Trabajo práctico 4: Relación entre dos variables

AL TERMINAR ESTA GUÍA

- Construir una tabla de contingencia y calcular los porcentajes en el sentido que contesta la pregunta.
- Reconocer una variable de confusión y la paradoja de Simpson.
- Calcular e interpretar la covarianza, la correlación de Pearson, la recta de mínimos cuadrados y el R^2, y no extrapolar.
- Calcular las mismas medidas con software sobre datos reales, y distinguir correlación de causalidad.

PARA EL DOCENTE

Los ejercicios 1 a 3 son para el encuentro práctico; el 3 lleva la mayor parte de la hora y conviene que el diagrama se haga en papel cuadriculado. Antes del 3 se puede jugar a adivinar r con el simulador de correlación del sitio.

El 4 y el 5 necesitan computadora: sirven para un encuentro virtual o como tarea. El 6 funciona como cierre, discutido en grupos.

EJERCICIO 1 Actividad física y edad

A mano

En una encuesta de salud a personas adultas de una ciudad se preguntó si hacen actividad física al menos tres veces por semana. Las respuestas, por grupo de edad:

Edad	Hace actividad	No hace
18 a 39	60	43
40 a 59	39	47
60 y más	18	55

a) Completá la tabla con los totales de cada fila y de cada columna. ¿Qué porcentaje del total hace actividad física?

RESOLUCIÓN

Edad	Hace actividad	No hace	Total
18 a 39	60	43	103
40 a 59	39	47	86
60 y más	18	55	73
Total	**117**	**145**	**262**

Los totales son las frecuencias marginales: la distribución de cada variable por separado. Hace actividad física 117 de 262 personas, el 44,7 %.

b) ¿Qué porcentaje de las personas de 60 años y más hace actividad física? ¿Qué porcentaje de las que hacen actividad física tiene 60 años y más? ¿Por qué no da lo mismo?

RESOLUCIÓN

$$P(\text{hace} \mid 60+) = \frac{18}{73} = 0{,}247 \qquad P(60+ \mid \text{hace}) = \frac{18}{117} = 0{,}154$$

Las dos usan las mismas 18 personas, pero dividen por totales distintos: la primera, por las 73 personas de 60 y más; la segunda, por las 117 que hacen actividad. Son preguntas distintas. La primera dice cómo se comporta un grupo de edad; la segunda, cómo se compone el grupo de los que hacen actividad, que depende también de cuántas personas hay de cada edad.

c) ¿La actividad física cambia con la edad? Calculá los porcentajes que permiten contestarlo y decí en qué sentido de la tabla los calculaste.

RESOLUCIÓN

Edad	Hace actividad	No hace	Total
18 a 39	58,3	41,7	100,0
40 a 59	45,3	54,7	100,0
60 y más	24,7	75,3	100,0
Total	**44,7**	**55,3**	**100,0**

Por fila: dentro de cada grupo de edad. Hace actividad el 58,3 % de los de 18 a 39, el 45,3 % de los de 40 a 59 y el 24,7 % de los de 60 y más. Las variables están asociadas: si no lo estuvieran, los tres porcentajes serían parecidos al 44,7 % del total.

La regla: los porcentajes se calculan dentro de las categorías de la variable que se piensa como explicativa —acá la edad— y se comparan entre ellas. Calcular por columna contesta otra pregunta y, con grupos de distinto tamaño, confunde.

EJERCICIO 2 ¿Qué hospital opera mejor?

A mano

Dos hospitales informan cuántos pacientes operaron de la misma cirugía el último año y cuántos sobrevivieron. El A es el hospital de referencia de la provincia, al que se derivan los casos complicados:

Hospital	Estado al ingresar	Operados	Sobrevivieron
A	Leve	52	49
A	Grave	209	141
B	Leve	209	185
B	Grave	44	26

a) Sin separar por estado, ¿qué porcentaje de los operados sobrevivió en cada hospital? ¿Cuál parece mejor?

> RESOLUCIÓN
>
> $A: \frac{49+141}{52+209} = \frac{190}{261} = 72{,}8\,\%$ $\qquad$ $B: \frac{211}{253} = 83{,}4\,\%$
>
> Mirado así, el B parece mejor.

b) Ahora calculalo por separado para los pacientes leves y para los graves. ¿Qué hospital tiene mejor resultado en cada grupo?

> RESOLUCIÓN
>
Estado	A	B
Leve	94,2 %	88,5 %
Grave	67,5 %	59,1 %
Todos	72,8 %	83,4 %
>
> El A tiene mejor sobrevida tanto con los leves como con los graves. Y sin embargo, juntando los dos grupos, sale peor.

c) ¿Cómo puede ser? Mirá qué proporción de pacientes graves recibió cada hospital.

> RESOLUCIÓN
>
> El A operó 80 % de pacientes graves; el B, 17 %. Los graves sobreviven mucho menos en cualquier hospital, y el A tiene muchos más: su porcentaje global queda tirado hacia abajo por la mezcla de pacientes, no por cómo opera.
>
> Es la paradoja de Simpson. El estado al ingresar es una variable de confusión: está asociada con el hospital (los graves van al A) y con el resultado (los graves sobreviven menos). Cuando una tercera variable hace eso, la comparación global puede dar vuelta la que se ve dentro de cada grupo.

d) Si tuvieras que operarte, ¿qué hospital elegirías? ¿Qué comparación es la justa?

RESOLUCIÓN

El A: sea cual sea el estado de un paciente, ahí le va mejor. La comparación justa es la que compara pacientes parecidos, o sea la de cada grupo por separado. El porcentaje global mezcla la calidad del hospital con el tipo de pacientes que recibe, y castiga al que acepta los casos difíciles.

EJERCICIO 3 Horas de estudio y nota

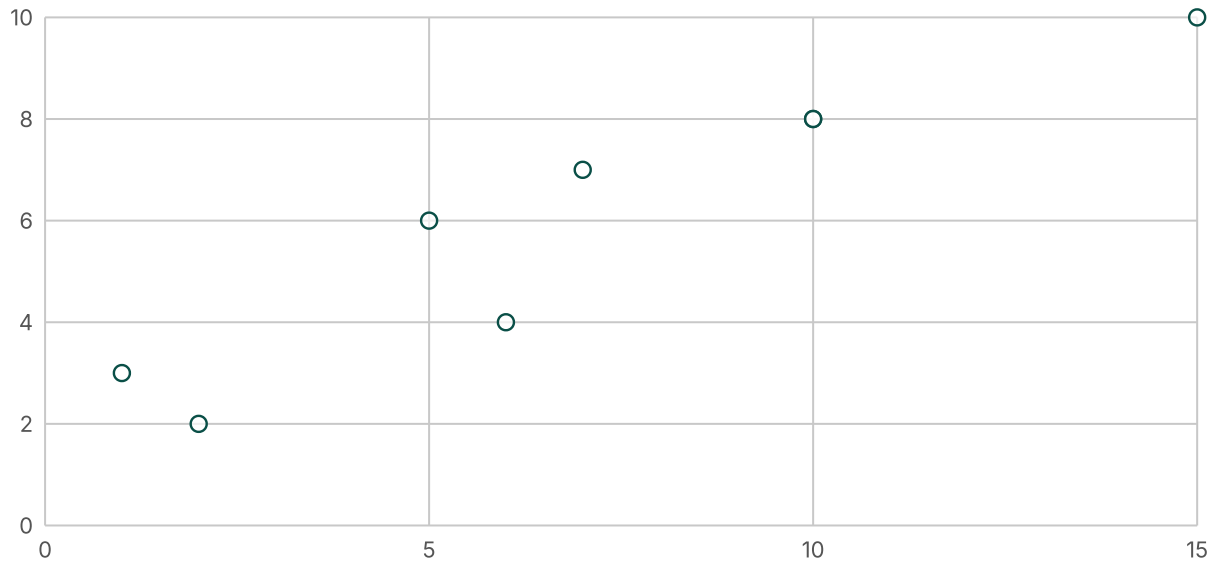
A mano

Ocho estudiantes anotaron cuántas horas estudiaron para el primer parcial. Después se registró la nota que sacaron:

Estudiante	1	2	3	4	5	6	7	8
Horas de estudio	6	1	10	15	2	5	10	7
Nota	4	3	8	10	2	6	8	7

a) Dibujá el diagrama de dispersión, con las horas en el eje horizontal. Antes de calcular nada, ¿cuánto dirías que vale el coeficiente de correlación?

RESOLUCIÓN



Eje horizontal: horas de estudio. Eje vertical: nota.

Los puntos suben de izquierda a derecha y siguen más o menos una línea: es una relación positiva y bastante fuerte. Una estimación a ojo entre 0,7 y 0,9 está bien. Adivinar antes de calcular sirve para detectar errores: si la cuenta da negativa o cerca de cero, algo se hizo mal.

b) Calculá la covarianza y el coeficiente de correlación de Pearson. Armá la tabla de desvíos y productos.

RESOLUCIÓN

Medias: $\bar{x} = 7$ horas y $\bar{y} = 6$ puntos.

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$
6	4	−1	−2	2	1	4
1	3	−6	−3	18	36	9
10	8	3	2	6	9	4
15	10	8	4	32	64	16
2	2	−5	−4	20	25	16
5	6	−2	0	0	4	0
10	8	3	2	6	9	4
7	7	0	1	0	0	1
Σ = 56	**48**	**0**	**0**	**84**	**148**	**54**

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{84}{7} = 12{,}00$$

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{84}{\sqrt{148 \cdot 54}} = 0{,}940$$

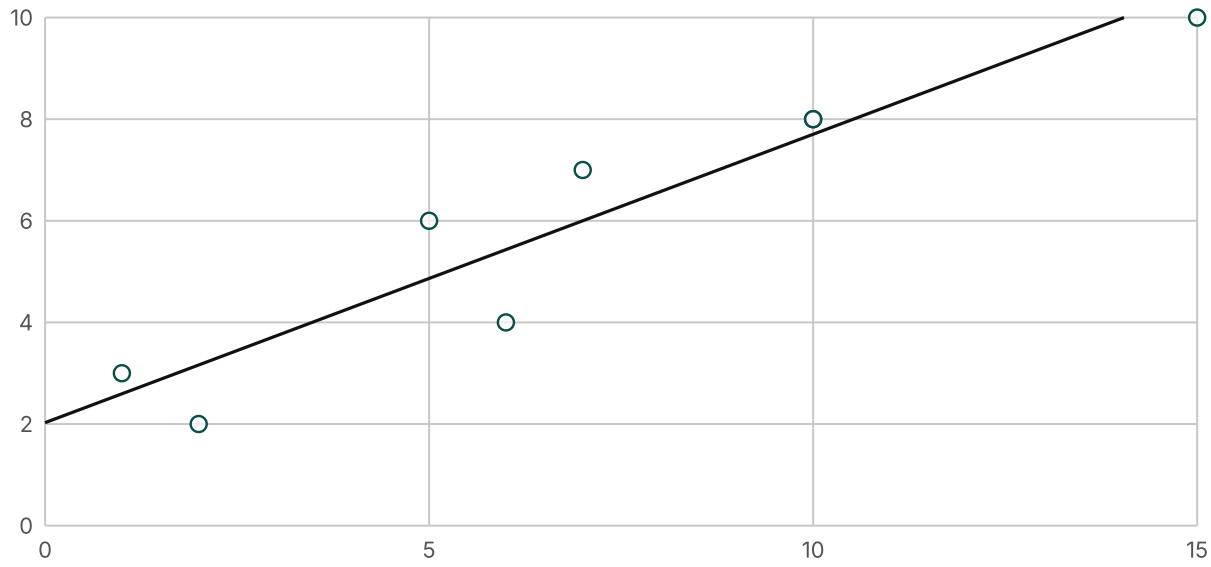
La covarianza es positiva, pero su tamaño depende de las unidades (horas por puntos) y no dice si la relación es fuerte. La correlación sí: no tiene unidades y va de −1 a 1.

c) Calculá la recta de mínimos cuadrados para predecir la nota a partir de las horas. ¿Qué significan la pendiente y la ordenada al origen?

RESOLUCIÓN

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{84}{148} = 0{,}568 \qquad a = \bar{y} - b\,\bar{x} = 6 - 0{,}568 \cdot 7 = 2{,}027$$

$$\hat{y} = 2{,}027 + 0{,}568\,x$$



Eje horizontal: horas de estudio. Eje vertical: nota. La línea es la recta de mínimos cuadrados.

> La pendiente dice que, en promedio, cada hora más de estudio se asocia con 0,57 puntos más de nota. La ordenada, 2,03, es la nota que predice la recta para cero horas: ningún estudiante de la muestra estudió cero horas, así que es un valor de la fórmula más que algo para interpretar.

d) Calculá el coeficiente de determinación e interpretalo. ¿Qué nota predice la recta para alguien que estudió 8 horas? ¿Y para alguien que estudió 40?

> RESOLUCIÓN
>
> $R^2 = r^2 = 0{,}940^2 = 0{,}883$
>
> El 88,3 % de la variación de las notas se explica por su relación lineal con las horas de estudio; el resto depende de otras cosas: lo que cada uno sabía antes, cómo estudió, el día del examen.
>
> $\hat{y}(8) = 2{,}027 + 0{,}568 \cdot 8 = 6{,}57 \qquad \hat{y}(40) = 24{,}73$
>
> Para 8 horas la predicción es razonable: está dentro del rango observado, de 1 a 15 horas. Para 40 horas la recta da 24,7, una nota imposible: es una extrapolación. Nada garantiza que la relación siga siendo lineal fuera de los datos, y en este caso seguro que no, porque la nota tiene techo.

EJERCICIO 4 Informalidad e ingreso en 32 aglomerados

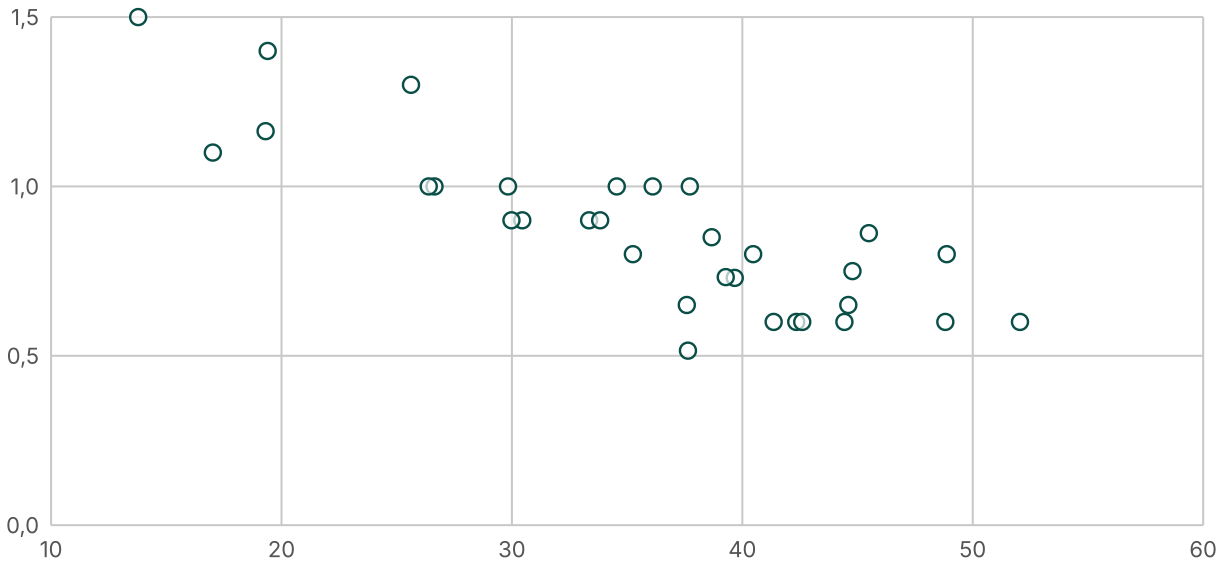
Con software

La base «Treinta y dos aglomerados por diez indicadores» resume la EPH del primer trimestre de 2026: una fila por aglomerado urbano del país, con tasas de empleo y desocupación, ingresos, educación, informalidad y edad. Se baja de estadistica.ar/datos/indicadores-aglomerados. El separador es la coma y el punto es el decimal.

Interesan no_registrado, el porcentaje de asalariados sin descuento jubilatorio, e ingreso_mediano, el ingreso mediano de la ocupación principal en pesos. Se puede usar R, Python, Excel, InfoStat o SPSS.

a) Hacé el diagrama de dispersión del ingreso mediano (eje vertical) contra el porcentaje de asalariados no registrados (eje horizontal). Calculá el coeficiente de correlación y describí la relación.

RESOLUCIÓN



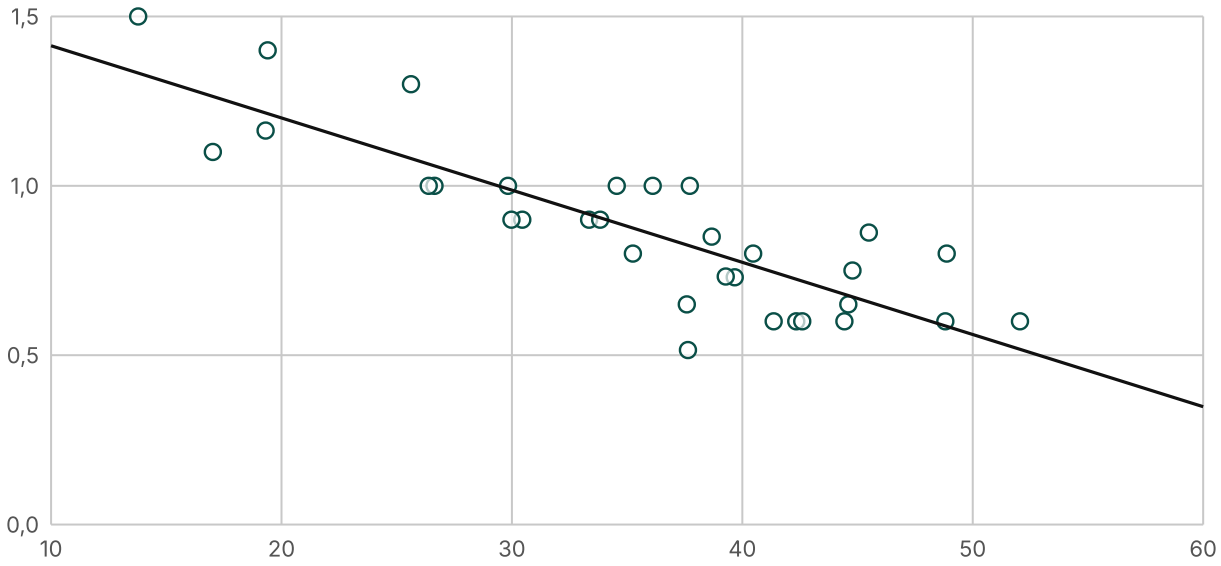
Eje horizontal: asalariados no registrados (%). Eje vertical: ingreso mediano, en millones de pesos.

r = −0,844: una relación negativa y fuerte. Los aglomerados con más informalidad tienen ingresos más bajos. En R: cor(d$no_registrado, d$ingreso_mediano); en Excel: COEF.DE.CORREL.

b) Calculá la recta de mínimos cuadrados y el R². Interpretá la pendiente en pesos.

RESOLUCIÓN

$\hat{y} = 1.626.616 - 21.311\,x \qquad R^2 = 0{,}713$



Eje horizontal: asalariados no registrados (%). Eje vertical: ingreso mediano, en millones de pesos. La línea es la recta de mínimos cuadrados.

Cada punto porcentual más de asalariados no registrados se asocia, en promedio, con $ 21.311 menos de ingreso mediano. La informalidad explica el 71,3 % de la variación del ingreso mediano entre aglomerados. En R: lm(ingreso_mediano ~ no_registrado, data = d).

c) Gran Catamarca es el aglomerado 22. ¿Qué ingreso mediano predice la recta para su nivel de informalidad? ¿Está por encima o por debajo de lo esperado?

> RESOLUCIÓN
>
> $\hat{y} = 1.626.616 - 21.311 \cdot 39{,}67 = 781.221$
>
> Con 39,67 % de asalariados no registrados, la recta predice $ 781.221. El ingreso mediano observado es $ 730.000: está por debajo de lo esperado, con un residuo de $ −51.221. El residuo es lo que la recta no explica: la diferencia entre lo observado y lo predicho.

d) Calculá también la correlación entre secundario_mas y desocupacion, y entre edad_media y prop_65_mas. ¿Qué dice cada una? ¿Alguna de las tres correlaciones dice algo sobre las personas?

> RESOLUCIÓN
>
> Educación y desocupación: r = −0,202, una relación débil: entre aglomerados, saber el nivel educativo casi no ayuda a anticipar la desocupación. Edad media y porcentaje de 65 y más: r = 0,961, casi perfecta, pero no es un hallazgo: las dos variables miden casi lo mismo.
>
> Ninguna habla de personas: la unidad de análisis es el aglomerado. Que los aglomerados con más informalidad tengan ingresos más bajos no prueba que cada trabajador informal gane menos que uno registrado (aunque sea probable). Sacar conclusiones individuales de correlaciones entre grupos es la falacia ecológica.

EJERCICIO 5 Varones, mujeres y trabajo en Gran Catamarca

Con software

Con la base de la EPH de Gran Catamarca (estadistica.ar/datos/eph-catamarca-2026t1), quedate con las personas de 10 años o más con entrevista: ESTADO igual a 1 (ocupado), 2 (desocupado) o 3 (inactivo). Interesa la relación entre CH04, el sexo, y ESTADO.

a) Armá la tabla de contingencia de sexo por condición de actividad, con frecuencias absolutas y totales.

> RESOLUCIÓN

Sexo	Ocupado	Desocupado	Inactivo	Total
Varón	347	17	209	573
Mujer	284	9	336	629
Total	**631**	**26**	**545**	**1202**

> En R: table(d$CH04, d$ESTADO); en Excel, una tabla dinámica con CH04 en filas y ESTADO en columnas.

b) Calculá los porcentajes por fila. ¿Qué diferencia hay entre varones y mujeres?

RESOLUCIÓN

Sexo	Ocupado	Desocupado	Inactivo	Total
Varón	60,6	3,0	36,5	100,0
Mujer	45,2	1,4	53,4	100,0

La diferencia grande está en la inactividad: hay muchas más mujeres inactivas que varones, y por eso menos ocupadas. La inactividad incluye a quienes estudian, se jubilaron o hacen trabajo doméstico no pago, que la EPH no cuenta como trabajo.

c) La tasa de desocupación es el porcentaje de desocupados entre los activos (ocupados más desocupados), no entre todos. Calculala para cada sexo. ¿Por qué no se divide por el total de la fila?

RESOLUCIÓN

$$\text{Varones: } \frac{17}{347+17} = 4{,}7\,\% \qquad \text{Mujeres: } \frac{9}{284+9} = 3{,}1\,\%$$

Porque la desocupación mide qué parte de los que quieren trabajar no consigue. Los inactivos no buscan trabajo: si se los pusiera en el denominador, la tasa de las mujeres saldría artificialmente baja, solo porque hay más inactivas. Las tasas de actividad son 63,5 % para varones y 46,6 % para mujeres. Con tan pocos desocupados en la muestra, la diferencia entre las dos tasas de desocupación es poco estable.

d) Un informe dice: «el 62 % de los inactivos son mujeres, así que las mujeres trabajan menos». ¿Qué porcentaje calculó el informe? ¿Es el que corresponde para esa conclusión?

RESOLUCIÓN

Calculó por columna: de los inactivos, 61,7 % son mujeres. Ese porcentaje describe cómo se compone el grupo de inactivos, y depende de cuántas mujeres hay en la población. Para comparar a varones y mujeres hay que calcular por fila, dentro de cada sexo: es la tabla de la consigna b. La conclusión puede ser cierta, pero el número que la respalda es el otro. Y «trabajan menos» no es lo que mide la tabla: el trabajo doméstico no pago no figura como ocupación.

EJERCICIO 6 ¿Una cosa causa la otra?

Para discutir

En cada caso hay una correlación real. Decidí si alcanza para afirmar que una variable causa la otra y, si no, qué otra explicación hay.

a) En los meses en que se venden más helados hay más personas ahogadas en ríos y piletas.

RESOLUCIÓN

Una tercera variable, el calor, aumenta las dos cosas: en verano se toman más helados y se nada más. Es una variable de confusión. Prohibir los helados no salvaría a nadie.

b) En los incendios a los que acuden más bomberos, los daños son mayores.

> RESOLUCIÓN
>
> La causalidad va al revés, o mejor, las dos dependen del tamaño del incendio: a un incendio grande se mandan más bomberos, y un incendio grande hace más daño. Los bomberos no causan los daños.

c) Los estudiantes que desayunan todos los días tienen mejores notas que los que no.

> RESOLUCIÓN
>
> Puede haber un efecto del desayuno, pero también confusión: los hogares donde se desayuna todos los días suelen tener otras condiciones —horarios ordenados, más recursos, más acompañamiento— que también influyen en las notas. Para saberlo haría falta un experimento, o al menos comparar estudiantes parecidos en esas otras condiciones.

d) En un ensayo clínico, se sorteó qué pacientes recibían un medicamento y cuáles un placebo. Los que recibieron el medicamento mejoraron más.

> RESOLUCIÓN
>
> Acá sí se puede hablar de causa. Como el sorteo decidió quién recibía el medicamento, los dos grupos son parecidos en todo lo demás, medido o no; la única diferencia sistemática es el tratamiento. Por eso los experimentos aleatorizados son la manera más firme de establecer causalidad.

e) Con los cuatro casos a la vista: ¿cuándo una correlación permite hablar de causa?

> RESOLUCIÓN
>
> Una correlación sola nunca alcanza. Antes de hablar de causa hay que descartar la confusión (una tercera variable que mueve a las dos), la causalidad inversa y el azar. El camino más firme es el experimento con asignación al azar; sin él, se necesita controlar las variables de confusión conocidas, que la relación tenga sentido y que se repita en otros estudios. Ni así queda probado del todo.

---

estadistica.ar/docentes/trabajos-practicos/semana-4 · Las lecturas de cada tema y la hoja de fórmulas están en el sitio.