Estadística Descriptiva

Introducción a la Estadística

Fiona Franco Churruarín
fionafch96@gmail.com

UTDT

Febrero 2022

¿Qué hacer con tantos datos?

La **estadística descriptiva** permite la agrupación, resumen y presentación de datos.

- **¿Para qué?** Entender mejor qué nos dicen, encontrar relaciones, tomar decisiones.
- **¿Cómo?** Tabulados, técnicas gráficas, medidas matemáticas (media, varianza, etc.).
- **¿Por qué?** La abundancia de información complica el análisis, resumirla y ordenarla ayuda.

Estadística Descriptiva

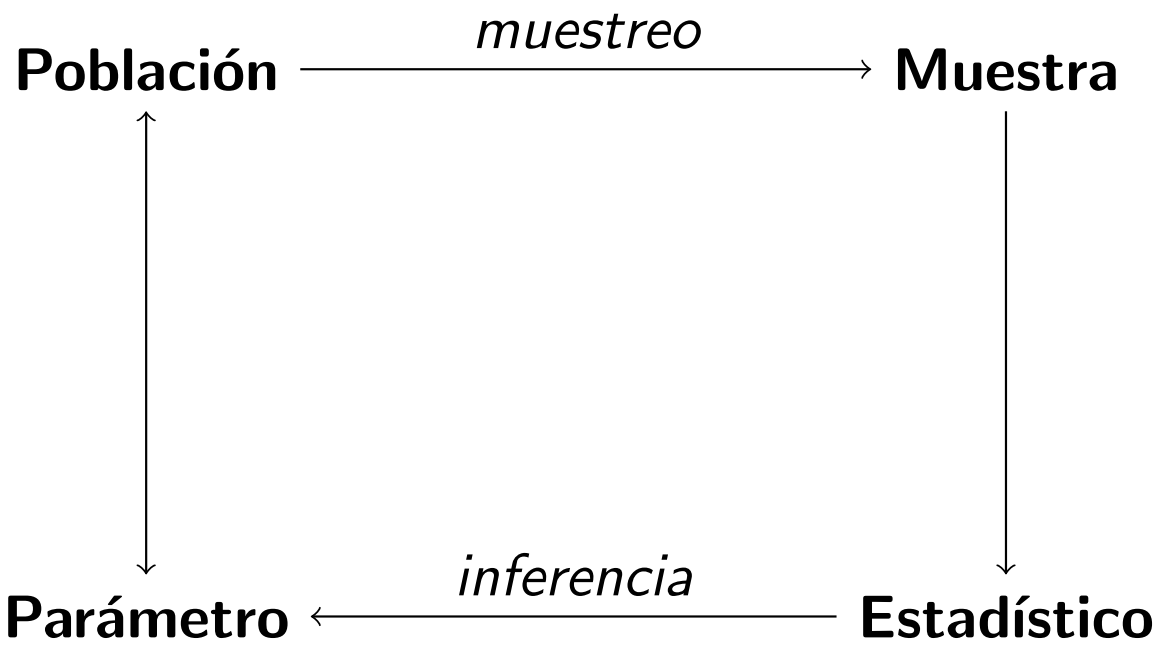
- Proceso de **recolectar**, **agrupar** y **presentar** datos de una manera tal que describa fácil y rápidamente la información.
- **Reducir** la información lo máximo posible, pero teniendo cuidado de que una reducción excesiva no nos lleve a omitir características importantes.
- Fundamental para el análisis de los datos y la toma de decisiones en distintas disciplinas.

Algunos conceptos básicos

- **Población:** o universo, es el conjunto completo de observaciones de interés para el investigador.
- **Parámetro:** es una cantidad descriptiva de la población total.
- **Muestra:** es una parte limitada (pero con suerte, representativa) o subconjunto de la población.
- **Estadístico:** elemento que describe una muestra, es función de éstos. Algunos de ellos sirven como *estimadores* de parámetros poblacionales.

Por ejemplo, se toma una muestra de 20 alumnos que cursan actualmente cierta maestría. Se quiere conocer la edad media (parámetro) de los alumnos de dicha maestría (población), para lo cual se calcula el promedio (estadístico/estimador) de las edades de los 20 alumnos de la muestra.

Población y Muestra



¿Cómo se obtienen los datos?

- **Encuestas:** se elige un subgrupo (muestra) de la población de interés. Existen muchas técnicas de muestreo (veremos algunas brevemente en el curso).
- **Censos:** se recolecta la información de toda la población en cierto momento del tiempo.
- **Datos administrativos:** bases de datos de compañías, organismos nacionales o internacionales, etc.
- **Estudios experimentales:** se obtienen los datos de un experimento buscando establecer "relaciones causales" (e.g. grupo de tratamiento y control)
- **Internet**: websites, redes sociales $\rightarrow$ Web scraping, Data mining.

Fuentes de datos

- **Primaria**: recolectada por el investigador (encuestas o experimentos)
- **Secundaria**: recolectada por otra persona (censos, bases de datos públicas, internet, etc.)

Fundamental (para considerar SIEMPRE)

- Saber cómo se obtuvieron los datos.
- Entender sus limitaciones (en la forma en cómo se obtuvieron o cómo se miden las variables).
- Evaluar si es consistente (¿faltan datos? ¿hay errores?).
- Entender si podemos (o no) decir algo válido y fundamentado sobre el universo o población en base a una porción limitada de muestra.

Ojo, aún estamos lejos de poder hacer esto último.

Tipo de bases de datos

- Datos de **corte transversal** (*cross-section*)
- Datos de **series temporales** (*time-series*)
- Datos en **panel** (*panel data*)

Creando una base de datos

Corte transversal:

$$\begin{array}{c} i \\ 1 \\ 2 \\ \vdots \\ n \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & \cdots & X_p \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \end{array}$$

Series temporales:

$$\begin{array}{c} t \\ 1 \\ 2 \\ \vdots \\ T \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & \cdots & X_p \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{T1} & x_{T2} & \cdots & x_{Tp} \end{pmatrix} \end{array}$$

Creando una base de datos

Datos en panel:

$$
\begin{array}{c}
it \\
11 \\
12 \\
\vdots \\
1T \\
21 \\
22 \\
\vdots \\
2T \\
n1 \\
n2 \\
\vdots \\
nT
\end{array}
\begin{array}{c}
\begin{array}{cccc} X & Y & \cdots & Z \end{array} \\
\left(\begin{array}{cccc}
x_{11} & y_{11} & \cdots & z_{11} \\
x_{12} & y_{12} & \cdots & z_{12} \\
\vdots & \vdots & & \vdots \\
x_{1T} & y_{1T} & \cdots & z_{1T} \\
x_{21} & y_{21} & \cdots & z_{21} \\
x_{22} & y_{22} & \cdots & z_{22} \\
\vdots & \vdots & & \vdots \\
x_{2T} & y_{2T} & \cdots & z_{2T} \\
x_{n1} & y_{n1} & \cdots & z_{n1} \\
x_{n2} & y_{n2} & \cdots & z_{n2} \\
\vdots & \vdots & & \vdots \\
x_{nT} & y_{nT} & \cdots & z_{nT}
\end{array}\right)
\end{array}
$$

Principales Estadísticos Descriptivos

Los principales estadísticos descriptivos pueden resumirse en las siguientes grandes familias (esencialmente contrapartidas muestrales de lo que vimos en variables aleatorias):

- **Medidas de tendencia central**
- **Medidas de dispersión**
- **Medidas de posición**
- **Medidas de forma**

Principales medidas de tendencia central

- Media
- Mediana
- Moda

Media aritmética y geométrica

Media aritmética:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Media geométrica:

$$\overline{X}_G = \sqrt[n]{\prod_{i=1}^{n} X_i} = \sqrt[n]{X_1 \cdot X_2 \cdots X_n}$$

Media geométrica

- **Ventaja**: es menos sensible que la media aritmética a los valores extremos.
- **Desventajas**: si algún valor de X_i=0, se anula. Es de significado estadístico menos intuitivo.
- El logaritmo de la media geométrica es igual a la media aritmética de los logaritmos de los valores de la variable.
- Si X_i toma valores positivos, su media geométrica es siempre menor o igual que la media aritmética:

$$\sqrt[n]{X_1 \cdot X_2 \cdots X_n} \leq \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Media geométrica - Ejemplo

Según datos del Banco Central de Venezuela, la tasa de inflación fue del 56% en 2013, del 69% en 2014 y del 181% en 2015. ¿Cuál fue la **tasa promedio de inflación** durante esos 3 años?

- ERROR: $(56\% + 69\% + 181\%)/3 = 102\%$
- **CORRECTO:**

$$(1+\pi)^3 = (1+\pi_{13}) \times (1+\pi_{14}) \times (1+\pi_{15})$$
$$(1+\pi)^3 = 1.56 \times 1.69 \times 2.81 = 7.408284$$
$$(1+\pi) = \sqrt[3]{7.408284} = 1.949422$$
$$\pi = 1.949422 - 1 = 94.9422\%$$

Media ponderada

Otorga diferente peso o importancia a las distintas observaciones.

ω: ponderador $(0 < \omega < 1)$

$$\overline{X} = \omega_1 X_1 + \omega_2 X_2 + ... + \omega_n X_n = \sum \omega_i X_i$$

donde

$$\omega_1 + \omega_2 + ... + \omega_n = \sum \omega_i = 1$$

Algunos usos:

- Algunos índices compuestos
- Bases de datos estratificadas
- Portafolios financieros

Mediana

La **mediana** es la observación que ocupa el lugar central cuando las observaciones están ordenadas en sentido ascendente (o descendente).

Año	2009	2010	2014	**2012**	2013	2008	2011
Beneficiarios	31	33	33	**34**	34	35	36

Esto es así si la cantidad de observaciones es impar (ej.,n=7).

Mediana

Considere n observaciones $x_1, x_2, ..., x_n$ que pueden ser ordenadas como $x(1) \leq x(2) \leq \cdots \leq x(n)$. El cálculo de la mediana depende de si la cantidad de observaciones n es par o impar.

$$\widetilde{X}_{0.5} = \begin{cases} x_{((n+1)/2)} & \text{si } n \text{ es impar} \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}) & \text{si } n \text{ es par} \end{cases}$$

Media vs Mediana

Si bien la media es más utilizada por su simplicidad, en algunas situaciones la mediana puede ser preferible. **¿Cuándo?**

La media es muy sensible a la presencia de **valores extremos**, mientras la mediana es una medida más robusta. En estos casos, la mediana puede ser preferible a la media como medida de tendencia central.

Moda

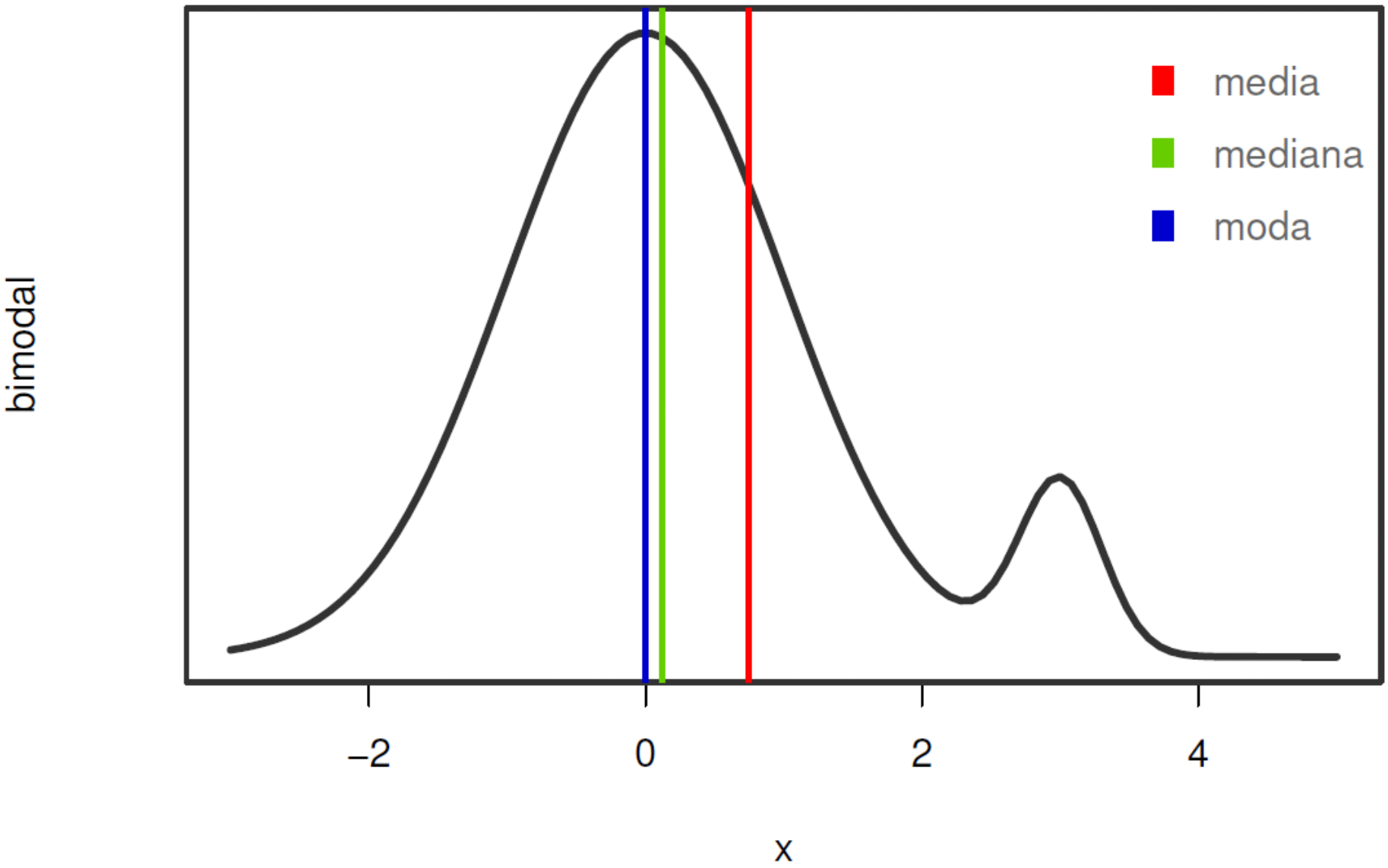
La **moda** (o el modo) es el valor más frecuente dentro del conjunto de observaciones.

$$\overline{X}_M = a_j \Leftrightarrow n_j = \max\{n_1, n_2, ..., n_k\}$$

Pros y Contras

Medida	**Pro**	Contra
Moda	-Cálculo sencillo -Interpretación clara -Puede calcularse en variables cualitativas	-Sensible a n -Puede haber más de una moda -No siempre está en el centro -Poco sentido para var. continuas
Mediana	-Fácil de calcular -Robusto a outliers	-Para var. numéricas ordenables -Sensible a n -Concepto no tan familiar
Media	-Fácil de entender -Poco sensible a n -Usa todos los datos	-Afectado por outliers -Puede caer fuera de los posibles -No es útil para variables discretas

Media, Mediana, Moda



Medidas de dispersión

Queremos comparar la cantidad de personas que ingresaron en cierto local por trimestre en 2016 y 2017 y observamos lo siguiente:

Trimestre	2016	2017
Primer trimestre	1200	1070
Segundo trimestre	1500	2694
Tercer trimestre	1350	6
Cuarto trimestre	1220	1500
Media	**1317.5**	**1317.5**
Mediana	**1285**	**1285**

La media y la mediana son las mismas en 2016 y 2017,

¿se puede decir que los dos años son parecidos?

Principales medidas de dispersión

Si nos limitáramos solamente a fijarnos en las medidas de tendencia central, no tendríamos una idea acabada de cómo se distribuyen los datos.

¿Son los datos muy dispersos?

Para responder a esta pregunta, veremos las siguientes medidas de dispersión muestral:

- **La varianza**
- **El desvío estándar**
- **La desviación absoluta**
- **El coeficiente de variación**
- **El rango o recorrido**

Varianza muestral

El promedio de los cuadrados de las diferencias nos proporciona la medida de la **varianza muestral**, denotada con S^2:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

Su desventaja es que resulta difícil de interpretar. Como en el caso poblacional, las unidades están elevadas al cuadro.

Una forma sencilla de resolverlo es tomando su raíz cuadrada (positiva): el **desvío estándar**.

$$S = \sqrt{S^2}$$

La desviación absoluta

La **desviación media absoluta** (o *mean absolute deviation*, *MAE* es el promedio de las distancias o diferencias absolutas entre cada observación y la media.

$$DMA = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}$$

Para un mismo conjunto de datos se da que:

$$DMA \leq S$$

¿Por qué?

Rango o recorrido

$$R = \max - \min$$

- La medida de dispersión más simple.
- Puede verse influido por valores extremos.
- No considera la dispersión del resto de las observaciones que no estén en los dos extremos.

Rango intercuartil

$$R = Q_3 - Q_1$$

- Sigue siendo simple.
- Menos influido por valores extremos que el recorrido.
- No considera la dispersión en el "centro" de la distribución.

Unidades

Hasta ahora estudiamos medidas de dispersión que dependen de las unidades de medición (centímetros, grados, dólares, años).

Con estas medidas, comparar la dispersión de conjuntos de datos donde las unidades son distintas no sería válido, ya que modificando las unidades de medición podríamos alterar a conveniencia las medidas de dispersión.

Coeficiente de Variación

- Se usa fundamentalmente para comparar la variabilidad de dos o más conjuntos de datos con distintas unidades de medida,
- o distinta media.

$$CV = \frac{S}{\overline{X}} \times 100$$

¿En qué caso hay más dispersión?

Semana	1	2	3	4	5	6	7	8	9	10	Media	Var.	Desvío
Temperatura (°C)	31	29	27	23	21	12	25	18	22	25	**23.3**	**30.5**	**5.5**
Cantidad de lluvia (mm)	12	11	6	6	6	14	6	3	3	2	**6.9**	**16.8**	**4.1**

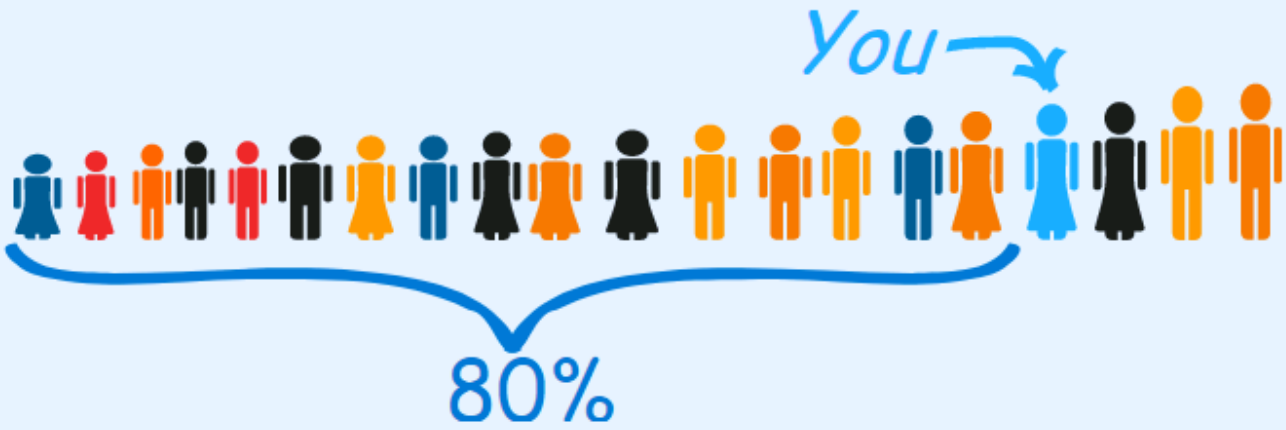
Medidas de posición - Cuantiles

Para datos en forma ascendente (o descendente) se dividen al conjunto de datos en distintos grupos con igual número de observaciones.

- Cuartiles: los 3 valores de la variable que dividen al conjunto de datos en 4 partes iguales (c/u representa el 25%).
- Quintiles: los 4 valores de la variable que dividen al conjunto de datos en 5 partes iguales (c/u representa el 20%).
- Deciles: los 9 valores de la variable que dividen al conjunto de datos en 10 partes iguales (c/u representa el 10%).
- Percentiles: los 99 valores de la variable que dividen al conj. de datos en 100 partes iguales (c/u representa el 1%).

Medidas de posición - Cuantiles

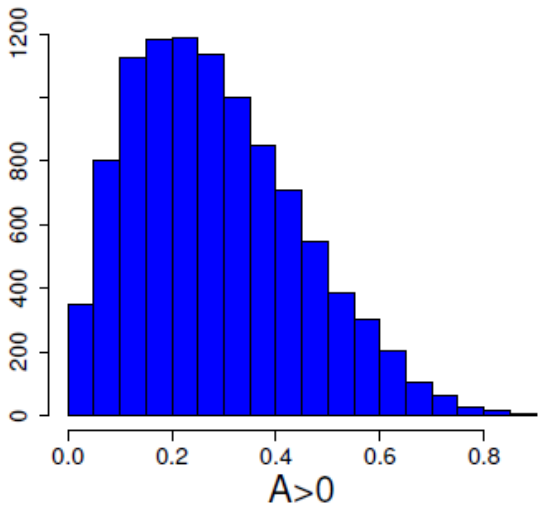
80% of people are shorter than you:

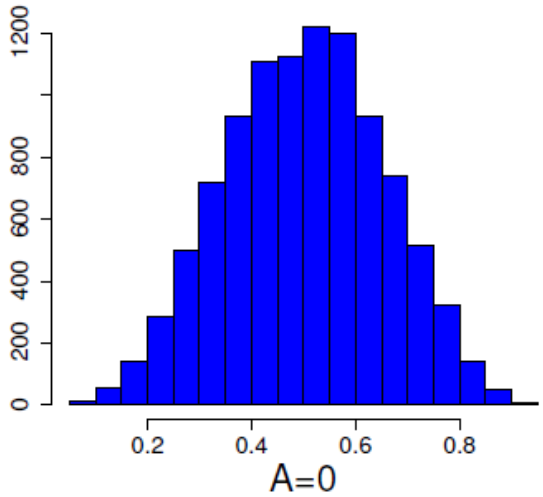


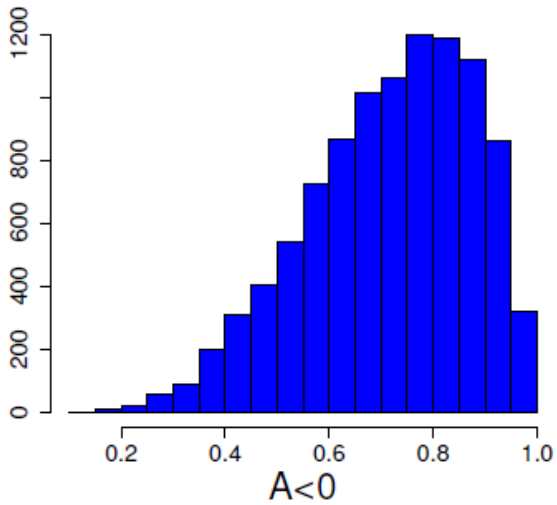
That means you are at the **80th percentile**.

Coeficiente de Asimetría

Permite establecer el grado de asimetría (o simetría) que presenta una distribución de datos.







Medidas de forma - Coeficiente de Asimetría

Si bien más adelante veremos cómo calcularla (à la Fisher), podemos ahora usar la medida del **coeficiente de asimetría de Bowley-Yule** que es bastante intuitiva:

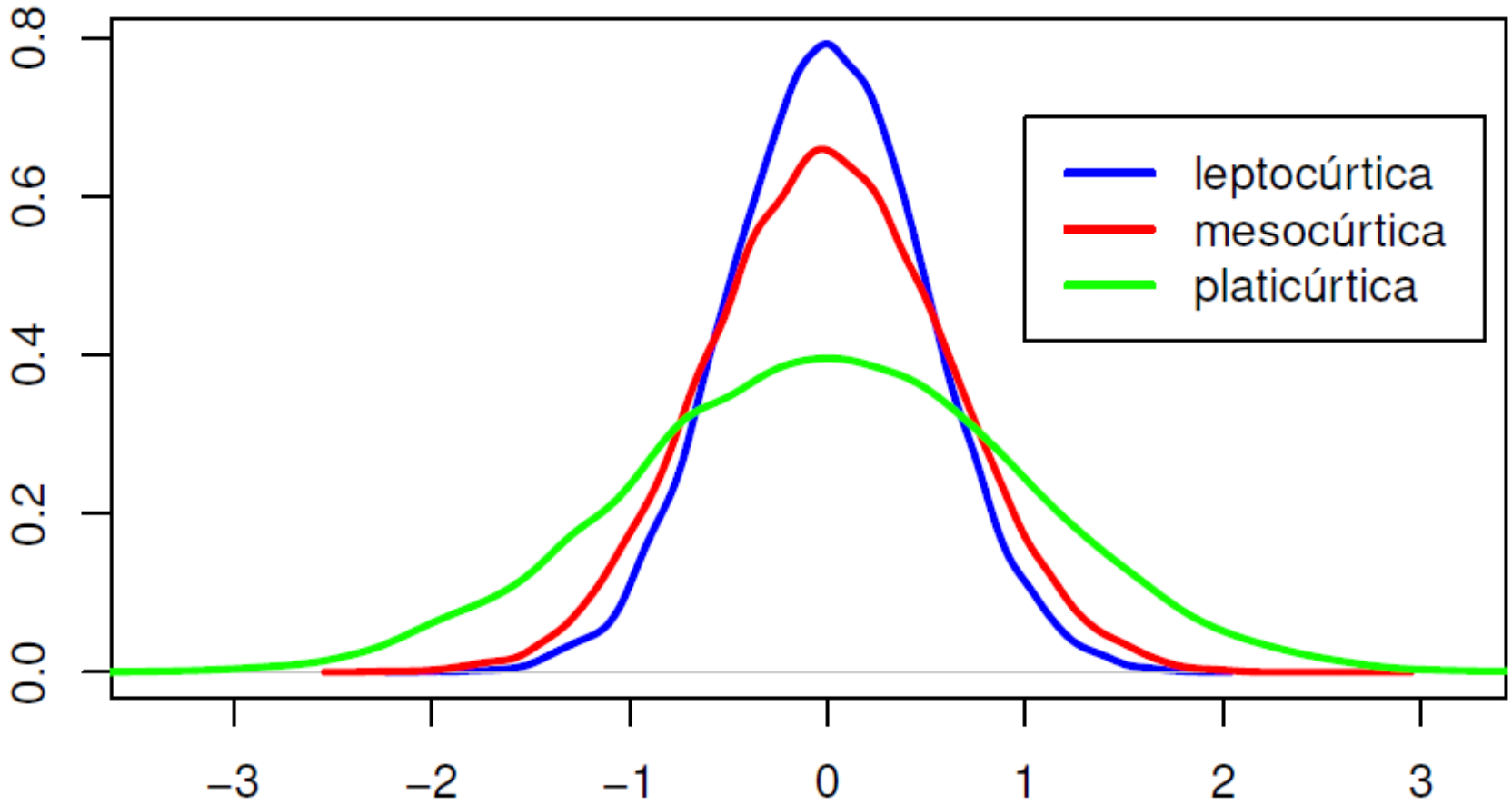
$$A_{BY} = \frac{Q_3 + Q_1 - 2Q_2}{Q3 - Q1}$$

Pensar:

- ¿Qué tiene que suceder para que $A_{BY} = 0$
- ¿Qué tiene que suceder para que $A_{BY} > 0$? ¿Y para que $A_{BY} < 0$?

Curtosis

Mide el grado de concentración de las valores entorno de la media.



Datos agrupados

En la práctica, la mayor parte de los conjuntos de datos contienen muchas observaciones, por lo que cuando la cantidad de información es muy grande resulta conveniente reducirla agrupando las observaciones en intervalos o rangos de valores.

Muchas veces, no solo se representan los datos en forma agrupada para reducir la cantidad de información sino también porque la información puede ser sensible a la no-respuesta. Es decir, es más probable que una persona en una encuesta conteste cierta información cuando se le pide que se ubique en un rango, que cuando se le pide el valor exacto (ejemplo: nivel de ingreso o edad).

Datos agrupados

Por ejemplo, la siguiente tabla tiene datos de ingreso mensual de 3200 individuos radicados en el conurbano bonaerense,

Ingresos	Nro de individuos
(150-450]	250
(450-750]	800
(750-1050]	1250
(1050-1350]	700
(1350-1650]	180
(1650-1950]	16
(1950-2250]	3
(2250-2550]	1
Suma	**3200**

El paréntesis excluye el valor y el corchete incluye el valor.

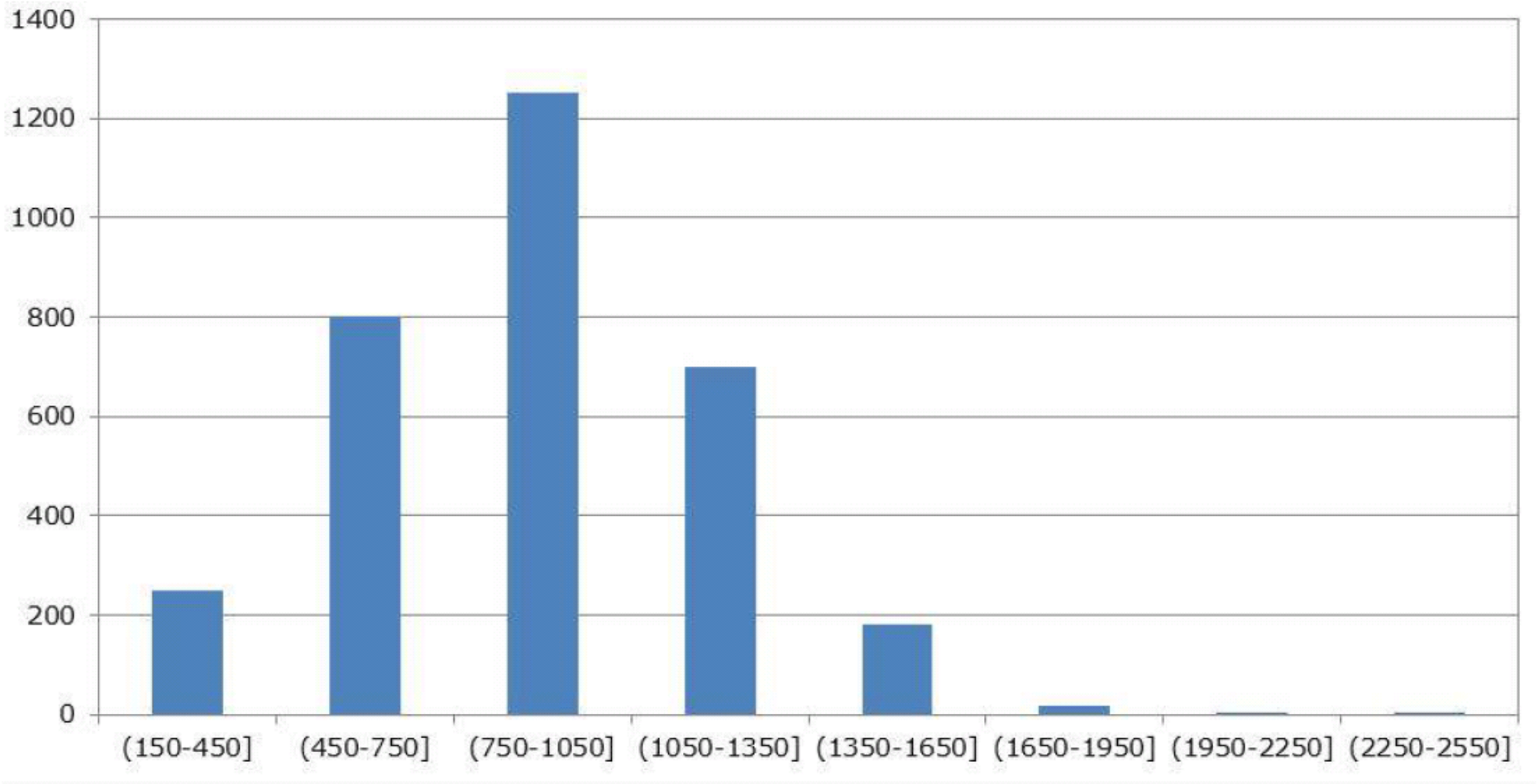
Frecuencias

Al número de observaciones de cada intervalo se lo llama **frecuencia**. La **frecuencia acumulada** es el número total de observaciones que hay en ese intervalo y en los anteriores.

Ingresos	Frecuencia	Frec. Acumulada
(150-450]	250	250
(450-750]	800	1050
(750-1050]	1250	2300
(1050-1350]	700	3000
(1350-1650]	180	3180
(1650-1950]	16	3196
(1950-2250]	3	3199
(2250-2550]	1	3200
Suma	**3200**	

Histograma

El **histograma** nos permite hacernos una idea visual rápida de la proporción de observaciones que se encuentran dentro de un determinado intervalo, relativo a los demás.



Histograma

Un histograma sirve para:

- Tener una primera vista de los datos, cómo se distribuyen
- Detectar casos extremos
- Detectar problemas con los datos
- Ver qué es lo que sucede más frecuentemente

Frecuencias Relativas

Si expresamos las frecuencias en términos de proporciones sobre el total las llamamos **frecuencias relativas**. Las **frecuencias relativas acumuladas** son las proporciones de observaciones que hay en un determinado intervalo o bien en alguno de los anteriores.

Ingresos	Frec.	Frec.Acum.	Frec.Rel.	Frec.Rel.Acum.
(150-450]	250	250	0.0781	0.0781
(450-750]	800	1050	0.2500	0.3281
(750-1050]	1250	2300	0.3906	0.7188
(1050-1350]	700	3000	0.2188	0.9375
(1350-1650]	180	3180	0.0563	0.9938
(1650-1950]	16	3196	0.0050	0.9988
(1950-2250]	3	3199	0.0009	0.9997
(2250-2550]	1	3200	0.0003	1
Suma	**3200**		**1**	

Relaciones entre variables (correlación)

Se profundizará más adelante, por ahora... una intuición del concepto.

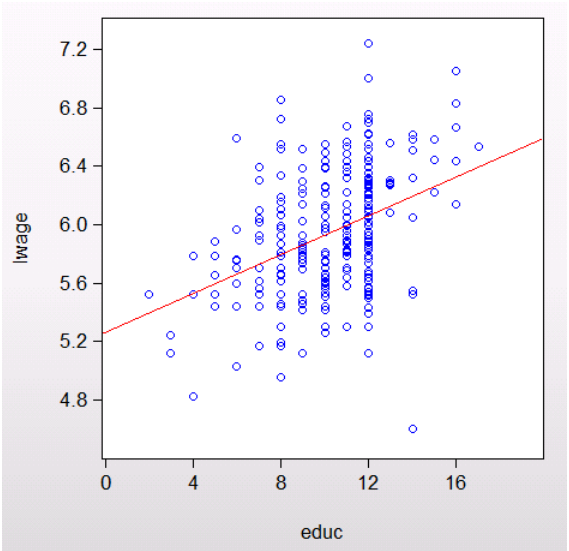
Hasta ahora hemos discutido cómo utilizar las medidas de tendencia central, variabilidad y posición para resumir un conjunto de datos.

- También estamos interesados en medir la fuerza de la relación entre dos conjuntos de datos.
- Por ejemplo, ¿cómo se relaciona el precio de una casa con el tamaño? ¿Cómo se relacionan los rendimientos de dos activos?
- La relación (lineal) entre dos conjuntos de datos se mide por la **covarianza** y **correlación**.

Diagramas de dispersión (Scatter Plot)

Se busca evaluar si:

- dos variables están asociadas (positivamente o negativamente) e para identificar comportamientos anómalos;
- dos variables se mueven linealmente en forma muy cercana (es decir, si están muy "correlacionadas").



Coeficiente de correlación de Pearson

Intenta medir cuán asociadas (linealmente) están dos variables.

$$r_{X,Y} = \frac{S_{X,Y}}{S_X S_Y}$$

donde,

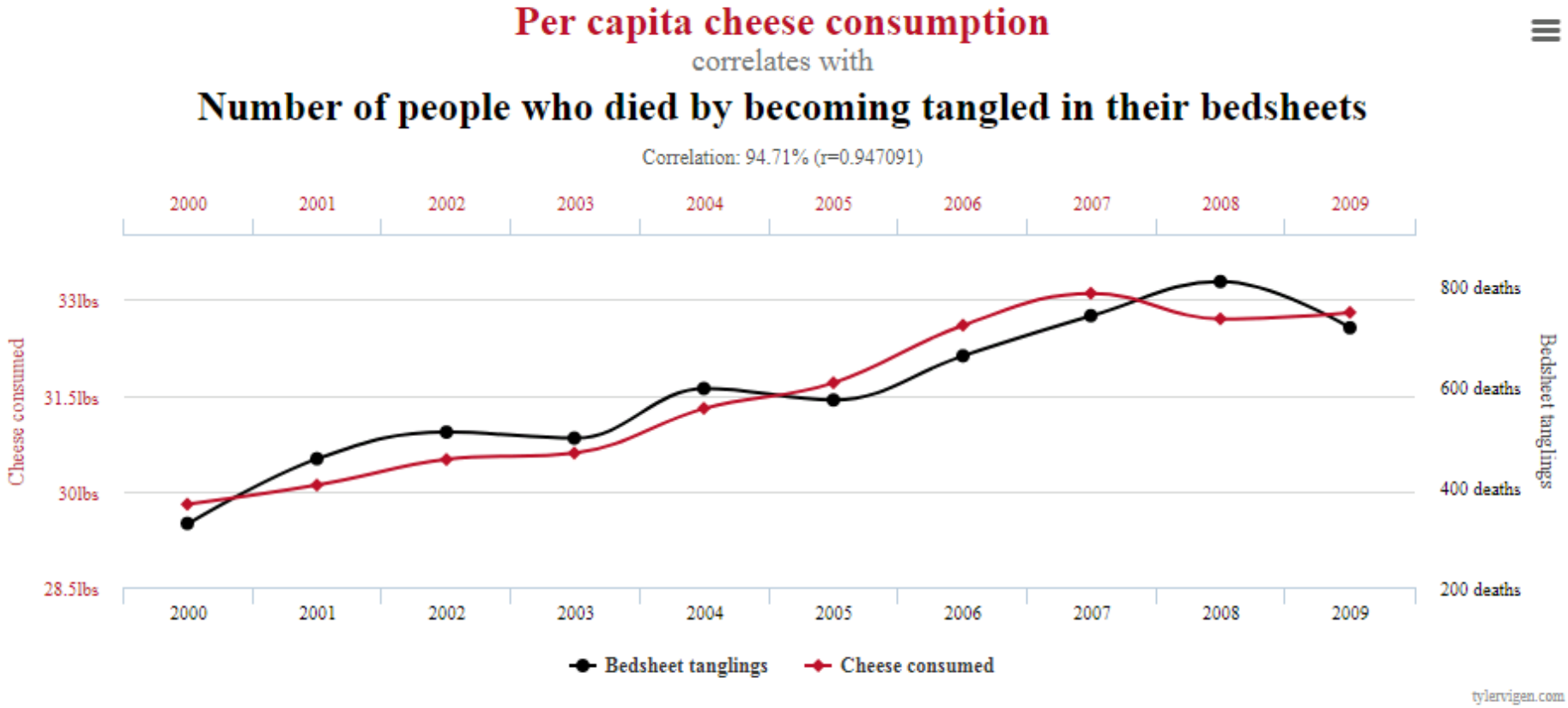
- $S_{X,Y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$ es la covarianza entre X e Y
- S_X es el desvío estándar de X
- S_Y es el desvío estándar de Y

Propiedades del coeficiente de correlación

- Libre de unidades de medida
- Invariante al cambio de unidades (ej. si paso se medir X en miles de USD a millones de USD, no cambia el coeficiente).
- $-1 \leq r \leq 1$, ya que está normalizado por los desvíos.
- Mide dependencia lineal, si los datos (X_i, Y_i) tienden a caer sobre una recta.
- Cuantifica la fuerza de la relación, pero no la forma de la recta (su pendiente y ordenada al origen).
- ¡Correlación **no** es sinónimo de causalidad!

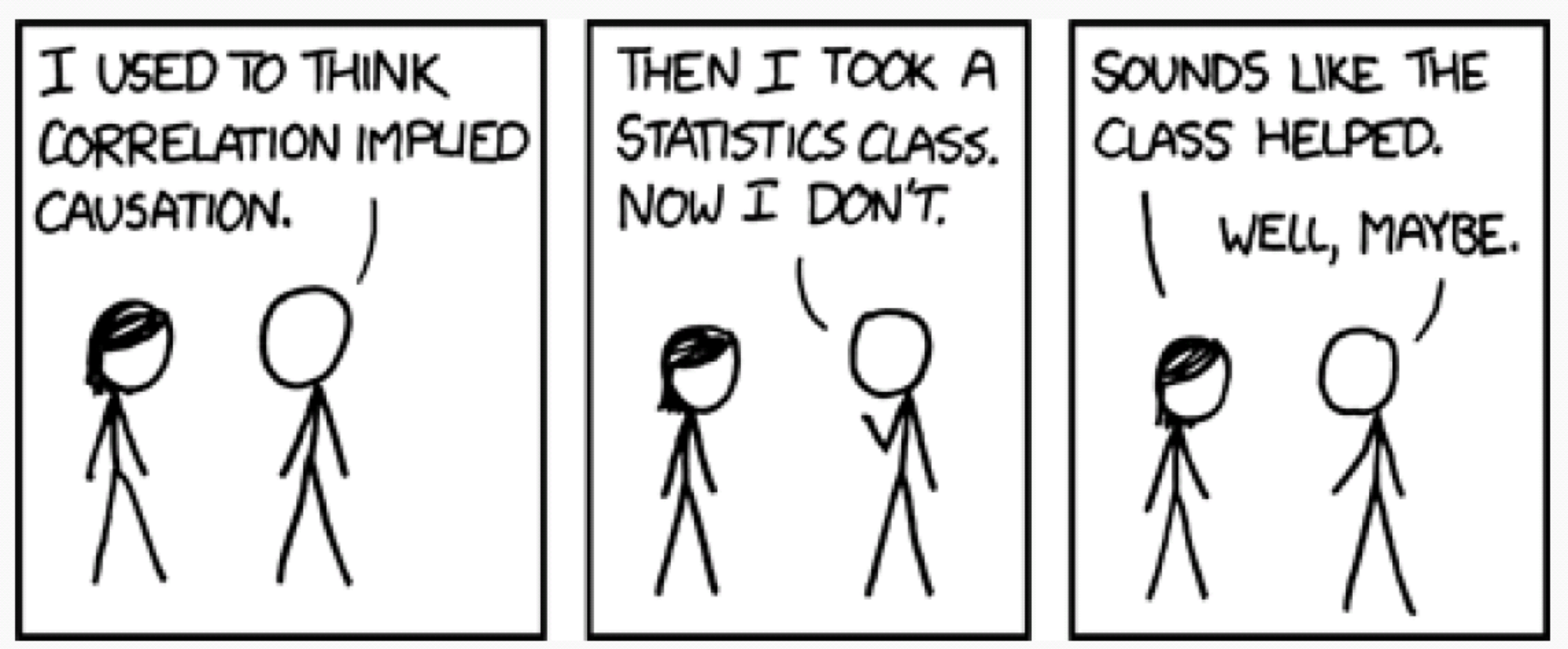
Correlacion espuria

Ver "*Spurious Correlations*" - Tyler Vigen



Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention

Correlación no implica causalidad



Misleading statistics - Paradoja de Simpson

Caso: UC Berkeley Gender Bias

- En 1973, las admisiones a UC Berkeley mostraban que los hombres tenían más chances de ser admitidos que las mujeres:

	Men		Women	
	Applicants	Admitted	Applicants	Admitted
Total	8442	**44**%	4321	35%

- Pero al analizar individualmente los departamentos, 6 de los 85 departamentos estaban significativamente sesgados contra los hombres y solo 4 contra las mujeres.
- De hecho, los datos agrupados y correctamente analizados mostraron "un pequeño sesgo estadísticamente significativo a favor de las mujeres".
- **¿Cómo puede ser esto?**

Misleading statistics - Paradoja de Simpson

- Bickel et al. (1975) probaron que las mujeres tendían a aplicar a departamentos competitivos con tasas bajas de admisión incluso entre solicitantes calificados (ej. Departamento de Inglés), mientras que los hombres tendían a presentarse en los departamentos menos competitivos con tasas de admisión altas entre los solicitantes calificados (ej. Departamentos de Ingeniería y Química).

P.J. Bickel, E.A. Hammel and J.W. O'Connell (1975). "Sex Bias in Graduate Admissions: Data From Berkeley". *Science*. 187 (4175): 398–404

Misleading statistics - Paradoja de Simpson

- Listado de los 6 departamentos más grandes:

	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	**825**	62%	108	**82%**
B	**560**	63%	25	**68%**
C	325	**37%**	**593**	34%
D	417	33%	375	**35%**
E	191	**28%**	**393**	24%
F	373	6%	341	**7%**

Casos como este sugieren que hay que ser precavido en el nivel de agregación de los datos que se utiliza a la hora de sacar conclusiones. A priori parecía que el en admisiones se favorecía a los hombres, pero *desagregando* por departamento resulta que la conclusión es opuesta.

Comentarios - missing data o valor perdido

Se refiere a cuando no hay un valor para una variable para un individuo particular (no contestó, se perdió, etc.).

- En distintos softwares aparecen como `NA`.
- A veces se codifican (típico de encuestas: ej. 99): ¡OJO! La variable podría tomar valores numéricos como 1, 2, 3, y un 'not available' podría codificarse con 4.
- Si los *missing values* son pocos y ocurren de forma aleatoria, se los obvia en los cómputos.
- Si no son al azar, obviarlos puede generar sesgos: por ejemplo si hay gente que por algún motivo relevante al estudio decidió no participar o no contestar una pregunta.

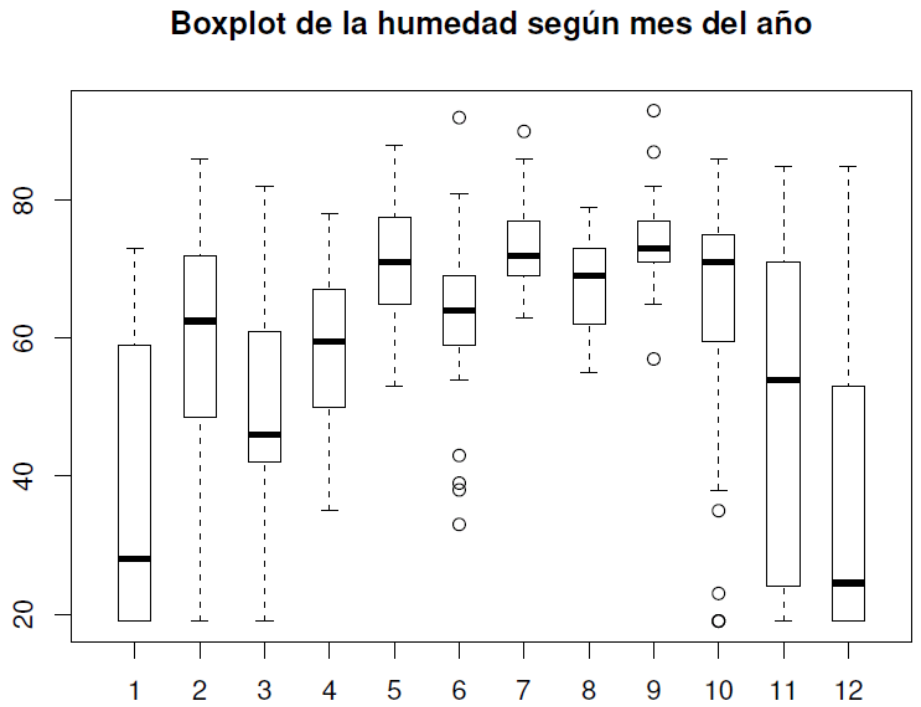
Comentarios - Outliers

Outlier se refiere a una observación que es "muy distinta" al resto (valor extremo), y esto se puede deber a un comportamiento atípico o a un error en los datos.

- Se identifican con gráficos o comparando media y mediana. No se descartan sin una buena razón.
- En la literatura hay posturas que defienden incluir observaciones como estas ya que puede aprenderse mucho de los casos extremos, mientras que otras apoyan la 'poda' de datos (usar una media *podada* al 5% por ejemplo).
- En la práctica, se definen como los valores que están más allá de $1.5 \cdot$ (rango intercuartil)

Box Plot o Diagrama de caja

Método para ilustrar gráficamente un conjunto de datos a través de sus cuartiles. Proporciona información sobre el mínimo, máximo, los cuartiles, la existencia de valores atípicos y la simetría de la distribución.



Aplicaciones: Medidas de Desigualdad del Ingreso

Existen distintas medidas de desigualdad del ingreso basadas en varias medidas de tendencia central y dispersión. Algunos ejemplos:

- **Los índices de Kuznets:** Simon Kuznets introdujo estos índices en su estudio pionero de la distribución del ingreso entre países desarrollados y en vías de desarrollo. Estos índices se refieren al cociente entre el ingreso que obtiene el x% más rico y el que obtiene el y% más pobre, donde x e y representan cifras como 10, 20, 40. (Concepto análogo al rango intercuartil)
- **El coeficiente de variación:** es el desvío estándar divido por la media, por lo que sólo son importantes los ingresos relativos: CV=S/$\overline{X}$.

Aplicaciones: Medidas de Desigualdad del Ingreso

- **Coeficiente de Gini:** medida muy utilizada para comparar desigualdades en el ingreso (entre países, regiones, etc.). Desarrollada por el estadístico Corrado Gini. Mide qué tan equitativamente se distribuye el ingreso en una escala de 0 a 1. Se puede calcular a nivel individual o a nivel de los hogares. No tiene ningún significado intrínseco, es una herramienta útil para comparar.

Gini $= 0 \rightarrow$ perfecta igualdad
Gini $= 1 \rightarrow$ perfecta desigualdad

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2 \overline{x}}$$

Muchas veces se habla del **índice de Gini** que básicamente es el coeficiente de Gini expresado en porcentaje (x 100).