $= 0.039 \pm 0.01568 = [0.0233,\ 0.0547]$.
(2) Sí: el IC del 95 % *no* contiene a 0, así que el efecto es significativo al 5 % (el seguro público aumenta la probabilidad de reportarse en buena salud).

Ejercicio 1.21 (Nube de puntos: tratamiento vs. control)**.** La figura muestra los resultados individuales (puntaje) de un RCT: círculos azules son el grupo de **control** y triángulos rojos el de **tratamiento**; las líneas punteadas marcan las medias de cada grupo.



(1) A partir de las líneas, ¿cuál es el efecto estimado $\hat{\tau} = \bar{Y}_t - \bar{Y}_c$?

(2) ¿Por qué hay *dispersión* dentro de cada grupo si todos recibieron lo mismo?

Solución

(1) $\hat{\tau} = 77 - 67 = 10$ puntos.
(2) Porque las personas difieren entre sí (variabilidad individual: aptitud previa, esfuerzo, suerte en el examen). El RCT no elimina la variación individual; la aleatorización solo garantiza que *en promedio* ambos grupos partan iguales, de modo que la diferencia de medias estime el efecto causal.

Ejercicio 1.22 (Aleatorización por bloques (*stratified*))**.** Para evaluar un material didáctico, una investigadora teme que el sexo influya en el resultado. En vez de sortear "a ciegas", forma bloques de hombres y de mujeres y *dentro de cada bloque* asigna la mitad a tratamiento.

(1) ¿Qué ventaja tiene esto frente a una aleatorización completamente al azar?

(2) ¿Sigue siendo válido estimar el efecto como diferencia de medias dentro del estudio? (idea)

Solución

(1) Garantiza que tratamiento y control queden *balanceados por sexo* por construcción (no solo en promedio); reduce la varianza del estimador al eliminar esa fuente de ruido.
(2) Sí: la asignación sigue siendo aleatoria (dentro de cada bloque), así que la comparación de medias (idealmente promediando los efectos por bloque) sigue estimando el efecto causal. Es el diseño estratificado (*stratified randomization*).

Ejercicio 1.23 (Comparación no paramétrica: prueba del signo)**.** Se compara el dolor (escala 0–10) *antes* y *después* de una terapia en 8 pacientes. En vez de suponer normalidad, se usa la *prueba del signo*: se cuenta en cuántos pacientes bajó el dolor. Cambios (después − antes): -2, -1, $+1$, -3, -2, 0, -1, -2.

(1) Bajo H_0 (la terapia no tiene efecto), cada paciente con cambio no nulo tiene prob. 0.5 de bajar. Descarta los empates (0) y cuenta n útiles y el número de bajadas (signos $-$).

(2) Con $n = 7$ ensayos Bernoulli(0.5), ¿es "raro" observar 6 bajadas? (idea: bajo H_0, esperaríamos ≈ 3.5).

Solución

(1) Se descarta el $0 \Rightarrow n = 7$ cambios útiles. Signos negativos (bajó): $-2, -1, -3, -2, -1, -2$ son 6; positivos: $+1$ es 1. Así 6 de 7 bajaron.
(2) Bajo H_0, el número de bajadas es $\text{Bin}(7, 0.5)$ con media 3.5. Observar 6 (o más) es bastante raro: $\mathbb{P}(\geq 6) = \dfrac{\binom{7}{6} + \binom{7}{7}}{2^7} = \dfrac{7+1}{128} = \dfrac{8}{128} = 0.0625$. La evidencia sugiere que la terapia reduce el dolor (la prueba del signo no supone ninguna distribución de los cambios).

Ejercicio 1.24 (Diferencia de medias con varianzas distintas (Welch))**.** Un RCT sobre horas de estudio da: tratamiento $\bar{Y}_t = 8.0$, $s_t^2 = 4.0$, $N_t = 50$; control $\bar{Y}_c = 7.0$, $s_c^2 = 9.0$, $N_c = 30$. Usa $\widehat{\text{Var}}(\hat{\tau}) = \dfrac{s_t^2}{N_t} + \dfrac{s_c^2}{N_c}$ (no se asume varianza común).

(1) Calcula $\hat{\tau}$, el error estándar y el estadístico $t = \hat{\tau}/\text{ee}$.

(2) Con muestra moderada usando $\mathcal{N}(0,1)$, ¿se rechaza $H_0 : \tau = 0$ al 5 %? Da el IC del 95 %.

Solución

(1) $\hat{\tau} = 8.0 - 7.0 = 1.0$. $\widehat{\text{Var}}(\hat{\tau}) = \dfrac{4}{50} + \dfrac{9}{30} = 0.08 + 0.30 = 0.38$. $\text{ee} = \sqrt{0.38} \approx 0.6164$. $t = \dfrac{1.0}{0.6164} \approx 1.622$.
(2) $|t| \approx 1.62 < 1.96$: **no** se rechaza H_0 al 5 %. IC: $1.0 \pm 1.96(0.6164) = 1.0 \pm 1.208 = [-0.21,\ 2.21]$, que contiene a 0 (coherente con no rechazar). Aquí fue clave *no* suponer varianzas iguales: el grupo de control es más variable y pequeño, lo que infla la incertidumbre.

Ejercicio 1.25 (Tamaño de muestra y potencia para un RCT (conexión Clase 6))**.** Para detectar un efecto δ en un RCT con grupos de igual tamaño n por brazo, desviación estándar común σ, nivel α y potencia $1-\beta$, una fórmula usual es $n = \dfrac{2\,(z_{1-\alpha/2} + z_{1-\beta})^2\,\sigma^2}{\delta^2}$ (por brazo). Toma $\sigma = 10$, $\delta = 4$, $\alpha = 0.05$ ($z_{0.975} = 1.96$) y $1 - \beta = 0.80$ ($z_{0.80} = 0.84$).

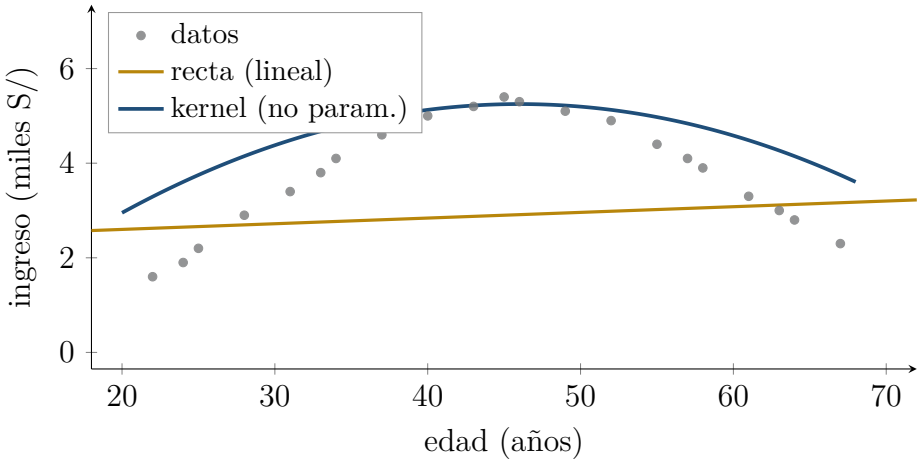
(1) Calcula n por brazo y el total $2n$.

(2) Si quieres detectar un efecto la mitad de grande ($\delta = 2$), ¿cómo cambia n?

Solución

(1) $(z_{0.975} + z_{0.80})^2 = (1.96 + 0.84)^2 = (2.80)^2 = 7.84$. $n = \dfrac{2 \cdot 7.84 \cdot 10^2}{4^2} = \dfrac{2 \cdot 7.84 \cdot 100}{16} = \dfrac{1\,568}{16} = 98$ por brazo. Total $2n = 196$.
(2) Con $\delta = 2$: $n = \dfrac{2 \cdot 7.84 \cdot 100}{2^2} = \dfrac{1\,568}{4} = 392$ por brazo (total 784). Detectar un efecto la mitad de grande *cuadruplica* el tamaño ($n \propto 1/\delta^2$), igual que en la Clase 6.

Ejercicio 1.26 (Regresión por kernel vs. recta (interpretar))**.** La figura compara, para datos de ingreso (en miles de soles) según la edad, un ajuste lineal (recta) y una regresión no paramétrica por *kernel* (suavizado local $\hat{g}(x) = \dfrac{\sum_i y_i\, K((x - x_i)/h)}{\sum_i K((x - x_i)/h)}$, promedio ponderado de los y cercanos a x).



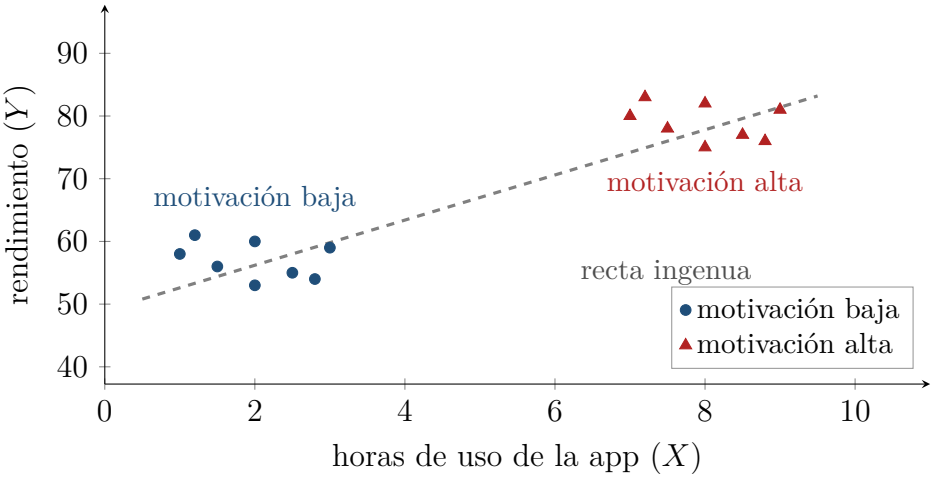
(1) ¿Por qué la recta da una conclusión engañosa sobre la relación edad– ingreso, y qué capta el kernel que la recta no?

(2) En la fórmula del kernel, ¿qué pasa con el *sesgo* y la *varianza* de $\hat{g}$ si el ancho de banda h es muy grande? ¿Y si es muy pequeño?

Solución

(1) La relación es **no lineal** (forma de joroba: el ingreso sube hasta la mediana edad y luego baja). La recta es casi plana y sugiere "poca o ninguna relación"; el kernel, al promediar localmente, *capta la curvatura* (el pico alrededor de los 45 años y el descenso posterior) sin imponer una forma funcional.
(2) h **grande**: mucho suavizado $\Rightarrow$ *más sesgo* (se pierden rasgos como el pico) pero *menos varianza* (curva muy estable). h **pequeño**: poco suavizado $\Rightarrow$ *menos sesgo* pero *más varianza* (la curva persigue el ruido, con muchos sube-y-baja). Elegir h es un balance sesgo–varianza (p. ej. por validación cruzada); teóricamente, si $h \to 0$ el sesgo $\to 0$ y si $nh \to \infty$ la varianza $\to 0$.

Ejercicio 1.27 (Confusión en una nube de puntos)**.** La figura muestra horas de uso de una app educativa (X) vs. rendimiento (Y). Los puntos están coloreados por un confusor: **motivación alta** (rojo) o **baja** (azul). La recta gris punteada es el ajuste *ignorando* la motivación.



(1) La recta ingenua tiene pendiente *positiva fuerte*. ¿Refleja eso un efecto causal de usar la app sobre el rendimiento? Explica usando los colores.

(2) ¿Qué verías *dentro* de cada grupo de motivación, y qué diseño permitiría estimar el efecto causal limpio?

Solución

(1) No necesariamente. La pendiente positiva fuerte surge porque los de **motivación alta** (rojo) usan más la app *y* rinden más por su motivación, mientras los de **motivación baja** (azul) usan poco *y* rinden menos. La app y el rendimiento están correlacionados vía el confusor *motivación*, no (necesariamente) por causalidad.
(2) *Dentro* de cada grupo (mismo color) la relación horas–rendimiento es casi plana o incluso ligeramente negativa: a igual motivación, usar más la app no eleva el rendimiento. Para estimar el efecto causal sin confusión: **aleatorizar** la asignación de uso de la app (RCT) —o, en datos observacionales, *controlar* por motivación (comparar a igual motivación). La aleatorización balancea la motivación entre grupos y rompe la correlación espuria.

Ejercicio 1.28 (Controlar por una variable cambia la conclusión (sabor a Simpson))**.** Un programa de empleo parece *reducir* la tasa de colocación cuando se comparan todos los participantes con todos los no participantes. Pero al separar por nivel educativo aparece lo contrario. Datos (colocados / total):

	Participan (trat.)	No participan (control)
Sin secundaria	36/40 (90 %)	9/10 (90 %)
Con secundaria	14/60 (≈ 23 %)	54/90 (60 %)
Total	50/100 (50 %)	63/100 (63 %)

(1) Verifica que el **total** sugiere que participar *baja* la colocación (50 % vs. 63 %). ¿Qué pasa *dentro* de cada nivel educativo?

(2) Explica por qué el nivel educativo es un confusor y cuál conclusión es la correcta para juzgar el programa.

Solución

(1) Total: tratamiento $50/100 = 50\,\% <$ control $63/100 = 63\,\%$. Pero por estrato: *sin secundaria*, $90\,\%$ vs. $90\,\%$ (igual o no peor); *con secundaria*, $23\,\%$ vs. $60\,\%$. Aquí el patrón agregado es engañoso: el grupo tratado está *sobrecargado de personas con secundaria que colocan poco* (¡revisa los datos: la relación se invierte si la composición difiere!).
(2) El nivel educativo afecta tanto la *probabilidad de participar* como la *colocación*: es un **confusor**. La comparación total mezcla el efecto del programa con la distinta composición de los grupos (la mayoría de tratados tiene secundaria, donde colocar es más difícil en estos datos). La conclusión válida exige *controlar* por educación (comparar dentro de cada estrato) o, mejor, **aleatorizar**: así la composición educativa quedaría balanceada y la diferencia de medias estimaría el efecto causal. Es el fenómeno de la paradoja de Simpson aplicado a causalidad.

Nivel 4 – Reto

Difícil: integra varios temas en un mismo problema.

Ejercicio 1.29 (Descomposición del sesgo de selección)**.** La diferencia de medias observada se descompone (Holland; Imbens–Rubin) como

$$\underbrace{\mathbb{E}[Y \mid W=1] - \mathbb{E}[Y \mid W=0]}_{\text{diferencia observada}} = \underbrace{\mathbb{E}[Y(1) - Y(0) \mid W=1]}_{\text{ATT: efecto sobre tratados}} + \underbrace{\mathbb{E}[Y(0) \mid W=1] - \mathbb{E}[Y(0) \mid W=0]}_{\text{sesgo de selección}}.$$

Usa la tabla del problema de auto-selección (Nivel 2, Núcleo): tratados $i = 1, 2, 3$, controles $i = 4, 5, 6$, con $Y(1) = (10, 9, 11, 6, 5, 7)$ y $Y(0) = (6, 5, 7, 5, 4, 6)$.

(1) Calcula la diferencia observada, el ATT y el sesgo de selección, y verifica que ATT + sesgo = diferencia observada.

(2) ¿Qué supuesto sobre la asignación hace que el sesgo de selección sea cero?

Solución

(1) Diferencia observada $= \bar{Y}_t - \bar{Y}_c = 10 - 5 = 5$ (como en Núcleo). ATT $= \mathbb{E}[Y(1) - Y(0) \mid W = 1]$: para los tratados los efectos son $4, 4, 4$, así ATT $= 4$. Sesgo $= \mathbb{E}[Y(0) \mid W = 1] - \mathbb{E}[Y(0) \mid W = 0] = \dfrac{6+5+7}{3} - \dfrac{5+4+6}{3} = 6 - 5 = 1$. Verificación: $4 + 1 = 5$. ✓
(2) Si la asignación es **aleatoria** (o "ignorable"), entonces $Y(0)$ es independiente de W, de modo que $\mathbb{E}[Y(0) \mid W = 1] = \mathbb{E}[Y(0) \mid W = 0]$ y el sesgo de selección es 0; la diferencia

observada estima el ATT (que además iguala al ATE).

Ejercicio 1.30 (Descomposición completa: diferencia observada = ATT + sesgo de selección)**.** Un programa de tutorías se ofrece en un colegio. La tabla muestra los *resultados potenciales* (puntaje en una prueba estandarizada) de 6 estudiantes: $Y_i(0)$ sin tutoría e $Y_i(1)$ con tutoría. Los estudiantes 1–3 *eligieron* inscribirse ($W_i = 1$) y los 4–6 no ($W_i = 0$). Recuerda que solo se observa $Y_i = Y_i(W_i)$ (problema fundamental de la inferencia causal). La descomposición clave es

$$\underbrace{\mathbb{E}[Y \mid W=1] - \mathbb{E}[Y \mid W=0]}_{\text{diferencia observada}} = \underbrace{\mathbb{E}[Y(1) - Y(0) \mid W=1]}_{\text{ATT (efecto sobre tratados)}} + \underbrace{\mathbb{E}[Y(0) \mid W=1] - \mathbb{E}[Y(0) \mid W=0]}_{\text{sesgo de selección}}.$$

Estudiante i	W_i	$Y_i(0)$	$Y_i(1)$	observado Y_i
1	1	8	12	12
2	1	9	14	14
3	1	10	16	16
4	0	5	6	5
5	0	6	7	6
6	0	7	8	7

(1) Calcula la **diferencia observada** $\mathbb{E}[Y \mid W = 1] - \mathbb{E}[Y \mid W = 0]$ usando solo la columna Y_i observado.

(2) Calcula el **ATT** y el **ATE** $= \mathbb{E}[Y(1) - Y(0)]$ usando las columnas de resultados potenciales. ¿Por qué difieren?

(3) Calcula el **sesgo de selección** y verifica numéricamente que diferencia observada = ATT + sesgo. Interpreta: ¿la diferencia observada *sobrestima* o *subestima* el efecto causal sobre los tratados?

Solución

(1) Tratados observan $Y = 12, 14, 16$, media $= 14$. Controles observan $Y = 5, 6, 7$, media $= 6$.

$$\mathbb{E}[Y \mid W=1] - \mathbb{E}[Y \mid W=0] = 14 - 6 = \boxed{8}.$$

(2) *ATT* (efecto sobre los tratados, filas 1–3): efectos $12-8=4$, $14-9=5$, $16-10=6$, promedio ATT $= \dfrac{4+5+6}{3} = \boxed{5}$.

ATE (los 6): efectos $4, 5, 6, 1, 1, 1$, promedio ATE $= \dfrac{4+5+6+1+1+1}{6} = \dfrac{18}{6} = \boxed{3}$.

Difieren porque el efecto es **heterogéneo**: los que se inscribieron son justamente los que más ganan con la tutoría (selección en ganancias). El ATT>ATE.

(3) *Sesgo de selección* $= \mathbb{E}[Y(0) \mid W=1] - \mathbb{E}[Y(0) \mid W=0]$. Para los tratados $\mathbb{E}[Y(0) \mid W=1] = \dfrac{8+9+10}{3} = 9$; para los controles $\mathbb{E}[Y(0) \mid W=0] = \dfrac{5+6+7}{3} = 6$.
Sesgo $= 9 - 6 = \boxed{3} > 0$.
Verificación: ATT + sesgo $= 5 + 3 = 8 =$ diferencia observada. ✓
Interpretación: los estudiantes que se inscriben ya rendían más *aunque no hubieran recibido tutoría* ($Y(0)$ mayor). Por eso la diferencia observada (8) **sobrestima** el verdadero efecto sobre los tratados (ATT = 5): le suma 3 puntos de sesgo de selección. Comparar inscritos contra no inscritos confunde el efecto del programa con quien decidió entrar. La aleatorización elimina ese sesgo porque hace $\mathbb{E}[Y(0) \mid W=1] = \mathbb{E}[Y(0) \mid W=0]$.

(Adaptado de MIT 14.310x, Clases 14–15: potential outcomes y selection bias; MIT 18.05.)

Ejercicio 1.31 (RCT estilo Duflo: ATE, error estándar de Welch, IC, prueba y potencia)**.** Un experimento aleatorizado evalúa un programa de transferencias condicionadas sobre el gasto mensual en alimentos (en cientos de soles). Se asignaron al azar $n_t = 50$ hogares al *tratamiento* y $n_c = 50$ al *control*. Resultados:

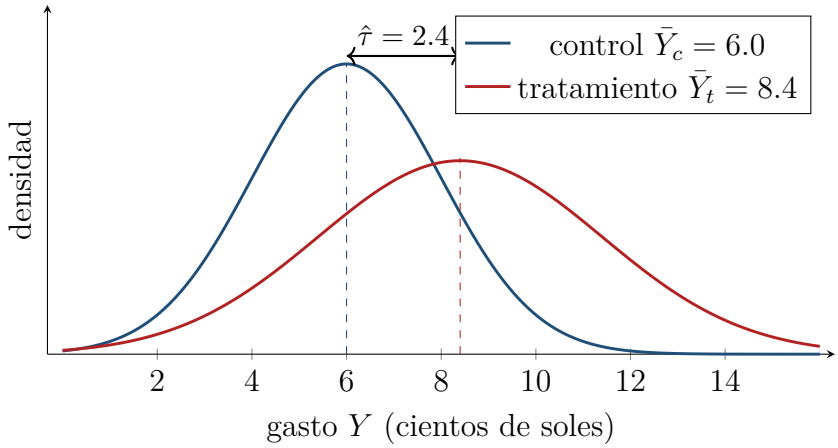
$$\bar{Y}_t = 8.4, \quad s_t = 3.0 \text{ (tratamiento)}; \qquad \bar{Y}_c = 6.0, \quad s_c = 2.0 \text{ (control)}.$$

Como las varianzas difieren, usa el error estándar de *Welch* $\mathrm{SD}(\hat{\tau}) = \sqrt{s_t^2/n_t + s_c^2/n_c}$ (no agrupado). Usa $z_{0.975} = 1.96$, $z_{0.80} = 0.84$.

(1) Estima el ATE con $\hat{\tau} = \bar{Y}_t - \bar{Y}_c$ y calcula su error estándar.

(2) Construye el IC del 95 % y prueba $H_0 : \tau = 0$ vs. $H_1 : \tau \neq 0$ al 5 %. ¿Es significativo el efecto?

(3) **Conexión con la Clase 6 (potencia).** Para un futuro estudio se quiere detectar un efecto de $\delta = 1$ (cien soles) con potencia 0.80 y $\alpha = 0.05$, suponiendo $\sigma \approx 2.5$ en cada brazo. La fórmula de tamaño por brazo es $n = \dfrac{(z_{1-\alpha/2} + z_{1-\beta})^2\, 2\sigma^2}{\delta^2}$. Calcula n.



Solución

(1) $\hat{\tau} = 8.4 - 6.0 = \boxed{2.4}$ (cien soles = 240 soles más de gasto en alimentos).
$\mathrm{SD}(\hat{\tau}) = \sqrt{\dfrac{3.0^2}{50} + \dfrac{2.0^2}{50}} = \sqrt{\dfrac{9}{50} + \dfrac{4}{50}} = \sqrt{0.26} \approx \boxed{0.51}$.
(2) $\mathrm{IC}_{95\,\%}$: $2.4 \pm 1.96 \cdot 0.51 = 2.4 \pm 1.00 = [1.40,\ 3.40]$.
Estadístico: $z = \dfrac{2.4}{0.51} \approx 4.71$. Como $|4.71| > 1.96$ (y el IC no contiene 0), **se rechaza** H_0 al 5 % ($p \approx 2.5 \times 10^{-6}$). El efecto es claramente significativo: el programa aumenta el gasto en alimentos. La aleatorización permite leer esto como *causal*, no solo correlacional.
(3) $n = \dfrac{(1.96 + 0.84)^2 \cdot 2 \cdot 2.5^2}{1^2} = \dfrac{(2.80)^2 \cdot 2 \cdot 6.25}{1} = \dfrac{7.84 \cdot 12.5}{1} = 98$. Se necesitan $\boxed{n = 98}$ hogares *por brazo* (196 en total). Igual que en la Clase 6, detectar efectos pequeños exige muestras grandes: $n \propto 1/\delta^2$.
(Adaptado de MIT 14.310x, Clase 15: randomized experiments; MIT 18.05, Class 18: two-sample tests y sample size.)

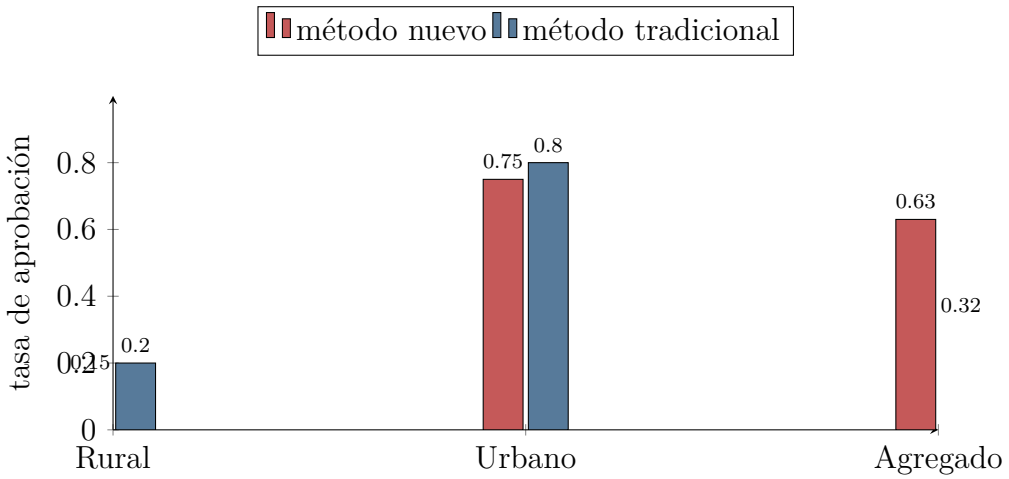
Ejercicio 1.32 (Confusión y paradoja de Simpson: estratificar invierte la conclusión)**.** Un nuevo método de enseñanza ($W = 1$) se compara con el tradicional ($W = 0$) en cuanto a la tasa de aprobación. Los datos agregados parecen favorecer al método nuevo, pero hay un *confusor*: el método nuevo se aplicó sobre todo en colegios **urbanos** (con mejor desempeño de base), y el tradicional sobre todo en **rurales**.

	n tratados (nuevo) / tasa	n control (tradic.) / tasa
Colegios rurales	20 / 0.15	80 / 0.20
Colegios urbanos	80 / 0.75	20 / 0.80

(1) Calcula la tasa de aprobación *agregada* del método nuevo y del tradicional (ponderando por n). ¿Qué método parece mejor a simple vista?

(2) Calcula la diferencia (nuevo − tradicional) *dentro de cada estrato* (rural y urbano). ¿Qué método es mejor controlando por zona?

(3) Explica la paradoja: ¿por qué la conclusión se invierte? ¿Cuál es el efecto causal correcto y por qué estratificar (controlar por el confusor) lo recupera?



Solución

(1) *Agregado nuevo* (tratados): 20 rurales al 0.15 y 80 urbanos al 0.75: $\dfrac{20 \cdot 0.15 + 80 \cdot 0.75}{100} = \dfrac{3 + 60}{100} = 0.63$.
Agregado tradicional (control): 80 rurales al 0.20 y 20 urbanos al 0.80: $\dfrac{80 \cdot 0.20 + 20 \cdot 0.80}{100} = \dfrac{16 + 16}{100} = 0.32$.
A simple vista el método **nuevo** parece muchísimo mejor (0.63 vs. 0.32, +31 puntos).

(2) Dentro de cada estrato:

- Rural: nuevo 0.15 − tradicional $0.20 = -0.05$.
- Urbano: nuevo 0.75 − tradicional $0.80 = -0.05$.

En *ambas* zonas el método nuevo es **peor** por 5 puntos. ¡La conclusión se invierte!

(3) **Paradoja de Simpson.** La zona es un confusor: afecta el resultado (urbanos aprueban más) *y* está correlacionada con el tratamiento (el método nuevo se usó sobre todo en urbanos). Al agregar, el método nuevo "hereda" la alta aprobación de las zonas urbanas; esa ventaja es del *contexto*, no del método. El efecto causal correcto es el de los estratos: -0.05 (el método nuevo perjudica ligeramente). Estratificar compara nuevo vs. tradicional *a igual zona*, eliminando la influencia del confusor; el agregado mezcla peras con manzanas. Un RCT (asignar la zona al azar respecto al método) rompería la correlación tratamiento–zona y evitaría la paradoja.

(Adaptado de MIT 14.310x, Clase 14: confounding; MIT 18.05, Class 14: conditional probability y Simpson's paradox.)

Ejercicio 1.33 (A/B testing con proporciones: diferencia, SE, IC y prueba)**.** Una plataforma de gobierno digital prueba dos diseños de un formulario de trámite. A $n_A = 1200$ usuarios se les muestra el diseño A (actual) y a $n_B = 1200$ el diseño B (nuevo), asignados al azar. Completan el trámite $x_A = 96$ usuarios con A y $x_B = 144$ con B.

(1) Estima las proporciones de éxito $\hat{p}_A, \hat{p}_B$ y la diferencia $\hat{p}_B - \hat{p}_A$.

(2) Construye el IC del 95 % para la diferencia usando el SE *no agrupado*

$$\mathrm{SD} = \sqrt{\hat{p}_A(1-\hat{p}_A)/n_A + \hat{p}_B(1-\hat{p}_B)/n_B}.$$

(3) Prueba $H_0 : p_A = p_B$ vs. $H_1 : p_A \neq p_B$ al 5 % usando el SE *agrupado* bajo H_0, con $\hat{p} = (x_A + x_B)/(n_A + n_B)$ y

$$\mathrm{SD}_0 = \sqrt{\hat{p}\,(1-\hat{p})\,(1/n_A + 1/n_B)}.$$

Interpreta la decisión en términos de negocio/política.

Solución

(1) $\hat{p}_A = \dfrac{96}{1200} = 0.08$, $\hat{p}_B = \dfrac{144}{1200} = 0.12$. Diferencia $\hat{p}_B - \hat{p}_A = 0.12 - 0.08 = \boxed{0.04}$ (+4 puntos porcentuales).

(2) SE no agrupado: $\mathrm{SD} = \sqrt{\dfrac{0.08 \cdot 0.92}{1200} + \dfrac{0.12 \cdot 0.88}{1200}} = \sqrt{\dfrac{0.0736 + 0.1056}{1200}} = \sqrt{0.0001493} \approx 0.0122$.

$\mathrm{IC}_{95\,\%}$: $0.04 \pm 1.96 \cdot 0.0122 = 0.04 \pm 0.024 = [0.016,\ 0.064]$. No incluye 0.

(3) Bajo H_0, $\hat{p} = \dfrac{96+144}{2400} = \dfrac{240}{2400} = 0.10$. $\mathrm{SD}_0 = \sqrt{0.10 \cdot 0.90 \cdot \left(\frac{1}{1200} + \frac{1}{1200}\right)} = \sqrt{0.09 \cdot \frac{2}{1200}} = \sqrt{0.00015} \approx 0.01225$.

$z = \dfrac{0.04}{0.01225} \approx 3.27$. Como $|3.27| > 1.96$, **se rechaza** H_0 ($p \approx 0.001$).

Interpretación: el diseño B aumenta la tasa de finalización del 8 % al 12 % (un +50 % relativo), efecto estadísticamente sólido. Como la asignación fue aleatoria, es plausible leerlo como *causal*: conviene desplegar B. (Conviene aún valorar costo de implementación y posibles efectos a largo plazo.)

(Adaptado de MIT 14.310x, Clase 15: A/B testing como RCT; MIT 18.05, Class 17–18: two-proportion z-test.)

Nivel 5 – Reto+

Muy difícil / fuera del scope típico: demostraciones y casos límite.

Ejercicio 1.34 (Prueba exacta de Fisher (hipótesis nula aguda))**.** Fisher prueba la *hipótesis nula aguda* $H_0 : Y_i(1) = Y_i(0)$ para *todo* i (el tratamiento no afecta a nadie). Bajo H_0 podemos rellenar el contrafactual de cada unidad (es igual a su valor observado) y, reasignando el tratamiento de todas las formas posibles, obtener la distribución del estadístico. En un mini-experimento con 3 unidades tratadas y 3 de control ($\binom{6}{3} = 20$ asignaciones), el estadístico observado (diferencia de medias) resultó el más extremo en valor absoluto: 16 de las 20 asignaciones dan un valor $\geq$ al observado.

(1) Da el p-valor exacto de Fisher, $p = \mathbb{P}(|T| \geq |T^{\text{obs}}|)$ sobre las 20 asignaciones.

(2) ¿Se rechaza H_0 al 5 %? ¿En qué se diferencia esta nula de "el ATE es cero"?

Solución

(1) $p = \dfrac{16}{20} = 0.80$.
(2) Como $p = 0.80 \gg 0.05$, **no** se rechaza la nula aguda: con tan pocos datos no hay evidencia de efecto. La nula aguda de Fisher dice "*ningúna* unidad tiene efecto" ($Y_i(1) = Y_i(0)\ \forall i$), más fuerte que la nula de Neyman "el efecto *promedio* (ATE) es cero", que permite efectos individuales que se cancelan. La prueba de Fisher no usa supuestos de muestreo: la incertidumbre viene de la aleatorización misma.

Ejercicio 1.35 (Regresión por kernel (Nadaraya–Watson) a mano y el bandwidth)**.** Se estudia cómo varía el salario y (en cientos de soles/hora) con los años de educación x. Hay 6 datos:

$$(x, y): \ (1,2),\ (3,5),\ (5,7),\ (7,8),\ (9,8),\ (11,7).$$

El estimador de Nadaraya–Watson predice en x_0 un *promedio ponderado local*:

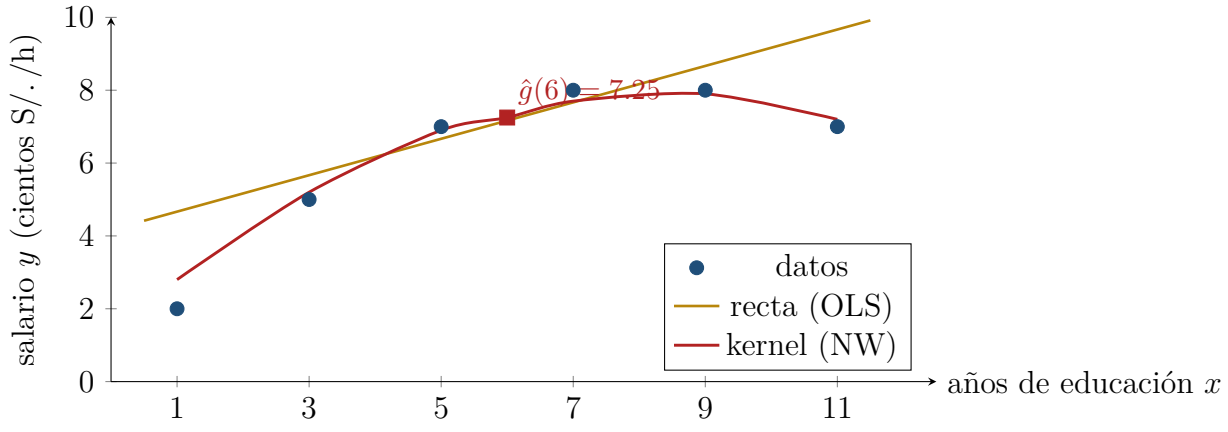
$$\hat{g}(x_0) = \frac{\sum_i K\left(\frac{x_i - x_0}{h}\right) y_i}{\sum_i K\left(\frac{x_i - x_0}{h}\right)},$$

donde h es el *bandwidth* y K el kernel. Trabaja en $x_0 = 6$.

(1) **Kernel uniforme** $K(u) = 1$ si $|u| \leq 1$ y 0 si no, con $h = 4$ (incluye los x con $|x - 6| \leq 4$). ¿Qué puntos entran y cuál es $\hat{g}(6)$?

(2) **Kernel triangular** $K(u) = \max(0,\ 1 - |u|)$, con $h = 4$. Calcula los pesos $K\left(\frac{x_i - 6}{4}\right)$ y obtén $\hat{g}(6)$. Compara con el uniforme.

(3) La regresión *lineal* (OLS) da $\hat{y} = 4.17 + 0.50\,x$, así $\hat{y}(6) = 7.17$. ¿Qué capta el kernel que la recta no? Discute el **compromiso sesgo–varianza** al elegir h: ¿qué pasa con h muy grande y con h muy pequeño?



Solución

(1) *Uniforme,* $h = 4$: entran los x con $|x-6| \le 4$, es decir $x \in \{3,5,7,9\}$ (sale $x=1$ con $|1-6| = 5 > 4$ y $x = 11$ con $|11-6| = 5 > 4$). Cada uno pesa 1:

$$\hat{g}(6) = \frac{5+7+8+8}{4} = \frac{28}{4} = \boxed{7.0}.$$

(2) *Triangular,* $h = 4$: $u_i = \dfrac{x_i - 6}{4}$ y $K(u) = \max(0, 1-|u|)$:

x_i	1	3	5	7	9 (y 11)
u_i	-1.25	-0.75	-0.25	0.25	0.75 (1.25)
$K(u_i)$	0	0.25	0.75	0.75	0.25 (0)

$$\hat{g}(6) = \frac{0.25\cdot 5 + 0.75\cdot 7 + 0.75\cdot 8 + 0.25\cdot 8}{0.25+0.75+0.75+0.25} = \frac{1.25+5.25+6.0+2.0}{2.0} = \frac{14.5}{2.0} = \boxed{7.25}.$$

El triangular da 7.25 (algo mayor que 7.0) porque pondera más los vecinos cercanos ($x = 5, 7$, ambos altos) y menos los lejanos.

(3) La recta impone una relación *monótona creciente* ($\hat{y}(6) = 7.17$) y no puede capturar que el salario **se aplana e incluso baja** para mucha educación (la "joroba": sube hasta ~ 8 y cae a 7). El kernel, al promediar localmente, *sigue la curvatura* sin imponer una forma global.

Compromiso sesgo–varianza:

- h *grande* $\Rightarrow$ se promedia sobre una ventana ancha: la curva es muy suave (poca varianza) pero **sesgada** (alisa demasiado, se parece a la media global y pierde la joroba). En el límite $h \to \infty$ es una constante.
- h *pequeño* $\Rightarrow$ pocos vecinos por punto: sigue cada dato (poco sesgo) pero es muy **ruidosa** (mucha varianza, sobreajuste). En el límite $h \to 0$ interpola los datos.

El h óptimo equilibra ambos y suele elegirse por validación cruzada.

(Adaptado de MIT 14.310x, Clase 16: nonparametric / kernel regression y bandwidth.)

Ejercicio 1.36 (Prueba no paramétrica del signo (idea de Wilcoxon) a mano)**.** Se mide el ingreso semanal (en soles) de 10 vendedores *antes* y *después* de una capacitación. No se quiere asumir normalidad, así que se usa la **prueba del signo** sobre las diferencias $d_i =$ después $-$ antes, que contrasta H_0 : "la mediana de d es 0" vs. H_1 : "la mediana $\neq 0$". Datos:

i	1	2	3	4	5	6	7	8	9	10
antes	20	22	19	25	30	18	24	21	27	23
después	24	21	23	28	33	20	24	26	30	25

(1) Calcula las diferencias d_i y sus signos. Descarta los empates ($d_i = 0$). ¿Cuántos signos positivos (S_+) y negativos hay, y cuál es el n efectivo?

(2) Bajo H_0, el número de positivos $S_+ \sim \text{Bin}(n, 0.5)$. Calcula el p-valor de dos colas usando $\mathbb{P}(S_+ \geq 8) + \mathbb{P}(S_+ \leq 1)$ con $n = 9$.

(3) ¿Se rechaza H_0 al 5 %? Explica una *ventaja* y una *desventaja* de la prueba del signo frente a una prueba t pareada.

Solución

(1) Diferencias d_i = después − antes:

$$4,\ -1,\ 4,\ 3,\ 3,\ 2,\ 0,\ 5,\ 3,\ 2.$$

Signos: $+, -, +, +, +, +, 0, +, +, +$. Hay un empate ($i = 7$, $d = 0$) que se descarta. Quedan $S_+ = 8$ positivos y 1 negativo, con $n = 9$ efectivo.

(2) Bajo H_0, $S_+ \sim \text{Bin}(9, 0.5)$. Las colas con ≤ 1 y ≥ 8:

$$\mathbb{P}(S_+ \leq 1) = \frac{\binom{9}{0} + \binom{9}{1}}{2^9} = \frac{1+9}{512} = \frac{10}{512}, \quad \mathbb{P}(S_+ \geq 8) = \frac{\binom{9}{8} + \binom{9}{9}}{2^9} = \frac{9+1}{512} = \frac{10}{512}.$$

$p = \dfrac{10+10}{512} = \dfrac{20}{512} \approx \boxed{0.039}$.

(3) Como $p \approx 0.039 < 0.05$, **se rechaza** H_0: hay evidencia de que la capacitación aumentó el ingreso (la mediana de d es > 0).

Ventaja: no asume normalidad ni forma de la distribución; solo usa signos, por lo que es **robusta** a valores atípicos y a colas pesadas. *Desventaja:* descarta la *magnitud* de las diferencias (solo el signo), así que tiene menos **potencia** que la t pareada cuando la normalidad sí se cumple. (La prueba de rangos con signo de Wilcoxon es un punto intermedio: usa el orden de las magnitudes; aquí daría un p aún menor.)

(Adaptado de MIT 14.310x, Clase 16: nonparametric tests; MIT 18.05, Class 17: distribution-free / sign & rank tests.)

Ejercicio 1.37 (Sesgo por variable omitida y la idea de variable instrumental (intro))**.** Se quiere estimar el efecto causal de los años de educación (W) sobre el salario (Y). El modelo verdadero es

$$Y = \beta\, W + \gamma\, Z + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid W, Z] = 0,$$

donde Z es la "habilidad" (no observada), que eleva el salario ($\gamma > 0$) y está correlacionada con la educación. Suponga que al regresar Z sobre W la pendiente es $\delta = \text{Cov}(W, Z)/\,\text{Var}(W)$.

(1) Se corre la **regresión simple** de Y sobre W *omitiendo* Z. Se puede mostrar que el coeficiente estimado converge a $\beta + \gamma\,\delta$. Identifica el **sesgo por variable omitida** y su *signo* si $\gamma > 0$ y $\delta > 0$.

(2) Con valores ilustrativos $\beta = 1$ (efecto causal real), $\gamma = 2$ y $\delta = 1.5$, calcula el coeficiente que daría la regresión simple. ¿Sobrestima o subestima el retorno a la educación?

(3) **Idea de variable instrumental (IV).** Suponga una variable Z_{inst} (p. ej. la distancia al colegio más cercano) que afecta la educación W pero *no* afecta directamente el salario ni se correlaciona con la habilidad Z. Explica, en palabras, por qué Z_{inst} permitiría recuperar β aunque no observemos la habilidad. Enuncia las dos condiciones que debe cumplir un buen instrumento.

Solución

(1) El estimador de la regresión simple converge a $\beta + \gamma\delta$, así el **sesgo por variable omitida** es

$$\text{sesgo} = \gamma \cdot \delta = (\text{efecto de } Z \text{ sobre } Y) \times (\text{relación } Z\text{–}W).$$

Si $\gamma > 0$ y $\delta > 0$ (la habilidad sube el salario y va de la mano con más educación), el sesgo es **positivo**: la regresión simple atribuye a la educación parte del efecto que en realidad es de la habilidad.

(2) Coeficiente estimado $\approx \beta + \gamma\delta = 1 + 2 \cdot 1.5 = 1 + 3 = \boxed{4}$. El verdadero retorno causal es $\beta = 1$, pero la regresión simple reporta 4: **sobrestima** fuertemente el retorno a la educación (cuadruplica). Las personas con más educación ganan más en parte porque *además* son más hábiles, no solo por estudiar.

(3) **Variable instrumental.** La distancia al colegio Z_{inst} "empuja" la educación (quien vive cerca estudia más) pero no tiene relación con la habilidad ni con el salario salvo *a través* de la educación. Entonces la variación en W *inducida por la distancia* es "como aleatoria" respecto a la habilidad: comparar salarios entre quienes vivieron cerca vs. lejos aísla el efecto causal de la educación, sin contaminación por habilidad. Las dos condiciones de un buen instrumento son:

- **Relevancia:** Z_{inst} sí afecta a W ($\text{Cov}(Z_{\text{inst}}, W) \neq 0$).
- **Exclusión / exogeneidad:** Z_{inst} no afecta a Y por otra vía que W y no se correlaciona con el error/habilidad ($\text{Cov}(Z_{\text{inst}}, \varepsilon) = 0$).

Esto es el preludio a las Clases 10–11 (variables instrumentales y modelos lineales para causalidad).

(Adaptado de MIT 14.310x, Clases 14–16: confounding; preludio a IV de Clases 10–11.)

Ejercicio 1.38 (Simulación: el RCT recupera el ATE; el observacional no)**.** Para *ver* la diferencia entre un experimento aleatorizado y un estudio observacional con confusor, se simula en R una población donde el efecto causal verdadero es $\tau = 5$. La variable Z (motivación de base, no observada) eleva el resultado y, en el mundo observacional, también hace más probable recibir el tratamiento.

Simulación: RCT vs. observacional

```
set.seed(123)
N   <- 2000
tau <- 5                        # efecto causal verdadero (ATE)
Z   <- rnorm(N)                 # confusor no observado (motivacion)
```

```
Y0  <- 10 + 3*Z + rnorm(N, 0, 2)   # resultado potencial sin tratamiento

## (A) RCT: asignacion al azar, INDEPENDIENTE de Z
W_rct <- rbinom(N, 1, 0.5)
Y_rct <- Y0 + tau*W_rct
ate_rct <- mean(Y_rct[W_rct==1]) - mean(Y_rct[W_rct==0])

## (B) Observacional: el tratamiento DEPENDE de Z (los mas motivados se tratan
   )
p     <- 1/(1 + exp(-1.5*Z))   # prob de tratamiento crece con Z
W_obs <- rbinom(N, 1, p)
Y_obs <- Y0 + tau*W_obs
ate_obs <- mean(Y_obs[W_obs==1]) - mean(Y_obs[W_obs==0])

c(RCT = ate_rct, Observacional = ate_obs)   # ~ 4.7  y  ~ 8.0
```

(1) En el bloque (A), ¿por qué `W_rct` es independiente de `Z`? ¿Qué garantiza eso sobre $\mathbb{E}[Y(0) \mid W = 1]$ vs. $\mathbb{E}[Y(0) \mid W = 0]$?

(2) El código devuelve ≈ 4.7 (RCT) y ≈ 8.0 (observacional), con $\tau = 5$. Explica por qué el RCT recupera el ATE (salvo error de muestreo) y el observacional lo *sobrestima*. ¿Cuánto vale aproximadamente el sesgo de selección en (B)?

(3) Propone *un* cambio al bloque (B) que haga que el estimador observacional coincida con el RCT, y *un* análisis que corrija el sesgo si Z sí fuera observada.

Solución

(1) `W_rct <- rbinom(N,1,0.5)` sortea el tratamiento con una moneda, sin mirar `Z` ni `Y0`. Por eso el tratamiento es **independiente de los resultados potenciales**: los grupos tratado y control tienen, en promedio, la misma motivación de base. Formalmente $\mathbb{E}[Y(0) \mid W = 1] = \mathbb{E}[Y(0) \mid W = 0]$, así que el **sesgo de selección es** 0.

(2) *RCT:* como no hay sesgo de selección, la diferencia de medias estima el ATE; el ≈ 4.7 (en vez de 5) es solo ruido de muestreo (con N mayor se acerca a 5).
Observacional: ahí $p = 1/(1 + e^{-1.5Z})$ hace que los de `Z` alta (más motivados, y con `Y0` más alto vía $+3Z$) se traten más. El grupo tratado ya tenía mejor $Y(0)$, así que la diferencia de medias mezcla el efecto causal (5) con la ventaja de base. Sesgo de selección $\approx 8.0 - 5 = \mathbf{3.0}$: el estudio observacional sobrestima.

(3) *Para igualar al RCT:* reemplazar la asignación dependiente por una al azar, p. ej. `W_obs <- rbinom(N,1,0.5)` (rompe la relación $Z \to W$). *Para corregir con Z observada:* ajustar por el confusor, p. ej. regresión `lm(Y_obs ~ W_obs + Z)` (el coeficiente de `W_obs` recupera ≈ 5), o estratificar/emparejar por niveles de `Z` y promediar los efectos dentro de estratos. La moraleja: sin aleatorización, solo se puede "descontar" la confusión que uno *observa y modela*.

(Adaptado de MIT 14.310x, Clases 14–15: causal inference y RCT; simulación en R estilo

MIT 18.05.)

Fuentes y atribución

Material de repaso **basado en MIT 14.310x – Data Analysis for Social Scientists** (Esther Duflo & Sara Fisher Ellison), MIT OpenCourseWare, licencia **CC BY-NC-SA 4.0** (uso no comercial, con atribución, compartir bajo la misma licencia), y en las fuentes de la bibliografía. Los enunciados se basan en el material del curso y en diversas fuentes abiertas; cuando un problema se adapta de una fuente específica, se cita en el enunciado.

Bibliografía.

- Duflo, E. & Ellison, S. F. *14.310x: Data Analysis for Social Scientists*. MIT OpenCourseWare (slides, lecture notes y problem sets). ocw.mit.edu.
- Larsen, R. J. & Marx, M. L. *An Introduction to Mathematical Statistics and Its Applications*, 6.ª ed., Pearson, 2018.
- DeGroot, M. H. & Schervish, M. J. *Probability and Statistics*, 4.ª ed., Pearson, 2012.
- Blitzstein, J. & Hwang, J. *Introduction to Probability* (Harvard Stat 110); W. Chen, *Probability Cheatsheet*.
- Diez, D., Çetinkaya-Rundel, M. & Barr, C. *OpenIntro Statistics*. openintro.org.
- Materiales abiertos de la comunidad del curso en GitHub: `gevorg-hunanyan/probability`; `hanmochen/Probability-Theory-2020`; `mynameisjanus/14310xDataAnalysis`.