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

Teoría

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

Índice

1. De qué trata esta clase **2**

2. ¿Qué es causalidad? Correlación no es causalidad **2**
2.1. Correlaciones espurias . . . 2
2.2. El marco de resultados potenciales (potential outcomes) . . . 3

3. Efecto promedio del tratamiento (ATE) y sesgo de selección **5**
3.1. El efecto promedio del tratamiento (ATE) . . . 5
3.2. ¿Por qué la diferencia de medias observada no es el ATE? . . . 6

4. Experimentos aleatorizados (RCT): la solución **7**
4.1. La aleatorización rompe la confusión . . . 7

5. Analizar un RCT: efecto, error estándar e intervalo de confianza **9**

6. Ejemplo completo: un RCT de 20 personas, paso a paso **11**
6.1. Contexto y datos . . . 11
6.2. Paso 1 — Medias de cada grupo . . . 12
6.3. Paso 2 — Efecto estimado (diferencia de medias) . . . 12
6.4. Paso 3 — Varianzas y desviaciones muestrales . . . 12
6.5. Paso 4 — Error estándar de la diferencia . . . 13
6.6. Paso 5 — Estadístico t y grados de libertad . . . 13
6.7. Paso 6 — p-valor (dos colas) . . . 14
6.8. Paso 7 — Intervalo de confianza al 95 % . . . 14
6.9. Paso 8 — Decisión e interpretación causal . . . 14
6.10. Otra ruta al p-valor: inferencia por permutación . . . 15

7. Comparaciones y regresión no paramétricas **17**
7.1. Comparar dos distribuciones sin asumir normalidad . . . 17
7.2. Regresión no paramétrica: suavizado por kernel . . . 17

8. Simular un RCT en R **18**

Resumen **20**

1. De qué trata esta clase

Hasta ahora aprendimos a *describir* datos, a *estimar* parámetros y a *cuantificar* la incertidumbre (intervalos de confianza, pruebas de hipótesis). Pero la pregunta que más le importa a un científico social rara vez es "¿cuál es el promedio?". Casi siempre es una pregunta **causal**:

- ¿El programa *Juntos* **causó** que más niños asistieran a la escuela, o esas familias ya eran distintas?
- ¿Un año más de educación **aumenta** el salario, o quienes estudian más ya tenían mejores perspectivas?
- ¿Repartir laptops **mejora** el aprendizaje, o solo lo reciben las escuelas que ya iban bien?

Esta clase —el corazón del trabajo de **Esther Duflo**, premio Nobel de Economía 2019 por su uso de experimentos para combatir la pobreza— da el marco para responderlas. Veremos:

- **Qué es causalidad**: el marco de *resultados potenciales* y por qué correlación $\neq$ causalidad.
- **El efecto promedio del tratamiento (ATE)** y por qué la diferencia de medias observada suele estar *sesgada* por confusión.
- **Los experimentos aleatorizados (RCT)**: cómo la aleatorización "arregla" el problema y vuelve creíble la inferencia causal.
- **Cómo analizar un RCT**: estimar el efecto, su error estándar y su intervalo de confianza (conecta con la Clase 6).
- **Comparaciones y regresión no paramétricas**: comparar grupos y estimar relaciones *sin* asumir una forma fija.

La estadística descriptiva nos dice *qué* pasó; la inferencia causal intenta decirnos *por qué*, y sobre todo *qué pasaría si* interveniéramos. Esa última pregunta —la del "¿y si?"— es la que guía las políticas públicas: no queremos saber solo que los pobres usan menos mosquiteros, sino *si* regalarlos reduciría la malaria. El gran descubrimiento de esta clase es que un experimento bien diseñado convierte una pregunta resbaladiza ("¿causó?") en un cálculo honesto (una simple diferencia de medias).

2. ¿Qué es causalidad? Correlación no es causalidad

2.1. Correlaciones espurias

Que dos variables "se muevan juntas" (estén correlacionadas) **no** significa que una cause a la otra. Hay tres explicaciones posibles para una correlación entre X e Y:

1. X **causa** Y (lo que solemos querer probar).

2. Y **causa** X (*causalidad inversa*).
3. Una tercera variable Z —una **variable de confusión (confounder)**— causa *ambas*, sin que X e Y se causen entre sí.

Ejemplo 2.1 (Helados y ahogamientos)**.** En verano se venden más helados *y* hay más ahogamientos: las dos series están fuertemente correlacionadas. ¿Comer helado provoca ahogarse? No. La variable de confusión es el **calor / la temporada de playa**: el verano sube tanto la venta de helados como el número de bañistas (y por tanto de ahogamientos). Prohibir los helados no salvaría a nadie.

Ejemplo 2.2 (Cigüeñas y bebés)**.** A nivel de país, el número de cigüeñas correlaciona con el número de nacimientos. Las cigüeñas *no* traen bebés: los países más grandes y rurales tienen más territorio (más cigüeñas) y más población (más nacimientos). De nuevo, una tercera variable explica la asociación. Estas correlaciones que engañan se llaman **espurias (spurious)**.

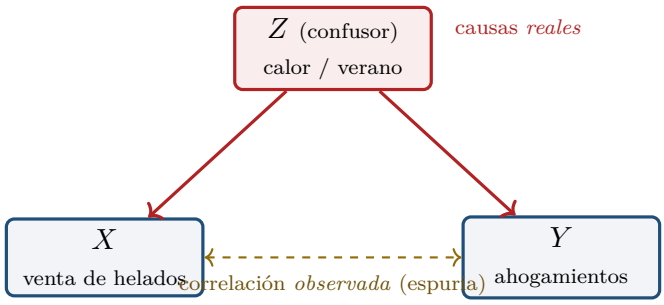

Figura 1: Una **variable de confusión** Z (el calor del verano) causa tanto X (venta de helados) como Y (ahogamientos). Eso genera una correlación *observada* entre X e Y (línea ámbar punteada) aunque *no* exista flecha causal entre ellas. Interpretar esa correlación como "X causa Y" es el error clásico.

"Correlación no implica causalidad" es cierto, pero incompleto: lo importante es *por qué*. La correlación puede reflejar causalidad directa, causalidad inversa o un confusor oculto. La tarea del científico social es *distinguir* cuál de los tres casos tenemos —y, como veremos, la aleatorización es la herramienta más limpia para lograrlo.

2.2. El marco de resultados potenciales (potential outcomes)

¿Cómo le damos un sentido *preciso* a la palabra "causa"? El marco más usado hoy, debido al estadístico Donald Rubin, piensa en términos de **contrafactuales (counterfactuals)**: ¿qué le habría pasado a esta persona *si* hubiera (o no) recibido el tratamiento?

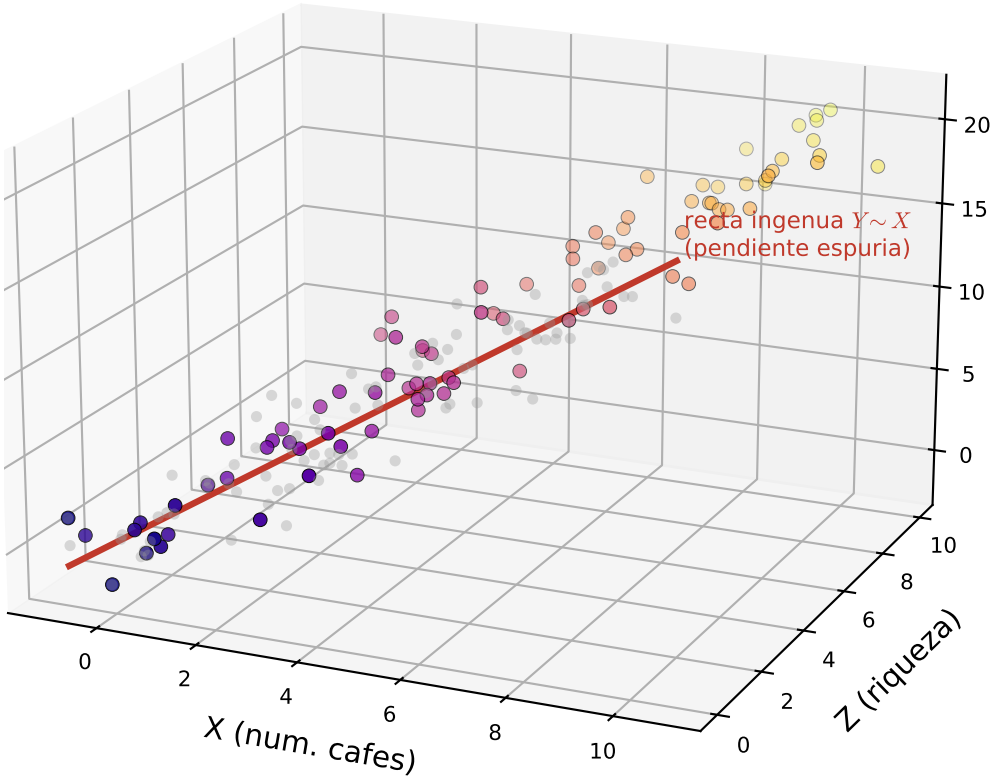

Figura 2: La confusión, vista en los *datos*. Un confusor Z (riqueza del distrito) impulsa **tanto** a X (número de cafeterías) **como** a Y (esperanza de vida). Si ignoramos Z y solo miramos el "piso" X–Y, aparece una recta de pendiente **positiva** (línea roja): más cafeterías parecen "predecir" más años de vida. Pero es una asociación **espuria**: al *controlar* por Z (comparar distritos de *igual* riqueza, mismo color), X deja de tener efecto sobre Y. El color codifica Z, el verdadero motor de ambas variables.

Definición 2.1 (Resultados potenciales y efecto causal individual)**.** Para cada unidad i (persona, hogar, escuela…) y un **tratamiento (treatment)** binario $T_i \in \{0, 1\}$, definimos *dos* **resultados potenciales**:

$$Y_i(1) = \text{resultado si } i \textit{ recibe} \text{ el tratamiento}, \qquad Y_i(0) = \text{resultado si } i \textit{ no} \text{ lo recibe}.$$

El **efecto causal del tratamiento sobre la unidad** i es la diferencia entre sus dos resultados potenciales:

$$\boxed{\tau_i = Y_i(1) - Y_i(0)}$$

La causalidad se define comparando *la misma unidad* en *dos mundos paralelos*: uno donde recibe el tratamiento y otro donde no, con todo lo demás igual. Si María toma una aspirina y su dolor de cabeza desaparece, el efecto causal es la diferencia entre "dolor de María con aspirina" y "dolor de María *sin* aspirina, ese mismo día". La causa no es "qué pasó" sino "qué pasó *comparado con lo que habría pasado*".

El problema fundamental de la inferencia causal

Para cada unidad solo podemos observar *uno* de los dos resultados potenciales: el que corresponde al tratamiento que efectivamente recibió. El otro queda como un contrafactual *inobservable*. Formalmente, el resultado observado es

$$Y_i^{\text{obs}} = \begin{cases} Y_i(1) & \text{si } T_i = 1, \\ Y_i(0) & \text{si } T_i = 0. \end{cases}$$

Como $\tau_i = Y_i(1) - Y_i(0)$ requiere *ambos*, **el efecto causal individual nunca se puede calcular directamente**. Es, en palabras de Holland (1986), un problema de *datos faltantes (missing data)*: siempre falta la mitad de cada par.

Unidad i	$Y_i(0)$	$Y_i(1)$	T_i	Y_i^{obs} **(observado)**
Ana	**?**	6	1	6
Beto	**?**	4	1	4
Carla	5	**?**	0	5
Diego	3	**?**	0	3

Figura 3: La **tabla de resultados potenciales**. Cada fila tiene *dos* resultados potenciales, pero solo observamos el de la columna correspondiente al tratamiento recibido T_i; el otro (en rojo, "?") es el contrafactual que *nunca* vemos. Por eso el efecto individual $\tau_i = Y_i(1) - Y_i(0)$ es incalculable a nivel de persona: a Ana le falta su $Y(0)$, a Carla le falta su $Y(1)$, etc.

El problema fundamental parece fatal: nunca vemos el efecto de una persona. La salida es *renunciar* al efecto individual y apuntar a un **promedio** sobre muchas personas. Si tuviéramos un grupo "tratado" y otro "no tratado" que fueran *comparables*, el promedio de uno aproximaría el contrafactual del otro. Toda la clase gira en torno a cómo conseguir esa comparabilidad —y la respuesta estrella será la aleatorización.

3. Efecto promedio del tratamiento (ATE) y sesgo de selección

3.1. El efecto promedio del tratamiento (ATE)

Definición 3.1 (Average Treatment Effect, ATE)**.** El **efecto promedio del tratamiento (ATE)** es el valor esperado del efecto individual sobre toda la población:

$$\boxed{\text{ATE} = \mathbb{E}\Big[Y(1) - Y(0)\Big] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]}$$

Aunque *nunca* veamos $Y_i(1)$ y $Y_i(0)$ juntos para una persona, el ATE es un objetivo más humilde y alcanzable: el efecto *promedio*. No nos dice si a Ana le ayudó el programa, sino si "en promedio" ayudó a personas como Ana. Para política pública eso suele bastar: ¿conviene escalar el programa a todo el país?

3.2. ¿Por qué la diferencia de medias observada no es el ATE?

Lo tentador es comparar el promedio de los tratados con el de los no tratados:

$$\widehat{\Delta} = \underbrace{\mathbb{E}[Y^{\text{obs}} \mid T=1]}_{\text{media de tratados}} - \underbrace{\mathbb{E}[Y^{\text{obs}} \mid T=0]}_{\text{media de controles}}.$$

El problema es que, cuando el tratamiento *no* se asignó al azar, los tratados y los controles pueden ser distintos *de entrada*. Descomponiendo esa diferencia:

Descomposición: efecto sobre los tratados + sesgo de selección

$$\mathbb{E}[Y^{\text{obs}} \mid T=1] - \mathbb{E}[Y^{\text{obs}} \mid T=0] = \underbrace{\mathbb{E}[Y(1) - Y(0) \mid T=1]}_{\text{efecto sobre los tratados (ATT)}} + \underbrace{\mathbb{E}[Y(0) \mid T=1] - \mathbb{E}[Y(0) \mid T=0]}_{\textbf{sesgo de selección}}.$$

El primer término es lo que queremos (un efecto causal); el segundo, el **sesgo de selección (selection bias)**, compara cómo *serían sin tratamiento* los dos grupos. Si los tratados ya eran distintos (mejores o peores) aun *sin* tratamiento, ese término *no* es cero y contamina la comparación.

El sesgo de selección aparece cuando una variable de confusión empuja a ciertas personas *hacia* el tratamiento y además afecta el resultado. La gente toma pastillas para el dolor *porque* le duele mucho; los que van a la universidad ya tenían familias, ingresos y habilidades distintas. Comparar sus resultados mezcla el efecto del tratamiento con esas diferencias previas. La diferencia de medias termina midiendo "efecto + confusión", no el efecto solo.

Ejemplo 3.1 (Confusión numérica: un programa de reforzamiento escolar)**.** Una ONG ofrece tutorías gratuitas. Para evaluarlas, alguien compara el puntaje promedio de los alumnos que se inscribieron (tratados) con el de los que no (controles). Supongamos —para poder "ver" la verdad oculta— que la población tiene dos tipos, y que el **verdadero efecto de la tutoría es** $+5$ **puntos para todos**:

Tipo	$Y(0)$	$Y(1)$	**efecto**	**% que se inscribe**
Alta motivación / NSE alto (100 alumnos)	70	75	+5	80 %
Baja motivación / NSE bajo (100 alumnos)	50	55	+5	20 %

El **ATE verdadero** es $\frac{1}{2}(5) + \frac{1}{2}(5) = \mathbf{5}$ puntos. Pero la *inscripción no es aleatoria*: los de alta motivación (que ya rinden más) se inscriben mucho más. Calculemos lo que vería el ingenuo:

Media de tratados (los inscritos, con su $Y(1)$):

$$\frac{80 \cdot 75 + 20 \cdot 55}{80 + 20} = \frac{6000 + 1100}{100} = 71.0.$$

Media de controles (los no inscritos, con su $Y(0)$):

$$\frac{20 \cdot 70 + 80 \cdot 50}{20 + 80} = \frac{1400 + 4000}{100} = 54.0.$$

Diferencia observada (ingenua): $71.0 - 54.0 = \mathbf{17}$ puntos.

¡El ingenuo concluiría un efecto de 17 puntos cuando el verdadero es 5! Los otros 12 puntos son **sesgo de selección**: el grupo tratado estaba lleno de alumnos que ya rendían más (su $Y(0)$ promedio era mayor). La confusión infló el efecto más de tres veces.

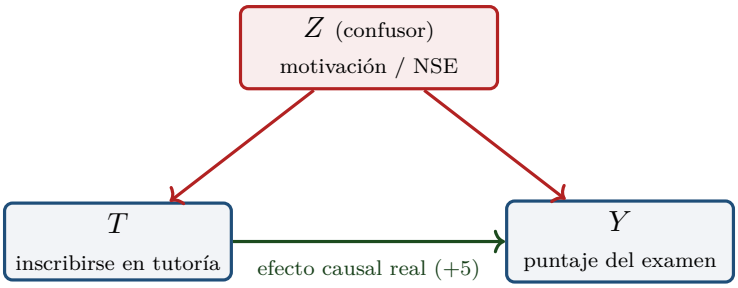

Figura 4: La motivación/NSE (Z) es un **confusor**: empuja a inscribirse ($Z \to T$) *y* sube el puntaje por su cuenta ($Z \to Y$). La flecha verde ($T \to Y$) es el efecto causal real (+5). La comparación ingenua de medias confunde la flecha verde con las dos rojas y reporta +17. **Cortar** la flecha $Z \to T$ (asignando T al azar) es justo lo que hará un RCT.

4. Experimentos aleatorizados (RCT): la solución

4.1. La aleatorización rompe la confusión

Definición 4.1 (Experimento completamente aleatorizado (RCT))**.** En un **ensayo controlado aleatorizado (randomized controlled trial, RCT)**, asignamos el tratamiento T_i *por sorteo*: cada unidad tiene la misma probabilidad de caer en el grupo de tratamiento o en el de control, *independientemente* de sus características o de sus resultados potenciales:

$$T \perp\!\!\!\perp \big(Y(0), Y(1)\big).$$

Por qué el RCT identifica el ATE

Si T es independiente de los resultados potenciales, entonces los tratados y los controles son —en promedio— *intercambiables*. En particular,

$$\mathbb{E}[Y(0) \mid T=1] = \mathbb{E}[Y(0) \mid T=0] \quad \Longrightarrow \quad \underbrace{\mathbb{E}[Y(0) \mid T=1] - \mathbb{E}[Y(0) \mid T=0]}_{\text{sesgo de selección}} = 0.$$

El sesgo de selección *se anula*, y por tanto la diferencia de medias observada **sí** estima el efecto causal:

$$\boxed{\mathbb{E}[Y^{\text{obs}} \mid T=1] - \mathbb{E}[Y^{\text{obs}} \mid T=0] = \mathbb{E}[Y(1) - Y(0)] = \text{ATE}}$$

La aleatorización es "magia" porque *equilibra todo a la vez*: lo observable (edad, ingreso, género) *y* lo inobservable (motivación, talento, salud previa). En promedio, el grupo de control se convierte en un *contrafactual válido* del grupo tratado: "cómo le habría ido a los tratados si no los hubiéramos tratado". El sorteo corta la flecha $Z \to T$ de la Figura 4: ya ninguna variable de confusión puede decidir quién entra al tratamiento.

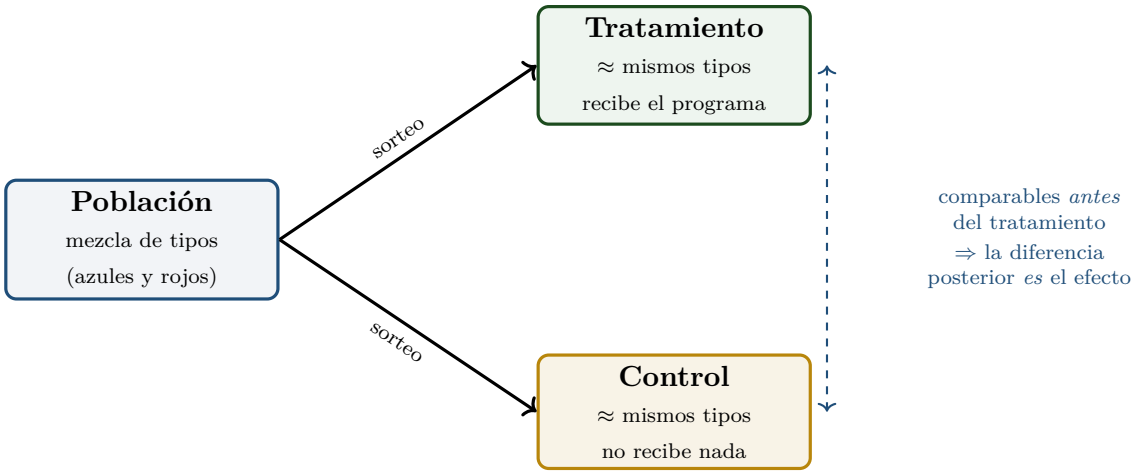

Figura 5: La **aleatorización** reparte a la población en dos grupos *estadísticamente idénticos antes* de tratar (misma mezcla de tipos, mismo perfil de confusores). Como solo difieren en que uno recibe el tratamiento, toda diferencia *posterior* en el resultado se atribuye al tratamiento: la diferencia de medias se vuelve un estimador insesgado del ATE.

Cuando una empresa tecnológica prueba dos versiones de su web ("A" y "B") mostrando cada una a usuarios elegidos al azar y comparando, por ejemplo, la tasa de clics, está haciendo **exactamente un RCT**. El *A/B testing* es un experimento aleatorizado: "A" es el control, "B" el tratamiento, y la aleatorización garantiza que las dos audiencias sean

comparables. Los mismos métodos de esta clase (diferencia de medias, error estándar, IC, prueba de hipótesis) son los que usan Netflix, Amazon o un ministerio que pilotea una política.

La independencia $T \perp\!\!\!\perp (Y(0), Y(1))$ vale *en esperanza*. En una muestra concreta, por azar, el grupo tratado puede quedar levemente más joven o con más mujeres: eso es ruido, no sesgo, y se refleja en el *error estándar*. Por eso un RCT sigue necesitando intervalos de confianza y pruebas de hipótesis (la Sección siguiente). Además, supone condiciones como *SUTVA* (el tratamiento de una unidad no afecta el resultado de otra), que pueden fallar (p. ej. vacunas, donde tratar a uno protege a su vecino).

5. Analizar un RCT: efecto, error estándar e intervalo de confianza

Una vez corrido el experimento, el análisis reutiliza *todo* lo de la Clase 6.

Estimador, error estándar y CI del efecto del tratamiento

Con N_t tratados y N_c controles, con medias muestrales $\bar{Y}_t, \bar{Y}_c$ y varianzas muestrales s_t^2, s_c^2:

- **Estimador del efecto** (diferencia de medias), insesgado para el ATE:

$$\hat{\tau} = \bar{Y}_t - \bar{Y}_c.$$

- **Error estándar** (varianza de una diferencia de grupos independientes = suma de varianzas):

$$\widehat{\mathrm{Var}}(\hat{\tau}) = \frac{s_t^2}{N_t} + \frac{s_c^2}{N_c}, \qquad \mathrm{SE}(\hat{\tau}) = \sqrt{\frac{s_t^2}{N_t} + \frac{s_c^2}{N_c}}.$$

- **Intervalo de confianza** al $(1-\alpha)$ y **prueba** de $H_0 : \mathrm{ATE} = 0$:

$$\hat{\tau} \pm t_{\mathrm{crit}} \cdot \mathrm{SE}(\hat{\tau}), \qquad t = \frac{\hat{\tau}}{\mathrm{SE}(\hat{\tau})}.$$

Con muestras grandes, $t_{\mathrm{crit}} = 1.96$ para 95 % y rechazamos H_0 si $|t| > 1.96$.

Ejemplo 5.1 (Un programa de transferencias y el gasto en alimentos)**.** Una ONG hace un RCT de un bono mensual en hogares rurales del Cusco. Mide el *gasto mensual en alimentos* (en soles) tras seis meses. Resultados:

Grupo	N	**media** $\bar{Y}$	**desv. est.** s
Tratamiento (recibe el bono)	50	420	80
Control (no recibe)	50	380	80

Efecto estimado:

$$\hat{\tau} = \bar{Y}_t - \bar{Y}_c = 420 - 380 = 40 \text{ soles.}$$

Error estándar:

$$\text{SE}(\hat{\tau}) = \sqrt{\frac{80^2}{50} + \frac{80^2}{50}} = \sqrt{128 + 128} = \sqrt{256} = 16 \text{ soles.}$$

Estadístico y prueba ($H_0 : \text{ATE} = 0$ vs. $H_1 : \text{ATE} \neq 0$, al 5 %):

$$t = \frac{\hat{\tau}}{\text{SE}} = \frac{40}{16} = 2.5.$$

Como $|t| = 2.5 > 1.96$, **rechazamos** H_0: el efecto es estadísticamente significativo. El p-valor (dos colas, aproximación normal) es $2\,(1 - \Phi(2.5)) \approx 0.012 < 0.05$.

Intervalo de confianza al 95 %**:**

$$40 \pm 1.96 \cdot 16 = 40 \pm 31.36 \implies [\,8.64,\ 71.36\,] \text{ soles.}$$

Interpretación: con 95 % de confianza, el bono incrementó el gasto en alimentos entre ~ 9 y ~ 71 soles mensuales; nuestra mejor estimación es 40 soles. Como el intervalo *no* incluye el 0, concluimos que hubo un efecto positivo. (Recuerda de la Clase 6: "significativo" no es lo mismo que "grande" —aquí la incertidumbre es amplia por N moderado.)

El estadístico $t = \hat{\tau}/\,\text{SE}$ será grande (efecto detectable) cuando: (i) el *efecto verdadero* sea grande, (ii) la *variabilidad* individual s sea baja, o (iii) la *muestra* N sea grande (porque $\text{SE} \propto 1/\sqrt{N}$). Por eso los experimentos planifican su tamaño con *cálculos de potencia* (Clase 6): para tener buena chance de detectar un efecto que valga la pena, antes de gastar en el estudio.

Un ejemplo real famoso: en Oregon no alcanzaba el presupuesto para dar seguro médico (Medicaid) a todos los que calificaban, así que lo **sortearon**. Eso creó un RCT natural: ganadores (tratamiento) vs. perdedores (control) de la lotería. Comparando ambos grupos, Amy Finkelstein y su equipo estimaron, por ejemplo, que tener el seguro subió la probabilidad de auto-reportar "buena salud" en 0.039 (con $\text{SE} = 0.008$). Un CI al 95 %: $0.039 \pm 1.96 \cdot 0.008 = [\,0.0233,\ 0.0547\,]$, que no incluye el 0: efecto significativo. Un sorteo administrativo se convirtió en evidencia causal de primera.

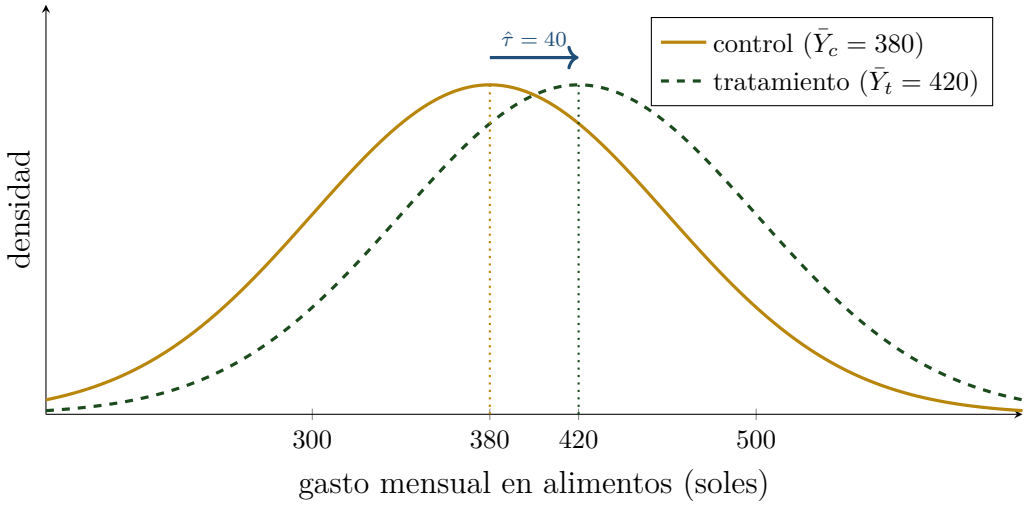

Figura 6: Distribuciones del gasto en los dos grupos del RCT. El grupo de tratamiento (verde) está *corrido* a la derecha respecto al control (ámbar): la diferencia de medias $\hat{\tau} = 40$ soles (flecha azul) es el efecto estimado. Aunque las distribuciones se *solapan* mucho (variabilidad individual alta), la diferencia de *promedios* es significativa porque el error estándar de la media es pequeño (SE = 16).

6. Ejemplo completo: un RCT de 20 personas, paso a paso

Para fijar todo lo anterior, hagamos *a mano* —sin atajos— el análisis completo de un experimento pequeño. Calcularemos las medias, el efecto, las varianzas, el error estándar, el estadístico t, el p-valor y el intervalo de confianza; y al final lo confirmaremos con un método distinto (*inferencia por permutación*).

6.1. Contexto y datos

Una ONG en Lima evalúa un **programa de capacitación laboral (job-training)** de tres meses para jóvenes que buscan empleo. Reclutó a 20 participantes y, por **sorteo**, asignó a 10 al grupo de **tratamiento** (T, reciben la capacitación) y a 10 al de **control** (C, no la reciben). Seis meses después mide el **ingreso mensual** (en soles) de cada persona. La pregunta causal es: ¿la capacitación *aumentó* el ingreso?

Tratamiento (T)		Control (C)	
id	**ingreso (S/)**	**id**	**ingreso (S/)**
1	1180	11	1095
2	1210	12	1110
3	1230	13	1125
4	1260	14	1155
5	1280	15	1185
6	1360	16	1215
7	1380	17	1245
8	1410	18	1275
9	1430	19	1290
10	1460	20	1305
suma	**13 200**	**suma**	**12 000**

Como el tratamiento se asignó *al azar*, los dos grupos son comparables "de entrada" (Sección anterior): no hay confusores que decidan quién se capacita. Por eso la diferencia de ingresos que veamos al final se podrá leer como **efecto causal** del programa, no como una diferencia preexistente.

6.2. Paso 1 — Medias de cada grupo

La media es la suma dividida entre el número de personas ($n_t = n_c = 10$):

$$\bar{Y}_t = \frac{13\,200}{10} = 1320 \text{ soles}, \qquad \bar{Y}_c = \frac{12\,000}{10} = 1200 \text{ soles}.$$

En palabras: en promedio, los capacitados ganan 1320 soles y los del control 1200.

6.3. Paso 2 — Efecto estimado (diferencia de medias)

El estimador del efecto del tratamiento es la diferencia de las dos medias:

$$\boxed{\hat{\tau} = \bar{Y}_t - \bar{Y}_c = 1320 - 1200 = 120 \text{ soles}.}$$

Interpretación: nuestra mejor estimación es que el programa subió el ingreso mensual en 120 soles en promedio. Falta saber si esa cifra es "real" o podría deberse al azar del sorteo: para eso necesitamos el error estándar.

6.4. Paso 3 — Varianzas y desviaciones muestrales

La **varianza muestral** mide cuánto se dispersan los datos alrededor de su media; se divide entre $n - 1$ (no entre n):

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2.$$

Grupo control (media 1200). Las desviaciones $Y_i - 1200$ son $\pm 15, \pm 45, \pm 75, \pm 90, \pm 105$. Sumando sus cuadrados:

$$\sum (Y_i - \bar{Y}_c)^2 = 2\,(15^2 + 45^2 + 75^2 + 90^2 + 105^2) = 2\,(225 + 2025 + 5625 + 8100 + 11025) = 54\,000.$$

$$s_c^2 = \frac{54\,000}{10-1} = \frac{54\,000}{9} = 6\,000, \qquad s_c = \sqrt{6\,000} \approx 77.46 \text{ soles.}$$

Grupo tratamiento (media 1320). Las desviaciones $Y_i - 1320$ son $\pm 40, \pm 60, \pm 90, \pm 110, \pm 140$. Sus cuadrados:

$$\sum (Y_i - \bar{Y}_t)^2 = 2\,(40^2 + 60^2 + 90^2 + 110^2 + 140^2) = 2\,(1600 + 3600 + 8100 + 12100 + 19600) = 90\,000.$$

$$s_t^2 = \frac{90\,000}{9} = 10\,000, \qquad s_t = \sqrt{10\,000} = 100 \text{ soles.}$$

El grupo tratado está algo más disperso ($s_t = 100$) que el control ($s_c \approx 77$): el programa pudo ayudar más a unos que a otros. La varianza es la materia prima del *error estándar*: a más dispersión individual, más incertidumbre sobre el promedio.

6.5. Paso 4 — Error estándar de la diferencia

Como los dos grupos son independientes, la varianza de la diferencia de medias es la *suma* de las varianzas de cada media:

$$\text{SE}(\hat{\tau}) = \sqrt{\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}} = \sqrt{\frac{10\,000}{10} + \frac{6\,000}{10}} = \sqrt{1\,000 + 600} = \sqrt{1\,600} = \boxed{40 \text{ soles.}}$$

Interpretación: si repitiéramos el experimento muchas veces (con otros sorteos), la diferencia de medias estimada variaría típicamente en unos ± 40 soles alrededor del valor verdadero. Es la "barra de error" de nuestro $\hat{\tau} = 120$.

6.6. Paso 5 — Estadístico t y grados de libertad

Comparamos el efecto con su propia incertidumbre:

$$t = \frac{\hat{\tau}}{\text{SE}(\hat{\tau})} = \frac{120}{40} = \boxed{3.0.}$$

El efecto estimado es **3 errores estándar** distinto de cero: bastante lejos del 0. Los **grados de libertad (degrees of freedom)** con la corrección de **Welch** (que no asume varianzas iguales) son

$$\text{df} = \frac{\left(\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}\right)^2}{\frac{(s_t^2/n_t)^2}{n_t - 1} + \frac{(s_c^2/n_c)^2}{n_c - 1}} = \frac{(1\,000 + 600)^2}{\frac{1\,000^2}{9} + \frac{600^2}{9}} = \frac{2\,560\,000}{151\,111.1} \approx 16.94.$$

(Con la versión de varianza combinada —*pooled*— serían $\text{df} = n_t + n_c - 2 = 18$; las conclusiones no cambian.)

6.7. Paso 6 — p-valor (dos colas)

El **p-valor** responde: *si el programa NO tuviera ningún efecto* (H_0 : ATE $= 0$), ¿qué tan probable sería ver una diferencia tan grande (o mayor) que la observada, en valor absoluto, sólo por el azar del sorteo? Con $t = 3.0$ y df ≈ 17, la tabla t (dos colas) da

$$p\text{-valor} \approx 0.0081.$$

Interpretación: habría sólo alrededor de 8 de cada 1 000 posibilidades de obtener una diferencia ≥ 120 soles (en magnitud) si la capacitación no sirviera de nada. Es muy improbable; eso es evidencia *en contra* de "efecto cero". (La aproximación normal grosera daría $2\,[1 - \Phi(3)] \approx 0.0027$; con n pequeño la t es más honesta y reporta 0.0081.)

6.8. Paso 7 — Intervalo de confianza al 95 %

Con la aproximación normal ($z_{0.975} = 1.96$):

$$\hat{\tau} \pm 1.96 \cdot \text{SE} = 120 \pm 1.96 \cdot 40 = 120 \pm 78.4 \implies [\,41.6,\ 198.4\,] \text{ soles.}$$

Si preferimos ser exactos con n pequeño usamos el valor crítico de la t de Welch ($t_{0.975,\,17} \approx 2.11$), que *ensancha* un poco el intervalo:

$$120 \pm 2.11 \cdot 40 = 120 \pm 84.4 \implies [\,35.6,\ 204.4\,] \text{ soles.}$$

Interpretación: con 95 % de confianza, el efecto verdadero de la capacitación está entre ~ 36 y ~ 204 soles mensuales; nuestra mejor estimación puntual es 120 soles.

6.9. Paso 8 — Decisión e interpretación causal

Conclusión del experimento

Al nivel $\alpha = 0.05$: como $|t| = 3.0 > 1.96$ (equivalentemente, $p \approx 0.008 < 0.05$, y el intervalo de confianza *no* contiene el 0), **rechazamos** H_0. Hay evidencia estadísticamente significativa de un efecto positivo. Y como el tratamiento se asignó **al azar**, esa diferencia se interpreta **causalmente**: la capacitación *causó* un aumento estimado de 120 soles en el ingreso mensual.

El intervalo es ancho (~ 36 a ~ 204 soles) porque la muestra es muy pequeña ($n = 20$). "Significativo" sólo dice "distinto de cero con confianza"; la magnitud exacta del efecto sigue siendo incierta. Con más personas, el SE bajaría (recuerda SE $\propto 1/\sqrt{n}$) y el intervalo se estrecharía.

6.10. Otra ruta al p-valor: inferencia por permutación

¿Y si no queremos confiar en la distribución t (que supone cierta normalidad), sobre todo con n chico? Hay un método que usa *sólo* la aleatorización que ya hicimos: la **inferencia por permutación (randomization / permutation inference)**.

La idea de la permutación

Bajo la hipótesis nula "el programa no tiene efecto", el ingreso de cada persona sería *el mismo* reciba o no la capacitación. Entonces las etiquetas T/C son **intercambiables (exchangeable)**: cualquier reparto de 10 etiquetas "T" entre las 20 personas es igual de probable. El procedimiento es:

1. Junta los 20 ingresos en una sola bolsa.
2. **Permuta** (baraja) las etiquetas: elige al azar 10 para ser "T" y 10 para ser "C", y recalcula la diferencia de medias.
3. Repite *muchas* veces. Esas diferencias forman la **distribución nula** (lo que veríamos "por puro azar" si no hubiera efecto).
4. El **p-valor de permutación** es la proporción de permutaciones cuya diferencia es, en magnitud, $\geq$ la observada $|\hat{\tau}| = 120$.

Con 20 personas hay exactamente $\binom{20}{10} = 184\,756$ formas de repartir las etiquetas, así que podemos recorrerlas *todas* (p-valor *exacto*). De esas, 1 684 dan una diferencia de magnitud ≥ 120 soles. Por lo tanto:

$$p_{\text{perm}} = \frac{1\,684}{184\,756} \approx 0.0091.$$

Es casi idéntico al p-valor de la prueba t (≈ 0.0081): **dos caminos muy distintos llegan a la misma conclusión**. La aleatorización no sólo justificó la interpretación causal: también nos dio, por sí sola, una prueba de hipótesis.

RCT de 20 personas: prueba t y prueba de permutación

```
# --- datos del experimento ---
y <- c(1180,1210,1230,1260,1280,1360,1380,1410,1430,1460,   # tratamiento
       1095,1110,1125,1155,1185,1215,1245,1275,1290,1305)   # control
grupo <- factor(c(rep("T",10), rep("C",10)))

# --- (1) diferencia de medias y prueba t de Welch ---
tau_obs <- mean(y[grupo=="T"]) - mean(y[grupo=="C"])
tau_obs                        # 120
t.test(y ~ grupo)              # t = 3.0, df ~ 16.9, p ~ 0.008, IC95 ~ [35.6,
   204.4]

# --- (2) inferencia por PERMUTACION (randomization inference) ---
set.seed(2026)
B <- 20000
```

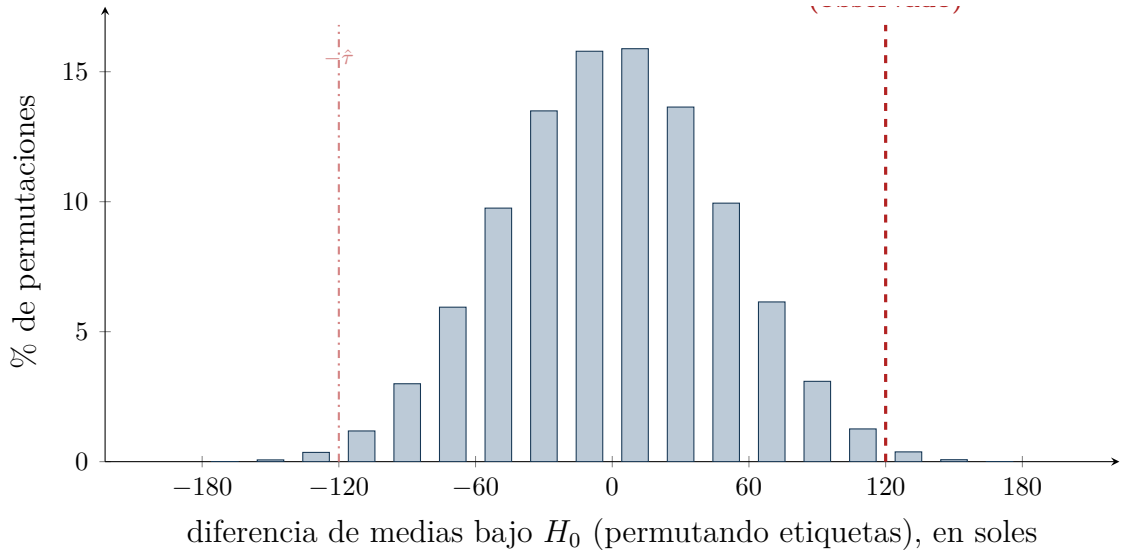

Figura 7: **Distribución nula por permutación** (exacta, las 184 756 formas de repartir las etiquetas T/C). Si el programa no tuviera efecto, la diferencia de medias se agruparía alrededor de 0. La línea roja marca el valor *observado* $\hat{\tau} = 120$: cae muy en la **cola**. La proporción de permutaciones tan extremas o más ($|\text{dif}| \geq 120$, ambas colas) es el p-valor de permutación ≈ 0.0091, coherente con la prueba t.

```
perm_dif <- replicate(B, {
  g <- sample(grupo)                    # baraja las etiquetas T/C
  mean(y[g=="T"]) - mean(y[g=="C"])     # diferencia bajo H0
})
p_perm <- mean(abs(perm_dif) >= abs(tau_obs))   # dos colas
p_perm                          # ~ 0.009  (coincide con la prueba t)

# (exacto, recorriendo las choose(20,10)=184756 particiones: p = 1684/184756 =
    0.0091)
```

La prueba t *supone* una forma (normal) para la distribución del estadístico; la permutación *no supone nada*: construye la distribución nula *a partir de los propios datos*, imitando el sorteo que de hecho hicimos. Por eso es especialmente confiable cuando la muestra es chica o los datos no parecen normales. Que ambos métodos coincidan ($p \approx 0.008$ vs. 0.009) nos da confianza extra en la conclusión: la capacitación tuvo un efecto causal positivo.

7. Comparaciones y regresión no paramétricas

Hasta aquí asumimos cosas: normalidad (para el t), o que el efecto se resume en una *diferencia de medias*. ¿Y si no queremos asumir una forma? Los métodos **no paramétricos (nonparametric)** dejan que "los datos hablen" sin imponer una familia de distribuciones ni una forma funcional (lineal, normal, etc.).

7.1. Comparar dos distribuciones sin asumir normalidad

A veces no basta con comparar *medias*: queremos saber si las *distribuciones completas* de $Y(1)$ y $Y(0)$ difieren (¿el tratamiento ayudó a todos, o solo a la cola?). La prueba **de Kolmogorov–Smirnov (KS)** compara las dos **funciones de distribución empíricas (empirical CDF)** sin suponer normalidad.

Definición 7.1 (Estadístico de Kolmogorov–Smirnov)**.** Dadas dos muestras con CDF empíricas F_n (grupo 1) y G_m (grupo 2), el estadístico KS es la **máxima distancia vertical** entre ambas curvas:

$$D_{nm} = \max_x \Big| F_n(x) - G_m(x) \Big|.$$

Se contrasta $H_0 : F = G$ (mismas distribuciones) contra $H_1 : F \neq G$; un D_{nm} grande es evidencia de que las distribuciones difieren. (En R: `ks.test`.)

Las pruebas **basadas en rangos / distribuciones** (Kolmogorov–Smirnov, Mann–Whitney/Wilcoxon, Fisher exacto) son *robustas*: no dependen de que los datos sean normales, y resisten valores atípicos porque usan posiciones (rangos) en vez de las magnitudes crudas. Son la opción prudente cuando n es chico o la forma de los datos es rara —algo común en encuestas de hogares.

7.2. Regresión no paramétrica: suavizado por kernel

Para dos variables continuas X e Y queremos $\mathbb{E}[Y \mid X = x] = g(x)$ *sin* suponer que g es una recta. La idea de la **regresión por kernel** (Nadaraya–Watson) es **local**: para predecir Y en un punto x, promediamos los Y_i de las observaciones *cercanas* a x, dando más peso a las más próximas.

Estimador de regresión por kernel

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} y_i \, K\Big(\frac{x - x_i}{h}\Big)}{\sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big)}$$

Es un *promedio ponderado* de los y_i: el **kernel** $K(\cdot)$ asigna pesos que decaen con la distancia, y el **ancho de banda (bandwidth)** h fija el "radio" del vecindario.

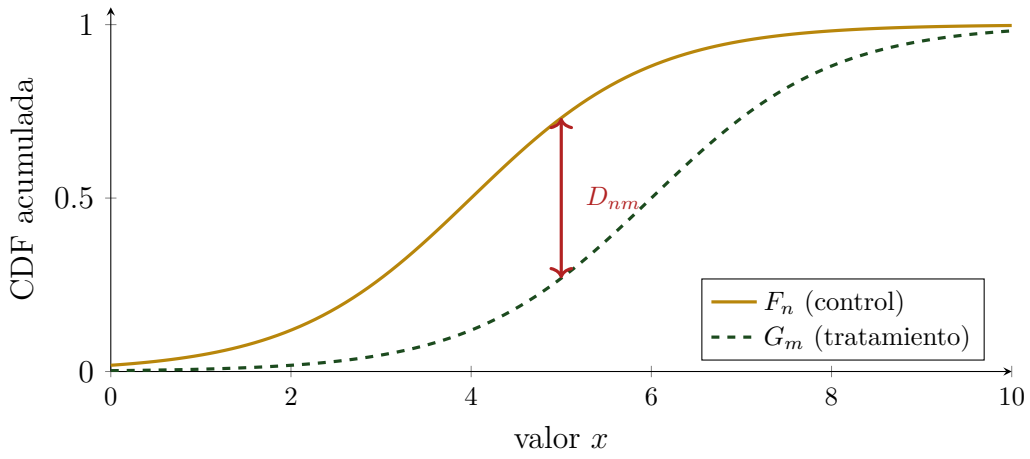

Figura 8: Dos CDF empíricas: control (ámbar) y tratamiento (verde, desplazada a la derecha porque el tratamiento subió los valores). El estadístico de Kolmogorov–Smirnov D_{nm} es la **mayor brecha vertical** entre ambas (flecha roja). Si las distribuciones fueran iguales, las curvas se solaparían y D_{nm} sería ≈ 0.

h controla cuánto suavizamos:

- h **grande** $\Rightarrow$ **mucho suavizado**: la curva es lisa pero puede *perder* estructura real (**sesgo** alto: "aplana" los detalles).
- h **pequeño** $\Rightarrow$ **poco suavizado**: la curva sigue cada subida y bajada, incluso el ruido (**varianza** alta: "sobre-ajusta").

Es el *compromiso sesgo–varianza* de la Clase 6, ahora en una curva. Se suele elegir h por **validación cruzada (cross-validation)**: el h que mejor predice los puntos dejados fuera. A medida que crece n se promete bajar h (más datos permiten vecindarios más finos).

Lo *paramétrico* (recta, normal) es eficiente *si* el supuesto es correcto; lo *no paramétrico* es flexible pero exige más datos (paga el precio de no asumir nada). En la práctica, la regresión por kernel es ideal en la fase **exploratoria**: antes de imponer un modelo, dibujamos la relación suavizada para *ver* su forma. Otras variantes son la regresión local lineal (*LOESS*), los *splines* y las series polinómicas.

8. Simular un RCT en R

La mejor forma de convencerse de que "la aleatorización identifica el ATE" es *verla* en una simulación, donde conocemos la verdad. Generamos resultados potenciales, asignamos el tratamiento al azar y comprobamos que la diferencia de medias recupera el ATE verdadero

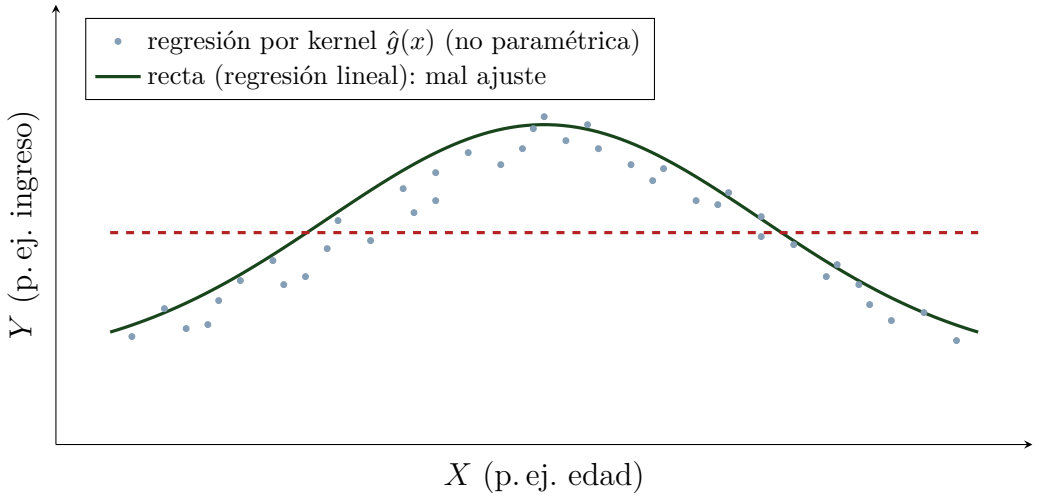

Figura 9: Una nube de puntos con relación *no lineal* entre X e Y (sube y luego baja, forma de joroba). La **recta** (rojo punteado) impone una forma equivocada y no captura nada: predice casi lo mismo para toda edad. La **regresión por kernel** (verde) sigue localmente a los datos y recupera la verdadera curva, sin que tuviéramos que adivinar su fórmula. Así "los datos hablan".

—algo imposible si dejáramos que un confusor decidiera quién se trata.

Simular un RCT: la diferencia de medias recupera el ATE

```
set.seed(2026)
N   <- 2000                  # numero de unidades (p.ej. hogares)
# --- Resultados potenciales (que en la realidad NO se ven juntos) ---
Z   <- rnorm(N, 50, 10)      # confusor latente (motivacion / NSE)
Y0  <- 300 + 1.5*Z + rnorm(N, 0, 40)   # resultado SIN tratamiento
tau_true <- 40                          # EFECTO causal verdadero (ATE = 40)
Y1  <- Y0 + tau_true                    # resultado CON tratamiento (efecto
    +40)

# ================= (A) RCT: tratamiento ALEATORIO =================
T_rct <- rbinom(N, 1, 0.5)              # sorteo: independiente de Z, Y0, Y1
Yobs  <- ifelse(T_rct == 1, Y1, Y0)     # solo vemos uno de los dos
dif_rct <- mean(Yobs[T_rct==1]) - mean(Yobs[T_rct==0])
dif_rct                                  # ~ 40  (recupera el ATE!)
t.test(Yobs[T_rct==1], Yobs[T_rct==0])   # IC y prueba del efecto

# ============ (B) NO experimental: confusor decide T =============
# Los de Z alto (mas motivados) se tratan mas -> sesgo de seleccion.
prob  <- plogis((Z - 50)/6)            # P(tratarse) sube con Z
T_obs <- rbinom(N, 1, prob)
Yobs2 <- ifelse(T_obs == 1, Y1, Y0)
dif_naive <- mean(Yobs2[T_obs==1]) - mean(Yobs2[T_obs==0])
```

```
dif_naive                                  # >> 40 : INFLADO por confusion
```

La diferencia de medias del **RCT** (`dif_rct`) cae cerca de 40, el ATE verdadero: la aleatorización funcionó. En cambio la diferencia **ingenua** (`dif_naive`), donde el confusor Z decide quién se trata, queda muy por encima de 40: el **sesgo de selección** infla el efecto, justo como en el ejemplo de las tutorías. La única diferencia entre (A) y (B) es *cómo se asignó el tratamiento* —y eso lo cambia todo.

Para "ver" una correlación espuria, genera dos variables que dependan de una tercera pero no entre sí: `Z<-rnorm(500); X<-Z+rnorm(500); Y<-Z+rnorm(500)`. Entonces `cor(X,Y)` será alta (porque ambas siguen a Z), pero X no causa Y. Al "controlar" Z —p. ej. con `lm(Y ~ X + Z)`— el coeficiente de X se desvanece: la asociación era obra del confusor.

Resumen

Resumen de la Clase 8

- **Correlación $\neq$ causalidad.** Una asociación puede deberse a causalidad directa, inversa o a una *variable de confusión*.
- **Resultados potenciales:** cada unidad tiene $Y_i(1)$ y $Y_i(0)$; el efecto individual es $Y_i(1) - Y_i(0)$. El *problema fundamental* es que solo observamos uno de los dos.
- **ATE** $= \mathbb{E}[Y(1) - Y(0)]$. La diferencia de medias observada = efecto + **sesgo de selección**; este último aparece cuando el tratamiento *no* es aleatorio.
- **RCT:** la aleatorización hace $T \perp\!\!\!\perp (Y(0), Y(1))$, anula el sesgo de selección y vuelve la diferencia de medias un estimador insesgado del ATE. Es el mismo principio del *A/B testing*.
- **Analizar un RCT:** $\hat{\tau} = \bar{Y}_t - \bar{Y}_c$, con SE $= \sqrt{s_t^2/N_t + s_c^2/N_c}$; de ahí salen el CI y la prueba de $H_0 : \text{ATE} = 0$ (todo de la Clase 6).
- **No paramétrico:** comparar distribuciones sin asumir normalidad (Kolmogorov–Smirnov) y estimar $\mathbb{E}[Y \mid X]$ *localmente* (regresión por kernel), con el *bandwidth* h gobernando el compromiso sesgo–varianza.

Mensaje central: un buen *diseño* (aleatorizar) hace la inferencia causal honesta, y un buen análisis (errores estándar, IC, métodos flexibles) la hace creíble. Esa combinación es la que llevó a Esther Duflo al Nobel.

Fuentes y atribución

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`.