

Estadística bivariada

Versión PDF

II-1120 Estadística para Ingeniería Industrial I

Steven García Goñi

steven.garciagoni@ucr.ac.cr

3 de agosto de 2026



Agenda

- Preguntas generadoras
- Introducción
- Causalidad
- Correlación
- Algunas medidas de asociación entre dos variables



Preguntas generadoras

- ¿Qué es causalidad?
- ¿Qué es correlación?
- ¿Cómo se relaciona y se diferencia la correlación de la causalidad?
- ¿Qué es covarianza?
- ¿Cuáles medidas de asociación entre dos variables existen?



Introducción

- Al inicio del curso hablamos de por qué se debe estudiar estadística.
- ¿Lo recuerdan?
- En esa misma clase introdujimos un término que no volvimos a usar: causalidad
- Repasemos...



Recordemos ...

- La estadística es una **herramienta poderosa**, pero su verdadero valor se obtiene cuando se aplica de **manera adecuada** y se interpreta dentro de un **contexto específico**.
- Sin contexto, los datos y los análisis estadísticos carecen de significado y utilidad práctica.

It's easy to lie with statistics, but it's hard to tell the truth without them - *Andrejs Dunkels*



Teoría, hipótesis y contexto

- Para Babbie (2000), una teoría es una explicación sistemática de los hechos y leyes observadas que se relacionan con un aspecto específico de la vida.
- Organizan las observaciones y les asignan un sentido. A menudo, las expectativas comprenden la idea de **causalidad** (Dado un evento A -> Ocorre un evento B).

Las teorías no deben confundirse con una opinión



Ejemplos

General



- Mediante una encuesta, se obtiene que el 90 % de las personas con baja educación son prejuiciosas y que el 30 % de las que tienen mayor educación también lo son. Es decir, que el 70 % no tiene prejuicios.
- La hipótesis sería que “algo” en la educación es la **causa** de que la persona sea prejuiciosa o no. La determinación de ese “algo” es la teoría asociada al contexto.

Específico



- En una fábrica de ensamble de relojes, se hipotetiza que dividir una tarea compleja en partes más pequeñas y asignar cada parte a una persona o equipo especializado aumenta la productividad.
- Hay teorías que sostienen esta afirmación y que pueden estudiarse/comprobarse con el uso de la estadística, mediante un experimento u observación, por ejemplo.



¿Por qué estudiar causalidad?



Causalidad

- La causalidad se estudia porque debemos darle sentido a los datos para guiar las acciones y políticas y aprender de los éxitos y los fracasos.
- Por ejemplo
 - El efecto de fumar -> Sobre el cáncer de pulmón
 - El efecto de la educación -> Sobre los salarios
 - El efecto de las emisiones de carbono -> Sobre el clima



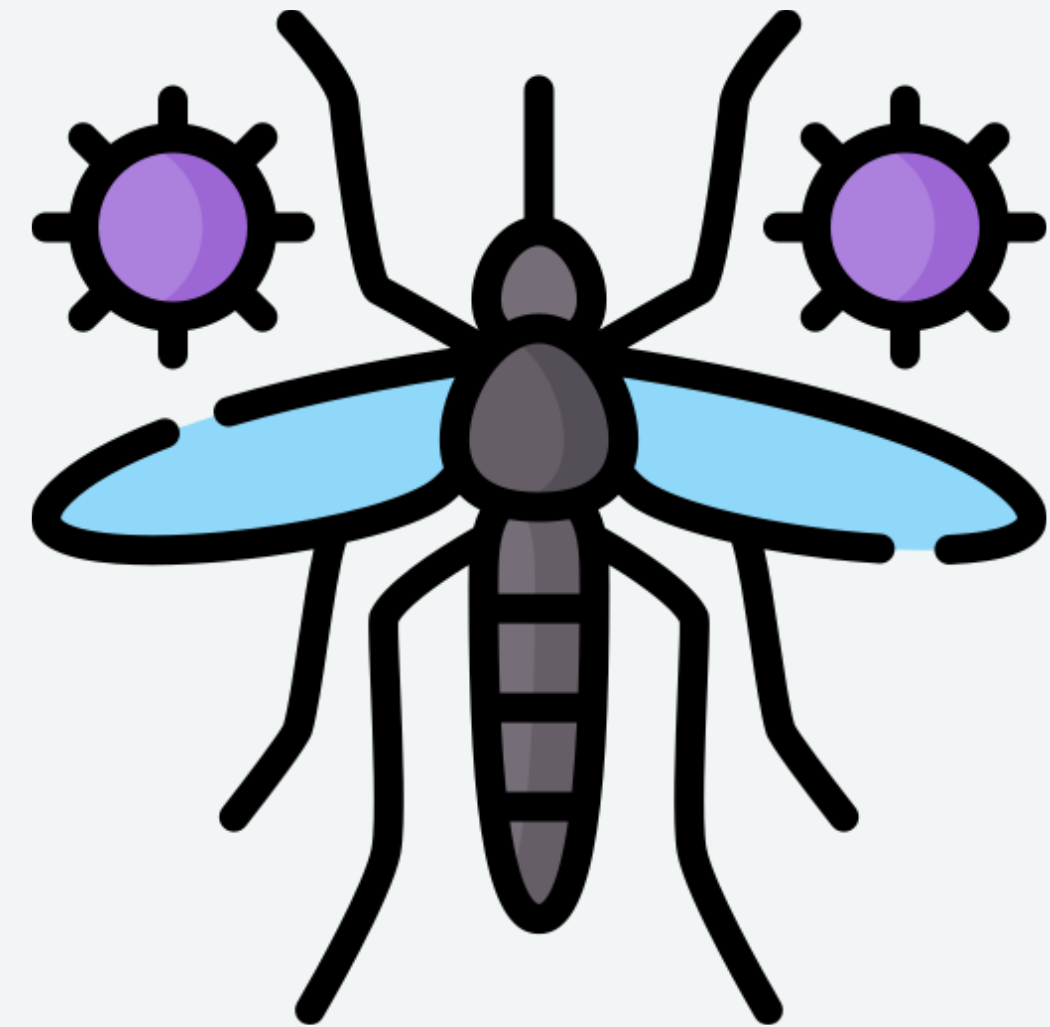
Causalidad

- De forma más ambiciosa, con estadística se busca entender cómo y por qué las causas influyen en los efectos.
- Por ejemplo, las creencias respecto al dengue de que los mosquitos solo pican de día o que el frío mata al mosquito pueden conducir a decisiones erróneas sobre los efectos.



Causalidad

- El nombre “malaria” proviene del italiano “mal aria”, que significa “mal aire”. Esta denominación se debe a la antigua creencia de que la enfermedad era transmitida por el aire viciado de los pantanos.
- Saber que la causa son los mosquitos y no el aire es el motivo por el cual se usan mosquiteros y no mascarillas al visitar este tipo de zonas.



Recordatorio

- En este curso, hemos trabajado múltiples veces con la idea de causalidad.
- Por ejemplo, en estadística descriptiva establecíamos relaciones como:
 - Si una máquina es más vieja, tendrá más fallos (causa -> efecto)
- En probabilidad:
 - El teorema de Bayes puede interpretarse dentro de algunos modelos causales. $P(A|B)$ es la probabilidad de A (efecto) dado B (causa). Nota: Bayes no implica necesariamente causalidad, solo se pretende ilustrar un ejemplo.
- Como pueden ver, es tema conocido, en esta clase se formaliza el concepto.



Causalidad



Formalmente

- La causalidad se refiere a una relación de causa y efecto entre dos variables, donde un cambio en una variable (la causa) provoca un cambio en otra (el efecto).
- Para que exista causalidad:
 - Debe existir consistencia, es decir, siempre se repite
 - No hay variables confusoras (espurias) que puedan intervenir.
 - Debe observarse asociación entre las variables.
 - Toda causa precede al efecto (causa -> efecto)
 - Coherencia



Confusores

- Una **variable confusora** (o confounding variable) es una **tercera variable** que distorsiona la relación entre dos variables que se están estudiando.
- Puede hacer parecer que hay una relación causal entre ellas cuando en realidad esa relación es **espuria o exagerada**.



Correlación



Correlación

- En ciencia e ingeniería a menudo las personas recolectan datos para determinar la relación que existe entre dos variables.
- **Correlación**: Indica que dos variables están relacionadas (se mueven juntas), pero no implica necesariamente que una cause a la otra.
- **Causalidad**: Implica una dirección clara del efecto. Si A causa B, entonces al intervenir en A, esperaríamos un cambio en B.

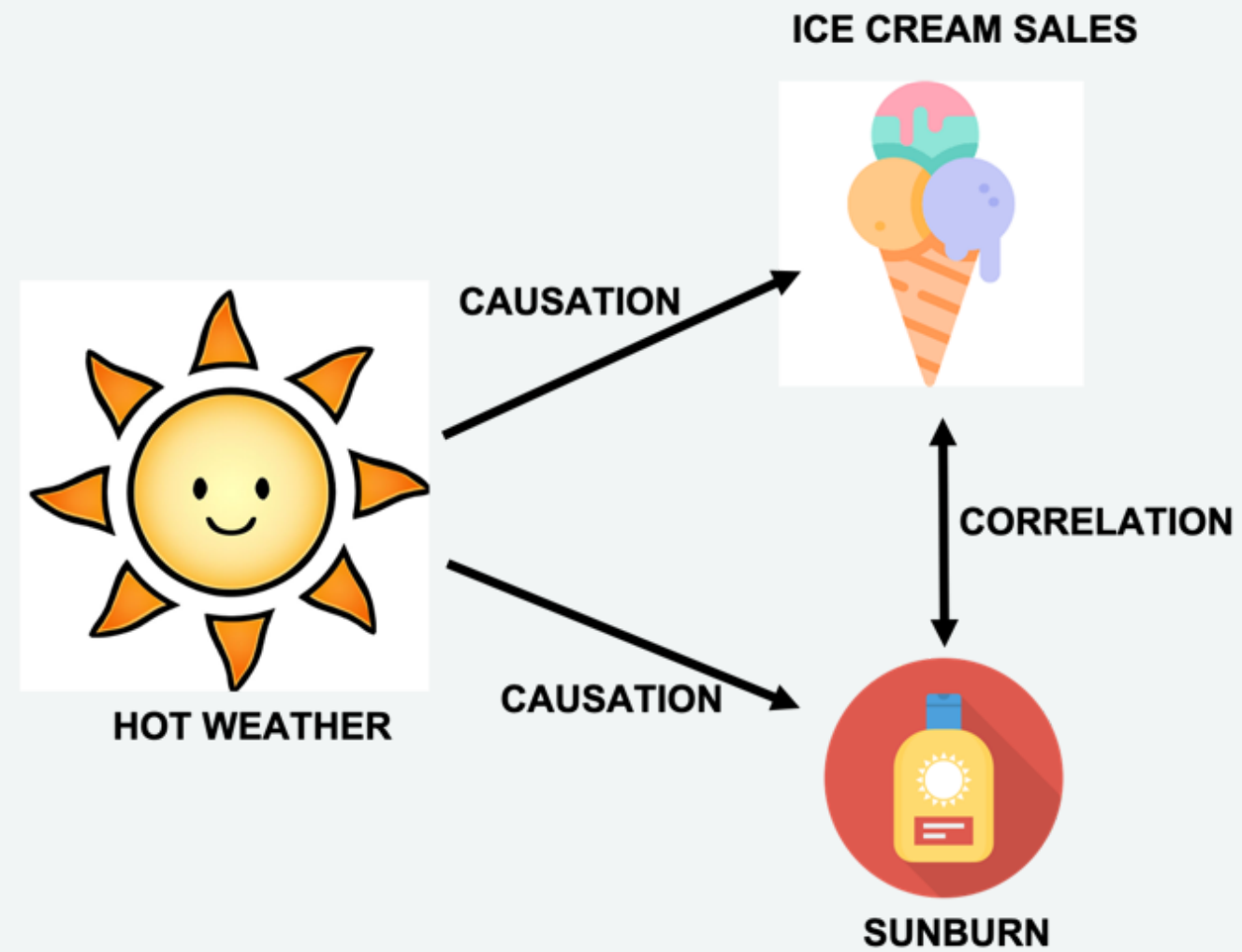


Correlación vs Causalidad

- Se pueden presentar relaciones espurias entre variables.
- **Relación espuria:** relación matemática de dos acontecimientos que NO tienen conexión lógica, aunque puede implicar que la tienen debido a un tercer factor no considerado aún.



Sesgo



Sesgo

- Construya un diagrama equivalente para esta situación: La habilidad lectora de un infante está correlacionada con la talla del zapato. ¿Cuál es la variable confusora?



¿Recuerda el concepto de sesgo?

- Los ejemplos anteriores se conocen como **sesgo de confusión**.
 - Donde un cambio producido en una variable (clima caliente) afecta, por separado, a otras dos variables (venta de helados y quemarse la piel), causando la **falsa** sensación de que un cambio en la venta de helados causa un cambio en las quemaduras solares (o viceversa)

Existen muchos tipos de sesgo

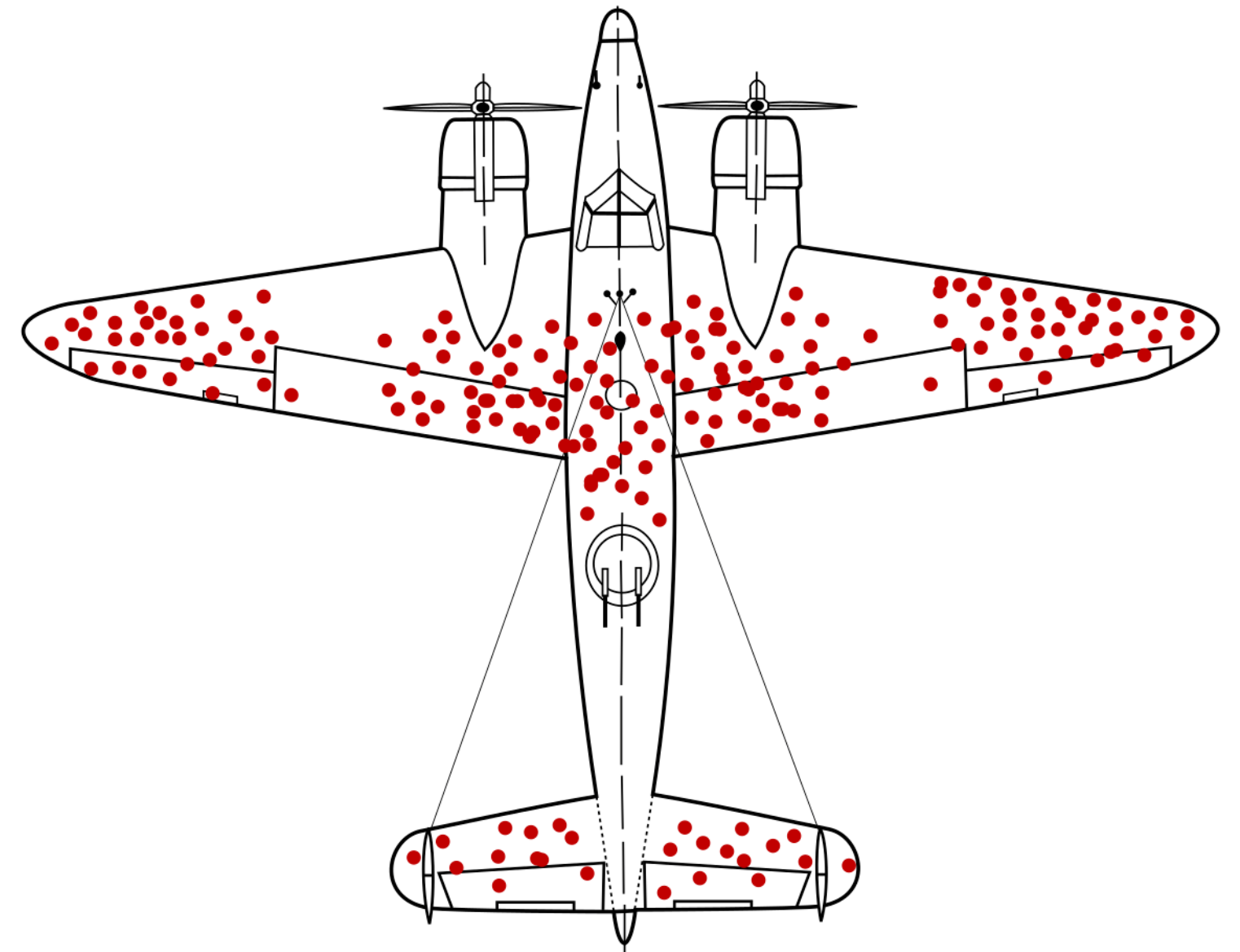
- De confusión
- Del colisionador
- De selección
- De medición
- De información
- De supervivencia
- Etc



Por ejemplo

- El sesgo del superviviente es uno muy famoso. Quizá algún día haya visto esta imagen.
- Es una falacia lógica que consiste en centrarse en aquellos que han logrado sobrevivir a un proceso, ignorando a aquellos que no lo lograron.
- En esa imagen, las partes dañadas de los aviones que regresan de la batalla muestran los lugares en los que pueden sufrir daños y aun así volver a casa; los que son alcanzados en los lugares no señalados no sobreviven. Centrarse en reforzar las partes dañadas de los aviones supervivientes es un sesgo de supervivencia.

Imagen de [Wikipedia](#)



Otro ejemplo

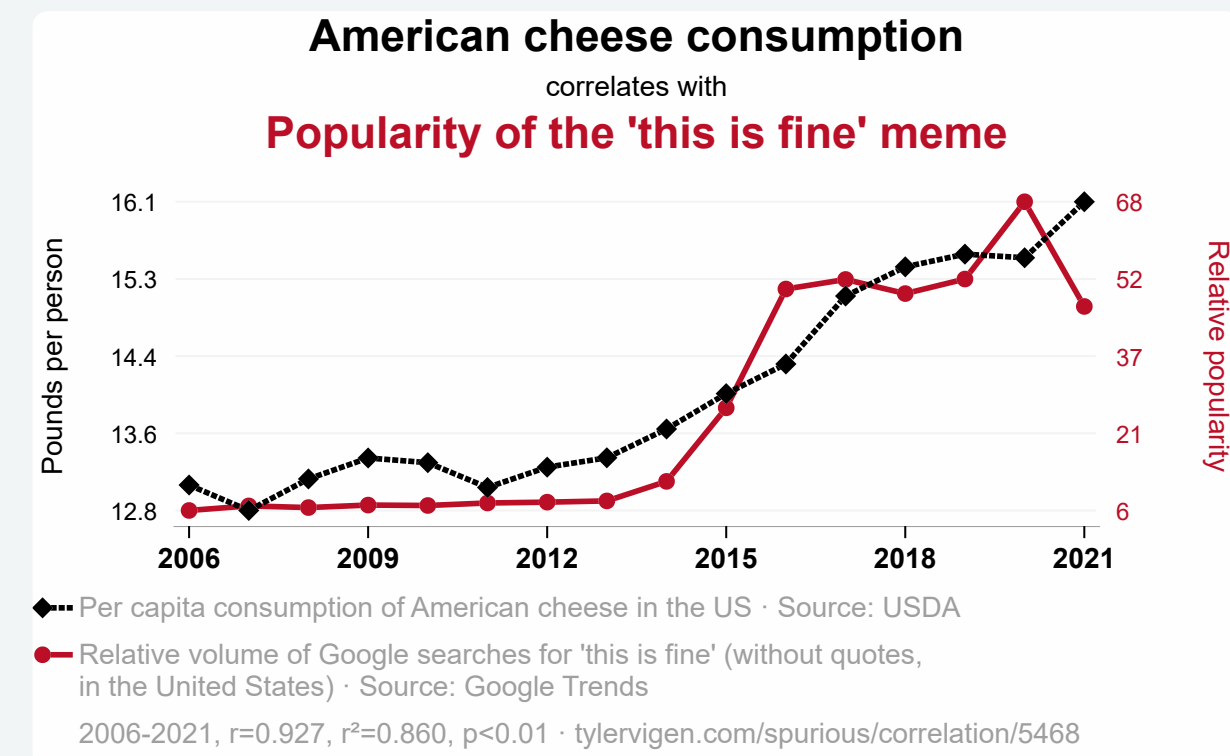


- Tomado de [@picanumeros](#)
- En 2020 se llevó a cabo un estudio para medir la evolución de la salud mental de la población estadounidense durante la pandemia de COVID-19.
- Dicho estudio se basaba en una encuesta longitudinal con cuatro oleadas (en abril, mayo, junio y septiembre de ese año).
- Sin embargo, un análisis de los datos en bruto nos daría un resultado sorprendente: que la salud mental mejoró durante el transcurso del estudio, a pesar de seguir en lo más duro de la pandemia.



¿Por qué volvemos a hablar de sesgo?

- Porque en estadística bivariada, asociar dos variables sin suficiente fundamento teórico nos puede llevar a conclusiones erradas.
- Observe el siguiente ejemplo:



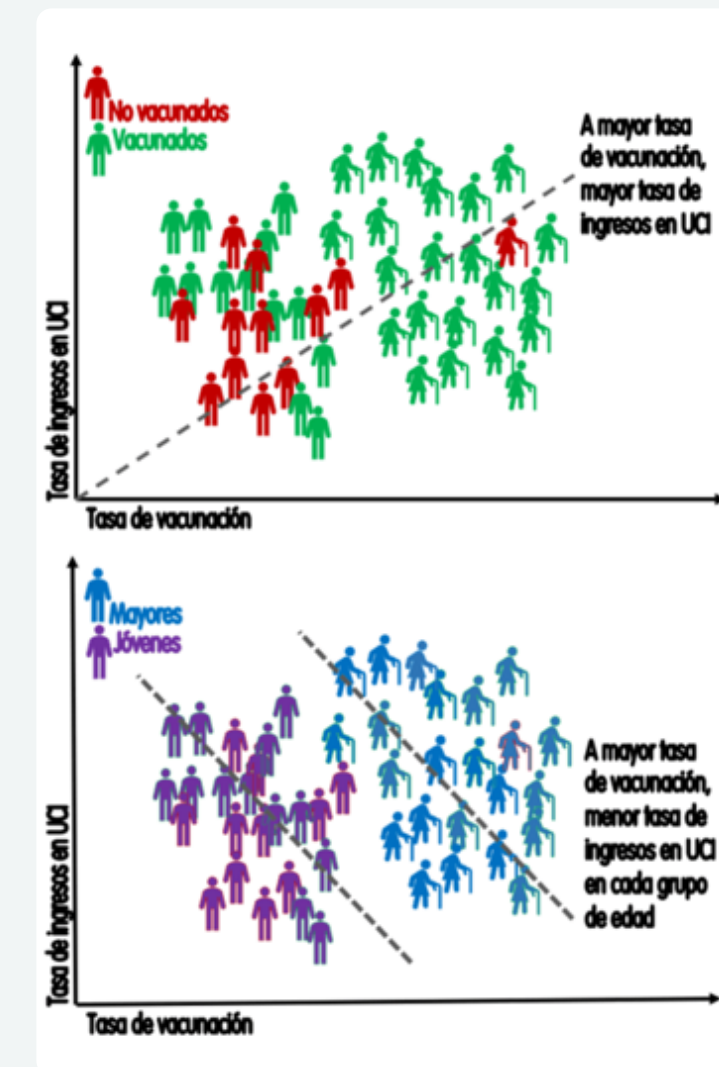
Otros problemas

- Además del sesgo, hay que cuidar la forma en la que analizan los datos
 - Agregados o desagregados, pues esta forma de organización puede introducir variables confusoras.
- Esto puede llevar a la paradoja de Simpson
 - La asociación que se sostiene en una población puede ser inversa a la que se obtiene en cada subgrupo de la población.



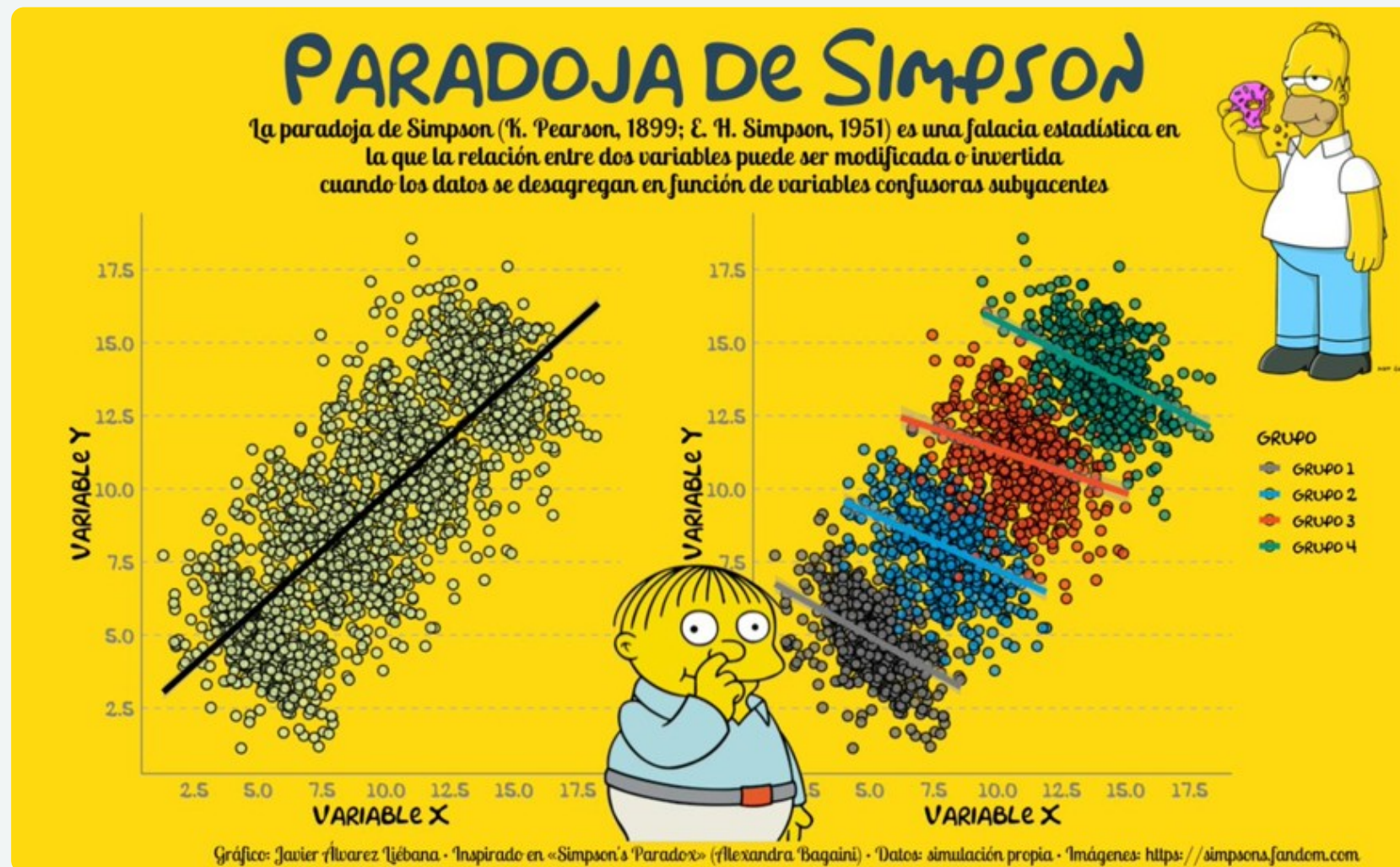
Paradoja de Simpson

- La paradoja de Simpson sucede cuando al analizar los datos desagregados por grupos resulta que las conclusiones estadísticas obtenidas entre las variables son **opuestas**.
- En esta falacia estadística cayeron muchas personas en la época de pandemia por COVID – 19. Sobre todo por parte de negacionistas.
 - Independientemente de si estamos o no de acuerdo, observemos el comportamiento de la paradoja a la derecha.
- Hay que tener cuidado con los análisis estadísticos cuando hay en juego variables ocultas.

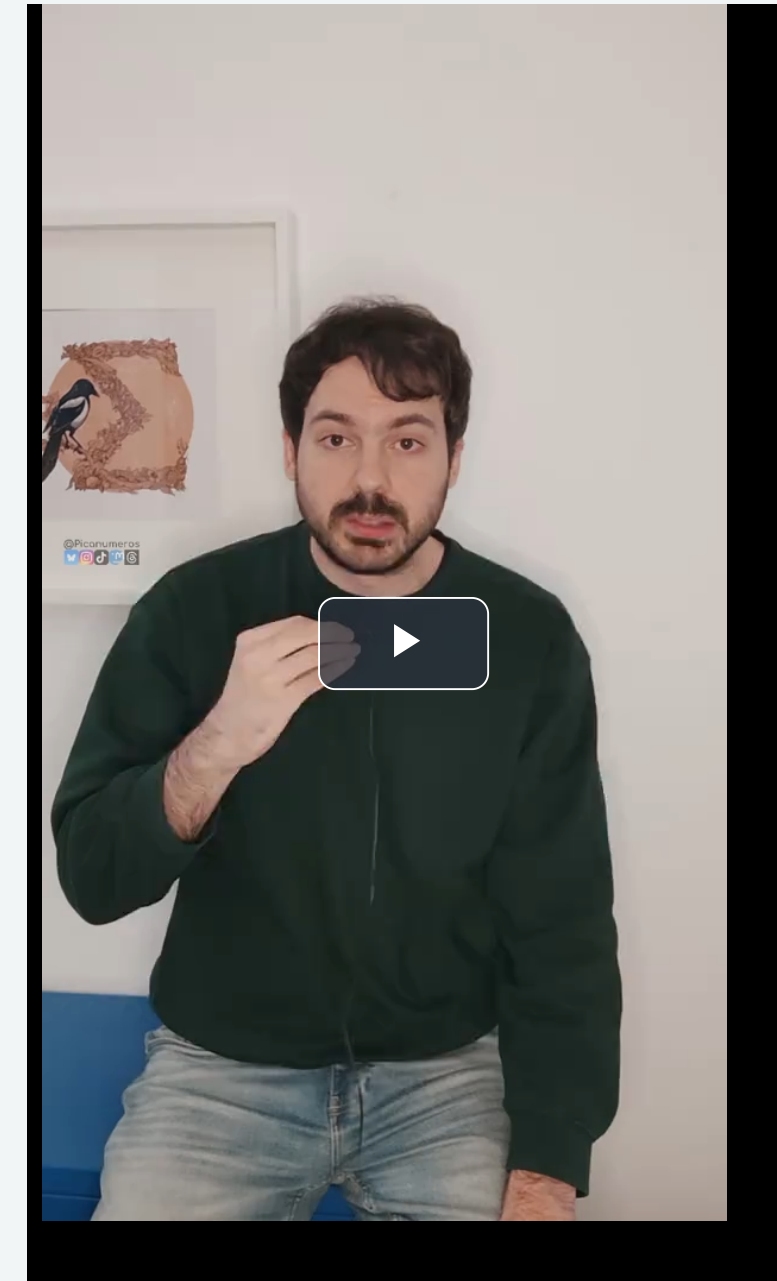


Paradoja de Simpson

Imagen de [Javier Álvarez Liébana](#)



Tomado de [@picanumeros](#)



Paradoja de Simpson

- Las conclusiones a las que se puede llegar son diferentes.

Imagen tomada del libro *Causal Inference in Statistics: A primer*

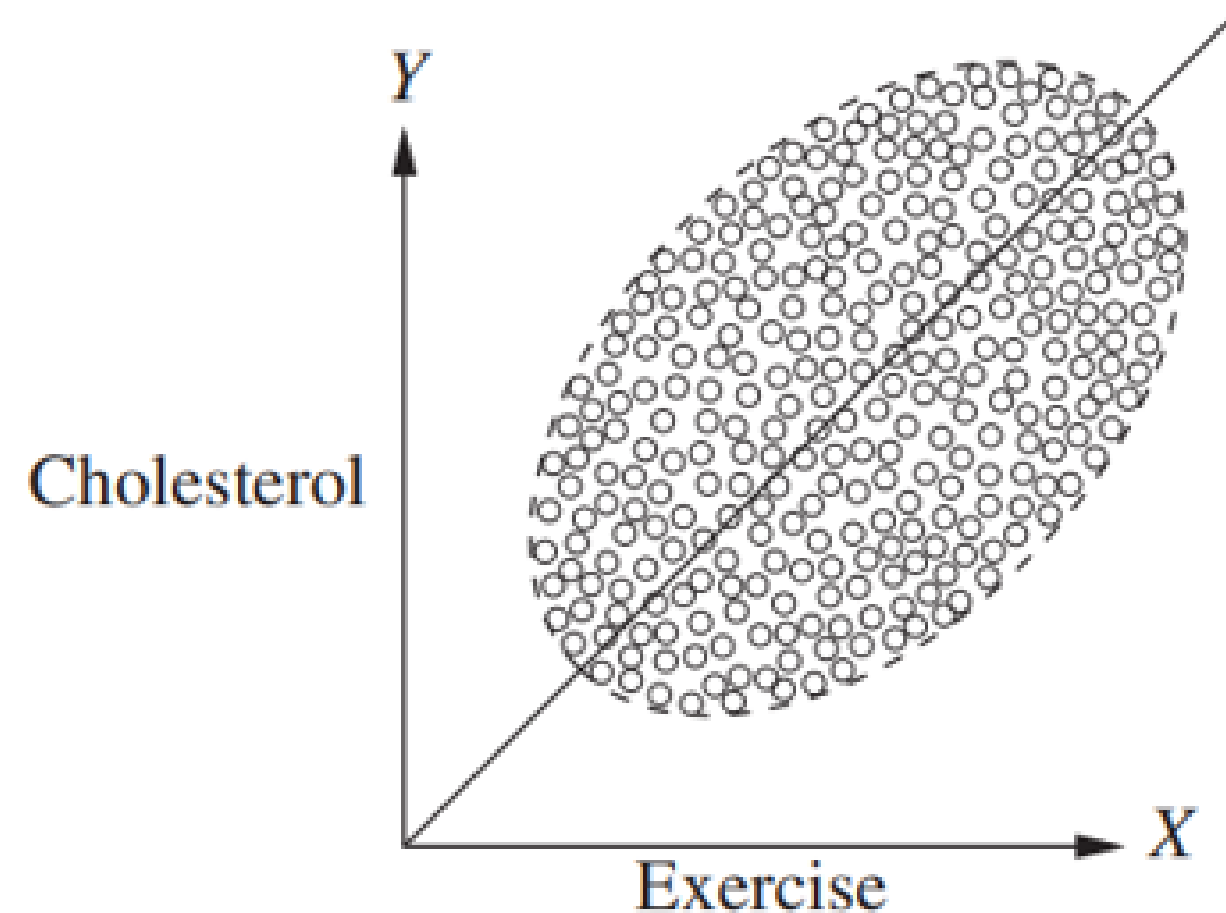
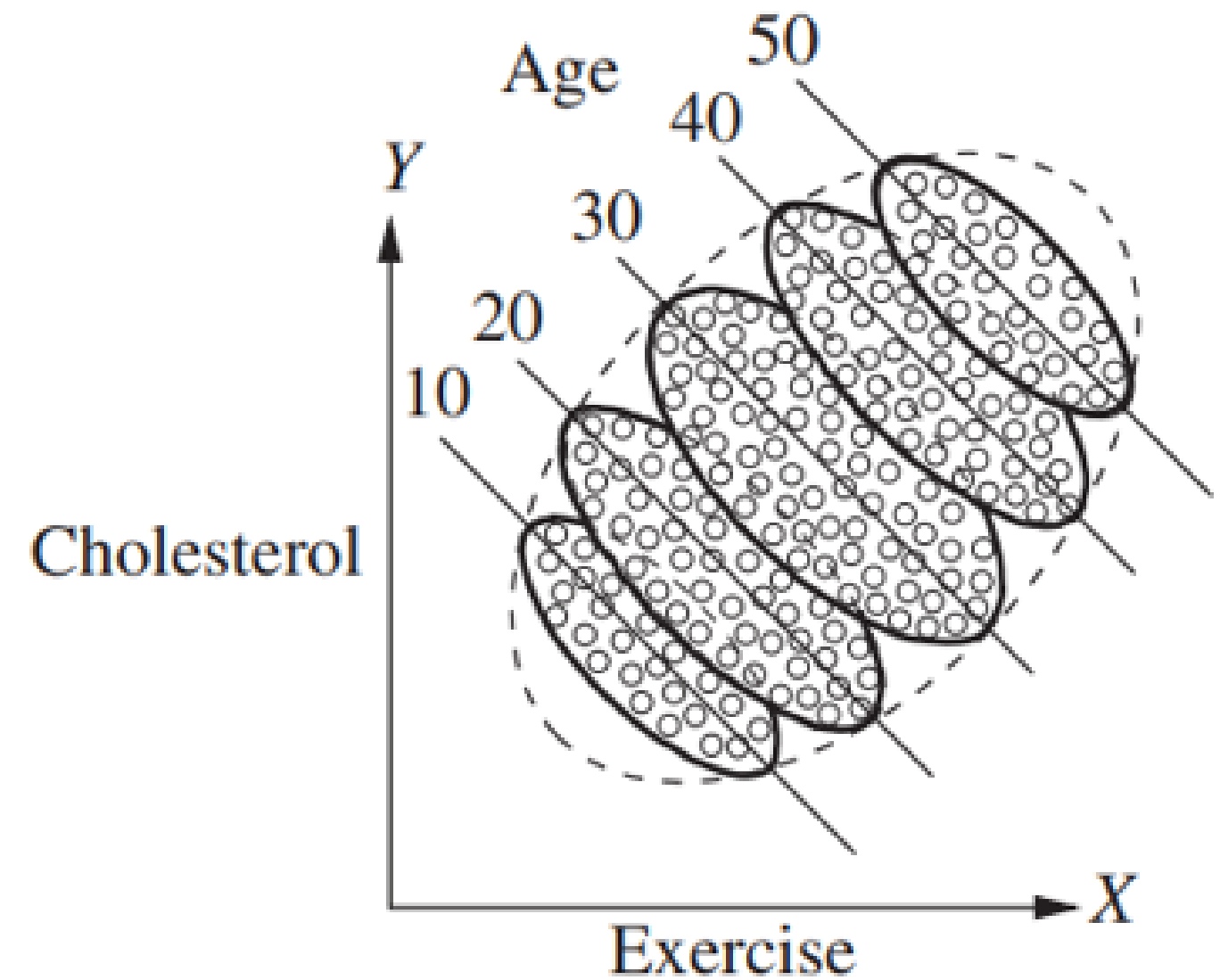


Imagen tomada del libro *Causal Inference in Statistics: A primer*



Medidas de asociación entre dos variables



Repaso

- Los datos se clasifican por niveles de medición. El nivel de medición y el tipo de variable **determinan los cálculos que se llevan a cabo** con el fin de resumir y presentar los datos.
- Es decir, que no todos los tipos de datos se analizan igual.

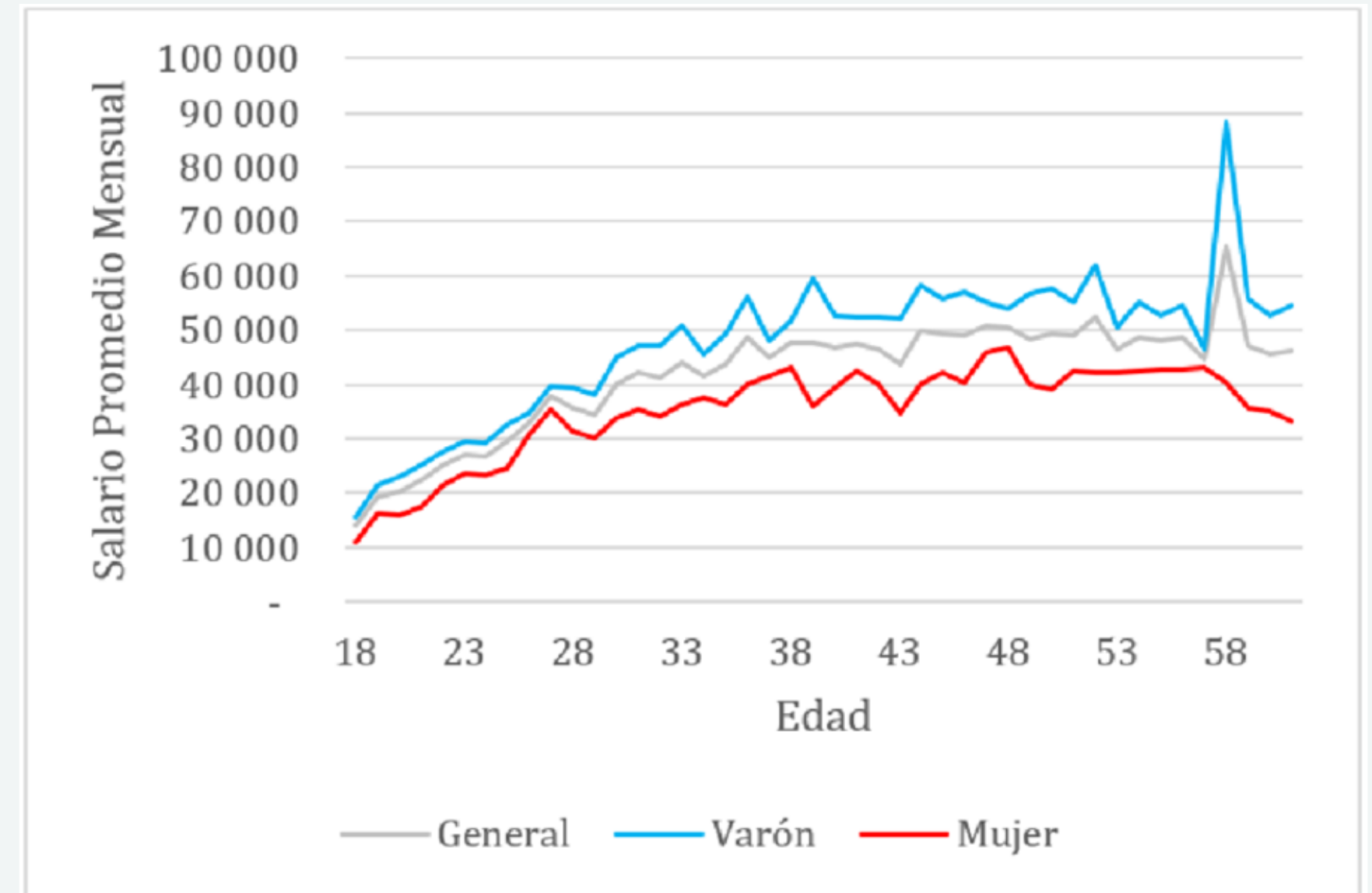
Niveles de medición

- Nominal
- Ordinal
- Intervalo
- Razón



Repaso

- En visualización de datos, estudiamos el gráfico de dispersión, ese en el que puede observar la relación entre dos variables al graficar una variable en el eje X y otra en el eje Y.
- Al realizar este tipo de análisis conviene en primer lugar **graficar los datos**.
- Por ejemplo, en el gráfico de la derecha puede encontrar la relación que existe entre el salario promedio mensual y la edad de las personas, clasificadas por género y en general.



Idea central

Toda causalidad implica asociación, pero no toda asociación implica causalidad.



Dos variables continuas



Covarianza

- La **covarianza** mide cómo **varían conjuntamente** dos variables cuantitativas. Indica si al aumentar una, la otra tiende a aumentar o disminuir.
- Como la covarianza NO está normalizada, se dificulta su interpretación, por eso se prefiere la **correlación**.
- No obstante, esta tiene múltiples aplicaciones y técnicas como el AFE (que se estudia en psicometría) prefieren el uso de la matriz varianza-covarianza sobre las correlaciones.

Poblacional

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

Muestral

$$S_{XY} = \frac{1}{n - 1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$



Interpretación

- $Cov(X, Y) > 0$: tienden a aumentar o disminuir juntos linealmente (X y Y)
- $Cov(X, Y) < 0$: si una aumenta, la otra disminuye, linealmente (X y Y)
- $Cov(X, Y) = 0$: no hay relación lineal aparente (X y Y)
- Recuerde que, como no está normalizada, no es sencillo decir si la relación es fuerte o débil.



Ejemplo 01

- Resuelva. Utilice los espacios en blanco para completar el ejercicio o como guía. A la derecha las respuestas.

i	x_i	y_i	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x}) \cdot (y - \bar{y})$
1	2	4			
2	4	2			
3	6	4			
4	8	6			

- Además, interprete los resultados obtenidos.

Los promedios son:

$$\bar{x} = 5$$

$$\bar{y} = 4$$

Y por tanto la covarianza es:

$$S_{XY} = \frac{0 + 2 + 0 + 6}{4 - 1} = 2.67$$



Correlación de Pearson (r)

- El coeficiente de correlación de Pearson, denotado como r , es una medida estadística que cuantifica **la fuerza y dirección de la relación lineal** entre dos variables **cuantitativas**.

$$r = \frac{Cov(X, Y)}{S_X \cdot S_Y}$$

Su valor está siempre entre -1 y 1

- $r = 1$: correlación lineal **positiva perfecta**
- $r = -1$: correlación lineal **negativa perfecta**
- $r = 0$: no hay relación lineal
 - Esto no implica ausencia de relación

Para aplicar la correlación de Pearson se asume un comportamiento **normal** entre ambas variables. Es importante no perder de vista este requisito, aún cuando en este curso no se aborda formalmente la distribución normal.



Ejemplo 02

- Calcule e **interprete** la correlación de Pearson a partir de los resultados del Ejemplo 01.
- $Cov(X, Y) = 2.67$
- $S_X = 2.58$
- $S_Y = 1.63$
- $r = \frac{2.67}{2.58 \cdot 1.63} = 0.63$



Correlación de Spearman (ρ)

- Es un prueba de índole no paramétrica que NO asume la normalidad.
 - Es una prueba menos sensible que Pearson.
 - Si se cumplen las condiciones para usar Pearson y se usa Spearman estaríamos incurriendo en un error.
- Permite el cálculo tanto con variables **continuas** como **discretas**.

$$\rho = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}$$

- Se deben numerar los datos según su orden de menor a mayor para la primera (i) y segunda columna (t).
- La diferencia al cuadrado de los órdenes es D^2 .



Correlación de Spearman (ρ)

- La interpretación es la misma que para Pearson, con la diferencia de que Spearman no evalúa la relación lineal, sino la relación **monótona** entre dos variables **continuas o discretas**.
- En una relación monótona, las variables tienden a cambiar al mismo tiempo, pero no necesariamente a un ritmo constante.



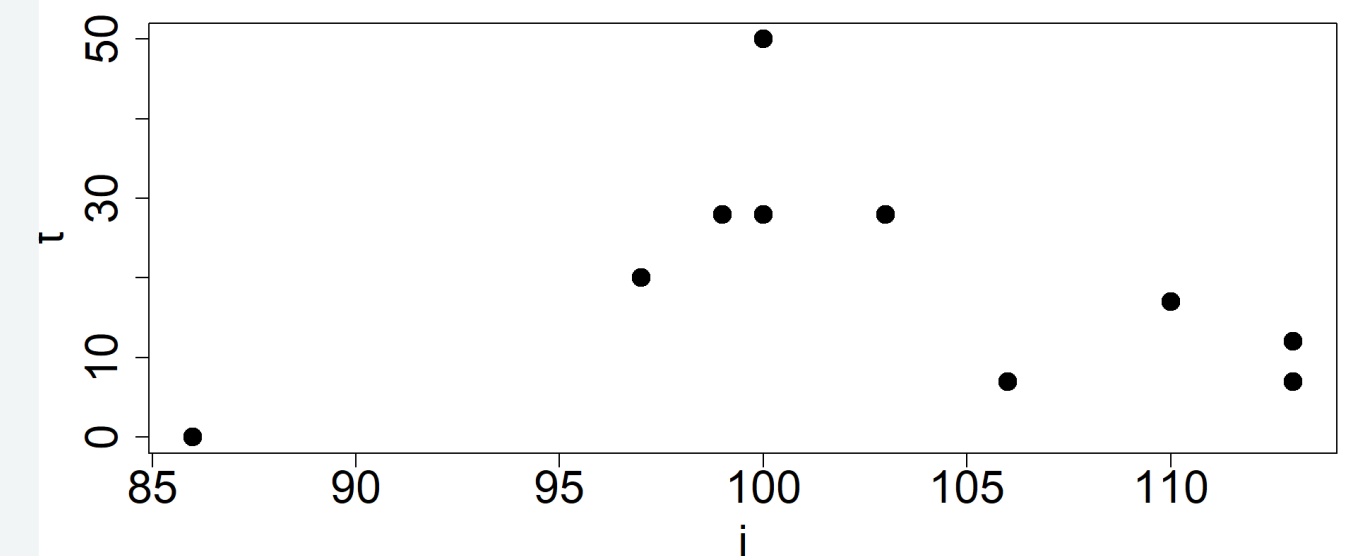
Ejemplo 03

- Calcule el coeficiente de correlación de Spearman entre el Coeficiente intelectual (CI), que se va a denotar por i y las horas de TV a la semana (t).
- Como parte de un ejercicio académico calcule también el coeficiente de correlación de Pearson, aunque no sea lo correcto.

— Compare ambos resultados

i	t
106	7
86	0
100	28
100	50
99	28
103	28
97	20
113	12
113	7
110	17

Observe el gráfico, ¿hay alguna relación lineal?



Ejemplo 03

i	t	$Orden_i$	$Orden_t$	D^2
106	7	7.0	2.5	
86	0	1.0	1.0	
100	28	4.5	8.0	
100	50	4.5	10.0	
99	28	3.0	8.0	
103	28	6.0	8.0	
97	20	2.0	6.0	
113	12	9.5	4.0	
113	7	9.5	2.5	
110	17	8.0	5.0	

Spearman (ρ)

$$\rho = 1 - \frac{6 \cdot 196}{10(10^2 - 1)} = -0.188$$

Pearson

$$r = -0.071$$



Correlación de Kendall (τ)

- Es una alternativa a Spearman, que se dice que es más robusto, pues es menos sensible a valores atípicos y se comporta mejor con muestras pequeñas.
- Tanto Kendall como Spearman se pueden usar sin mayor inconveniente con datos continuos, pero siendo consciente que esos datos continuos se transforman a rangos.

$$\tau = \frac{n_c - n_d}{\frac{1}{2} \cdot n \cdot (n - 1)}$$

Siendo n_c el número de pares concordantes y n_d el número de pares discordantes. Desde luego n es el número total de observaciones.



Ejemplo 04

- Se realiza el mismo ejercicio que en el ejemplo 03, pero se resuelve con τ de Kendall.
- Por su complejidad de cálculo, se recomienda realizar la correlación de Kendall con software estadístico.
 - Si $n = 5$ se harían 10 comparaciones, si $n = 10$ son 45 comparaciones y si $n = 20$ son 190 comparaciones.
- En el caso particular de R puede usar este código `cor(x, y, method = "kendall")`.

- $$\tau = -0.167$$



Una variable continua y una dicotómica



Punto - Biserial (r_{pb})

- Se recomienda cuando la variable dicotómica (1 y 0) es natural:
 - Hombre / Mujer
 - Vivo / Muerto
- No obstante también puede usarse cuando la variable dicotómica es generada artificialmente (pero NO se recomienda):
 - Reprobado / Aprobado (A partir de un corte como 67.5)
 - Baja / Alta productividad (A partir de un estándar como 90 %)
- Es similar a la correlación de Pearson, pero se adapta a este tipo específico de variables.

$$r_{pb} = \frac{\bar{X}_1 - \bar{X}_0}{S_X} \cdot \sqrt{\frac{n_1 \cdot n_0}{n \cdot (n - 1)}}$$

- $\bar{X}_0$ y $\bar{X}_1$ es la media de X para el grupo 0 y 1 respectivamente. S_X desviación estándar de X . n_1 y n_0 es el tamaño para grupo y $n = n_0 + n_1$.

Esta correlación también se conoce como **índice de discriminación**, que es un concepto muy usado en encuestas.



Ejemplo 05

El tiempo de secado al que se expone una unión de dos dispositivos médicos se cree está relacionado con la producción de piezas buenas (0) y malas (1).

Se toman los datos que se muestran a la derecha. Calcule e interprete.

Tiempo	Pieza
2	0
3	1
16	0
17	1
5	1
6	1
17	0
7	1
15	1
8	0
9	1
10	1
11	0
12	0
13	0

Medida	Valor
X_1	9.00
X_0	11.29
S_X	4.93
n_1	8.00
n_0	7.00
n	15.00

$$r_{pb} = \frac{9 - 11.29}{4.93} \sqrt{\frac{8 \cdot 7}{15 \cdot (15 - 1)}} = -0.239$$

- A mayor tiempo de secado, menor tendencia a obtener piezas defectuosas, aunque la relación es débil.



Correlación biserial (r_b)

- Esta se recomienda cuando se tiene una variable continua y una dicotómica artificial.
 - Es decir, la dicotomía surge una variable continua dividida (aprobar o no, un curso).

$$r_b = \frac{\bar{X}_1 - \bar{X}_0}{s_x} \cdot \frac{p \cdot q}{y}$$

Donde y es un factor de corrección calculado como:

$$y = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

con z asociado al punto de corte q entre ambas categorías. La estimación de z se aborda en otros cursos.



Ejemplo 06

Suponga la variable **Pieza** del Ejemplo 05 es construida de forma artificial. Si no soporta una presión de 500 psi, es defectuosa. Resuelva el Ejemplo 05 con la correlación biserial.

Utilice esta calculadora para obtener el valor de z .

Calculadora de z

Ingrese q : **Calcular**

$z =$

Complete los valores faltantes.

Medida	Valor
X_1	9.000
X_0	11.290
S_X	4.930
p	0.533
q	0.467
z	
y	

El resultado es:

$$r_b = -0.291$$

La interpretación es equivalente a la punto-biserial.



Dos variables dicotómicas



Coeficiente Phi (Φ)

Es una medida estadística para determinar la relación que existe entre dos variables dicotómicas naturales. Se basa en tablas de contingencia.

$$\phi = \frac{a \cdot d - b \cdot c}{\sqrt{(a + b) \cdot (c + d) \cdot (a + c) \cdot (b + d)}}$$

	Y=1	Y=0
X=1	a	b
X=0	c	d



Ejemplo 07

En el proceso de envasado y etiquetado de una planta de producción de ungüentos y cremas se cree que si la etiqueta está “torcida” (1) aumenta el chance de que el marcado de la fecha de vencimiento resulte ilegible (1).

	Ilegible=1	Legible=0
Torcida=1	a=4	b=1
Recta=0	c=1	d=4

$$\phi = \frac{4 \cdot 4 - 1 \cdot 1}{\sqrt{(4 + 1) \cdot (1 + 4) \cdot (4 + 1) \cdot (1 + 4)}}$$
$$\phi = 0.6$$



Correlación tetracórica (ρ_t)

- Se utiliza cuando hay dos variables dicotómicas artificiales
- Dado que son artificiales, se asume que devienen de variables latentes continuas.
 - En la siguiente lección se definirá **variable latente**.
- No se recomienda estimarla manualmente, sino su estimación con software.



Dos variables ordinales



Correlación policórica (ρ_p)

- Se usa en presencia de dos variables ordinales, bajo el supuesto de continuidad latente.
- Es muy común cuando se requiere estudiar si las personas responden similar a dos preguntas en una encuesta:
 - Califique del 1 al 5 que tan de acuerdo está con estas afirmaciones:
 - Me siento motivado en mi trabajo actual
 - ¿Con qué frecuencia siente que su trabajo es valorado por sus superiores?



Correlación policórica (ρ_p)

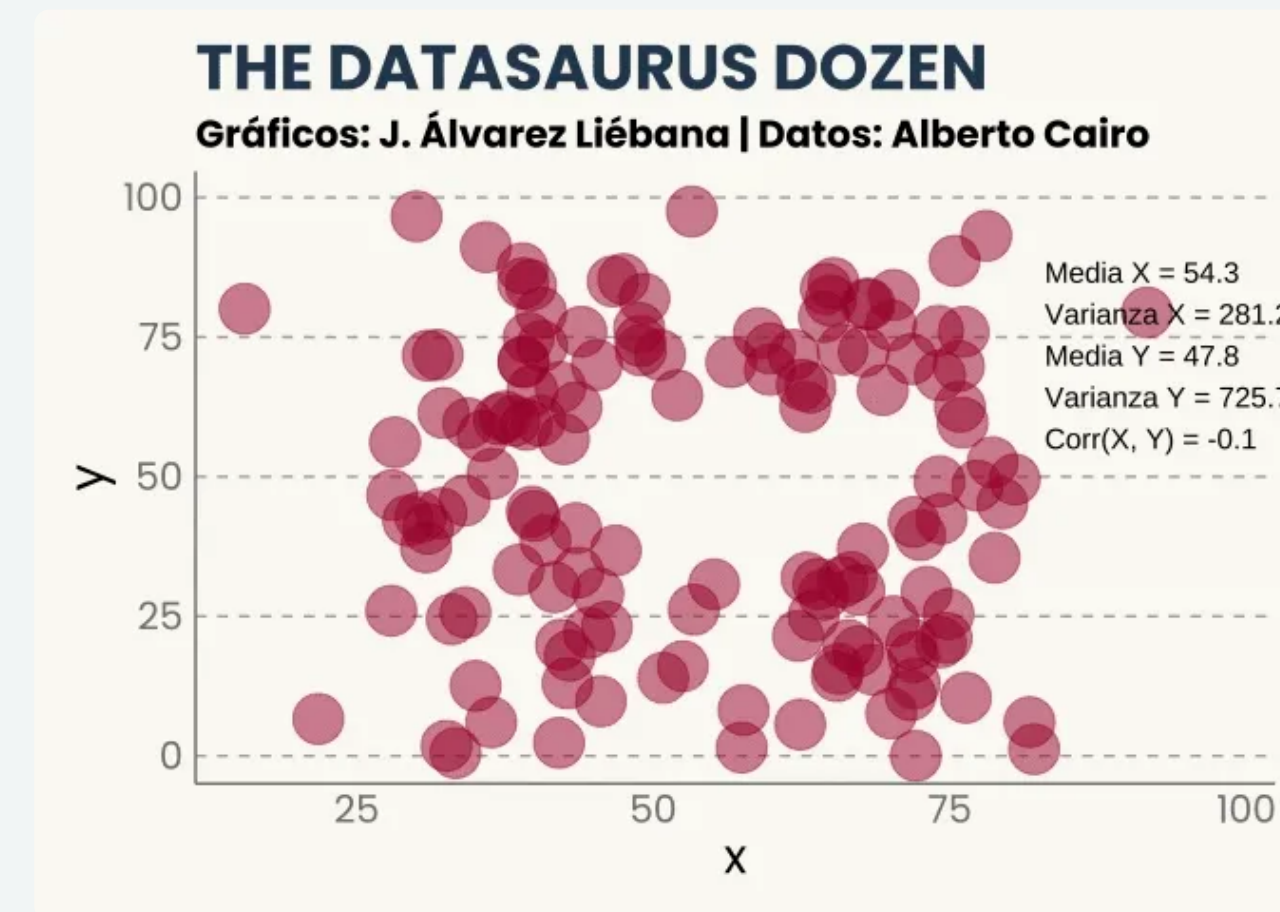
- La tetracórica es un caso especial de esta correlación, y por ende, en esta tampoco se recomienda su cálculo a mano.
- Para ambos casos se le provee de [código en R](#) para solventar estos ejercicios de ser necesario.
- Además, cuenta con **otros** ejemplos en [Excel](#) para todas las correlaciones a excepción de estas.



La importancia de la visualización

- Existe un conjunto de datos famoso, conocido como el **Datasaurus**, donde una serie de pares de puntos representando algunas formas (círculos, elipses y cruces, entre otros).
- Todos ellos muestran patrones diferentes pero las principales características numéricas son las mismas.

Tomado de [Javier Álvarez Liébana](#)



Resumen

Nombre de la correlación	Tipo de variables involucradas	Supuestos
Pearson	Continua + Continua	Linealidad, normalidad, métrica
Spearman	Ordinal + Ordinal o Continua no normal	Basada en rangos (no paramétrica)
Point-Biserial	Continua + Dicotómica natural	La dicotómica representa una categoría real
Biserial	Continua + Dicotómica artificial	Se asume variable continua subyacente, distribución normal
Phi (ϕ)	Dicotómica + Dicotómica naturales	Ambas categorías reales (no transformadas)
Tetracórica	Dicotómica + Dicotómica artificiales	Ambas variables reflejan continuas latentes normales
Policórica	Ordinal + Ordinal	Ambas reflejan variables continuas latentes, categorías ordenadas



Bibliografía

- Las fuentes consultadas para esta sesión son muy variadas y profundizan en exceso en temas que no competen enteramente a este curso. Razón por la cual se omite voluntariamente.
- Fuentes específicas fueron señaladas a lo largo de la presentación.



Estadística bivariada

II-1120 Estadística para Ingeniería Industrial I

Gracias por su atención
Steven García Goñi

steven.garciagoni@ucr.ac.cr

Dudas o correcciones requeridas pueden solicitarse al correo

