DATATHON 2022
#Oddatathon
- Producción y consumo responsable
- Aspectos medioambientales
- Cultura

Iniciación a la estadística

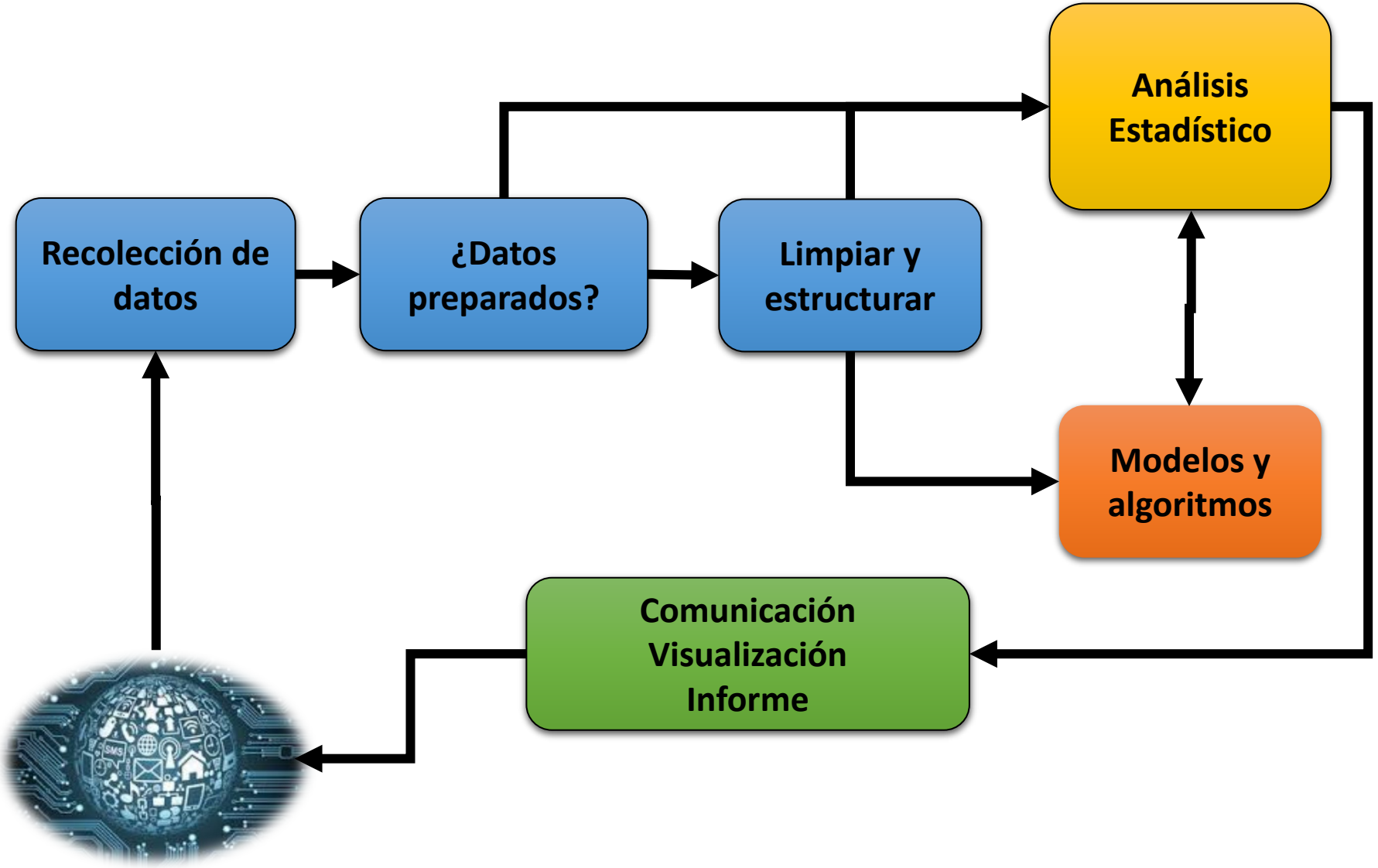
Elena Vázquez Barrachina
evazquez@eio.upv.es

Ángeles Calduch Losa
mcalduch@eio.upv.es

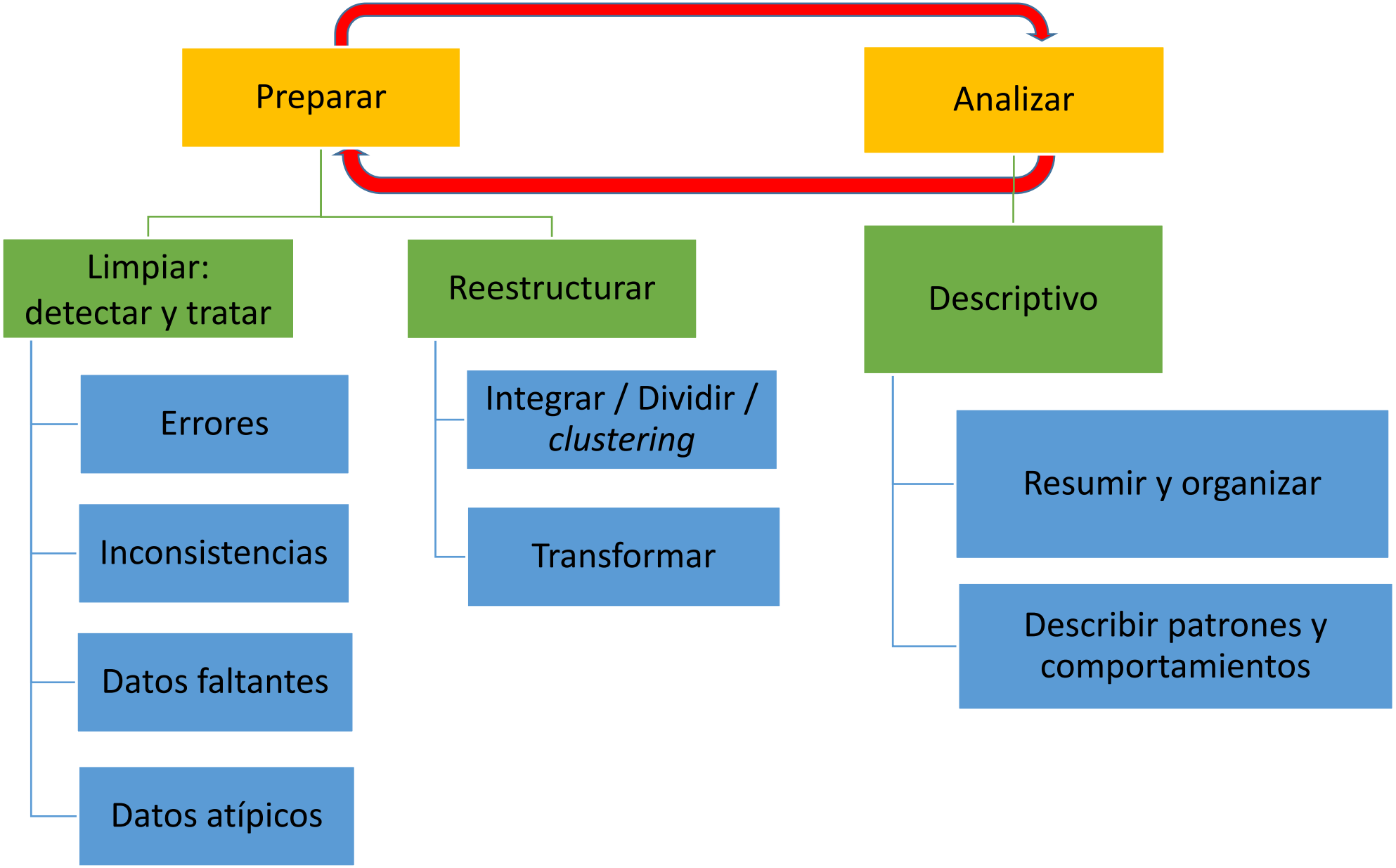
¿Para qué limpiar y preparar?

- Las técnicas de análisis de datos se centran en el modelado, predicción e inferencia, asumiendo que los datos son correctos, completos y en perfecto estado para el análisis.
- En la práctica, especialmente en el ámbito del Big Data, *web scraping*, etc, los datos crudos o *raw data* necesitan un proceso previo hasta que pueden considerarse ***técnicamente correctos***.
- Mediante la **preparación** se obtienen **datos sin errores, perfectamente codificados y consistentes para poder realizar el análisis estadístico**.
- Los resultados de la preparación pueden afectar significativamente a los resultados del análisis estadístico de los datos, por lo que puede considerarse como una fase más de éste.

Fases de un estudio de análisis de datos



1er Paso: Análisis Exploratorio de Datos



Antes de seguir…
un poco de estadística descriptiva

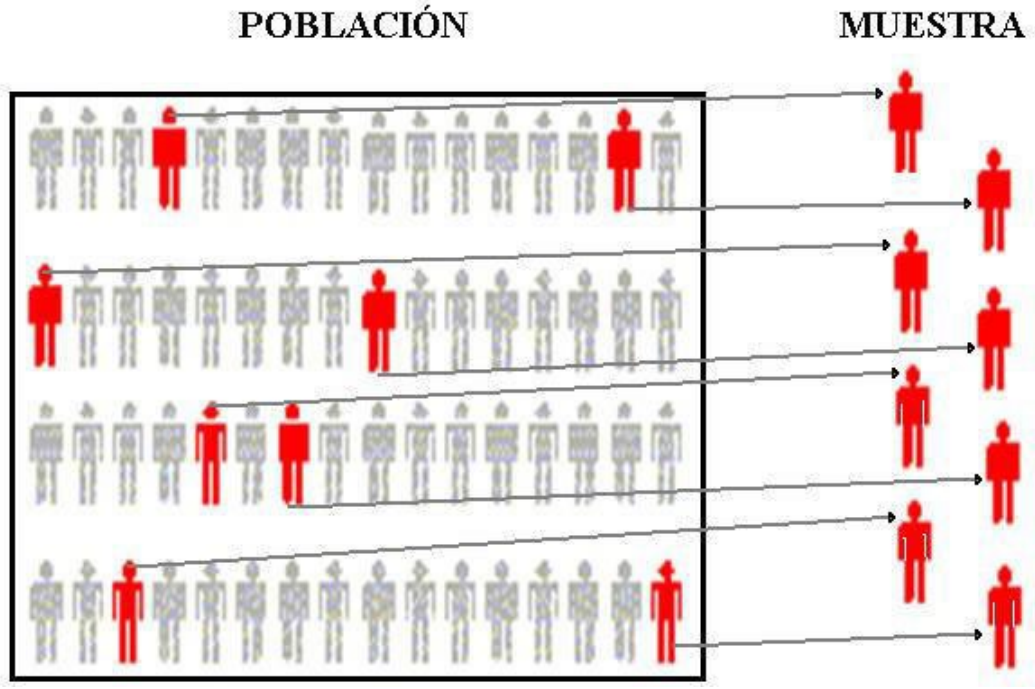
Análisis Exploratorio de Datos (AED)

El AED es el primer paso de cualquier estudio estadístico y consiste en la aplicación sistemática de un conjunto de técnicas estadísticas descriptivas y gráficas cuya finalidad es conseguir un entendimiento de la estructura básica de los datos, así como la identificación y tratamiento de errores e inconsistencias → Calidad

- **Resumir y organizar los datos**:
  - Representaciones gráficas
  - Cálculo de estadísticos descriptivos (parámetros) y otras herramientas

- Poner de manifiesto la **estructura subyacente básica de los datos**
  - Características
  - Regularidades
  - Errores e inconsistencias de los datos (duplicados, valores incorrectos, etc)
- **Preparar los datos para hacerlos accesibles**

Población y muestra

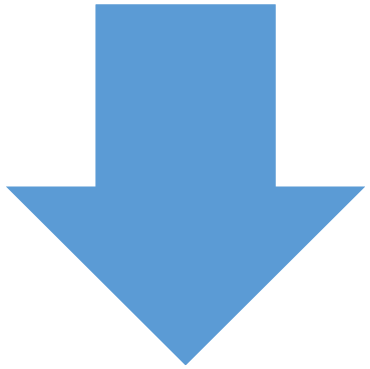


La **población** es el conjunto de todos los individuos o entes que constituyen el objeto de un determinado estudio y sobre los que se desea obtener ciertas conclusiones

La **muestra** es el subconjunto de individuos de la población sobre los que se recogen los datos a estudiar

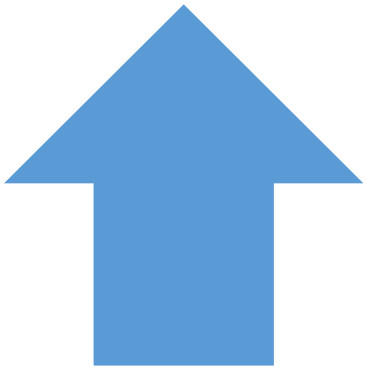
¿Por qué necesitamos una muestra?

- En muchas ocasiones **NO disponemos de toda la población** y necesitamos el muestreo porque:
  - Ensayos destructivos
  - Ensayos caros
  - Ensayos lentos
  - Ensayos difíciles/complejos
  - Población no accesible
  - Población infinita



Coste:
Temporal
Económico
Computacional
Validez resultados

Tamaño de
la muestra



¿Cómo debe ser la muestra?

La muestra debe ser **representativa de toda la población** y con un **error de muestreo conocido y aceptable**, de modo que permita extraer conclusiones razonablemente válidas sobre la toda la población.

Características aleatorias

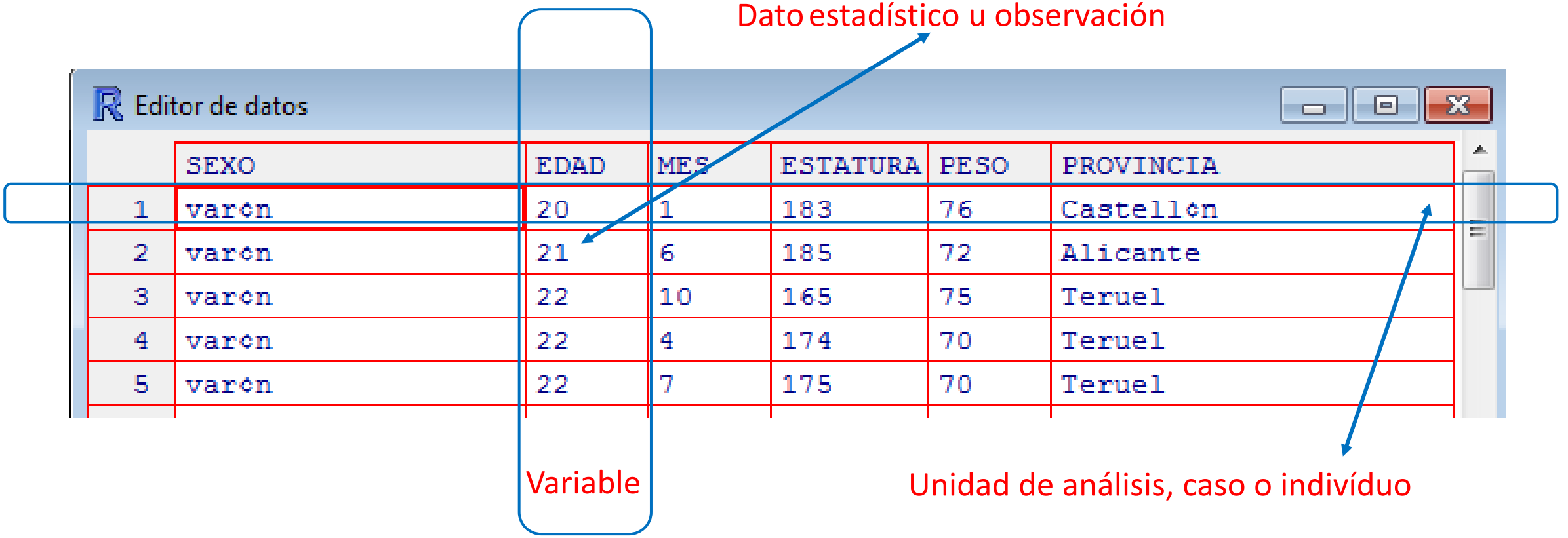
Una **característica o variable aleatoria** es cualquiera que puede constatarse en cada individuo de la población.

¡Toda población tiene VARIABILIDAD en sus características!

- La temperatura diaria en una ciudad a lo largo de un año.
- El número de asignaturas en las que se matricula un alumno también varía
- El partido votado en unas elecciones autonómicas
- ...

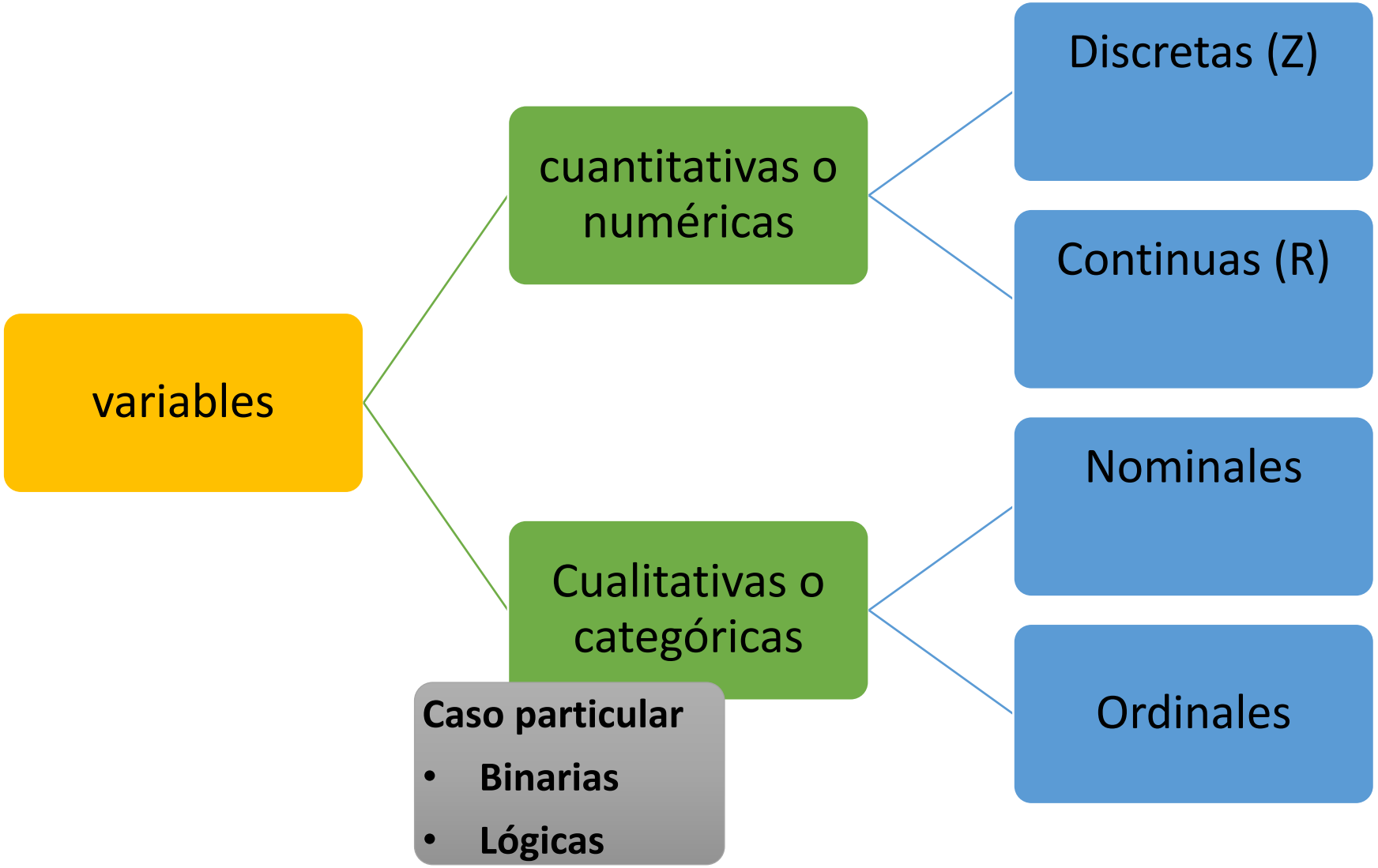
Datos

Los **datos** se obtienen a partir de una muestra de la población de interés y son los **valores observados** de la o las **variables** que se quieren analizar.



	SEXO	EDAD	MES	ESTATURA	PESO	PROVINCIA
1	var¢n	20	1	183	76	Castell¢n
2	var¢n	21	6	185	72	Alicante
3	var¢n	22	10	165	75	Teruel
4	var¢n	22	4	174	70	Teruel
5	var¢n	22	7	175	70	Teruel

Tipos de datos (o *variables*) por su naturaleza



Ejemplos

Editor de datos

	SEXO	EDAD	MES	ESTATURA	PESO	PROVINCIA
1	varón	20	1	183	76	Castellón
2	varón	21	6	185	72	Alicante
3	varón	22	10	165	75	Teruel
4	varón	22	4	174	70	Teruel
5	varón	22	7	175	70	Teruel

PROVINCIA: categórica nominal

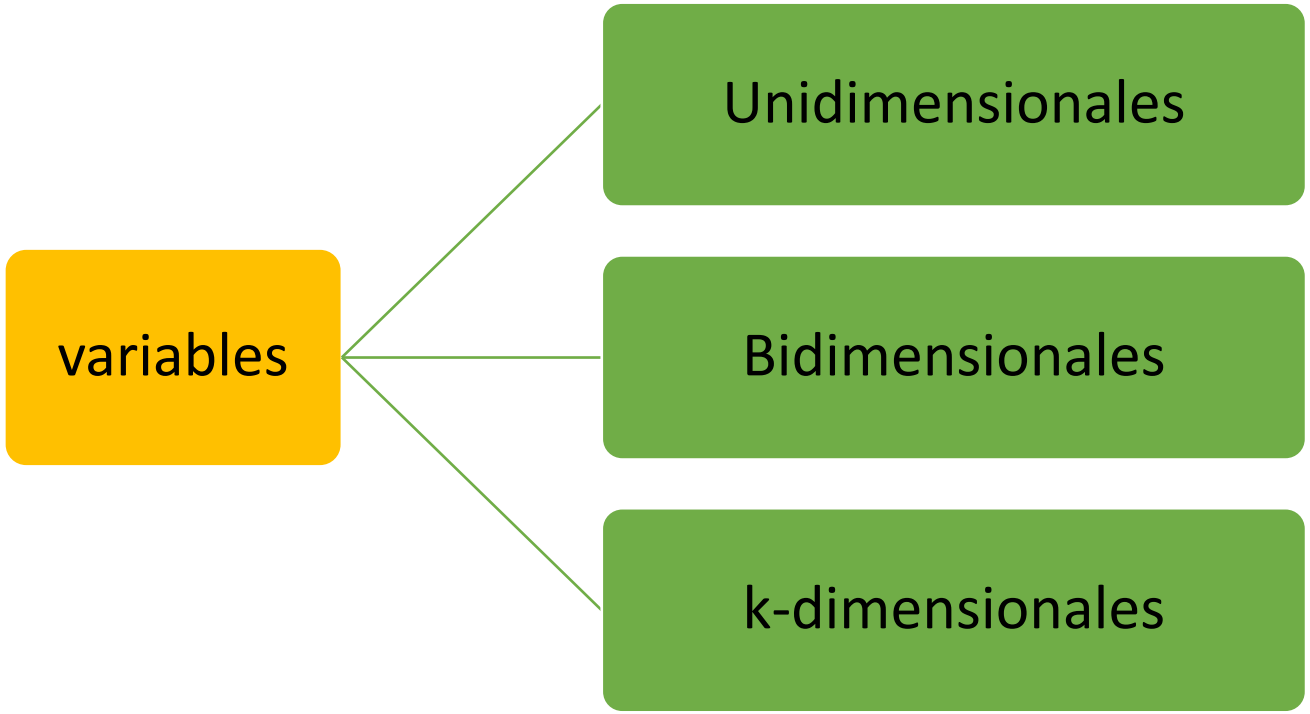
SEXO: categórica nominal binaria

EDAD: cuantitativa discreta

MES: categórica ordinal

ESTATURA: cuantitativa continua

Tipos de datos (o *variables*) por su dimensión



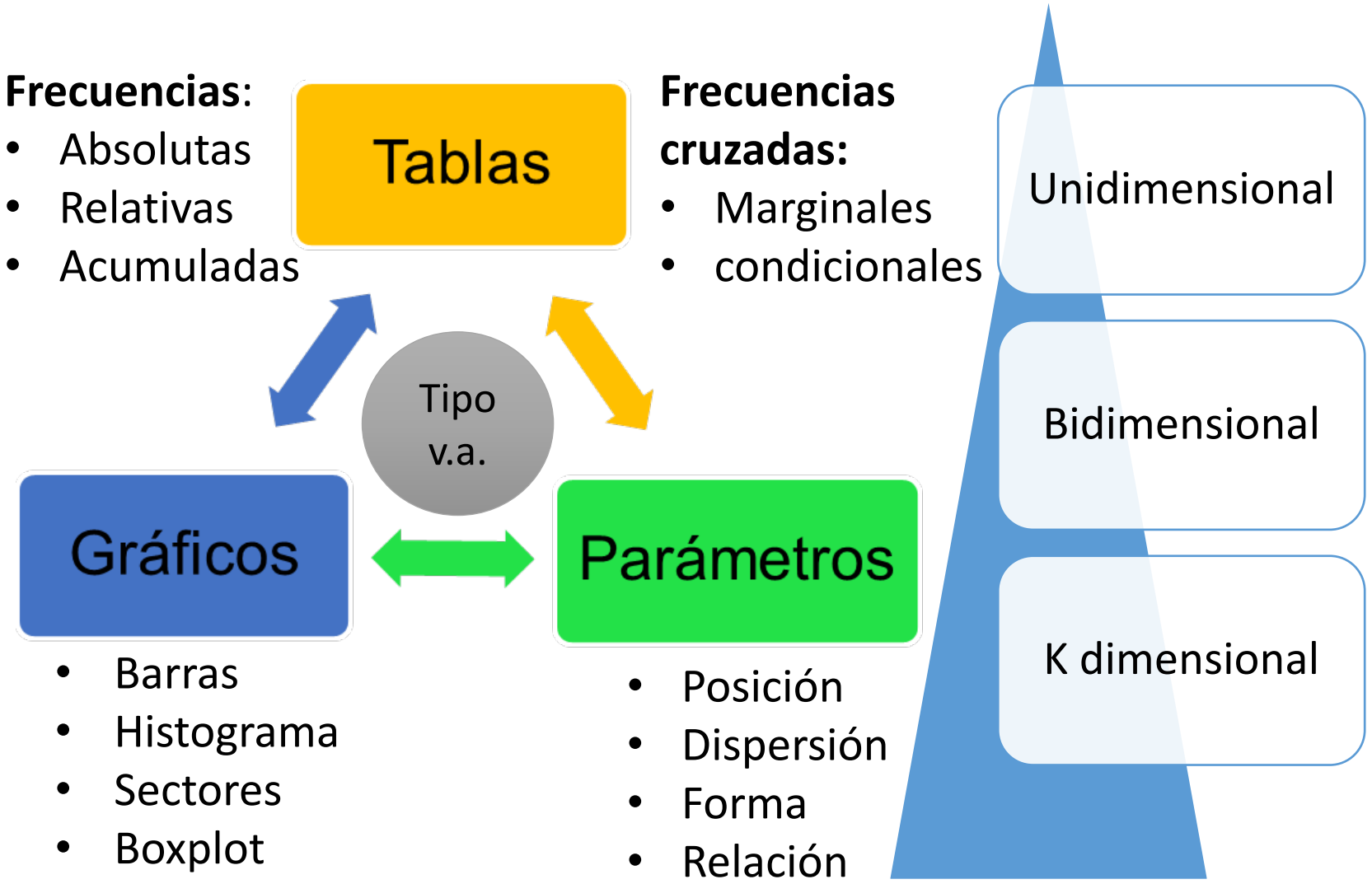
Importancia del tipo de datos

- El tipo de datos DETERMINA el método de análisis apropiado y válido y cada método de análisis estadístico es específico para un cierto tipo de datos.
- Entre las muchas clasificaciones que se pueden hacer de las técnicas estadísticas, a titulo meramente orientativo, se muestra la clasificación de Stevens[1] a continuación:

Variables dependientes	Variables independientes			
	Cualitativas		Cuantitativas	
	Una variable	Dos o más variables	Una variable	Dos o más variables
Sin variables dependientes	Test X^2 de bondad de ajuste	Medidas de asociación. Modelos Loglienales. Test X^2 de independencia. Análisis de correspondencias.	T-test. Estadísticos descriptivos. Test de normalidad.	Matriz de correlaciones. Componentes principales. Análisis Cluster.
Una variable cualitativa	X^2 Test. Test exacto de Fisher.	Regresión logística. Modelos Loglineales.	Análisis Discriminante. Regresión logística. Estadísticos univariantes de dos muestras.	Análisis discriminante. Regresión logística.
Más de una variable cualitativa	Modelos LogLineales. T-test. Análisis de varianza. Análisis de supervivencia.	Modelos LogLineales. Análisis de varianza. Análisis de Clasificaciones Múltiples. Análisis de supervivencia.	Regresión lineal. Análisis de correlaciones. Análisis de supervivencia.	Análisis de Regresión Múltiple. Análisis de supervivencia.
Una variable cuantitativa	T-Test. Análisis de varianza. Análisis de supervivencia.	Análisis de varianza. Análisis de clasificaciones. Análisis de supervivencia.	Regresión lineal. Análisis de correlaciones. Análisis de supervivencia.	Análisis de Regresión múltiple. Análisis de supervivencia.
Más de una variable cuantitativa	Análisis multivariado de la varianza. Análisis de varianza en componentes principales. T^2 de Hotelling.	Análisis Multivariado de la Varianza. Análisis de varianza en componentes principales.	Análisis de correlaciones canónicas.	Modelos de ecuaciones estructurales.

Stevens, S. S. (1951) Handbook of Experimental Psychology. Nueva York, NY: Wiley.

Análisis estadístico descriptivo



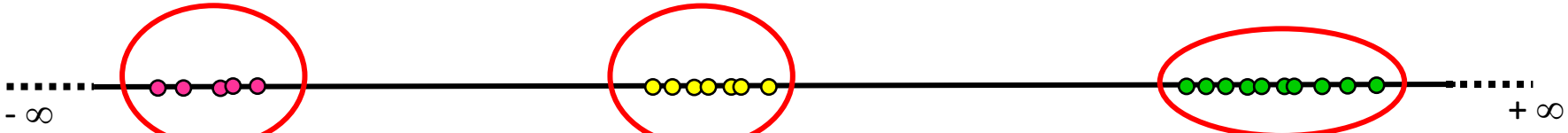
Parámetros muestrales o estadísticos de resumen

Las **medidas de resumen**, también llamadas **estadísticos** o **parámetros muestrales** son medidas de fácil interpretación que resumen y reflejen las **características cuantitativas** más relevantes de los datos de la muestra.

Se **clasifican** en parámetros que caracterizan:

- La **Posición** (**Centralización**) de las observaciones
- La **Dispersión** de las observaciones
- La **Forma** de las observaciones

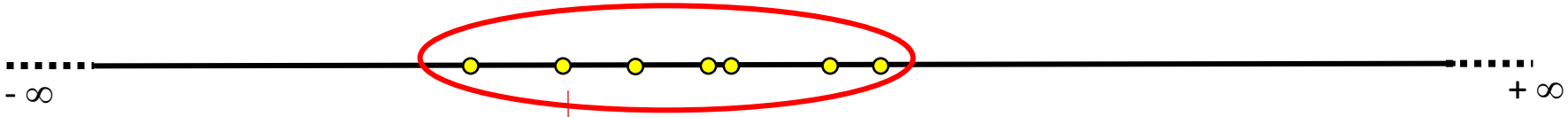
Centralización y Dispersión



Puntos alrededor de -200: posición 1

Puntos alrededor del 0: posición 2

Puntos alrededor de 300: posición 3



Puntos alrededor del 0: posición 2, pero los puntos (valores) están más alejados unos de otros: distinta dispersión

Parámetros de posición

- Permiten cuantificar y caracterizar, mediante un número, la **posición** de las observaciones
- Indican “alrededor” de qué valor están las observaciones.
- **Hay diversos estadísticos y hay que determinar cuál es el más adecuado**
- **Centralización o tendencia central:**
  - **Media** (Aritmética o Promedio, Geométrica, Recortada,…)
  - **Mediana**
  - **Moda**
- **Otros:**
  - **Cuantiles** (Cuartiles, Percentiles)

Media ($\overline{X}$)

- **Media aritmética** de los datos, **promedio** o **media**

$$\text{Media} = \overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

- Sintetiza la información existente en la totalidad de los datos en un número que da una idea clara sobre la tendencia central de los mismos.
- También puede ser **ponderada**
- Ventaja:
  - Es el parámetro de **centralización** (o posición) más utilizado
  - Recoge la información existente en la totalidad de los datos
- Inconveniente: se ve muy afectada por valores extremos

Moda (Mo)

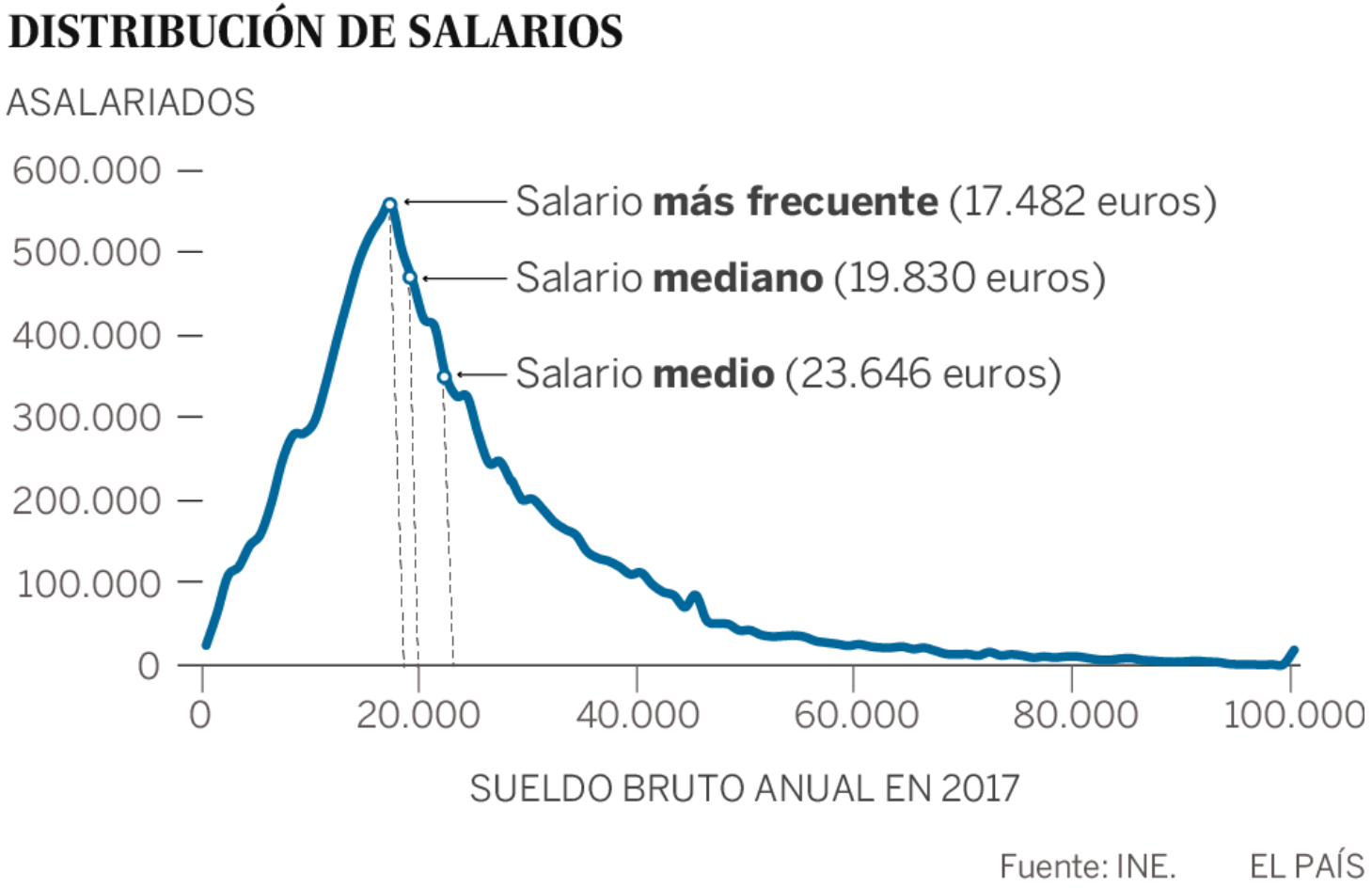
- Se **define** como aquel valor de la variable al que corresponde mayor frecuencia.

- **Ejemplo**: En el conjunto de datos: 4,5,6,6,3,6,4,5 la Mo=6

- **Propiedades**:
  - No es necesariamente única (puede haber varias modas)
  - Ventaja:
    - Se puede calcular con datos en escala nominal
    - No se ve afectada por valores extremos
  - Inconveniente: en su cálculo no intervienen todos los elementos

¿Qué parámetro es el adecuado?

¿Qué parámetro o estadístico consideráis que será una medida adecuada para representar la centralización de los salarios de los españoles en 2017?



Parámetros robustos de centralización

- En ocasiones la **media no** es un **buen parámetro de posición**.
- Cuando **tenemos** unos pocos **valores extremos** o **datos muy asimétricos** que pueden influir excesivamente en la media, esta medida resulta engañosa.
- Los **parámetros robustos o resistentes no se ven afectados por los valores extremos, ni por la asimetría de los datos.**

- **Parámetros robustos de centralización:**
  - **Mediana**
  - **Media Truncada**
  - **Media Winsorizada**
  - Otros parámetros robustos

Mediana (M_e o C_2 o Q_2)

Es el valor que :

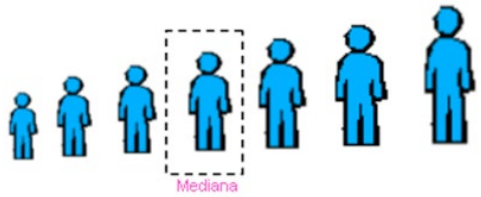
- deja a su izquierda el 50% de los datos
- deja a su derecha el 50% de los datos

Cálculo

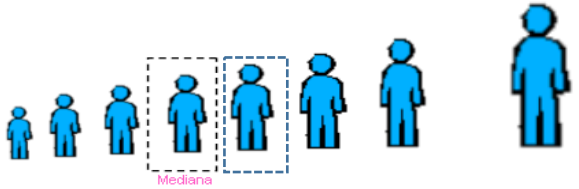
1. Se ordenan las observaciones de menor a mayor.

2. La mediana es el valor que:

➔ Ocupe la posición (N+1)/2 , si N es impar



➔ Media entre los valores que ocupan las posiciones N/2 y (N/2)+1, si N es par



Propiedades de la Mediana

- **Ventajas**:
  - No se ve alterada por:
    - Errores grandes de medida o transcripción
    - Valores extremos o asimetría de los datos
  - Se puede calcular con datos ordinales

- **Inconvenientes**: No utiliza toda la información contenida en la muestra, pues se basa en la posición de las observaciones.

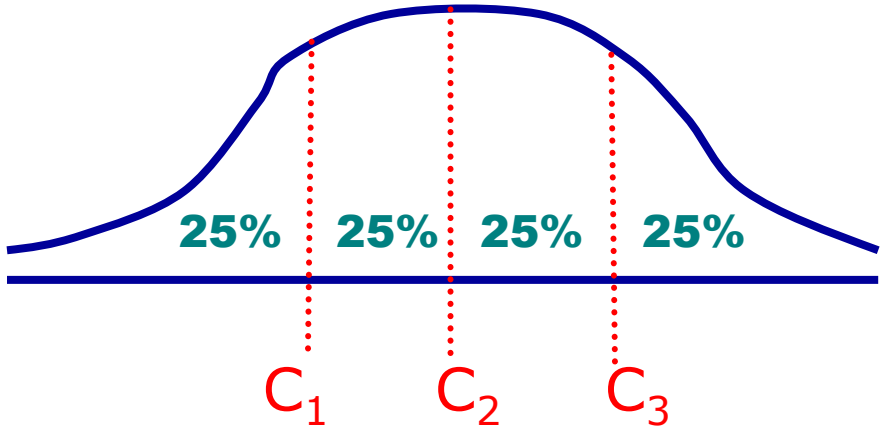
- **Conclusión:**
  - Se recomienda hallar los valores de ambas medidas (Media y Mediana)
  - Los dos parámetros difieren bastante de la media si la distribución es muy asimétrica, lo que sugiere heterogeneidad en los datos.

Cuantiles o Q_u o C_u

- Estadísticos que permiten conocer diversos puntos característicos de la distribución que no sean necesariamente valores centrales.
- Estas medidas se obtienen mediante un método que determina la ubicación de los datos de la muestra que dividen el conjunto de observaciones en partes iguales.
- Los más usados son:
  - los **cuartiles**, que dividen la muestra en cuatro partes
  - los **sextiles**, que dividen la muestra en seis partes
  - los **deciles**, que dividen la muestra en diez partes
  - los **centiles** o **percentiles**, que dividen la muestra en cien partes.
- Los cuartiles, como los deciles y los percentiles, son en cierta forma una extensión de la mediana:
  - Me = 2º cuartil
  - Me = 5º decil
  - Me = Percentil 50

Cuartiles: C_1, C_2, C_3

- **Cuartiles:**
  - **Primer cuartil** (C_1): el **25%** de los **datos** son **menores o iguales** a éste.
  - **Tercer cuartil** (C_3): el **75%** de los **datos** son **menores o iguales** a éste.
  - El **segundo cuartil** (C_2) es la **mediana**.
- Entre el Primer (C_1)y el Tercer cuartil (C_3) se encuentra comprendido el **50% central** de los datos



Percentiles

- El **percentil P** es aquel dato de la muestra de forma que el **P%** de los **datos** son **menores o iguales** a éste.

- Es aquel valor que, después de ordenar los datos de menor a mayor, **ocupa** la **posición i**, donde i se calcula como:

$$i = \frac{PN}{100}$$

p: percentil deseado

n: nº de observaciones

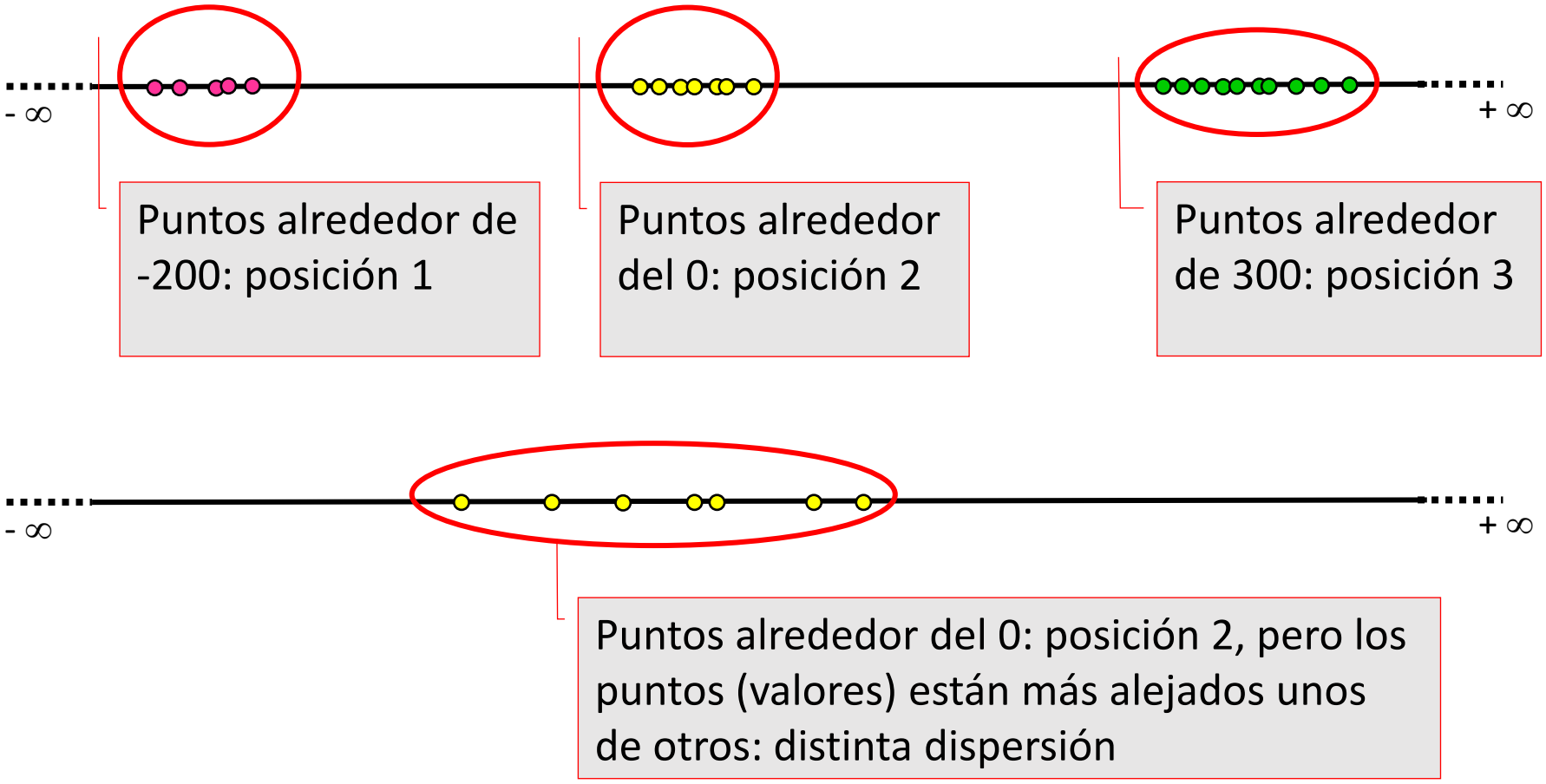
- **Ejemplo**: si decimos que el percentil 90 de la estatura de los alumnos de la UPV es 170 cm, quiere decir que, según nuestros datos el 90% de los alumnos tienen una estatura inferior o igual a 170 cm.

Para pensar…

- Para una persona que no sabe nadar, ¿es suficiente saber que la profundidad media de un lago es 1,40 m para lanzarse al baño en el mismo?
- ¿Aclararía mucho la decisión el conocer además la profundidad mediana del lago?

Fuente: ejemplo extraído de Romero Villafranca, Rafael y Zúnica Ramajo, Luisa (2013) *Métodos estadísticos para ingenieros*. Valencia: Universidad Politécnica de Valencia (Ref.: 4)

Centralización y Dispersión



Para describir un conjunto de datos no es suficiente con disponer de una medida de su posición ➔ es preciso también cuantificar el grado de dispersión que hay en ellos.

Parámetros de dispersión

- Permiten cuantificar, mediante un número el **grado de separación de los datos de una muestra con respecto a las medidas de tendencia central** consideradas.
- Las medidas de dispersión son de dos **tipos**:
  - **Medidas de dispersión absoluta**: como recorrido, desviación media, varianza y desviación típica, que se usan en los análisis estadísticos generales.
  - **Medidas de dispersión relativa**: que determinan la dispersión de la distribución estadística independientemente de las unidades en que se exprese la variable. Se trata de parámetros más técnicos y utilizados en estudios específicos, y entre ellas se encuentran el recorrido relativo, el coeficiente de variación (índice de dispersión de Pearson) y el índice de dispersión mediano, etc.
- Como en el caso de la centralización. También hay **parámetros robustos y no robustos**.

Parámetros de dispersión más relevantes

- **Recorrido o Rango**
- **Varianza y Desviación típica**
    - S Geométrica
    - Sigma Winsorizada
- **Coeficiente de Variación**
- **Recorrido intercuartílico**
- **Rango intersextil**
- Desviación absoluta media (Mean Absolute Deviation (DM))
- **Desviación absoluta mediana (Median Absolute Deviation (MAD))**
- Error estándar
- Sbi

Recorrido o Rango (R)

$$R = x_{Max} - x_{min}$$

- **Ventaja**: es más sencillo de calcular y da una Idea intuitiva de la dispersión
- **Inconveniente**: ignora gran parte de los datos
- **Útil** en muestras pequeñas (N≤10)
    - Por ejemplo, en Control Estadístico de Procesos (N=5)

Ejemplo

Horas de conexión a Internet al mes en niños de 5 a 11 años

Horas = {30, 38, 45, 47, 48, 50, 50, 52, 62}

R = Max – Min = 62 – 30 = 32 horas

Representativo

Horas´ = {30, 38, 45, 47, 48, 50, 50, 52, 150}

R = Max – Min = 150 – 30 = 120 horas

NO Representativo

Varianza (S^2)

Dado que la media es, en general un buen parámetro de tendencia central parece lógico tomar como parámetro de dispersión alguno que esté relacionado con ella:

$$\text{Varianza} = S^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$$

Suma de las diferencias de cada valor con respecto a la media al cuadrado

$S_x^2 = ((50-48{,}18)^2 + (38-48{,}18)^2 + (45-48{,}18)^2 + (30-48{,}18)^2 +$

$+ (47-48{,}18)^2 + (50-48{,}18)^2 + (48-48{,}18)^2 + (62-48{,}18)^2 +$

$+ (55-48{,}18)^2 + (53-48{,}18)^2 + (52-48{,}18)^2) / (11-1) = 72{,}76\ h^2$

¡OJO! Las unidades de la varianza están al cuadrado

Nota: Teóricamente, al dividir por (N-1) lo que obtenemos es la cuasivarianza muestral.

Varianza (S^2)

Tiempo	Diferencias con respect a la media	Cuadrados de las diferencias
50	1,82	3,31
38	-10,18	103,67
45	-3,18	10,12
30	-18,18	330,58
47	-1,18	1,40
50	1,82	3,31
48	-0,18	0,03
62	13,82	190,94
55	6,82	46,49
53	4,82	23,21
52	3,82	14,58

Media 48,18 727,64 **S. de cuadrados**

N 11 72,76 **Varianza**

Desviación Típica o Estándar (S)

- El más utilizado
- La raíz cuadrada de la varianza
- Más fácil de interpretar puesto que viene expresada en las mismas unidades que los datos originales.

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

- Siguiendo con el ejemplo de las horas de conexión a Internet:

$$S_X = \sqrt{72,76} = 8,53 \text{ horas}$$

Parámetros robustos de dispersión

- Al igual que ocurre con la media en el caso de las medidas de centralización, en ocasiones la **desviación típica no** es un **buen parámetro de dispersión**.
- Cuando **tenemos** unos pocos **valores extremos** o **datos muy asimétricos** que pueden influir excesivamente en la S, esta medida no resulta representativa de la dispersión de los datos.
- Los **parámetros robustos o resistentes no se ven afectados por los valores extremos, ni por la asimetría de los datos.**
- **Parámetros robustos de dispersión:**
  - **MAD**
  - **Recorrido Intercuartílico**
  - Desviación Típica Winsorizada
  - Sbi
  - …

Desviación Absoluta Mediana (MAD o DAM)

- **Parámetro robusto** de dispersión
- Se calcula como la mediana de las desviaciones (en valor absoluto) de cada valor con respecto a la Mediana

$$MAD\ o\ DAM = Mediana(|x_i - Me|)$$

Ejemplo: X = { 50, 38, 45, 30, 47, 50, 48, 62, 55, 53, 52}

- Me = 50
- Se calculan las diferencias de cada observación respecto a Me despreciando el signo:

0	12	5	20	3	0	2	12	5	3	2

- Se calcula la mediana de los valores calculados: 3

MAD = 3 h

Rango o Intervalo Intercuartílico (II o RI)

- **Parámetro robusto** de dispersión
- Se calcula como la diferencia entre el 3er y 1er cuartil:

$$II = C_3 - C_1$$

X = { 50, 38, 45, 30, 47, 50, 48, 62, 55, 53, 52}

$S_X = 8{,}53$ h $\approx$ II_X = C3 -C1 = 53 -45 = 8 h

X´ = { 50, 38, 45, 30, 47, 50, 48, 150, 55, 53, 52}

$S_{X'} = 31{,}94$ h $II_{X'}$ = C3 -C1 = 53 - 45 = 8 h

$S_X << S_{X'}$ $II_X = II_{X'}$

Coeficiente de Variación (CV)

- Indicador de **dispersión relativo**, no robusto
- Da una idea del “tamaño” de la dispersión respecto a la media.
- Como es **adimensional**, **permite comparar la variabilidad** de variables de naturaleza diferente

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

- Se suele expresar en %
- También se llama “variabilidad relativa”

Ejemplo

X = {Tamaño ficheros (Kb)} ≈ (media$_X$=20 Kb; S_X=10 Kb)

Y = {Tiempo ejecución (seg)} ≈ (media$_Y$ =180 seg; S_Y=36 seg)

¿ Qué variable tiene mayor VARIABILIDAD ?

reales

(S_Y=36 seg) >>(S_X=10 Kb) ¡No se deben comparar!

[**CV_Y**=36/180=**0,2**] << [**CV_X**=10/20=**0,5**]

Mayor dispersión

Tukey's Five Numbers

- Los **cinco números** resumen de un conjunto de datos o ***Tukey's Five Numbers*** consisten en la observación mínima, el primer cuartil, la mediana, el tercer cuartil y la observación máxima, escritos en orden de menor a mayor. De forma simbólica son:

Mín Q1 Me Q3 Máx

- Estos cinco números proporcionan una descripción razonablemente completa del centro y la dispersión

Ejemplo: Horas de conexión a Internet (X)

X = {50, 38, 45, 30, 47, 48, 62, 55, 53, 52} N=10

Mín	Q1	Me	Q3	Máx
30	**45**	**50**	**53**	**62**

Parámetros de forma

- Los **coeficientes** o índices de **Asimetría** y **Curtosis** son **parámetros de forma**.
- Además de los parámetros de posición y dispersión, éstos permiten **completar la caracterización de un conjunto de datos** sin necesidad de construir ningún gráfico.
- La caracterización se lleva a cabo en términos de lo **frecuente o infrecuente que es cada uno de los valores del conjunto de datos en referencia a algún valor central**.
- La forma nos da una idea acerca de la distribución de probabilidad de la v.a. y, por tanto, dependiendo de ella podemos usar unas técnicas estadísticas u otras.
- Gran parte de las técnicas de estadística inferencial están basadas en la hipótesis de que nuestros datos se asemejan suficientemente a una “campana de Gauss” (distribución Normal) y resulta de gran utilidad identificar esta forma.

Parámetros de forma

- Pautas de comportamiento que se alejan sensiblemente de la Normal exigen:
  - Revisión y corrección de datos anómalos, si procede o
  - Transformación de datos o

Preparación de Datos

  - Recurso a modelos o tratamientos estadísticos especiales.

- Tiene gran importancia detectar la forma de una v.a., especialmente conocer si puede considerarse “normal”

Mediante los parámetros de Asimetría y Curtosis podemos detectar la “no normalidad” de los datos (u otra forma) y obrar en consecuencia.

Coeficiente de Asimetría o Sesgo (Skewness)

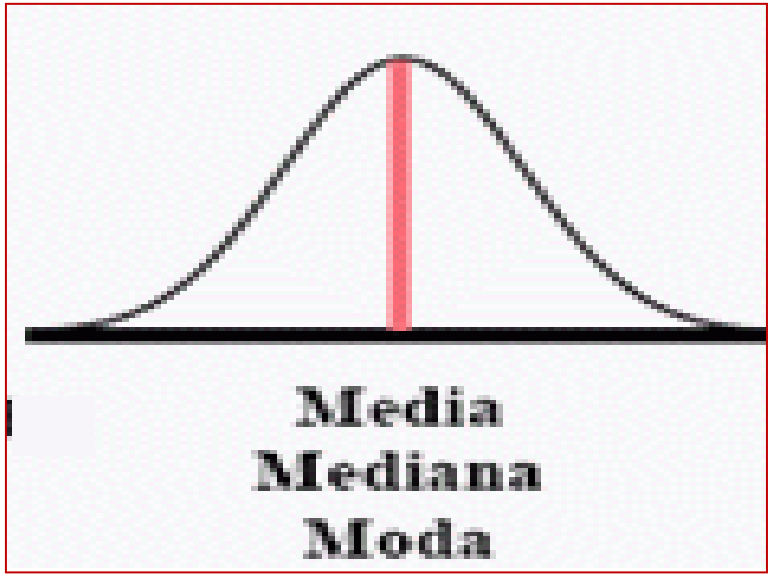
- **Permite cuantificar hasta qué punto (grado) las observaciones están dispuestas simétrica o asimétricamente respecto a un valor central de los valores de la v.a..**
- Hay diferentes coeficientes según el parámetro de centralización tenido en cuenta
- Todos los coeficientes que se utilizan son **números abstractos y, por tanto, adimensionales**
- Es el más usado es el de **Fisher** (si solo se aparece "Coeficiente de Asimetría", suponemos por defecto que se trata de éste) y su cálculo está basado en la diferencia de los datos sobre la media, elevando las diferencias al cubo para mantener los signos de las diferencias. Se divide por el cubo de la S para obtener un parámetro adimensional:

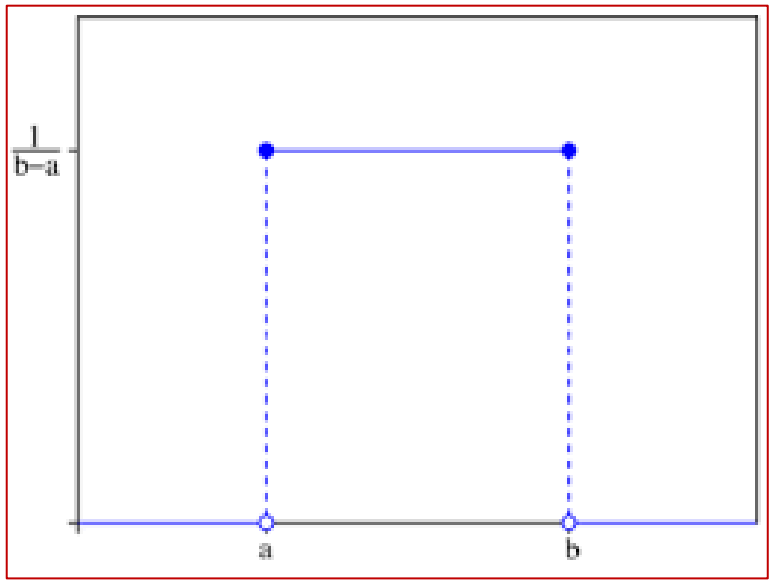
$$CA = \frac{\sum(X_i - \bar{X})^3 / (N-1)}{s^3}$$

- Si la distribución es simétrica CA será próximo a 0 (al contrario no es siempre cierto).
- Si la distribución es asimétrica positiva, CA será mayor que 0
- Si la distribución es asimétrica negativa, CA será menor que 0

Inconveniente: muy influida por valores extremos (ya que la media, que sirve como referencia, no es un parámetro robusto).

Simetría



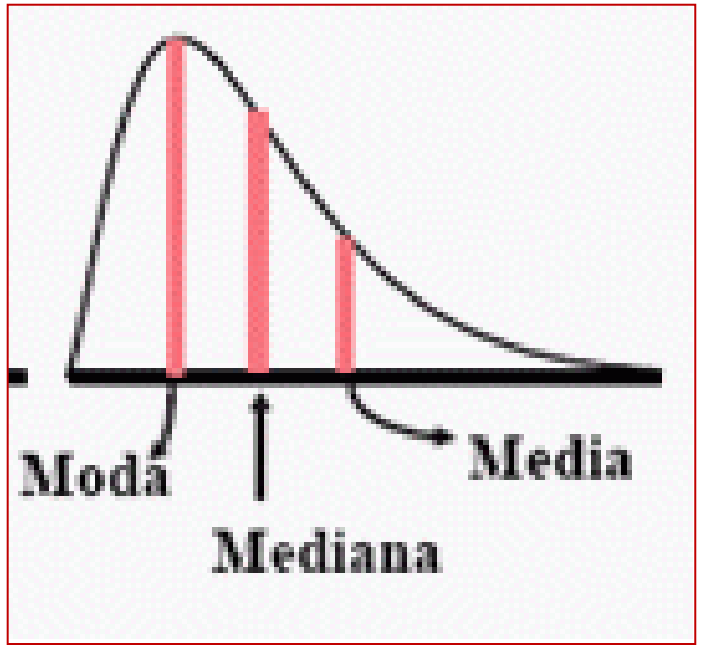


Se da cuando en una distribución hay, aproximadamente, los mismos valores distintos alejados de la media por la derecha de ésta que por la izquierda. El comportamiento de los datos es similar a ambos lados de la media.

No tiene alargamiento o sesgo o asimetría.

La media, mediana y moda son iguales (si hay una sola moda)

Asimetría positiva o por la derecha



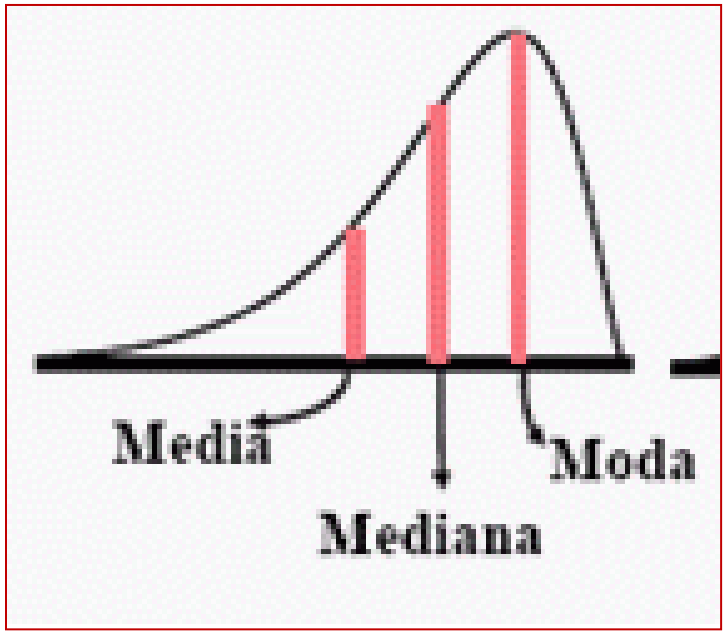
Ejemplos:

- Calificación de los alumnos en un examen difícil.
- Salarios de una profesión
- Tiempos de Reacción

Se da cuando en una distribución hay más valores distintos alejados de la media por la derecha de ésta que por la izquierda. La distribución presenta un alargamiento o sesgo hacia la derecha, es decir**, la distribución de los datos tiene a la derecha una cola más larga que a la izquierda**.

En este caso el valor de la **media aritmética es mayor que la mediana y éste a su vez es mayor que la moda**.

Asimetría negativa o por la izquierda



Ejemplo:

- Calificación de los alumnos en un examen fácil.

Se da cuando en una distribución hay más valores distintos alejados de la media por la izquierda de ésta que por la derecha. Este tipo de distribución presenta un alargamiento o sesgo hacia la izquierda, es decir, **la distribución de los datos tiene a la izquierda una cola más larga que a la derecha**.

El valor de la **media aritmética es menor que la mediana** y este valor de la **mediana, a su vez, es menor que la moda**.

Coeficiente de Asimetría Estandarizado de Fisher

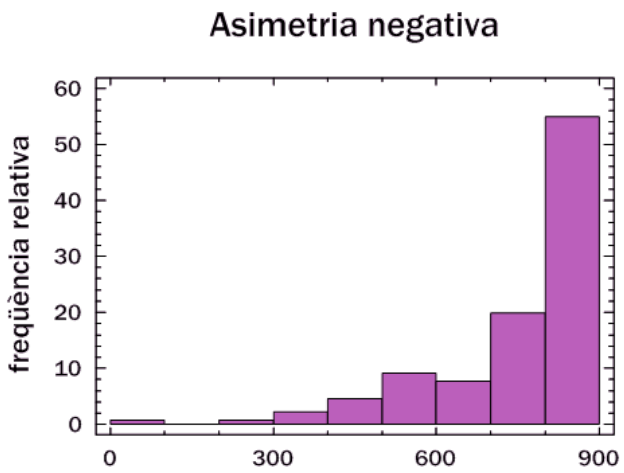
Aunque el CA es un parámetro adimensional, no está acotado y puede dificultar su interpretación.
Por ejemplo, CA = 0,2 ¿Es cercano a 0 y la distribución es simétrica?

Para evitar este problema, se utiliza el Coeficiente de Asimetría Estandarizado o Normalizado (CAE):

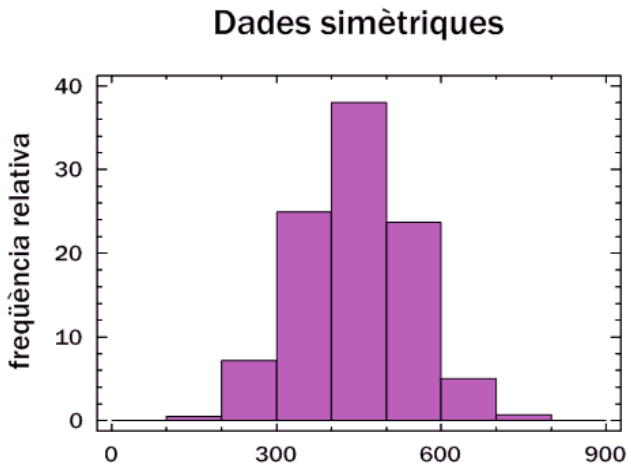
$$CAE = \frac{CA}{\sqrt{6/N}}$$

CAE $\in$ [-2, 2] → Datos simétricos.
CAE > 2 → Asimetría positiva (cola por la derecha)
CAE < -2 → Asimetría negativa (cola por la izquierda)

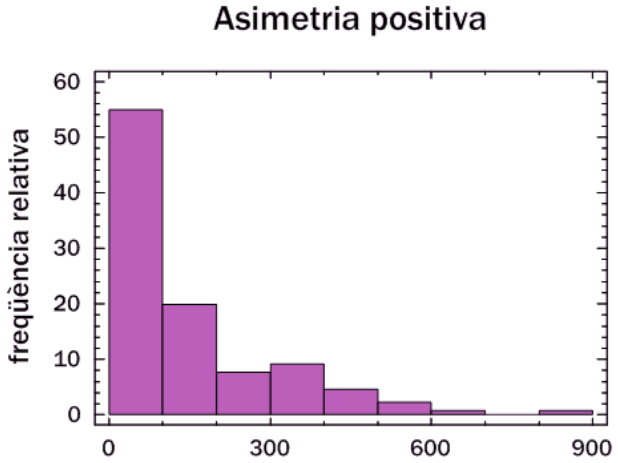
Ejemplo



CAE < -2



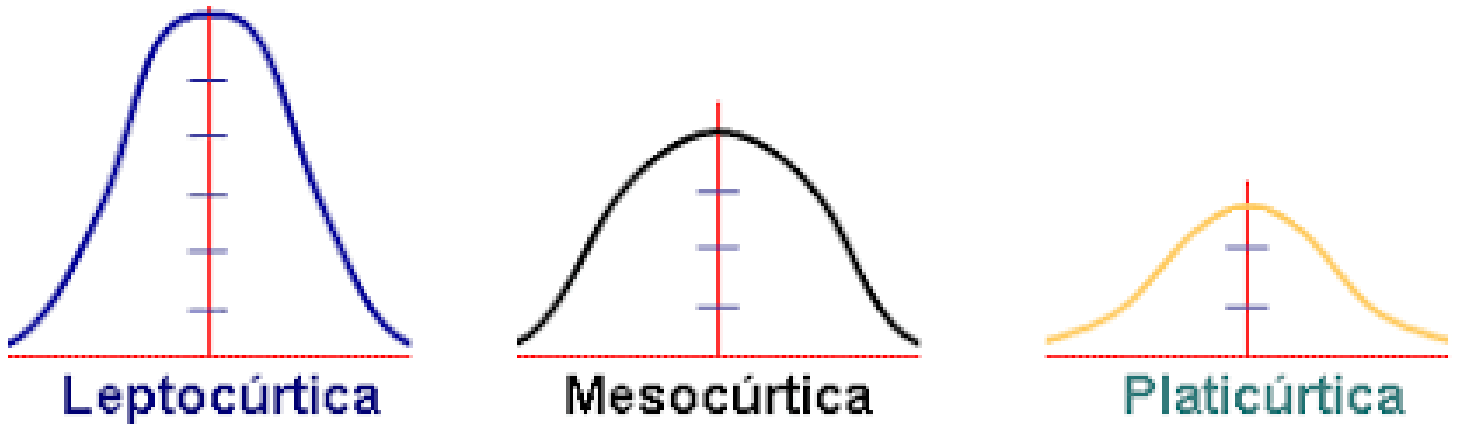
CAE ∈ [-2, +2]



CAE > +2

Coeficiente de Curtosis o Apuntamiento (*Kurtosis*)

- Dependiendo del número de observaciones que haya en la zona central de la distribución y del que haya en las zonas alejadas en dos distribuciones con la misma varianza los datos pueden tener perfiles distintos, con mayor o menor forma "de punta".
- **Al mayor o menor "apuntamiento" que puede tener una distribución, con independencia del valor que tome su varianza, se le llama Curtosis**, **Apuntamiento** o **Concentración Central** (***Kurtosis***)



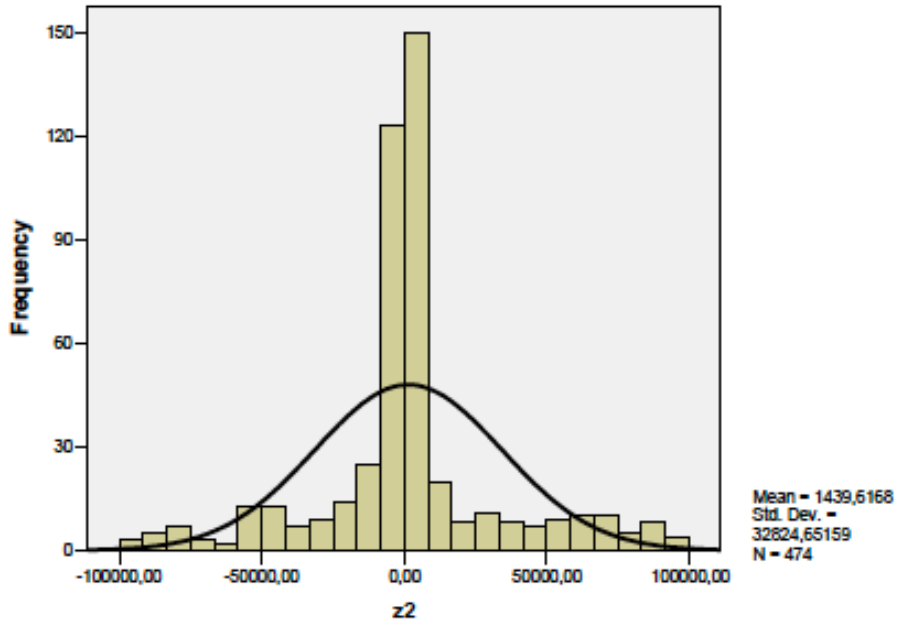
IMPORTANTE: Curtosis es independiente de la variabilidad (en el sentido de “varianza”)

Coeficiente de Curtosis o Apuntamiento (*Kurtosis*)

- Las medidas de apuntamiento tratan de estudiar la distribución de frecuencias de la zona central
- La mayor o menor concentración central respecto a la concentración de valores en los extremos de la distribución dará lugar a un mayor o menor apuntamiento o curtosis.
- **Estas medidas se deberían de aplicar únicamente a distribuciones simétricas, o con ligera asimetría**.
- En el caso de asimetría estaba claro cuándo una distribución era simétrica o no, pero en el caso de apuntamiento se suele tomar una distribución de referencia para comparar si el apuntamiento es mayor o menor.
- En general **la distribución de referencia suele ser la distribución Normal o Campana de Gauss**.

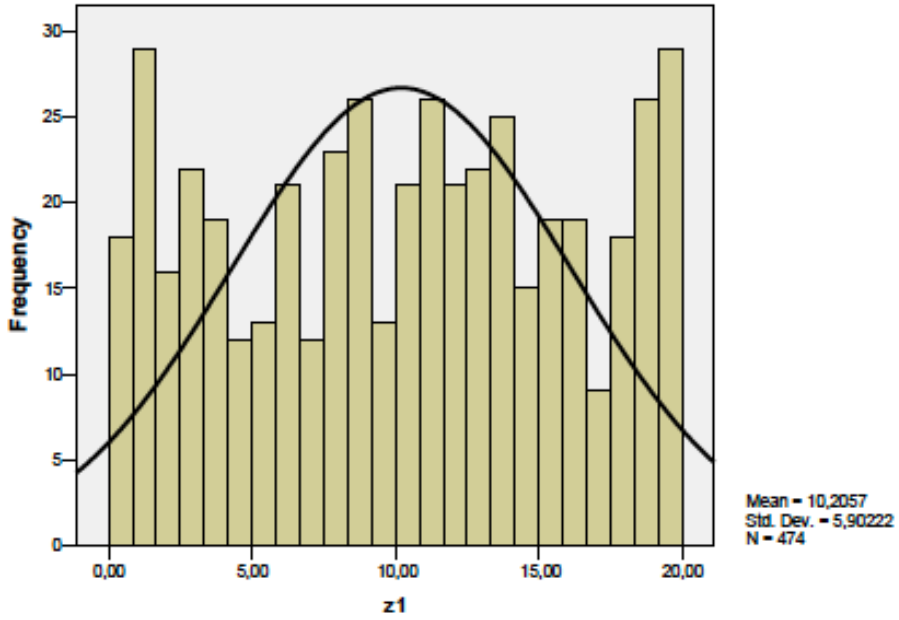
Datos Leptocúrticos

- **Presenta valores muy alejados de la media con mayor frecuencia de la que cabría esperar para unos datos normales que tuvieran la misma desviación típica**.
- Para compensar gráficamente estos valores extremos un histograma de datos leptocúrticos es más apuntado en las cercanías de la media que lo que lo sería el de unos datos normales con la misma desviación típica.
- Suele ser síntoma de observaciones anómalas: errores de transcripción o algún individuo perteneciente a una población distinta de la estudiada, etc.



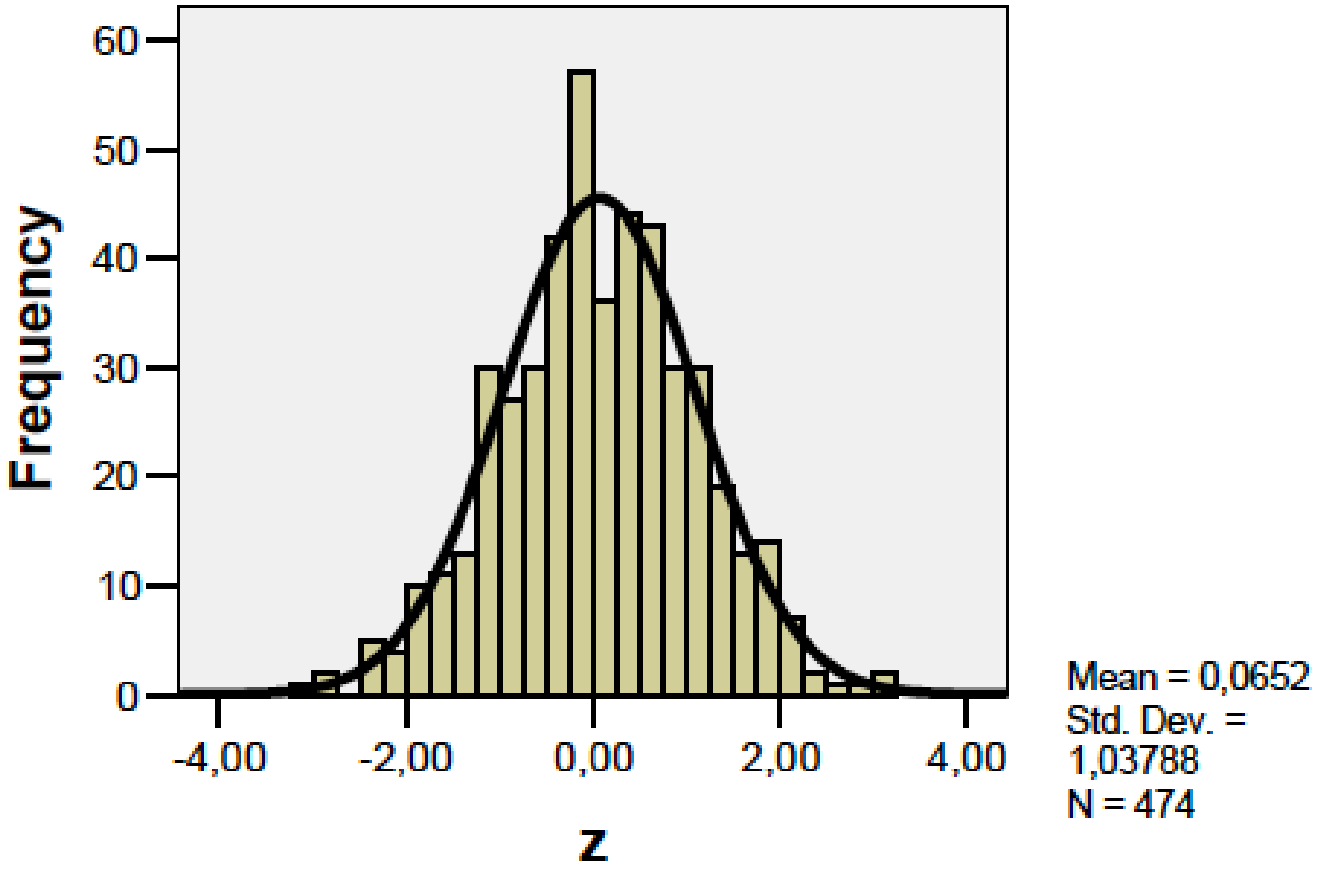
Datos Planicúrticos

- **Valores alejados de la media aparecen con una frecuencia menor que la que cabría esperar si los datos siguieran una distribución normal con la misma desviación típica**.
- Para compensar este hecho, el histograma de unos datos planicúrticos aparece más plano en el entorno de la media que lo que lo sería el de unos datos normales con idéntica varianza.
- Suele ser síntoma de que los datos han sido artificialmente *censurados* para eliminar los valores considerados extremos.



Datos Mesocúrticos

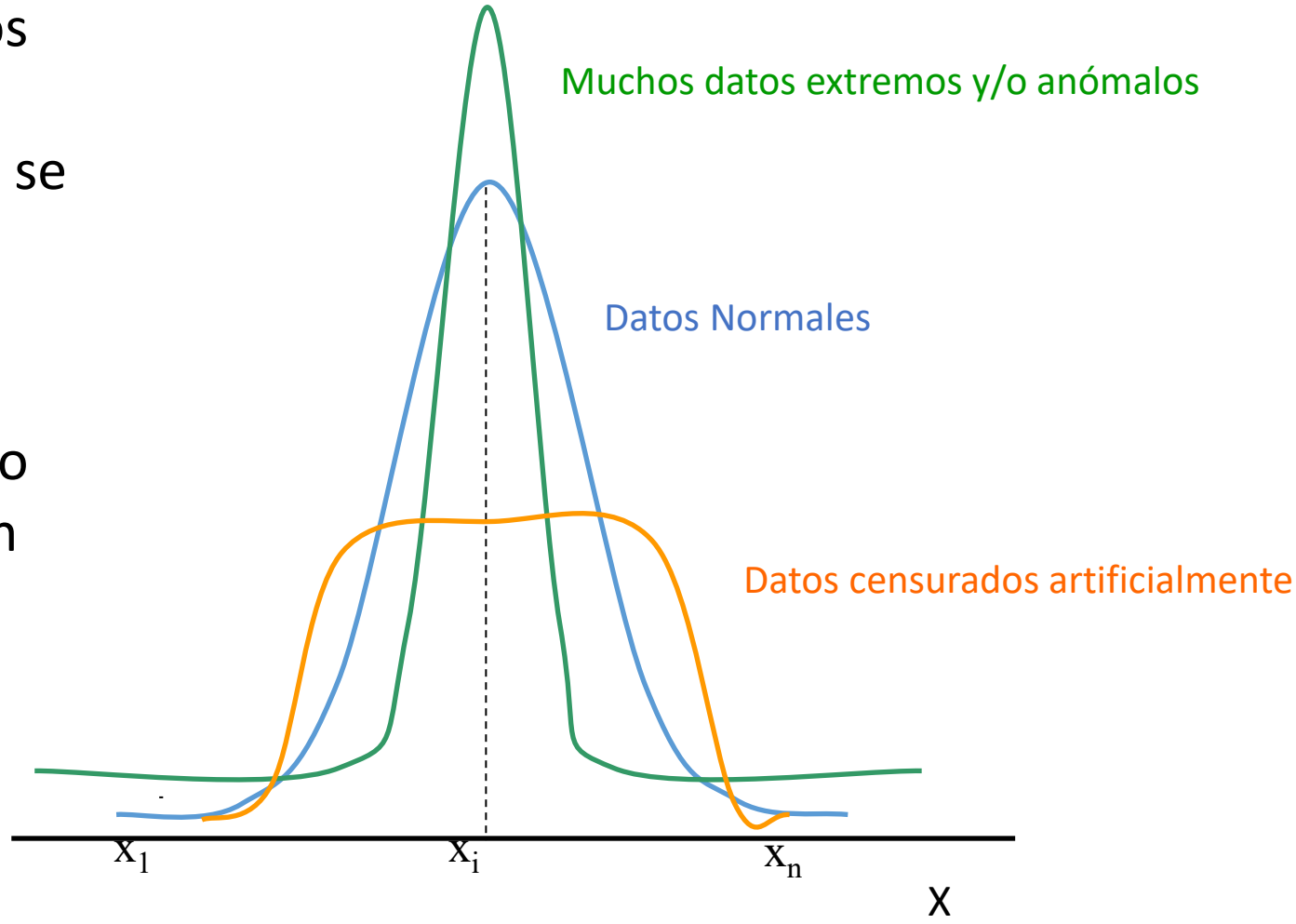
- **Presenta un grado de concentración medio alrededor de los valores centrales de la variable que debería ser el mismo que presenta una distribución normal con la misma desviación típica.**



Coeficiente de Curtosis o Apuntamiento (*Kurtosis*)

Si se representan tres grupos de datos con diferente curtosis en el mismo gráfico se vería así:

Estas diferencias de forma no se aprecian si se representan individualmente.



Coeficiente de Curtosis o Apuntamiento (*Kurtosis*)

- Cálculo[1]:

$$CC = \frac{\sum(X_i - \bar{X})^4 / (N - 1)}{s^4} - 3$$

Como nos interesa comparar (ponderadamente) el número de observaciones cercanas a la media con el número de observaciones lejanas (con independencia del signo de su distancia a la media), para medir la curtosis, consideramos las distancias a la media elevado a un número par; pero como la curtosis es el mayor o menor apuntamiento con independencia de la varianza, deberemos considerar las distancias a la 4.

Además, si queremos disponer de una medida valida para la comparación universal, el hecho de que las distancias a la cuarta dependan de las unidades es un inconveniente, por lo que dividimos por la desviación típica elevado a 4.

Por último, para que un apuntamiento normal tenga un CC de cero se le resta el 3.

Coeficiente de Curtosis o Apuntamiento (*Kurtosis*)

Aunque el CC es un parámetro adimensional, no está acotado y puede dificultar su interpretación, como ocurría con el CA.

Para evitar este problema se utiliza el Coeficiente de Curtosis Estandarizado o Normalizado (CCE):

$$CCE = \frac{CC}{\sqrt{24/N}}$$

CCE $\in$ [-2, 2] → datos "normales" o Mesocúrticos
CCE > 2 → más "apuntados" de lo normal o Leptocúrticos
CCE < -2 → datos "aplanados" o Planicúrticos

Normalidad

• Desde un punto de vista meramente descriptivo, podemos comprobar si nuestros datos se asemejan suficientemente a una "campana de Gauss" (distribución Normal) cuando

- Son **simétricos** → **CA = 0** ó **Std. Skewness ∈ [-2, 2]**

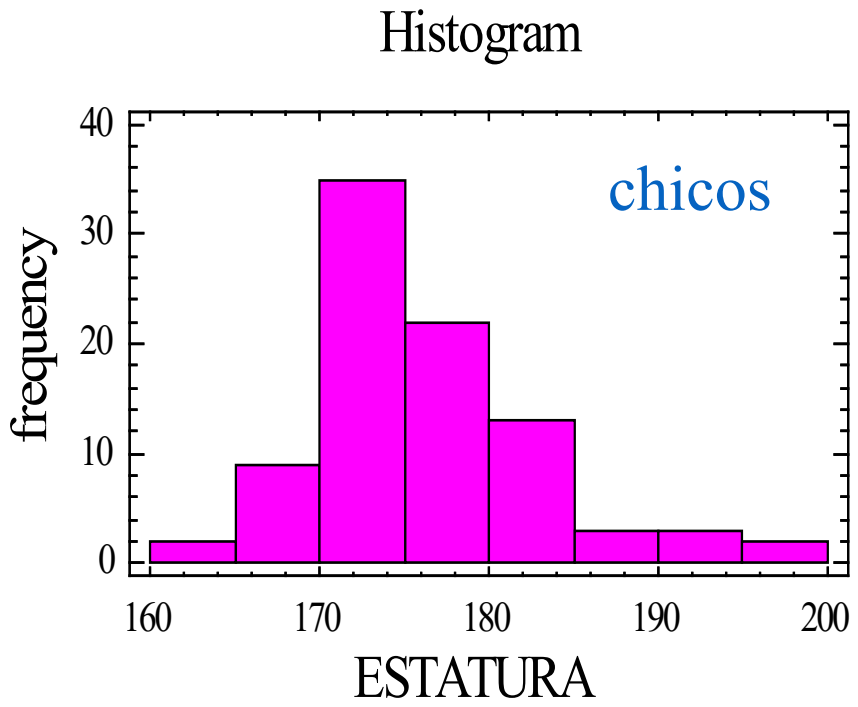
y

- Son **mesocúrticos** → **CC = 3** ó **CC = 0** ó **Std. Kurtosis ∈ [-2, 2]**

• Además, podemos estudiar el histograma o el diagrama Box & Whisker y comparar los parámetros de posición y dispersión.

Ejemplo

Estatura de los alumnos de 1er curso de Ciencia de Datos



CA-estandarizado = 3,38412
CC-estandarizado = 1,88624

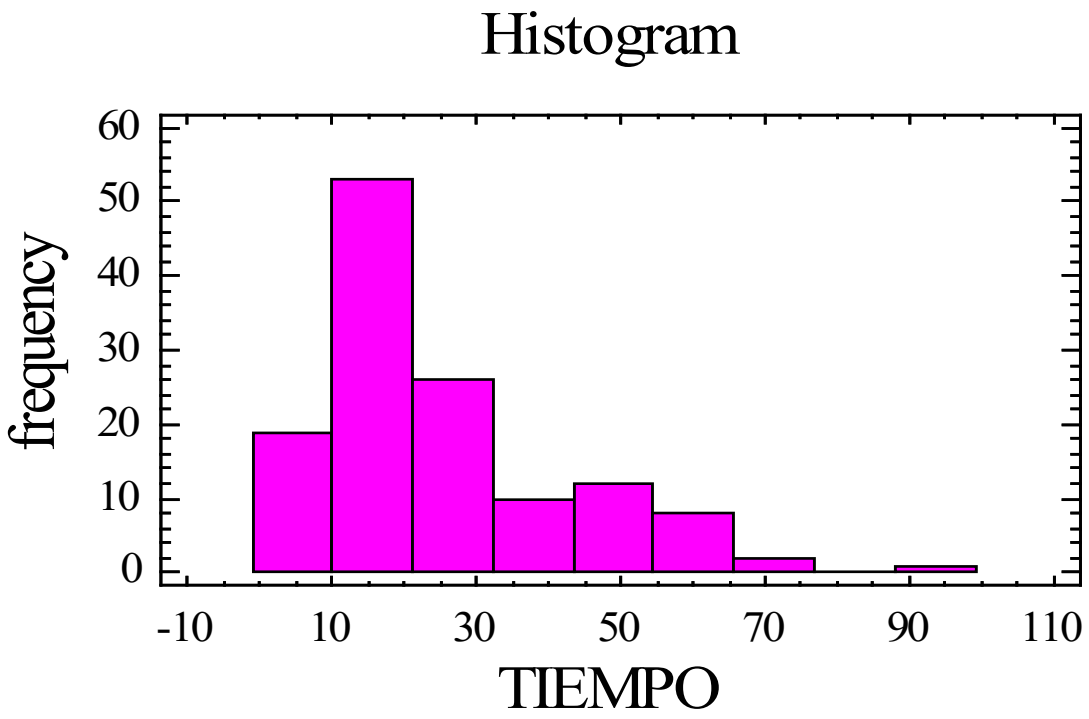
CA-estandarizado > 2 Asimetría positiva

CC-estandarizado $\in$ [-2, 2] Datos normales

No Distribución Normal

Ejemplo

Tiempo de bus para llegar a la UPV de los alumnos de 1er curso de Ciencia de Datos



CA-estandarizado = 5,90912
CC-estandarizado = 3,31496

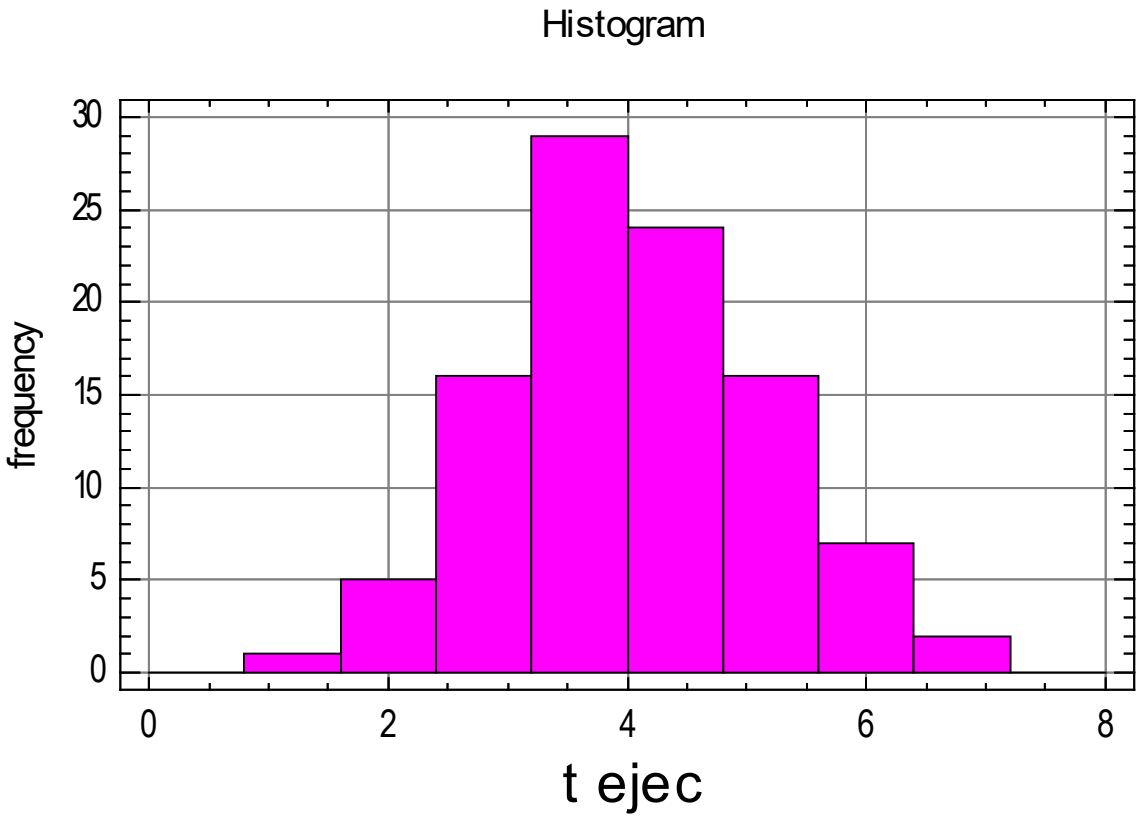
CA-estandarizado > 2 Asimetría positiva

CC-estandarizado > 2 Datos extremos

No Distribución Normal

Ejemplo

Distribución Normal



CA-estandarizado = 0,587782
CC-estandarizado = -0,39427

CA-estandarizado $\in$ [-2, 2] Simetría

CC-estandarizado $\in$ [-2, 2] Datos normales

Coeficiente de correlación lineal r

Parámetro adimensional que toma valores entre -1 y 1 y que mide el grado de asociación lineal entre las dos componentes X e Y de una v.a. bidimensional

$r_{XY} \in [-1,1]$

Si $r_{XY} \approx 0 \Leftrightarrow$ No hay relación LINEAL

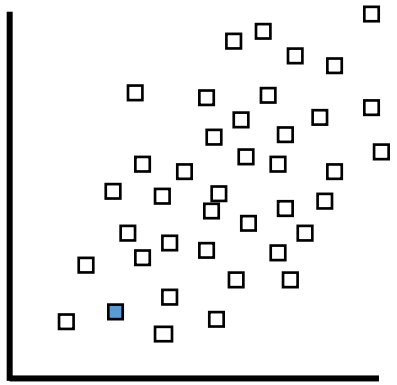
¡ puede haberla de otro tipo !

Si $r_{XY} \approx 1 \Leftrightarrow$ relación LINEAL directa

Si $r_{XY} \approx -1 \Leftrightarrow$ relación LINEAL inversa

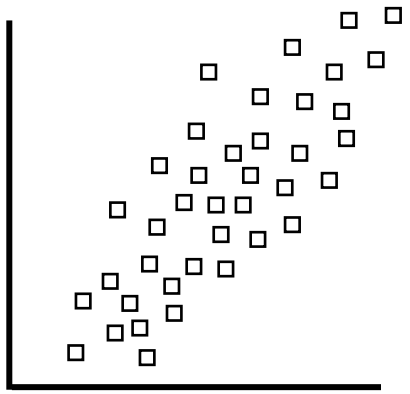
Diagramas de Dispersión y r_{XY}

$r_{XY} \in]0, 0{,}3]$



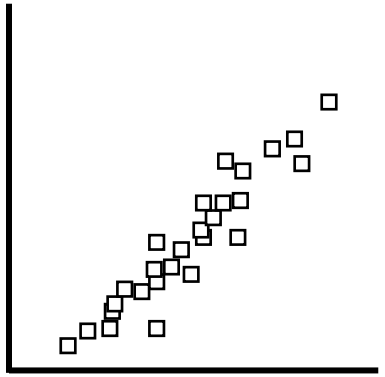
Relación lineal Débil o
No relación

$r_{XY} \in]0{,}3, 0{,}8[$



Relación lineal
Intermedia

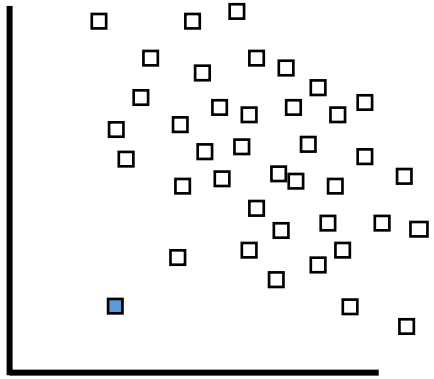
$r_{XY} \in [0{,}8, 1[$



Relación lineal
Fuerte

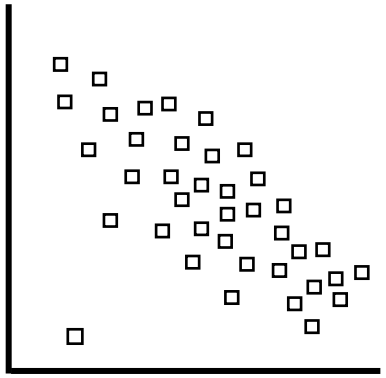
Relación positiva

$r_{XY} \in [-0{,}3, 0[$



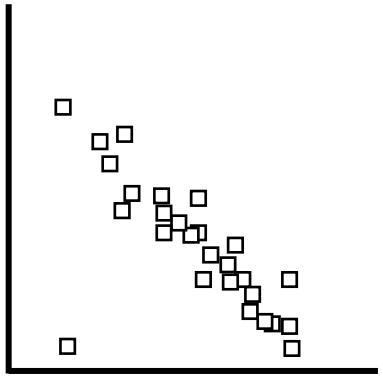
Relación lineal Débil o
No relación

$r_{XY} \in]-0{,}8, -0{,}3[$



Relación lineal
Intermedia

$r_{XY} \in]-1, -0{,}8]$

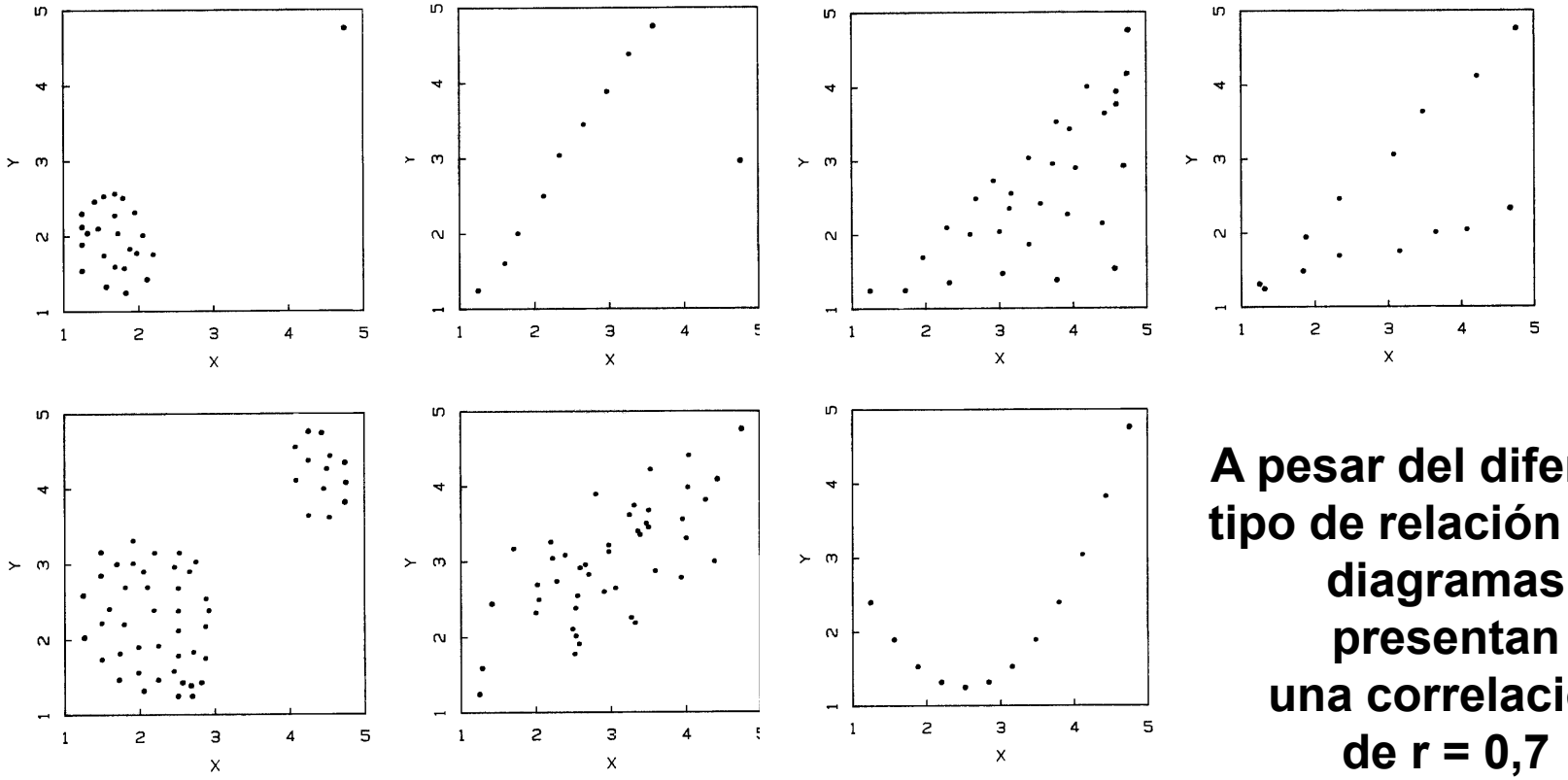


Relación lineal Fuerte

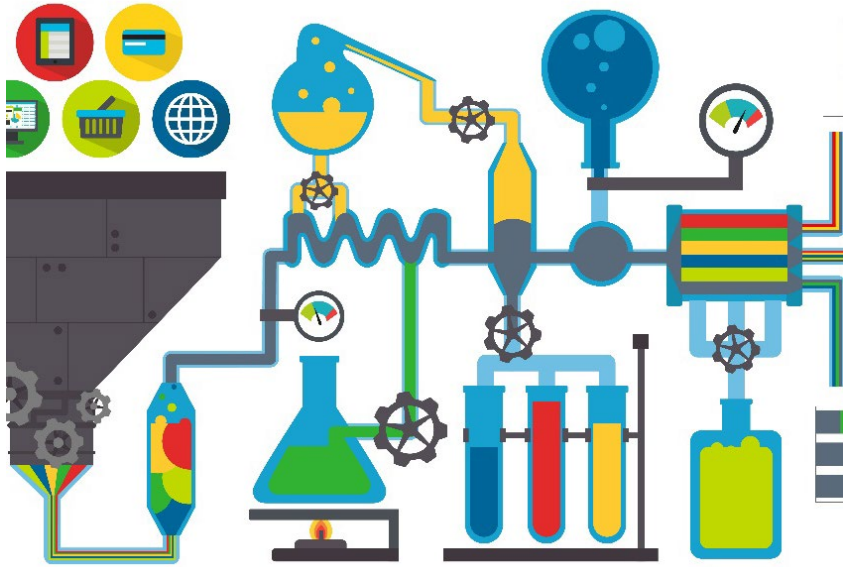
Relación negativa

Diagramas de Dispersión y r_{XY}

El valor de r no sustituye la información del diagrama bivariante



A pesar del diferente tipo de relación los 7 diagramas presentan una correlación de r = 0,7



Limpieza y preparación de datos

Datos técnicamente correctos

Datos técnicamente correctos

- Eliminar registros duplicados
- Separar varios valores contenidos en el mismo campo
- Homogeneizar valores posibles de una variable
  - Ej va. cualitativa: “peso”, “PESOS”, “kilos” → “PESO”
  - Homogeneizar mayúsculas / minúsculas
  - Unificar unidades en variables cuantitativas
- Revisar tipos de variables: numéricas, categóricas
- **Ejemplo**

Transformaciones

Propósito

Clarificar y simplificar la descripción de los datos tanto como sea posible:

- Homogeneizar dimensiones o unidades de las v.a. para facilitar su comparación
- Adecuar la escala de valores de las v.a. para mejorar su interpretación y la de las posibles relaciones entre éstas
- Corregir la forma de la distribución
- Recodificar
- Binarizar atributos categóricos
- Manejo de escalas Likert

Tipos

Según tipo de variable

- Variables cuantitativas
- Variables cualitativas

Según dimensión variable

- Univariantes
- Multivariantes

Según función utilizada

- Lineales
- No lineales

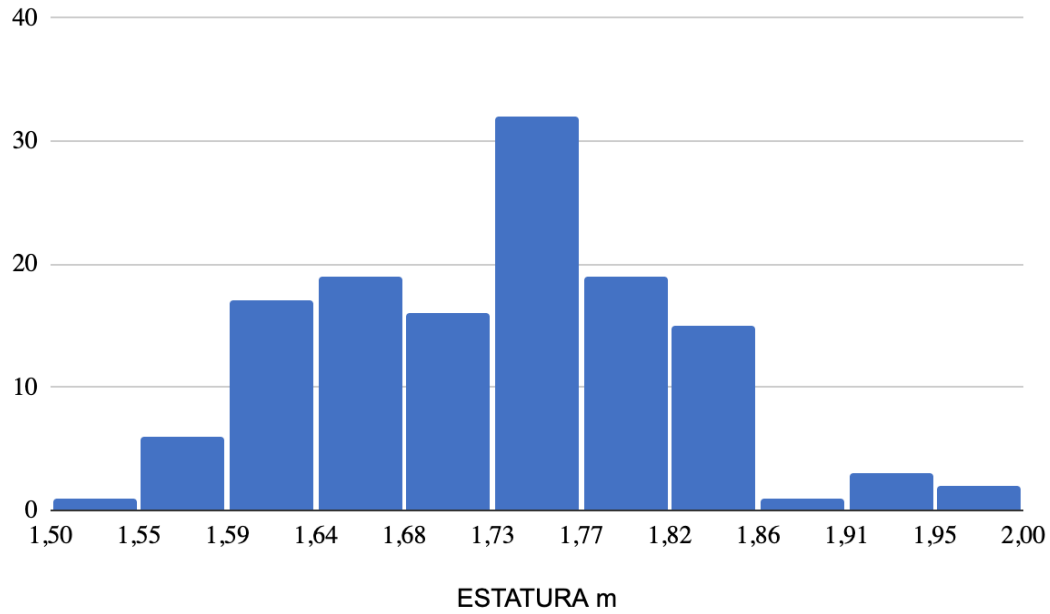
Transformaciones lineales para variables cuantitativas

- **Función general**: $\boxed{Y = a + bX}$ siendo X la variable a transformar

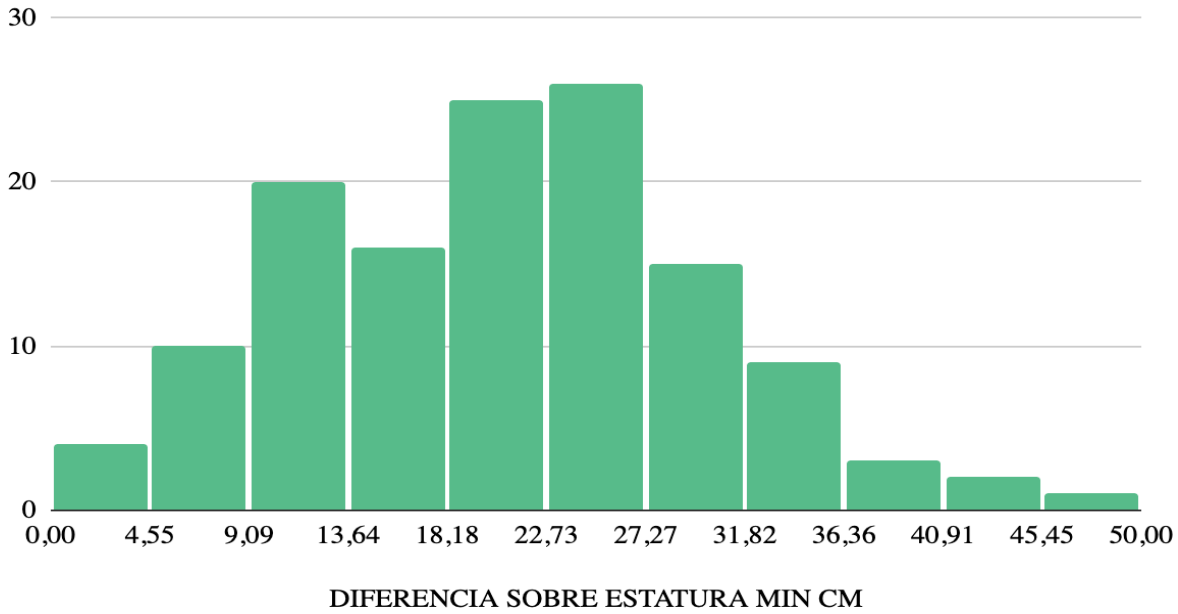
- **Utilidad**:
  - Cambio de escala.
  - Normalización de unidades.

- **Ejemplo**: variable **ESTATURA** de alumnos medida en metros con valores de 1, 52 a 1,98 (1,83; 1,85; 1,65; …)
  - En general, en la descripción inicial de los datos conviene representarlos con únicamente 2 o 3 dígitos, escogiendo apropiadamente las unidades.
  - Si calculamos y representamos las diferencias sobre el mínimo, medidas en cm, que conduce a valores 31; 33; 13; …
  - Además de conseguir una representación más simple, se aumenta la precisión de los resultados.

Ejemplo



DIFERENCIAS= $100(ESTATURA\ m - 1,52)$



Tipificación, normalización o estandarización

- Es una **transformación lineal** de enorme utilidad **para comparar variables que están expresadas en unidades diferentes**.
- **Función**:

$$Z = \frac{X - \bar{X}}{S_X}$$

- A cada valor de la variable original X se le resta la media aritmética y se divide por la desviación típica, de esta forma se obtiene una nueva variable cuya media aritmética de 0 y su desviación típica de 1.
- **Interpretación**:
  - **Signo**: el valor de Z es positivo si el valor de la variable original se sitúa por encima del promedio y negativo si lo hace por debajo
  - **Valor absoluto**: indica la dispersión, en relación con la desviación típica.
  - **Ej**: Si un valor z_i es -1,05, quiere decir que en la v.a. estudiada el individuo i se sitúa por debajo de la media y además, lo hace en 1,05 veces la desviación típica de dicha característica.

Ejemplo: Factores salud materna-neonatos

ID	PESOM kg	TALLAM cm	SEM	PASM mmHg	PADM mmHg
1	59	160	39	150	90
2	65	166	38	160	100
3	68	173	38	100	55
4	71	176	40	95	50
5	56	165	37	115	60
6	46	155	39	90	60
7	65	170	41	150	70
8	68	162	39	170	110
9	48	152	36	105	55
10	68	174	37	100	65
Media	**61,40**	**165,30**	**38,40**	**123,50**	**71,50**
Desv Típ	**8,82**	**8,12**	**1,51**	**30,46**	**20,96**

Datos originales

ID	PESOM	TALLAM	SEM	PASM	PADM
1	-0,27	-0,65	0,40	0,87	0,88
2	0,41	0,09	-0,27	1,20	1,36
3	0,75	0,95	-0,27	-0,77	-0,79
4	1,09	1,32	1,06	-0,94	-1,03
5	-0,61	-0,04	-0,93	-0,28	-0,55
6	-1,75	-1,27	0,40	-1,10	-0,55
7	0,41	0,58	1,73	0,87	-0,07
8	0,75	-0,41	0,40	1,53	1,84
9	-1,52	-1,64	-1,59	-0,61	-0,79
10	0,75	1,07	-0,93	-0,77	-0,31
Media	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**
Desv Típ	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**

Datos normalizados

Transformaciones no lineales para variables cuantitativas

- **Funciones habituales**:

$$Y = X^2$$

$$Y = \sqrt{X}$$

$$Y = ln\ X$$

$$Y = \frac{1}{X}$$

…

siendo X la variable a transformar

- **Utilidad**:
  - Integrar en una misma variable valores de una característica que se ha medido de diferente modo y la relación que liga las distintitas formas de medirla no es lineal.
  - Obtener distribuciones simétricas.
  - Analizar tasas de variación (estudios de temperaturas, rentas, …)

Ejemplos

- **Ejemplo 1**: **UNIDADES**
  - variable **CONSUMO de gasolina de un automóvil**
  - Se tienen datos de
    - Coches europeos, medido en litros / 100 km (X)
    - Coches norteamericanos, medido en km / 1 litro o galón (Y)
  - Para integrar los valores en una misma variable se debe usar una única forma de medir y en este caso, por ejemplo

$$Y = \frac{100}{X} = CONSUMO_{USA} = \frac{100}{CONSUMO_{CE}}$$

- **Ejemplo 2**: **TASAS DE VARIACIÓN**
  - Para comparar el crecimiento de consumo de energía en distintos países se puede usar las diferencias absolutas (C_t-C_{t-1}), pero es más relevante estudiar las relativas:

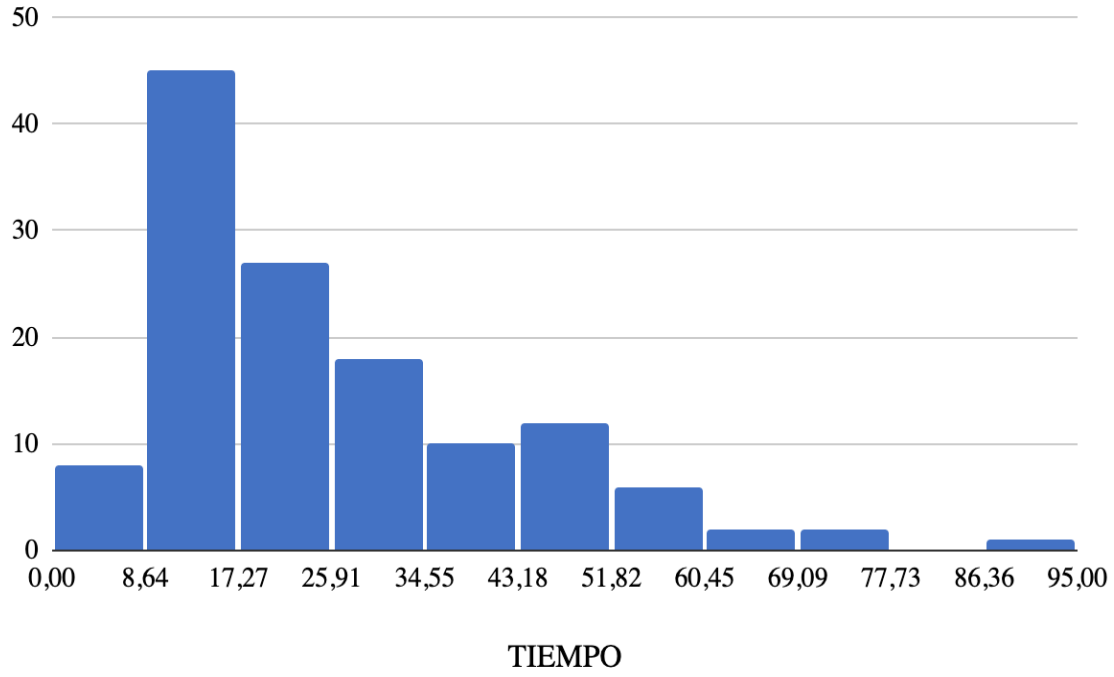
    ((C_t-C_{t-1})/C_{t-1}) o ((C_t-C_{t-1})/C_t)

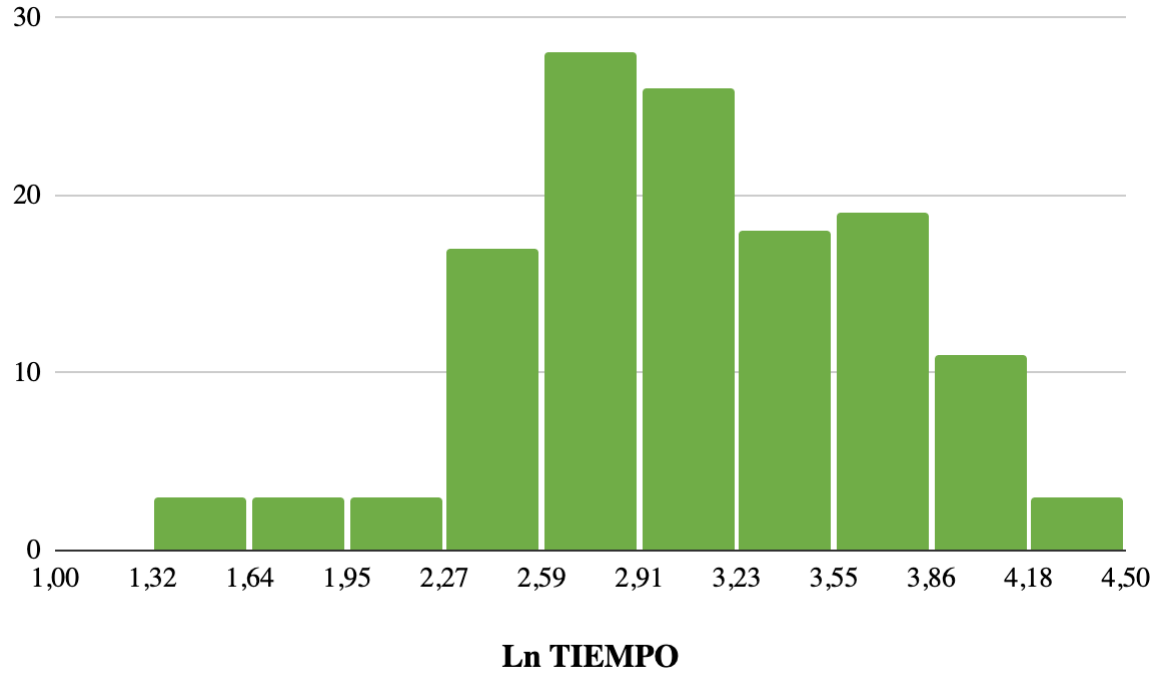
  - Si previamente expresamos el consumo en logaritmos, las diferencias absolutas se corresponden con el promedio de las dos formas de diferencias relativas

$$\ln C_t - \ln C_{t-1} \qquad \frac{C_t - C_{t-1}}{C_t} \leqslant \ln \frac{C_t}{C_{t-1}} \leqslant \frac{C_t - C_{t-1}}{C_{t-1}}$$

Ejemplos

- **Ejemplo 3**: **FORMA DISTRIBUCIÓN**
  - Conseguir que una variable asimétrica tenga una forma normal





Datos originales con Asimetría +

Datos transformados con log().
Más simétricos

Transformaciones más utilizadas

- Si se tienen distribuciones de frecuencias con **asimetría negativa** (frecuencias altas hacia el lado derecho de la distribución), es conveniente aplicar una transformación que comprima la escala para valores pequeños y la expande para valores altos:

$$Y = X^2$$

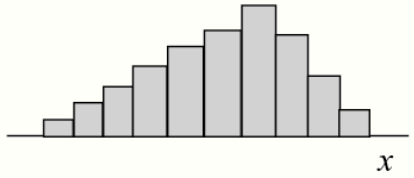
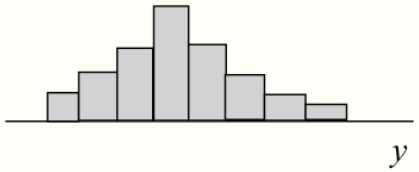
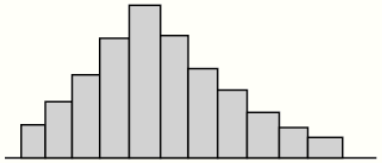
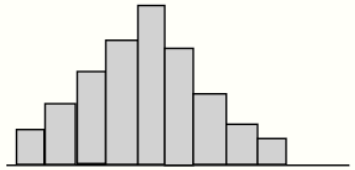
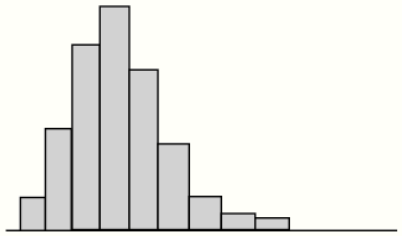
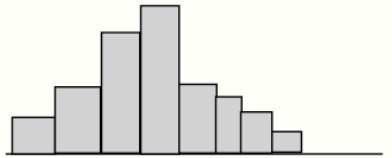
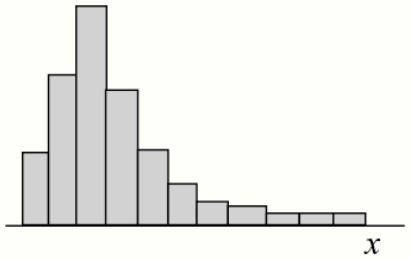
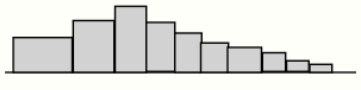
- Para distribuciones con **asimetría positiva** se usan las transformaciones que compriman los valores altos y expanden los pequeños:

$$Y = \sqrt{X} \qquad Y = \ln(X) \qquad Y = {}^{1}/_{X}$$

- El efecto de estas transformaciones está en orden creciente: menos efecto √x, más ln(x) y más aún 1/x.

- La transformación logarítmica también se usa con frecuencia para describir **variables que representan "tamaños"**: ciudades en el mundo, tamaño de empresas, distribuciones de rentas, consumos de electricidad,…

Transformaciones más utilizadas

Histograma inicial	Transformación	Histograma transformado
	$y = x^2$	
	$y = \sqrt{x}$	
	$y = \ln x$	
	$y = \frac{1}{x}$	

Eficacia de una transformación no lineal

- El **efecto** de una transformación depende del rango de los datos, si éste no cumple ciertas condiciones las transformaciones puede ser inapreciables, ya que cualquier transformación no lineal es aproximadamente lineal en un rango suficientemente pequeño

  - **Regla general 1**:

$$C1 = \frac{Max(X)}{Min(X)}$$

    - Si C1 < 2 → transformación no apreciable, no efecto
    - Si **2 ≤ C1 ≤ 10 → transformación eficaz** (perceptible)
    - Si C1 > 10 → transformación con efecto acusado

  - **Regla general 2**:

$$C2 = \frac{\bar{X}}{S}$$

    - Si C2 ≥ 4 → transformación no apreciable, no efecto
    - Si **C2 < 4 → transformación eficaz** (perceptible)

Discretización de variables cuantitativas

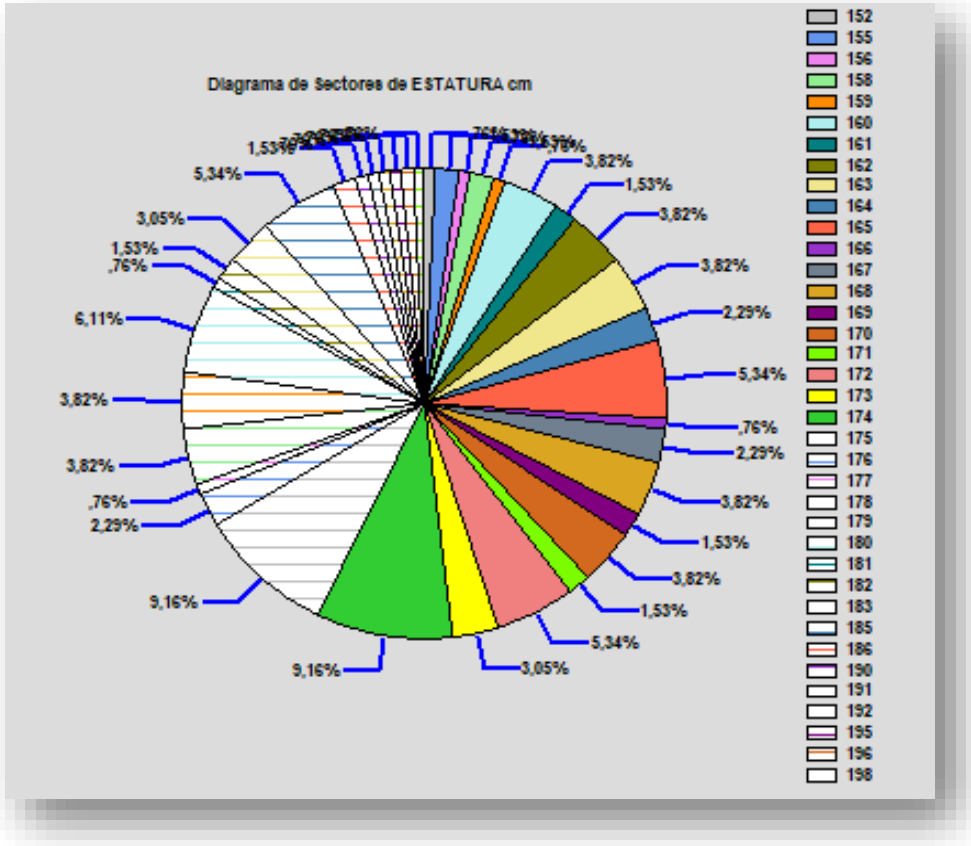
- **Utilidad**:
  - Transformar una variable cuantitativa en una cualitativa para poder usar las herramientas exploratorias correspondientes, dividiendo el rango de valores en tramos y asignándole a cada tramo o intervalo un valor o cadena de caracteres.

- **Ejemplo**:
  - variable **ESTATURA** de alumnos medida en cm con valores de 151 cm a 198 cm
  - La dividimos en 4 tramos y le asignamos un nuevo valor, por ejemplo:

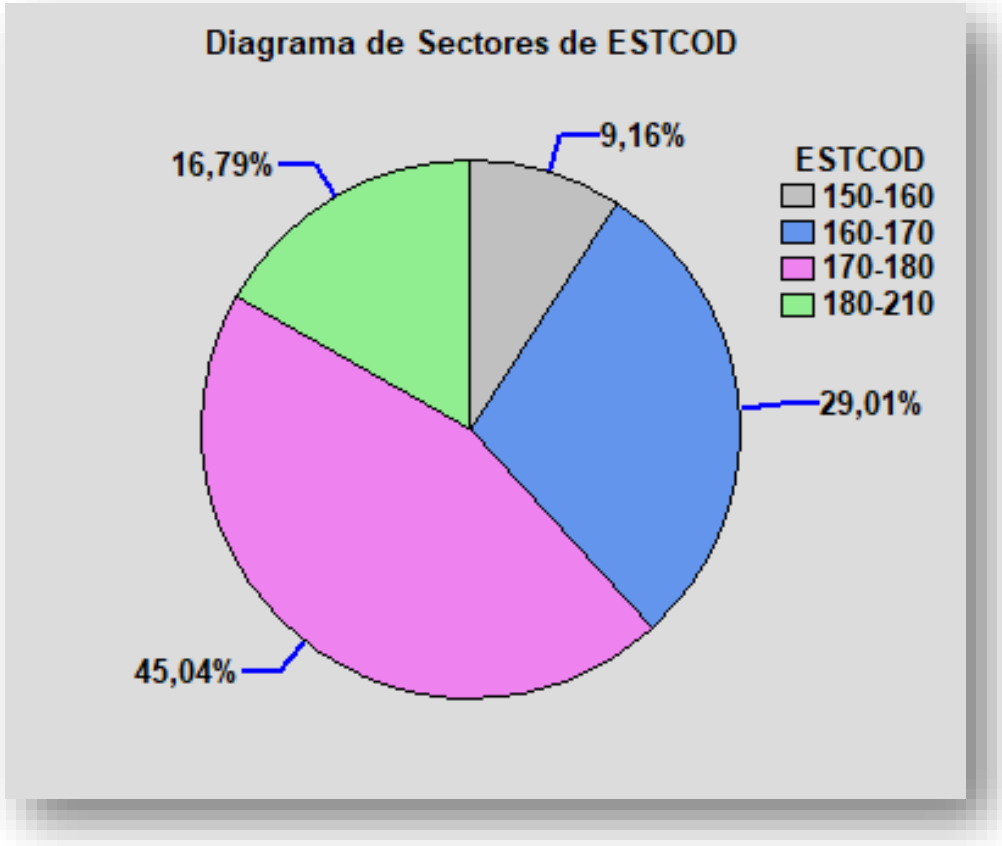
Límite inferior	***Límite superior***	***Nuevo valor***
150	*160*	*150-160*
160	*170*	*160-170*
170	*180*	*170-180*
180	*210*	*180-210*

  - La nueva variable pasa a ser cualitativa categórica con 4 valores posibles

Ejemplo



Variable original



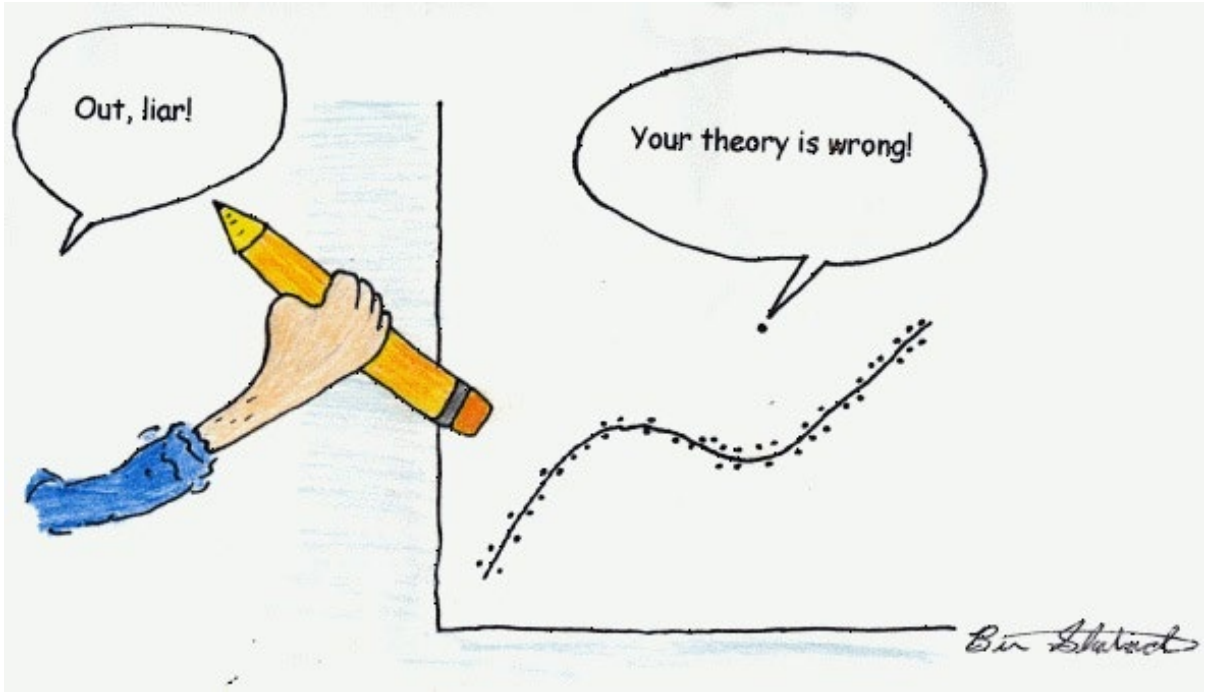
Variable discretizada

Consideraciones

La transformación puede cambiar la interpretación de las variables y de los resultados del análisis.

Hay que asegurarse la exploración de de todas las posibles interpretaciones de las variables transformadas.

Valores atípicos



Valores atípicos, anómalos, “extremos” o *outliers*

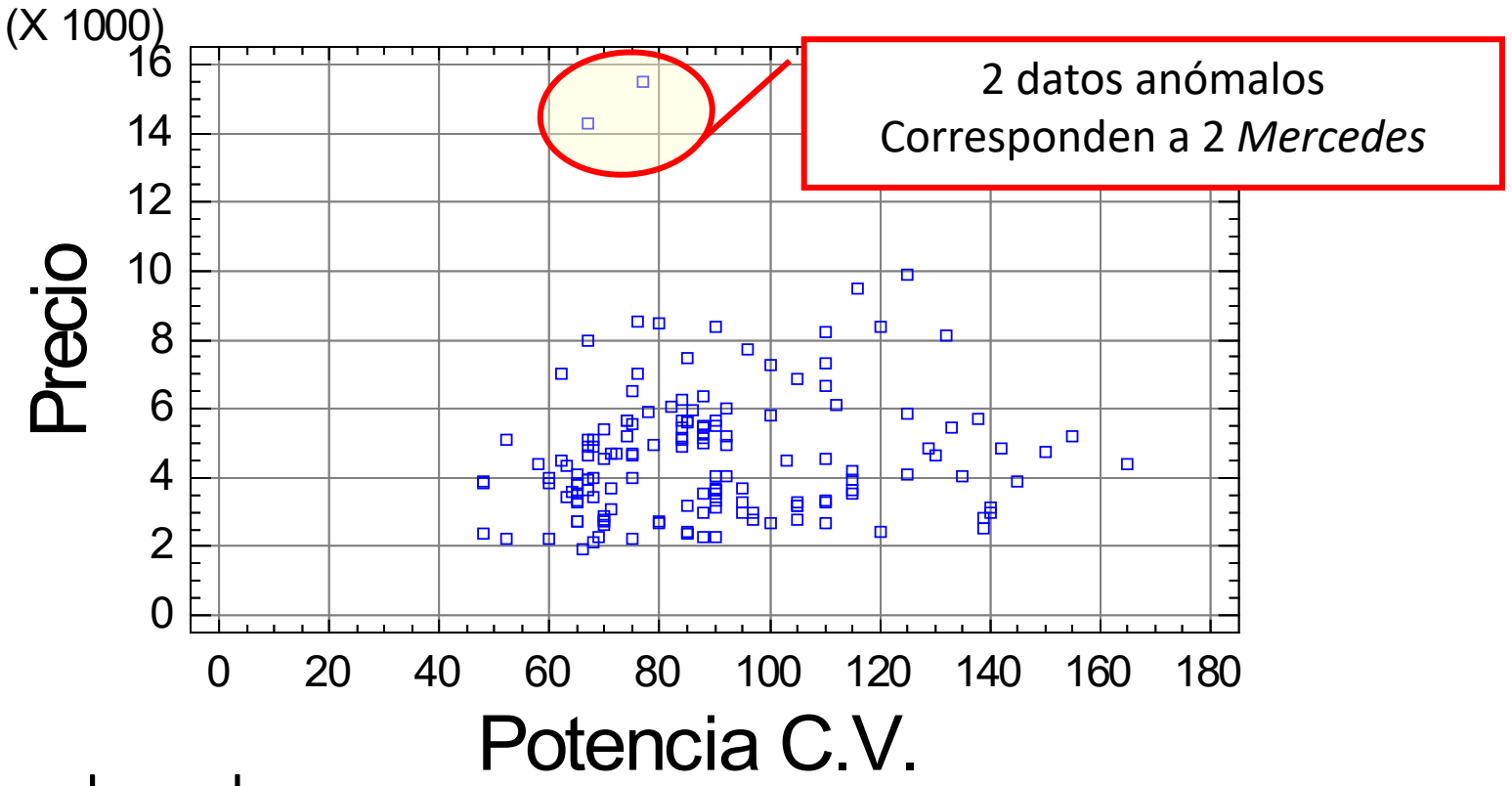
1. Qué son, cómo se originan y cuál es su impacto
2. Tipos de valores atípicos
3. Detección de valores atípicos
4. Tratamiento

Objetivo:

- Destacar la importancia de la existencia de valores atípicos y su impacto
- Conocer la terminología y conceptos fundamentales
- Conocer los tipos de valores atípicos
- Conocer y aplicar algunos de los procedimientos para su identificación y tratamientos más sencillos

¿Qué son?

Los **datos atípicos, anómalos** o ***outliers*** son aquellas observaciones con unas características muy diferentes, en todas o en algunas de las variables analizadas, al resto de observaciones.



Muestra de coches

¿Existen de verdad?

- Según Peña, D. (2014)[1], en datos recogidos con un estrecho control suelen aparecer entre un 1% y un 3% de valores atípicos, sin tanto control la **proporción puede llegar a ser del 5% o superior**.

- **Otras denominaciones** según el ámbito de aplicación en el que se desarrolla el análisis **:**

  - **aberraciones**
  - **extremos**
  - **ruido**
  - anomalías
  - observaciones discordantes
  - excepciones
  - fallos
  - defectos
  - errores
  - daños
  - sorpresas
  - etc

[1]Peña, D. (2014). Fundamentos de estadística. Madrid: Difusora Larousse - Alianza Editorial

Impacto

Los valores atípicos no son “malos” en si mismos.

- **El valor atípico puede formar parte de la variabilidad intrínseca de las variables estudiadas**.

- **Su presencia puede ayudar a obtener información útil**: fraudes con tarjetas de crédito, fallos de un sistema, alarmas meteorológicas, etc. En este caso su presencia es el objetivo del estudio.

- **Puede ser problemático** cuando:
  - Su **detección NO es el objetivo del análisis y**
  - El **valor atípico no es representativo de la población**:
    - Sesgan los resultados de los análisis estadísticos distorsionando estimaciones de parámetros y modelos
    - Impiden la aplicación de alguna técnica estadística por incumplimiento de alguna hipótesis como la normalidad

Evaluación del impacto

Precisar el impacto de los datos ausentes es esencial para determinar su tratamiento: eliminar o mantener

- **Criterios cuantitativos**
  - **Cuantificación de la desviación del valor** potencialmente atípico respecto a los datos considerados no anómalos.
  - **Establecimiento de límites** de valores basados en distintas reglas

- **Criterios cualitativos**
  - ¿Qué los origina? ¿De qué tipo son?
  - ¿Cuál es el objetivo del análisis?
  - ¿Cuál es el campo de aplicación del análisis?
  - ¿Afectan a algún parámetro que necesitamos en el análisis sobre el cual vamos a extraer conclusiones?

Tipos

Según dimensión

Unidimensionales

Multidimensionales

Según causa

Errores de procedimiento

Acontecimientos extraordinarios

Valores extremos

Causas desconocidas

Tipos según causa

1. **Errores de procedimiento**
   - En una variable en la que se mide el peso de un alumno en kg aparece el valor 900.
   - Se revisan las notas y nos damos cuenta de que debía aparecer el valor 90.
   - *Son errores*

2. **Acontecimientos extraordinarios**
   - En un estudio de comparación de condiciones meteorológicas en dos regiones se mide, entre otras cosas, la precipitación diaria a lo largo de varios años.
   - En una de las regiones un año determinado llueve de manera inusitada produciendo valores de precipitación extraordinariamente altos.
   - *No forman parte de la variabilidad intrínseca de la v.a., según su comportamiento habitual*.

Tipos según causa

3. Valores extremos

- En un estudio sobre tabaquismos se mide el número de cigarrillos consumidos a diario.
- En la muestra aparece el valor 60 porque hay un fumador que fuma sesenta cigarrillos al día, que es mucho más de lo que fuma el resto de la muestra, pero puede ser representativo de ésta.
- *Forman o pueden formar parte de la variabilidad intrínseca de la v.a., según su comportamiento habitual*.

4. Causas desconocidas

Detección

Consiste en utilizar reglas o criterios comúnmente conocidos y aceptados que permiten clasificar una observación como sustancialmente alejada del resto.

- **Procedimientos**:
  - **Herramientas gráficas exploratorias**
  - **Reglas o criterios mediante los que se obtienen límites** basados en el rango de valores esperado de una variable
    - **Límites internos**: atípicos leves
    - **Límites externos**: atípicos extremos
    - **Otros límites intermedios**
- **Dependen de si la detección se realiza a nivel**:
  - **Univariante**
  - **Bivariante**
  - **Multivariante**

Detección

No son reglas exactas, son criterios que ayudan a tomar decisiones.

Cualquiera que sea el procedimiento empleado siempre debe estar acompañado de una evaluación cualitativa

En este curso nos centramos en la detección cuantitativa univariante fundamentalmente y, someramente, en la bivariante

En cualquier caso, se usarán exclusivamente herramientas exploratorias, no de inferencia estadística

Detección univariante

- **Herramientas gráficas genéricas el AED**

  De más a menos adecuado:

  1. **Gráfico de Aberrantes o Atípicos**
  2. **Box & Whisker**
  3. **Gráfico de probabilidad Normal** o Gráficos Q-Q, …
  4. **Histograma**

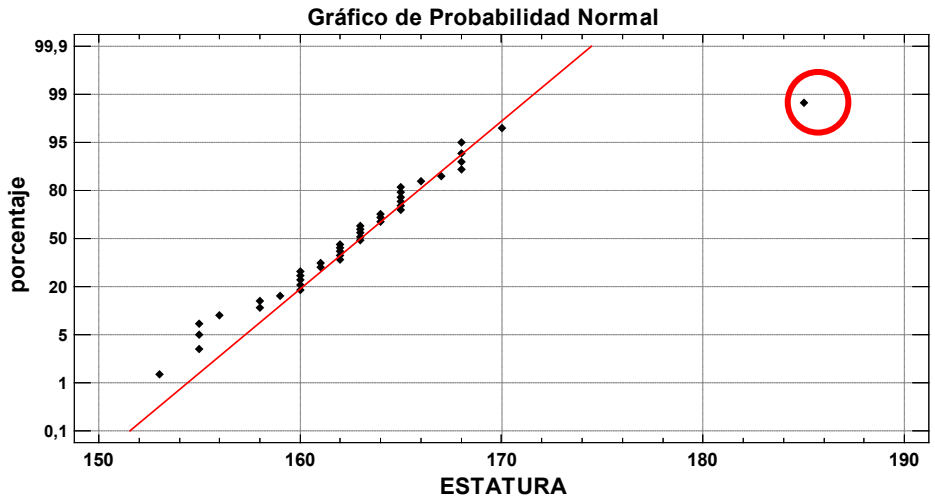
  Consejo: utilizar varias representaciones gráficas, cada una de ella proporciona información complementaria.

- **Reglas basadas en parámetros muestrales**

  1. Basadas en la **media** y la **desviación típica**
  2. Basadas en la **mediana** y la **desviación absoluta mediana**
  3. Basadas en los **cuartiles** y el **RI (*Prueba de Tukey*)**

  • **Estas reglas también se suelen representar gráficamente**

Herramientas gráficas



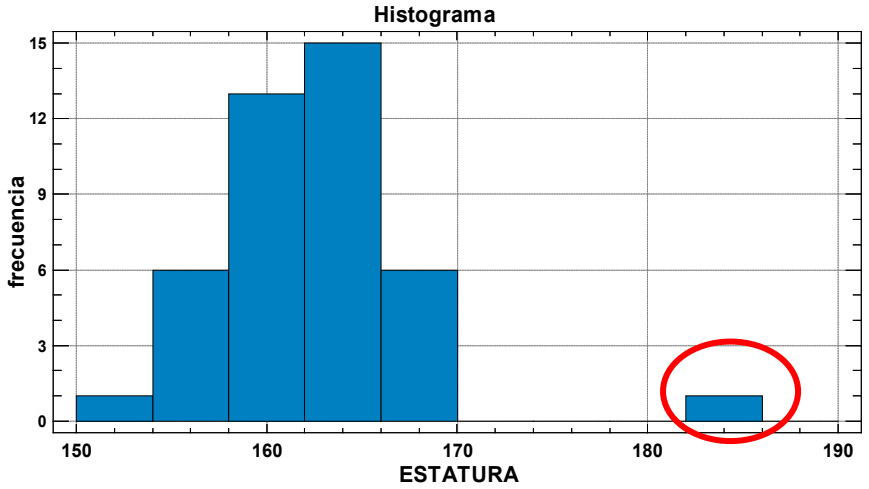
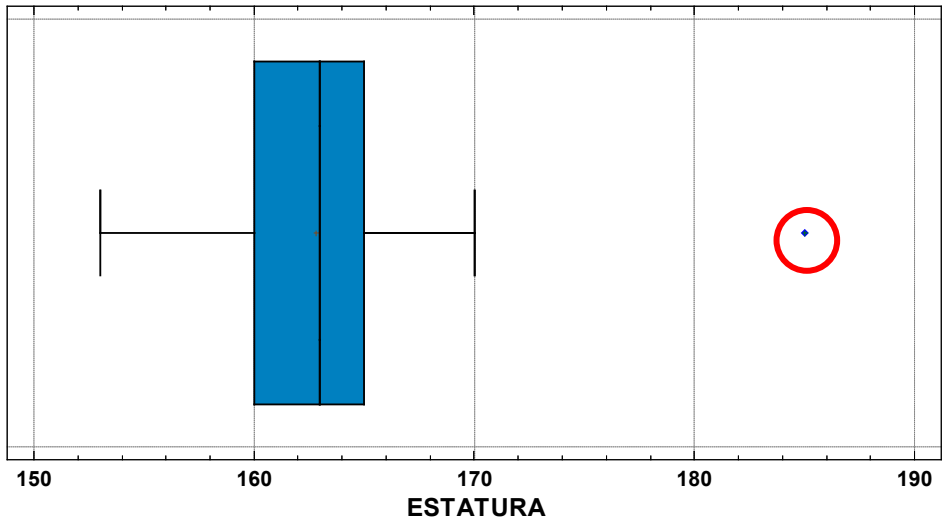


Gráfico *Box & Whisker*



Ejemplo

- **Horas de conexión a Internet al mes en niños de 5 a 11 años**

v.a. **X** = {50, 38, 45, 30, 47, 50, 48, 62, 55, 53, 52}

Datos ordenados

30	38	45	47	48	50	50	52	53	55	62

Regla basada en $\overline{X}$ y S

Se consideran sospechosas aquellas observaciones alejadas de la Media más de un numero determinado de Desviaciones Típicas:

$$\boxed{X_i \geq \overline{X} \mp kS}$$

- La **justificación** de esta regla deriva de la Desigualdad de *Tchebychev*, que relaciona la $\overline{X}$ y la S*:*
  - Entre la media y tres desviaciones típicas se encuentran, al menos, el 89% de los datos de la muestra de cualquier distribución (no solo normal).
- El **valor de K depende**:
  - **Tamaño muestra**. En muestras pequeñas, un dato fuera de este rango es muy poco frecuente y, por tanto, sospechoso. En muestras grandes no lo sería tanto.
  - Los límites de **identificación (atípicos leves o extremos)**

Criterios

- **Límites internos**: **k = 2**

Se consideran sospechosas leves aquellas observaciones alejadas de la media más de dos desviaciones típicas:

$$\boxed{X_i \geq \overline{X} \mp 2S}$$

- **Límites externos**: **k = 3 o 4**

Se consideran sospechosas extremas aquellas observaciones alejadas de la media más de tres o cuatro desviaciones típicas:

$$\boxed{X_i \geq \overline{X} \mp 3S \quad o \quad X_i \geq \overline{X} \mp 4S}$$

- **Otros límites**: **k = 1**

Independientemente de la consideración de leve o extremo, en muestras grandes (N ≥ 80) se suele utilizar k = 3 o 4 y en muestras pequeñas K=2

Ejemplo

Horas de conexión a Internet al mes en niños de 5 a 11 años

30	38	45	47	48	50	50	52	53	55	62

- **Límite interno**: $\bar{X} \mp 2S = 48{,}2 \mp 2x8{,}53 = [31,14\ ;\ \ 65,26]$

Cualquier valor fuera de este intervalo es sospechoso leve.

Tendríamos 1 observación atípica leve: el valor 30 está en la frontera

- **Límite externo**: $\bar{X} \mp 3S = 48{,}2 \mp 3x8{,}53 = [22,61;\ \ 73,79]$

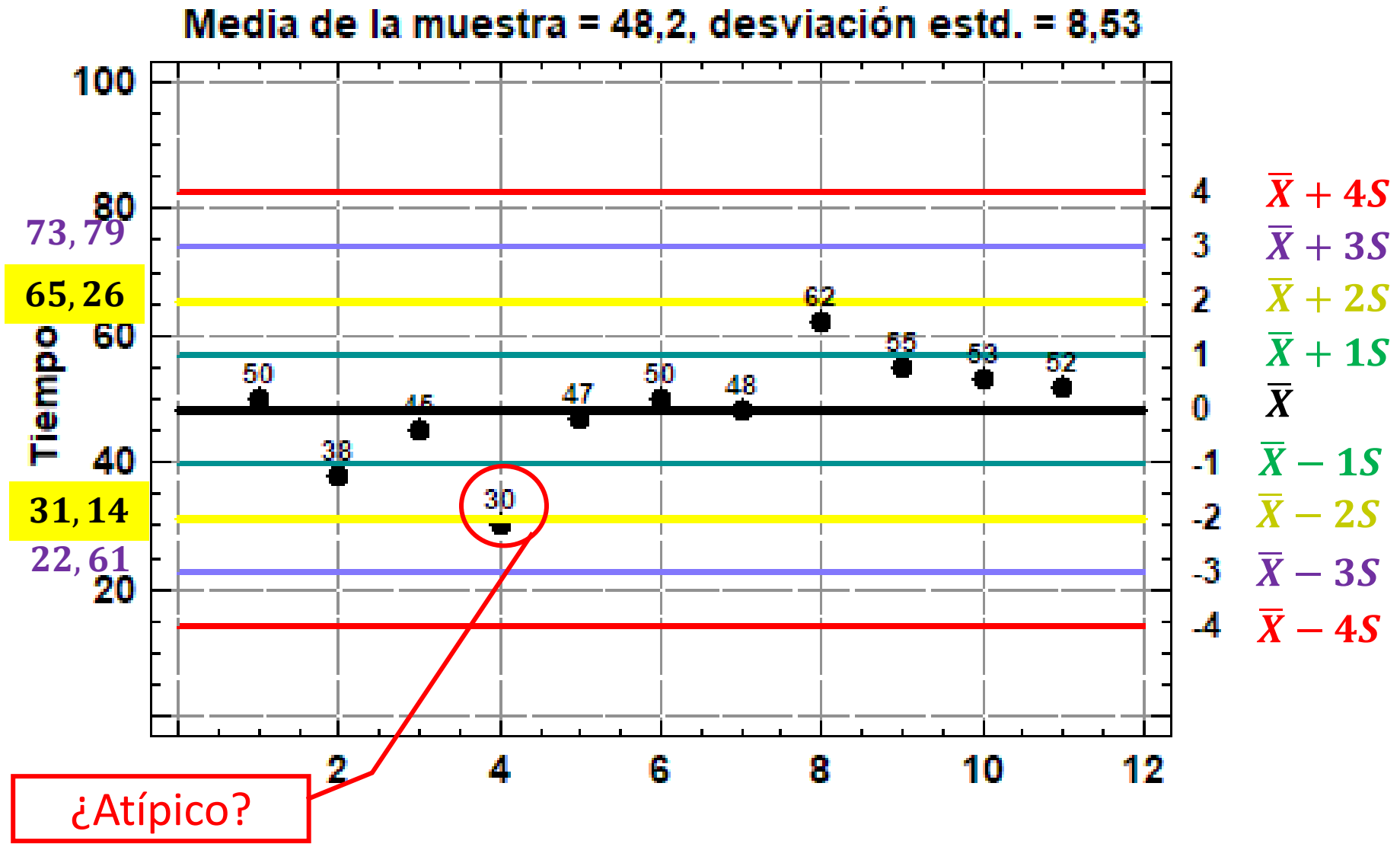
Cualquier valor fuera de este intervalo es sospechoso extremo.

No tendríamos ninguna observación atípica extrema

Podríamos concluir que no hay valores atípicos en estos datos o es leve.

Nota: para los límites externos, se podría haber usado K=4, si N fuera grande.

Gráfico de aberrantes



Consideraciones

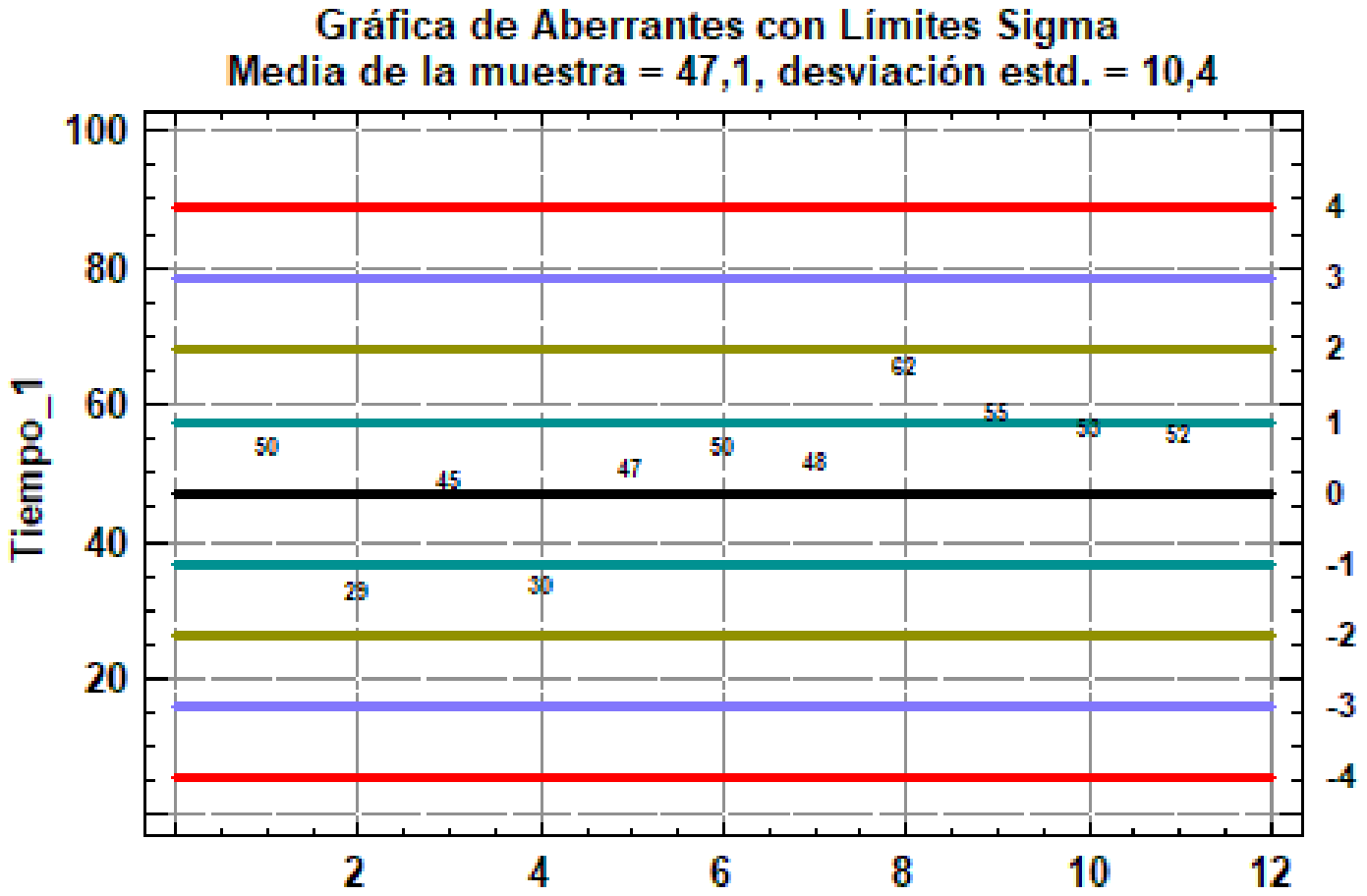
- **Inconveniente**:
  - Este criterio funciona bien mientras las distribuciones sean normales o, en todo caso, simétricas
  - **Si la distribución tiene valores muy extremos, la detección queda enmascarada, ya que ni $\overline{X}$, ni *S* son robustos**.

- Veamos un **ejemplo**
  - Supongamos que en el conjunto de datos que estamos usando la observación con el valor 38 tuviera el valor 29

v.a. **X** = {50, ~~38~~ 29, 45, 30, 47, 50, 48, 62, 55, 53, 52}

  - ¿Cambiaría la clasificación de dato atípico leve asignada al valor 30?

Ejemplo con otro dato extremo



Como hay un dato extremo adicional, la media y la desviación típica también cambian y el valor 30 queda enmascarado por la presencia de ese otro valor más extremo → **parámetros robustos**

Regla robusta basadas en Me y DAM[1]

- **Límites internos:** se consideran sospechosas leves aquellas observaciones alejadas de la Mediana más de 4,5 veces la Desviación Absoluta Mediana:

$$\boxed{X_i \geq Me \mp 4,5DAM}$$

- **Límites externos:** se consideran atípicas extremas aquellas observaciones alejadas de la Mediana más de 8 veces la Desviación Absoluta Mediana:

$$\boxed{X_i \geq Me \mp 8DAM}$$

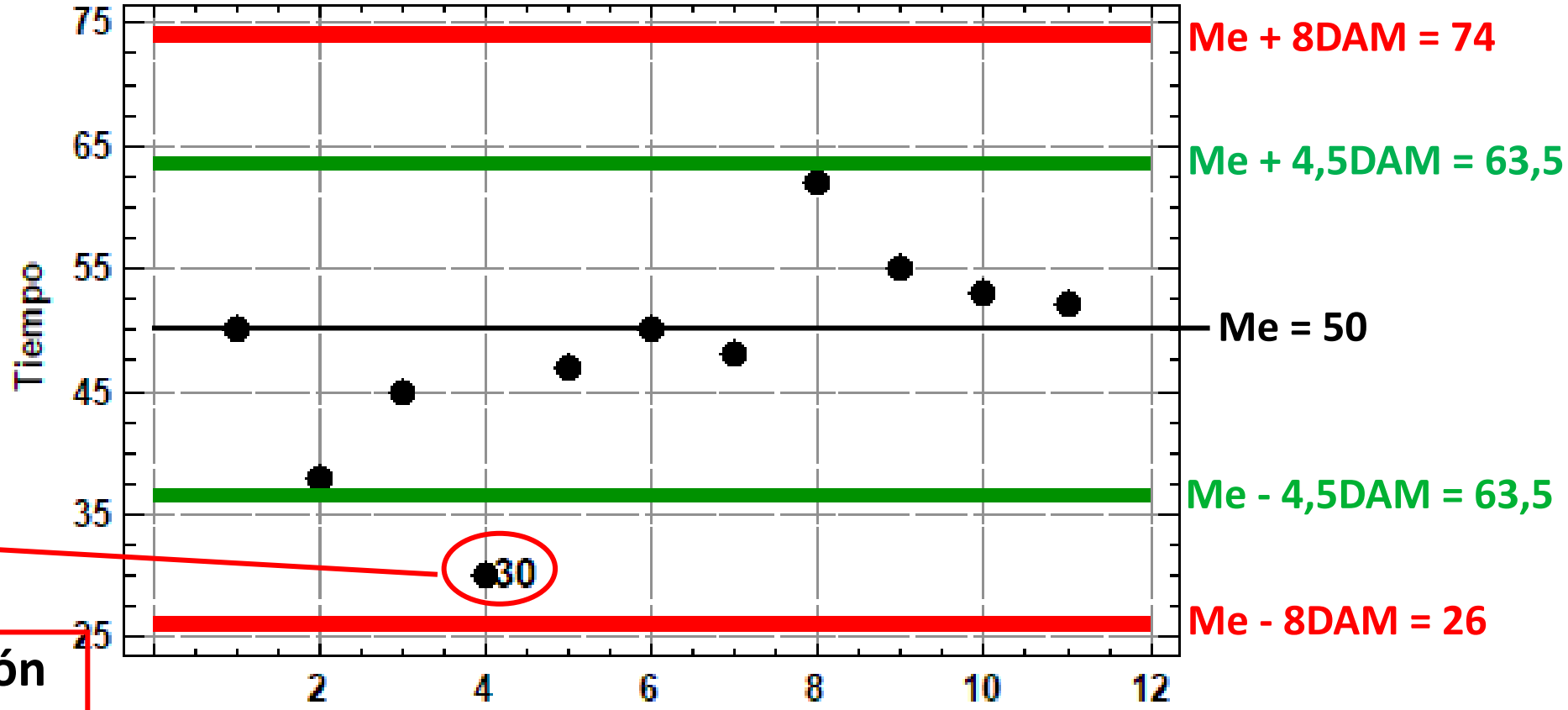
- La justificación es similar a la de la Regla anterior, pero usando **parámetros de centralización y dispersión robustos**, de modo que se tienen en cuenta los valores extremos que pueden hacer el criterio engañoso.
- **Inconveniente**: no tiene en cuenta la asimetría de la distribución

[1]DAM = MAD = Desviación Absoluta Mediana

Ejemplo

Horas de conexión a Internet al mes en niños de 5 a 11 años

30	38	45	47	48	50	50	52	53	55	62



observación atípica leve

Consideraciones

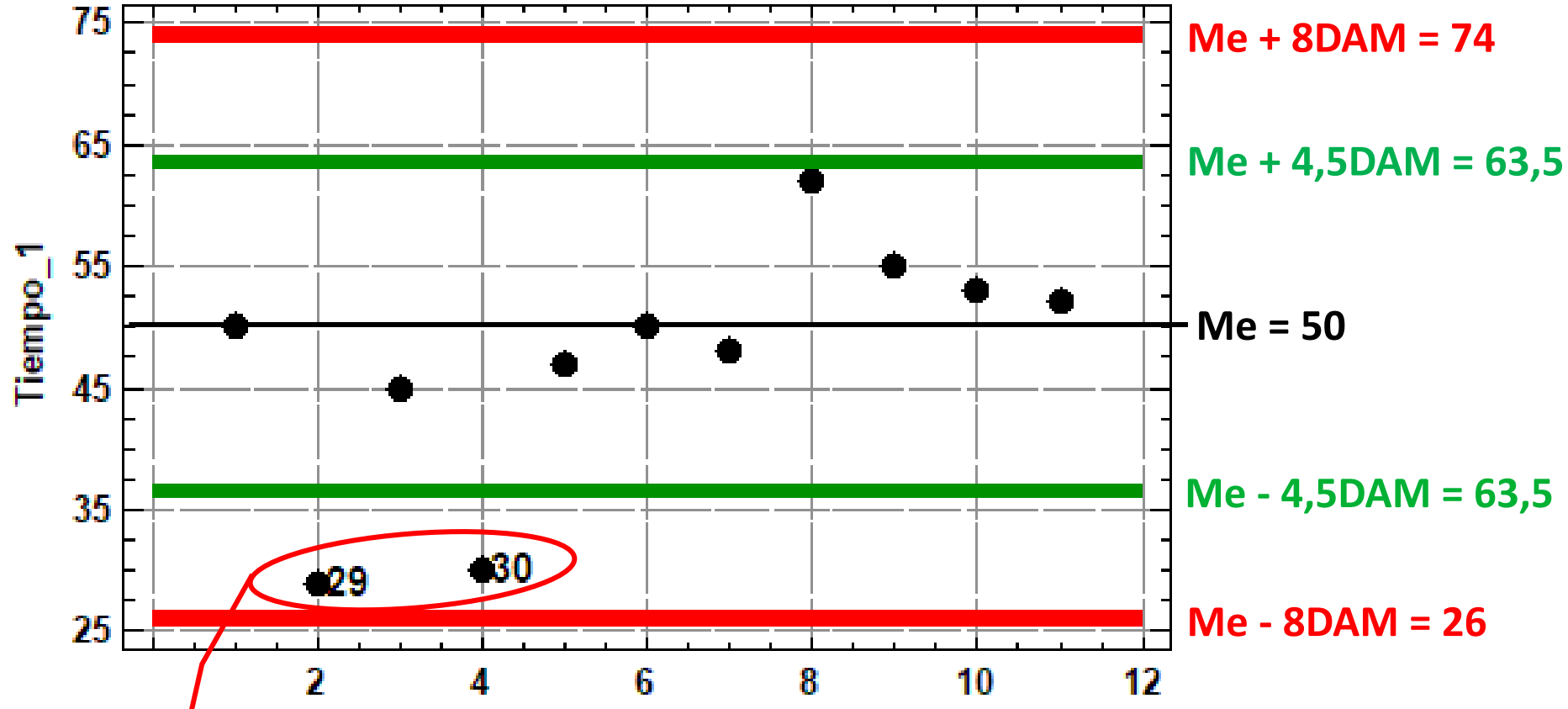
- **Inconveniente**:
  - **No tiene en cuenta la asimetría de la distribución.**
  - Este criterio funciona bien mientras las distribuciones sean normales y/o simétricas o incluso con algún valor extremo
  - La asimetría puede generar valores aparentemente atípicos, que, en realidad, serían “extremos” propiamente dichos.

- **Ejemplo con valores extremos:**
  - Supongamos que en el conjunto de datos que estamos usando la observación con el valor 38 tuviera el valor 29

    v.a. **X** = {50, ~~38~~ 29, 45, 30, 47, 50, 48, 62, 55, 53, 52}

  - ¿Cambiaría la clasificación de dato atípico leve asignada al valor 30?

Ejemplo con datos extremos



observaciones atípicas leves

La Me y la DAM NO cambian con ese dato extremo adicional. El dato extremo (30), NO queda enmascarado por la presencia de ese otro valor más extremo (29)

Prueba de Tukey

- **Límites internos:** se consideran sospechosas o atípicas leves aquellas observaciones alejadas del 1er y 3er cuartil, respectivamente, más de 1,5 veces el Recorrido Intercuartílico:

$$\boldsymbol{X_i \leq Q_1 - 1,5RI \ \ o \ \ X_i \geq Q_3 + 1,5RI}$$

- **Límites externos**: se consideran atípicas extremas aquellas observaciones alejadas del 1er y 3er cuartil, respectivamente, más de 3 veces el Recorrido Intercuartílico:

$$\boldsymbol{X_i \leq Q_1 - 3RI \ \ o \ \ X_i \geq Q_3 + 3RI}$$

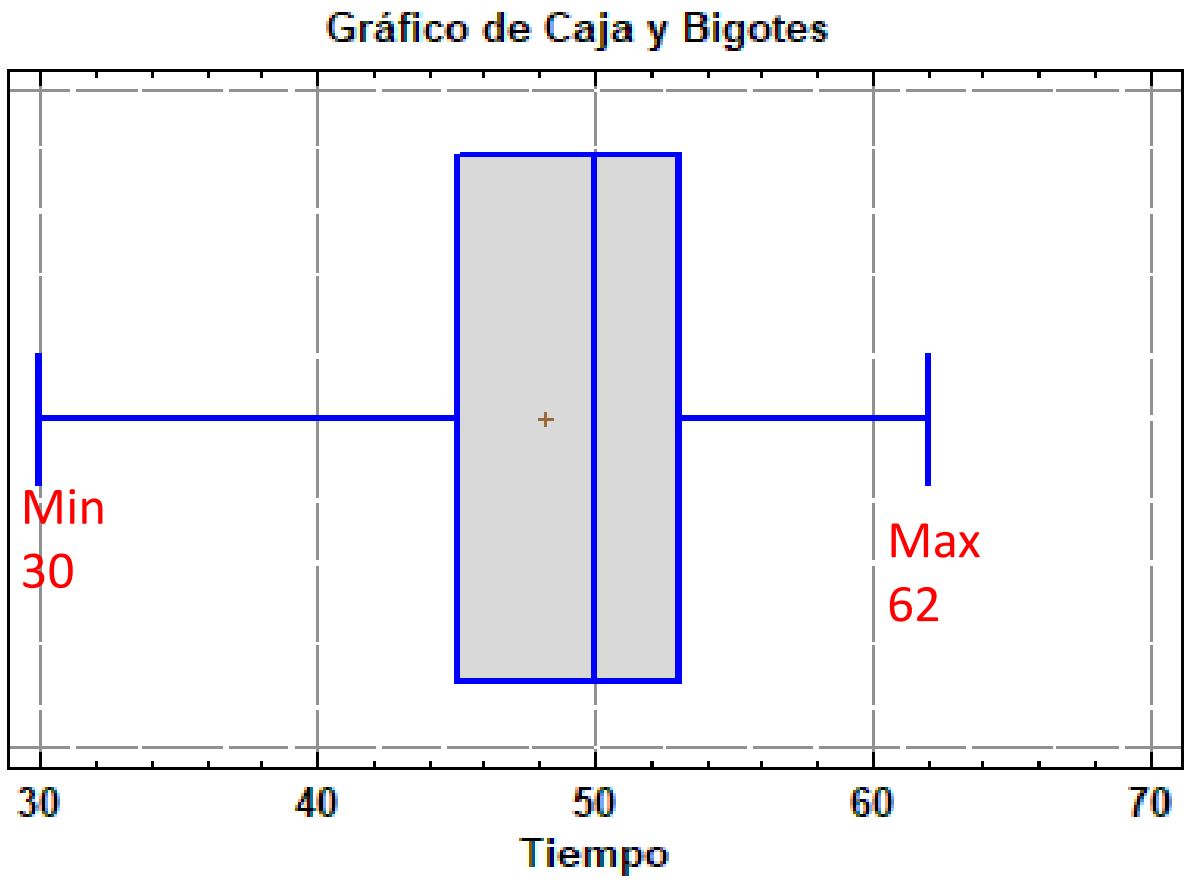
- Es el criterio utilizado en la construcción del gráfico de Caja y Bigotes
- Este criterio se puede ajustar usando otros valores distintos de 1,5. El *Statgraphics* no dispone de esta posibilidad.

Ejemplo

Comprobad si, según la regla anterior, existen valores atípicos en los datos de tiempo de conexión a Internet.

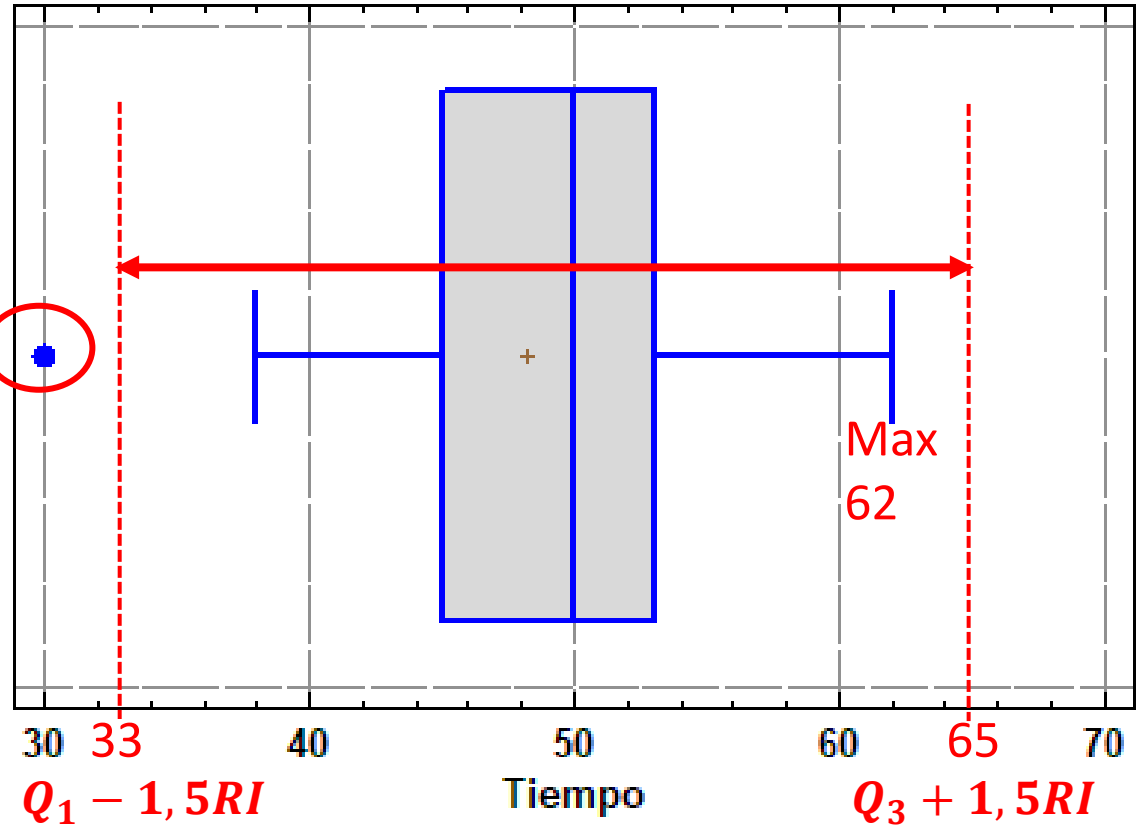
X = {50, 38, 45, 30, 47, 50, 48, 62, 55, 53, 52}

Sin usar los límites



Ejemplo

Usando los límites

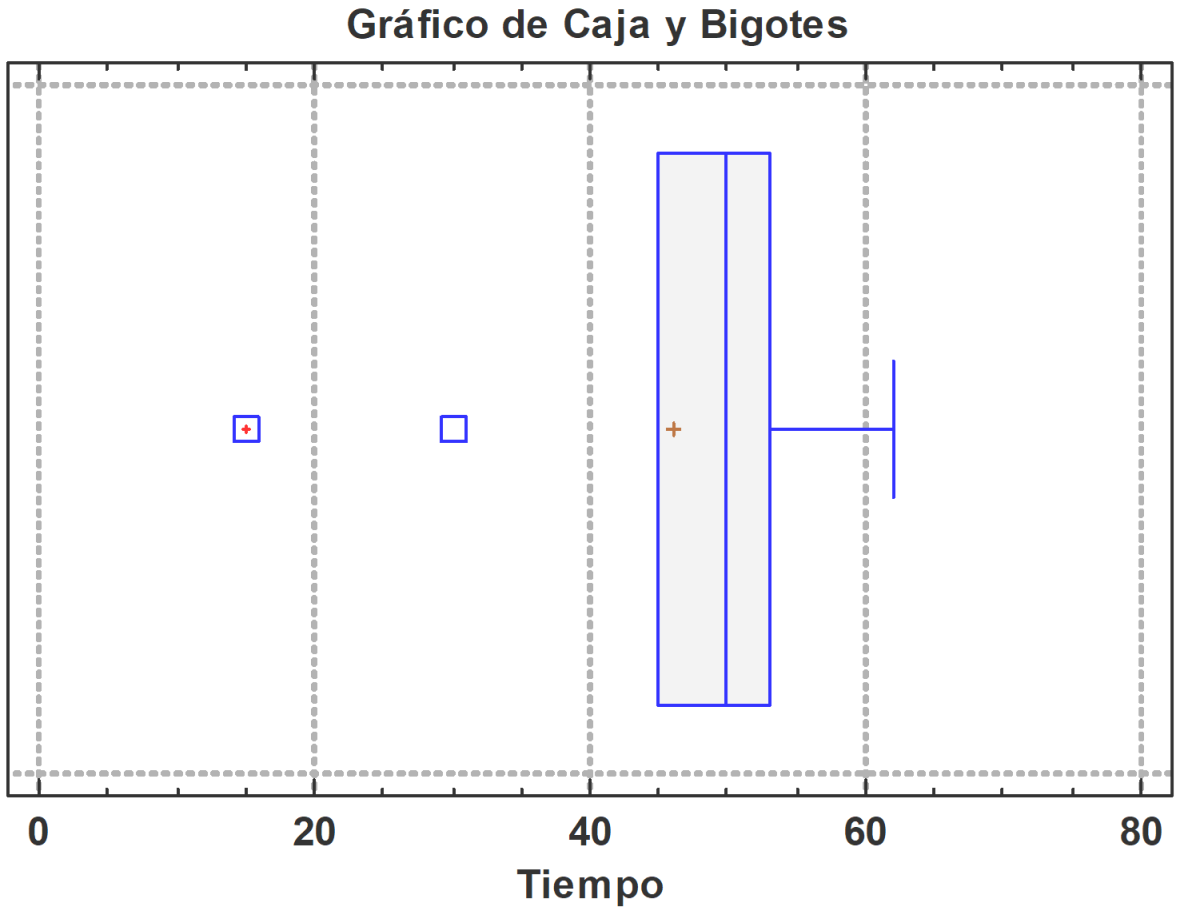


observación atípica leve

Ejemplo con datos extremos

Comprobad si, según la regla anterior, existen valores atípicos en los datos de tiempo siguientes:

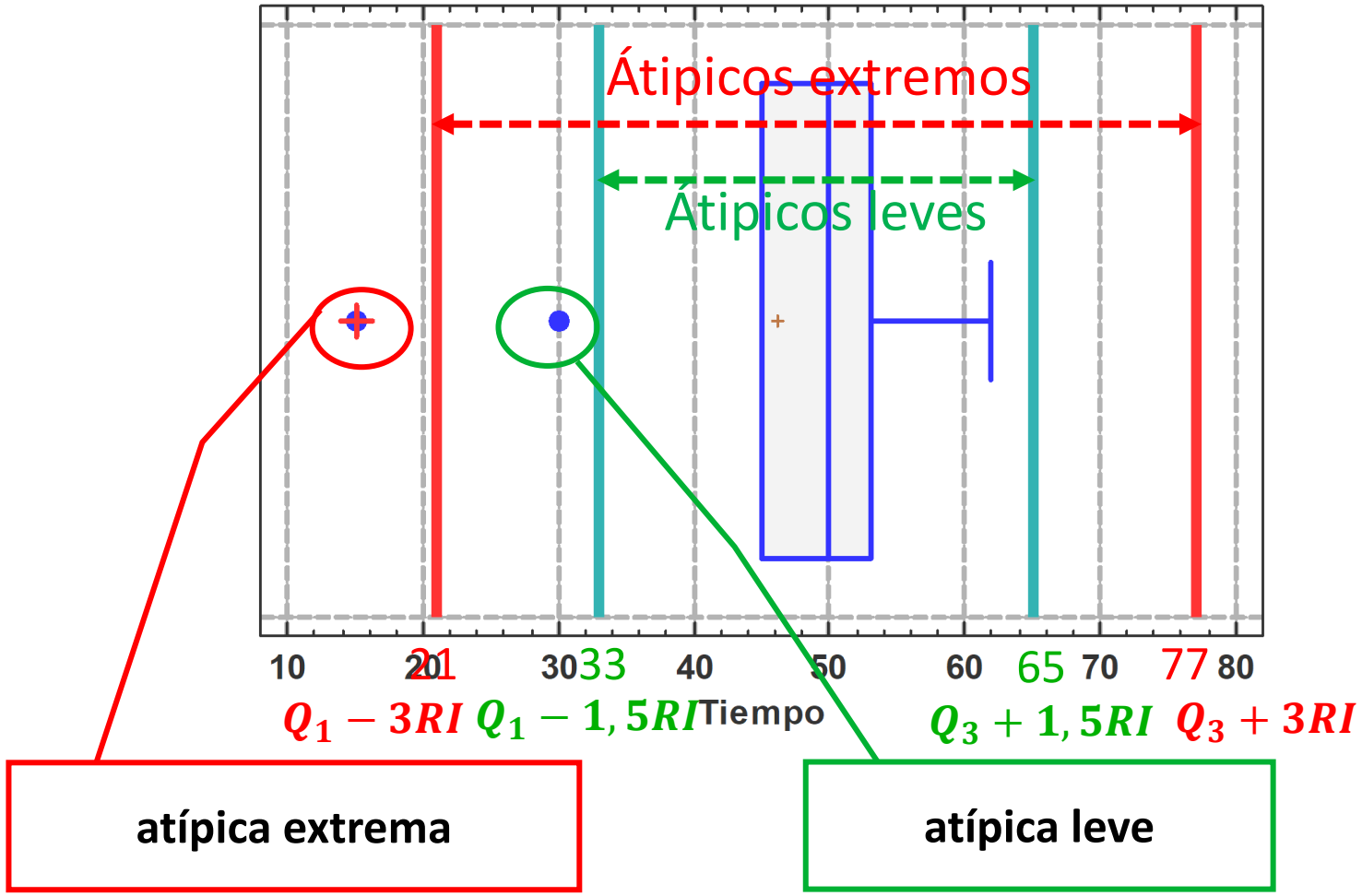
X = {50, ~~38~~ **15**, 45, 30, 47, 50, 48, 62, 55, 53, 52}



Ejemplo con datos extremos

Comprobad si, según la regla anterior existen valores atípicos en los datos de tiempo siguientes:

X = {50, ~~38~~ **15**, 45, 30, 47, 50, 48, 62, 55, 53, 52}



Regla basada en Valores Tipificados o Estandarizados

Es como la primera regla, pero tipificando previamente los valores.

- **Los valores tipificados (z_i) miden el número de desviaciones típicas a las que cada valor se encuentra de la media muestral y se calculan:**

$$z_i = \frac{x_i - \bar{X}}{S}$$

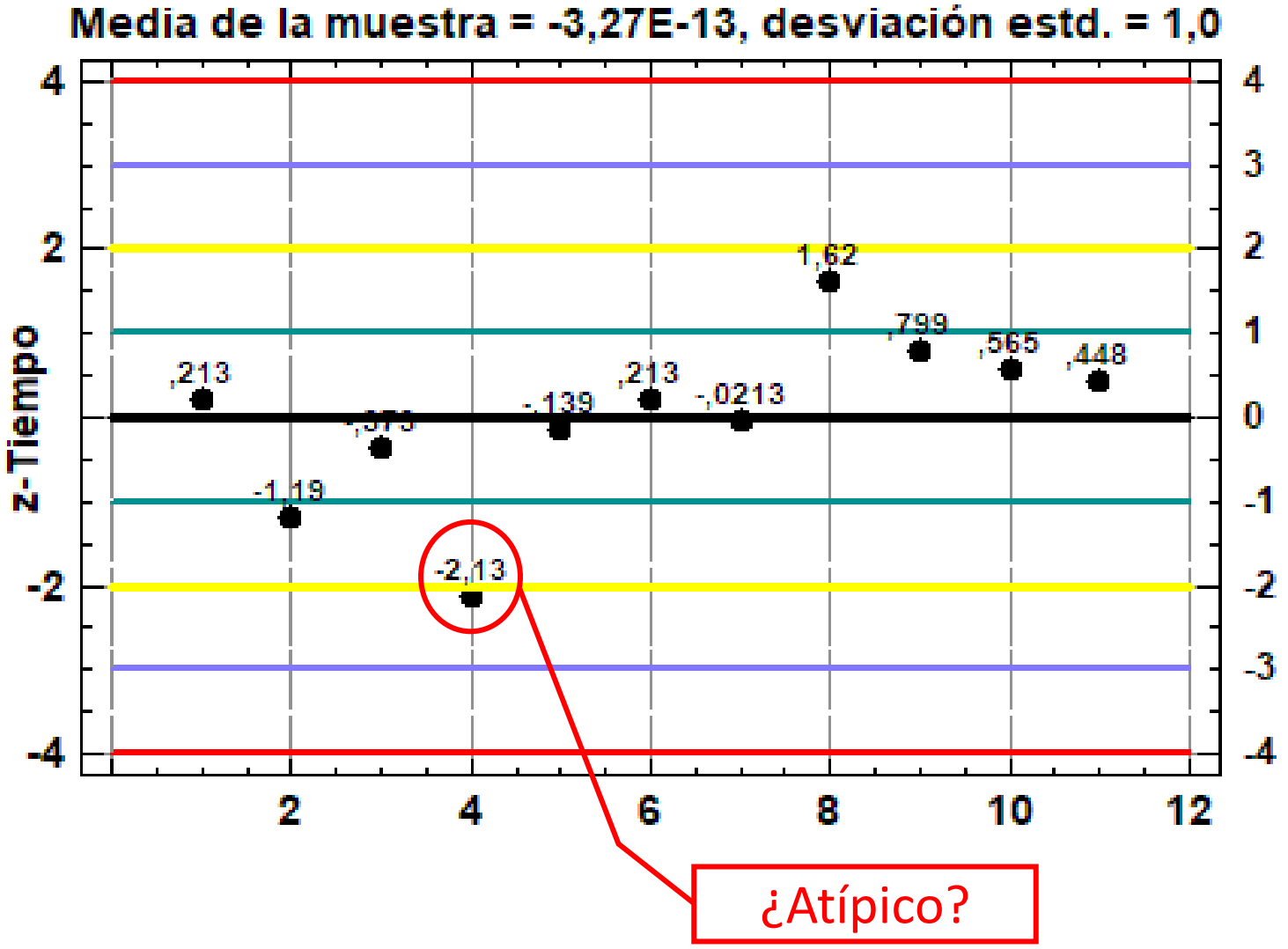
- Para detectar un valor como atípico, solo hay que comprobar si los valores tipificados están en los intervalos:

$\boldsymbol{z_i \in [-2, +2]}$	***Límites internos***
$\boldsymbol{z_i \in [-3, +3]}$	***Límites intermedios***
$\boldsymbol{z_i \in [-4, +4]}$	***Límites externos***

- Cuando se analiza más de una variable es más conveniente estandarizar previamente, así los resultados son comparables.

Ejemplo: Valores tipificados

$$\bar{X} - kS = 0 - 1k \qquad \bar{X} + kS = 0 + 1k$$



Regla basada en Valores Estudentizados

Es como la regla anterior, pero para cada valor de los datos (X_i) se tipifica usando la media ($\bar{X}_{[sin\ xi]}$) y la desviación típica ($S_{[sin\ xi]}$) obtenidas sin X_i, usando los *n* -1 valores restantes.

- **Los valores estudentizados (ze_i) con supresión miden el número de desviaciones típicas a las que cada valor se encuentra de la media muestral cuando ese valor no está en la muestra y se calculan:**

$$ze_i = \frac{x_i - \bar{X}_{[sin\ xi]}}{S_{[sin\ xi]}}$$

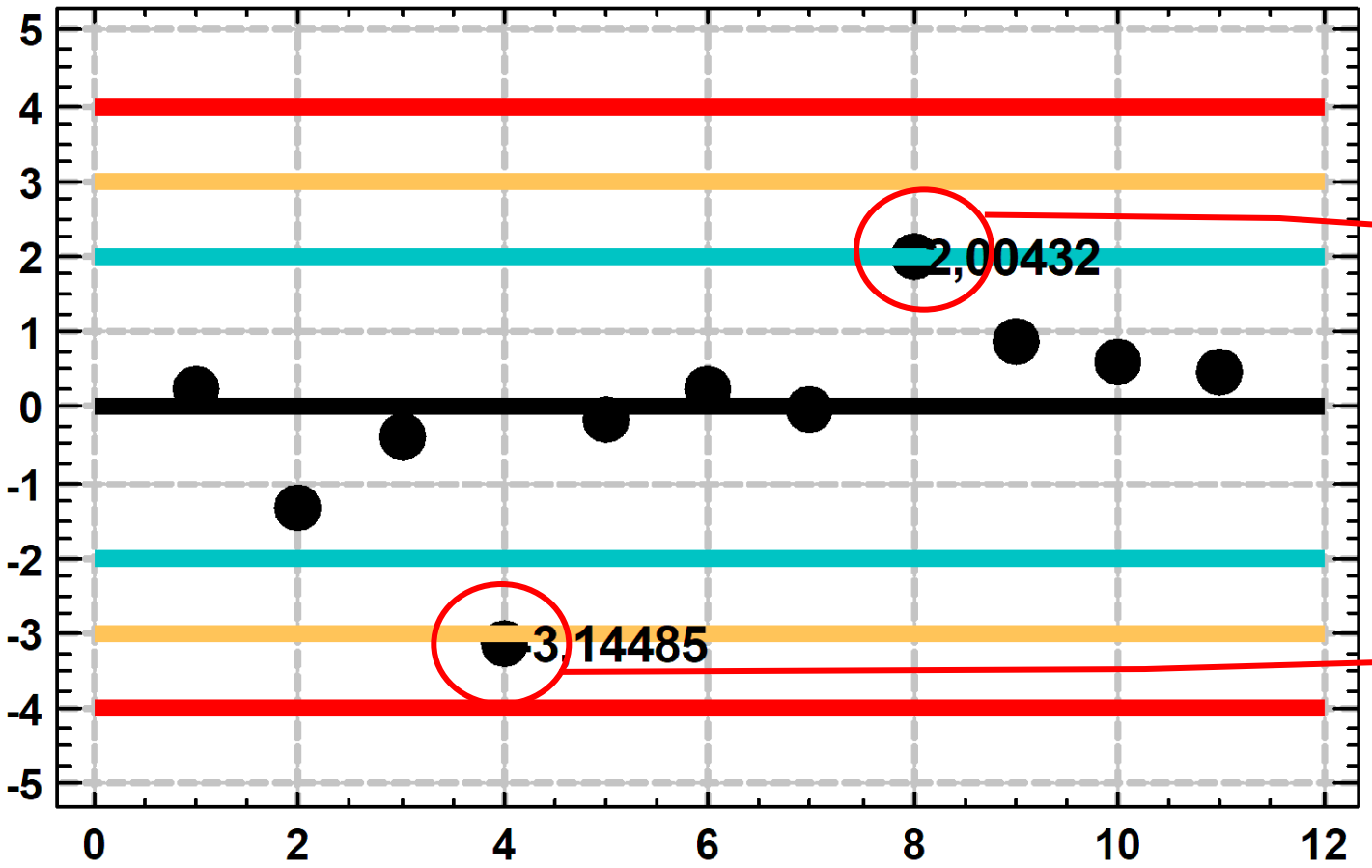
- Para detectar un valor como atípico , solo hay que comprobar si los valores tipificados están en los intervalos:

$\boldsymbol{ze_i \in [-2, +2]}$	***Límites internos***
$\boldsymbol{ze_i \in [-3, +3]}$	***Límites intermedios***
$\boldsymbol{ze_i \in [-4, +4]}$	***Límites externos***

Regla basada en Valores Estudentizados

- Al excluir el valor evaluado del calculo de $\overline{X}$ y **S**, para un valor X_i, si X_i fuera extremo (o atípico) el valor estandarizado no se vería tan afectado por la posible desviación de $\overline{X}$ y S.
- Es una manera de ”robustecer” la regla basada en $\overline{X}$ y S
- Lo **importante** no es el propio valor estudentizado con supresión, sino que **de la comparación de los valores estudentizados con y sin supresión se obtiene información de la influencia de un valor sobre $\overline{X}$ y S**, aunque puede quedar enmascarado por otro valor más atípico.

Ejemplo: Valores estudentizados



¿Atípico leve?

¿Atípico extremo?

Los valores estudentizados con supresión son todos más cercanos a la media (0) que los estudentizados sin supresión.

Otras pruebas de detección

- **Inferenciales**
  - **Prueba de *Grubbs*** o ESD (***E**xtreme **S**tudentized **D**eviate*)
  - **Prueba de *Dixon***
  - **Regresión Lineal**
  - **…**

Detección bivariante

- **Herramientas gráficas**:
  - Diagrama de dispersión (2D, 3D)
  - Gráfico de tela de araña
  - Etc

  Cualquier representación gráfica que incluya límites internos y externos en cada eje.

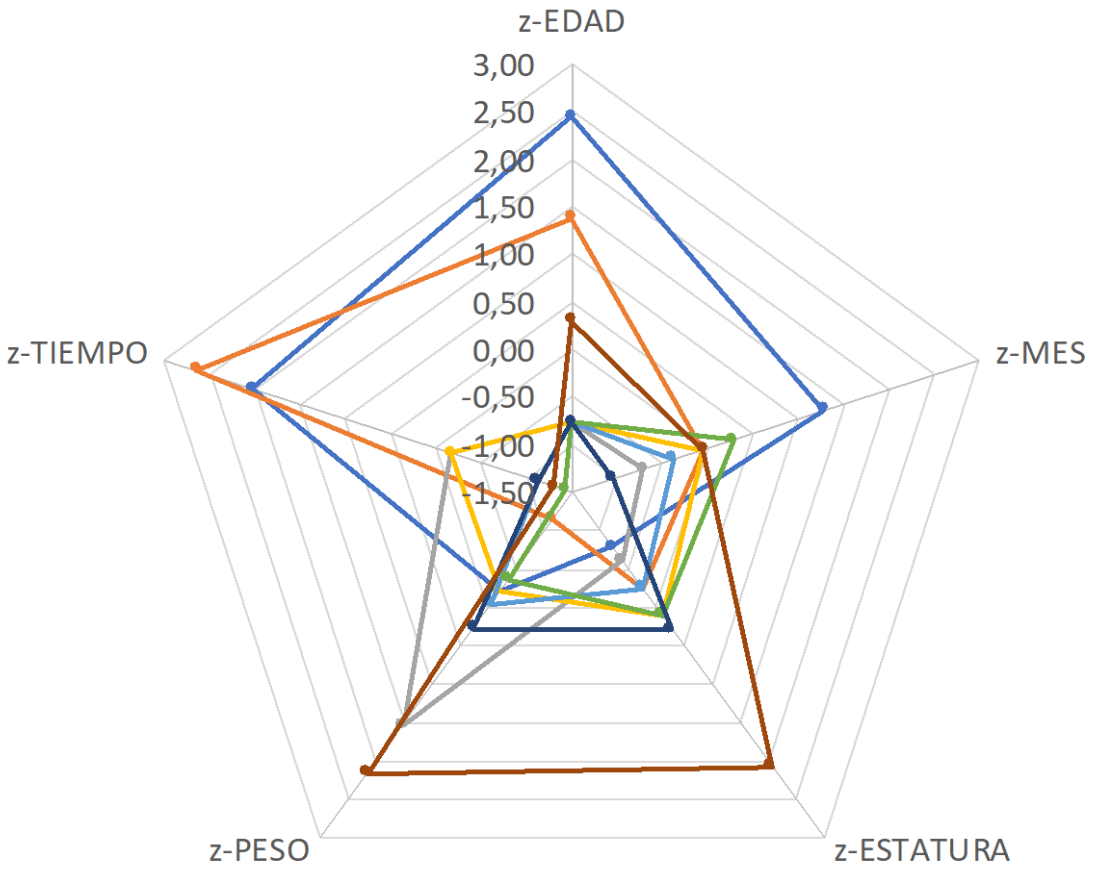
- **Tests o pruebas:**
  - Distancia de Mahalanobis
    - medida de la distancia de cada observación en un espacio multidimensional respecto del centro medio de las observaciones
  - Inferenciales
    - Las mismas que para la detección multivariante

Herramientas gráficas

Gráfico de Araña de los valores tipificados

- Cuando hay muchas variables y pocos casos o los casos son resúmenes de casos
- Permiten ver cómo de atípico es un valor en cada una de las variables.

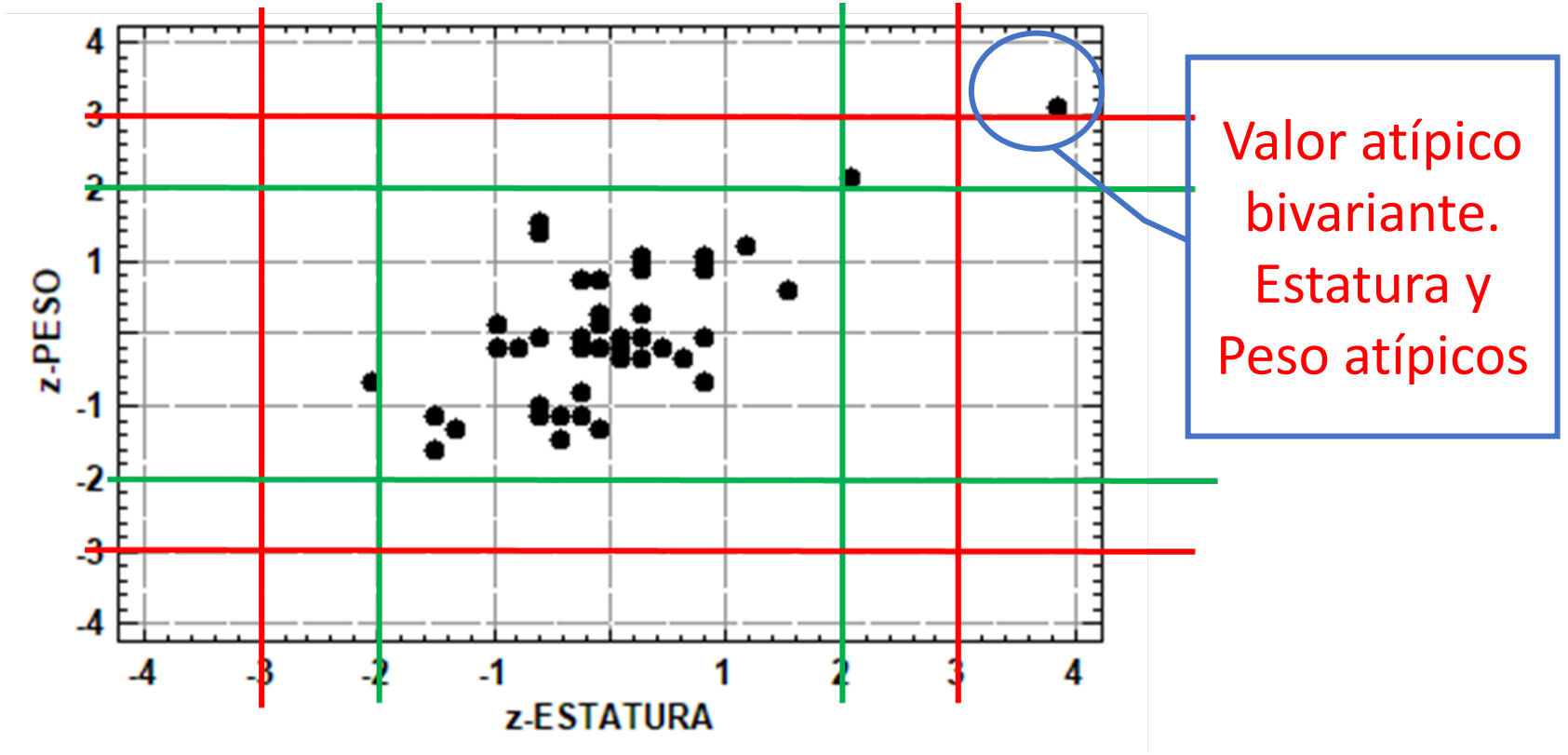
Cada línea de un color es un caso



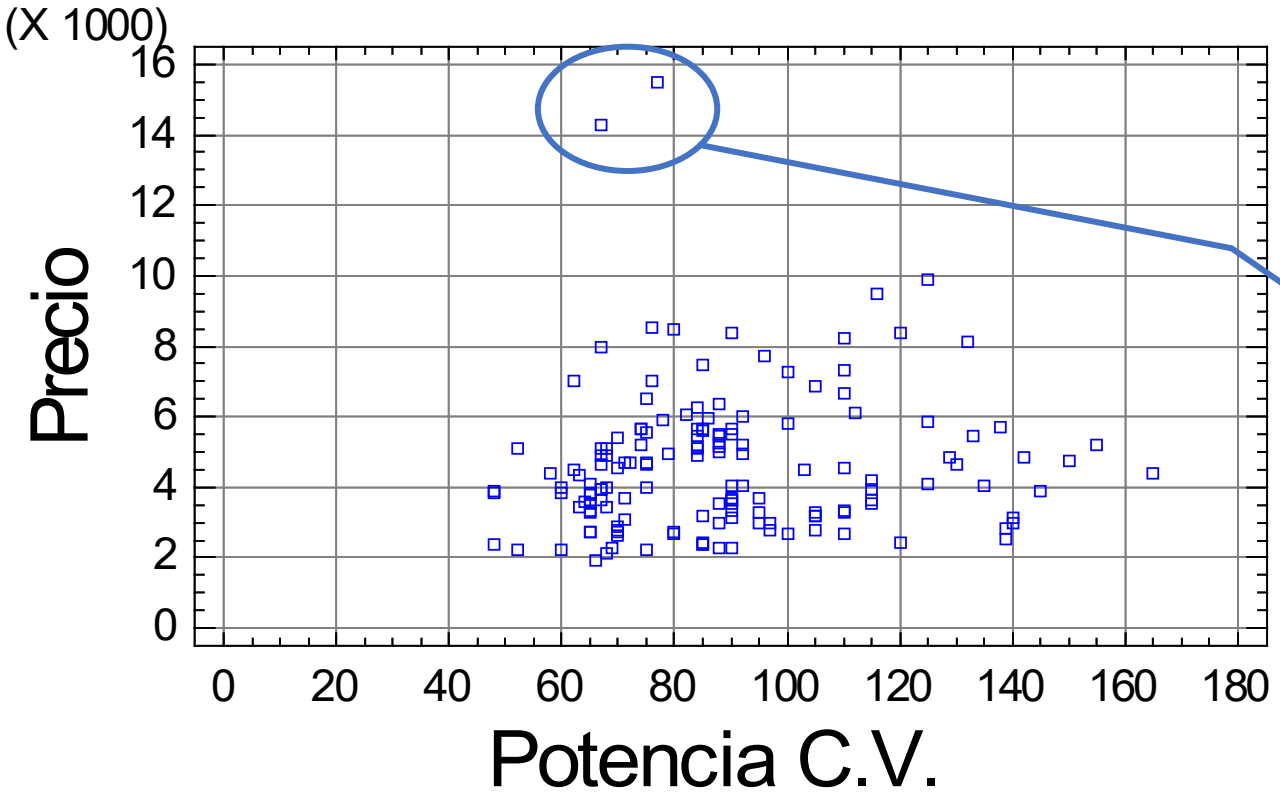
Herramientas gráficas

Diagrama de dispersión de valores tipificados

- Para evaluar dos variables
- Permiten ver cómo de atípico es un valor en cada una de las variables dependiendo de las reglas usadas para establecer diferentes límites internos y externos



Herramientas gráficas



Valor atípico univariante. Solo atípico para PRECIO, no para POTENCIA

Tratamiento

Errores de procedimiento

- **Corrección, si es posible**
- **Eliminación → Ausente**

Acontecimientos extraordinarios

- **Eliminación → Ausente**
- **Mantenerlo**
  - **Uso de parámetros robustos**, si es necesario
  - **Transformación**, si es necesario

Valores extremos

- **Mantenerlo**
  - **Uso de parámetros robustos**, si es necesario
  - **Transformación**, si es necesario

Causas no conocidas

- **Eliminación → Ausente**
- **Mantenerlo**
  - **Uso de parámetros robustos e investigaciones adicionales**

Consideraciones

- No son reglas exactas, son criterios que ayudan a tomar decisiones.

- Cualquiera que sea el procedimiento empleado siempre debe estar acompañado por una evaluación cualitativa

- En este curso nos hemos centrado en la detección (y tratamiento) cuantitativo univariante fundamentalmente y, someramente, en el multivariante.

Valores faltantes

Valores ausentes o faltantes o *missing data*

1. Qué son, cómo se originan y cuál es su impacto
2. Tipos de datos faltantes
3. Detección e identificación del patrón de datos ausentes
4. Tratamiento (para v.a. cuantitativas con faltantes MCAR)

¿Qué son?

Los **datos faltantes**, **ausentes**, **perdidos** o ***missing data*** son aquellas observaciones para las que el valor o dato está ausente en la muestra

	PESOM	TALLAM	SEM	PASM
	Peso de la madre	Talla de la madre	Semanas de gestación	Presión arterial sistólica de la madre
1	59	160	39	150
2	65	166	38	160
3	68	173	38	100
4	71	176	40	95
5	56	165	37	115
6	46	155	39	90
7	65	?	41	150
8	68	162	?	170
9	48	152	?	105
10	68	174	37	100
11	57	160	37	170
12	64	169	38	95
13	63	167	39	120
14	65	171	38	110
15	52	161	36	145

Impacto

- **La presencia de los datos faltantes tiene gran impacto sobre**:
  - Los **procesos de estimación**, cuando se usan técnicas de inferencia
    - La generalidad de los resultados puede verse afectada por los posibles sesgos
  - El **tamaño de la muestra**
    - Tras eliminar los casos y variables que contienen datos faltantes, la muestra resultante puede ser inadecuada
- El **tratamiento dependerá** del **proceso de datos ausentes:**
  - De las **pautas**: ¿Están los datos faltantes distribuidos aleatoriamente entre las observaciones o existen pautas?
  - De la **relevancia:** ¿Cuántos hay y dónde están?

Datos ausentes prescindibles

- **Resultado de procesos que se encuentran bajo control del analista y se pueden identificar explícitamente**:
  - **Observaciones** de una población que **no forman parte de la muestra**
    - Ej): Las observaciones asociadas a individuos de la población que no están en la muestra son en realidad datos faltantes
  - **Datos censurados**: observaciones incompletas como consecuencia del proceso de obtención de datos seguido en el análisis
    - Ej): En un estudio sobre causas de fallecimiento, los datos de un individuo vivo son ausentes
- **No es necesario una solución específica,** la ausencia es inherente a los procedimientos o técnicas estadísticas usadas:
  - Técnicas de Inferencia
  - Técnicas para datos censurados

Datos ausentes no prescindibles

- **Resultado de procesos que son ajenos al analista y/o no se pueden identificar explícitamente**:
  - **Errores de introducción de datos** que producen datos inválidos y que hay que eliminar
  - **Observaciones costosas o imposibles de generar o recopilar**
    - Una concentración de nutrientes que no se puede medir porque no se dispone del instrumento apropiado
    - Una respuesta inaplicable en un cuestionario como los años de matrimonio en individuos solteros
    - Un valor atípico que se ha eliminado
  - **Omisiones de respuesta intencionadas** en una encuesta
    - Ej: renuncia a contestar una pregunta sensible

Soluciones

Si los datos ausentes son PRESCINDIBLES:

Solo se puede actuar sobre:

- **Prevención de la generación de datos ausentes**
  - Diseño adecuado del experimento
  - Diseño de la encuesta apropiado
  - Procedimientos de control en la recogida de datos
  - …
- **Incorporación de los datos ausentes en las técnicas usadas**
  - Uso de **técnicas estadísticas conformadas** para analizar conjuntos de datos con valores ausentes
    - Como parte de un error muestral en inferencia
    - Técnicas específicas para datos censurados
    - …

Soluciones

Si los datos ausentes NO son PRESCINDIBLES:

- Se puede llevar a cabo un **tratamiento** **previo a la utilización de cualquier otra técnica estadística**.
- El tratamiento **dependerá** de la **pauta** y de la **relevancia**.
- La **pauta** se refiere al **grado de aleatoriedad** presente en los datos ausentes, o lo que es lo mismo, de **si existen o no patrones sistemáticos en el proceso** que puedan sesgar los resultados.
- Según el **grado de aleatoriedad** encontramos los siguientes **tipos de datos ausentes**:
  - **MNAR** (***M**issing **N**ot **A**t **R**andom*): **sistemáticos** o no aleatorios
  - **MAR** (***M**issing **A**t **R**andom*): **aleatorios**
  - **MCAR** (***M**issing **C**ompletely **A**t **R**andom*): **completamente aleatorios**

Operación con datos ausentes no prescindibles

1 Identificación

- ¿Existen datos faltantes no prescindibles?

2 Localización

- Determinar la extensión de datos ausentes

3 Evaluación del patrón de aleatoriedad

4 Tratamiento

- Eliminación casos/variables, Casos completos, Imputación

1) Identificación y 2) Localización

- Consiste en **evaluar la magnitud del problema derivado de la presencia de datos faltantes, analizando el porcentaje de datos ausentes totales, por variables y por casos**.

  - Si existen **casos con un alto porcentaje de datos ausentes** se deberían excluir del problema.

  - Si existe una **variable con un alto porcentaje de datos ausentes** su exclusión dependerá de la importancia teórica de la misma y la posibilidad de ser reemplazada por variables con un contenido informativo similar.

Ejemplo[1]

[1]Hair, J., Anderson, R., Tatham, R., & Black, W. (1999). Análisis Multivariante

Supongamos el siguiente conjunto de datos

Caso Id	V1	V2	V3	V4	V5
1	1,3	9,9	6,7	3	2,6
2	4,1	5,7			2,9
3		9,9		3	
4	0,9	8,6		2,1	1,8
5	0,4	8,3		1,2	1,7
6	1,5	6,7	4,8		2,5
7	0,2	8,8	4,5	3	2,4
8	2,1	8	3	3,8	1,4
9	1,8	7,6		3,2	2,5
10	4,5	8		3,3	2,2
11	2,5	9,2		3,3	3,9
12	4,5	6,4	5,3	3	2,5
13					2,7
14	2,8	6,1	6,4		3,8
15	3,7			3	
16	1,6	6,4	5		2,1
17	0,5	9,2		3,3	2,8
18	2,8	5,2	5		2,7
19	2,2	6,7		2,6	2,9
20	1,8	9	5	2,2	3

Ejemplo[1]

[1]Hair, J., Anderson, R., Tatham, R., & Black, W. (1999). Análisis Multivariante

Muestra original

Variables	5
Casos	20
Observaciones	100
N	77
Caos completos	5
	25%

¿Eliminar alguna variable y/o caso?

Caso Id	V1	V2	V3	V4	V5	Ausentes por caso Número	%
1	1,3	9,9	6,7	3	2,6	0	0
2	4,1	5,7			2,9	2	40
3		9,9		3		3	60
4	0,9	8,6		2,1	1,8	1	20
5	0,4	8,3		1,2	1,7	1	20
6	1,5	6,7	4,8		2,5	1	20
7	0,2	8,8	4,5	3	2,4	0	0
8	2,1	8	3	3,8	1,4	0	0
9	1,8	7,6		3,2	2,5	1	20
10	4,5	8		3,3	2,2	1	20
11	2,5	9,2		3,3	3,9	1	20
12	4,5	6,4	5,3	3	2,5	0	0
13					2,7	4	80
14	2,8	6,1	6,4		3,8	1	20
15	3,7			3		3	60
16	1,6	6,4	5		2,1	1	20
17	0,5	9,2		3,3	2,8	1	20
18	2,8	5,2	5		2,7	1	20
19	2,2	6,7		2,6	2,9	1	20
20	1,8	9	5	2,2	3	0	0
Ausentes por variable Número	2	2	11	6	2	Total ausentes 23	Número
%	10	10	55	30	10	23,0%	%

Ejemplo[1]

[1]Hair, J., Anderson, R., Tatham, R., & Black, W. (1999). Análisis Multivariante

Muestra original SIN V3

Variables	4
Casos	20
Observaciones	80
N	68
Casos completos	12
	60%

Eliminar
- Variable V3

		Caso Id	V1	V2	V4	V5	Ausentes por caso	
							Número	%
		1	1,3	9,9	3	2,6	0	0
		2	4,1	5,7		2,9	1	25
		3		9,9	3		2	50
		4	0,9	8,6	2,1	1,8	0	0
		5	0,4	8,3	1,2	1,7	0	0
		6	1,5	6,7		2,5	1	25
		7	0,2	8,8	3	2,4	0	0
		8	2,1	8	3,8	1,4	0	0
		9	1,8	7,6	3,2	2,5	0	0
		10	4,5	8	3,3	2,2	0	0
		11	2,5	9,2	3,3	3,9	0	0
		12	4,5	6,4	3	2,5	0	0
		13				2,7	3	75
		14	2,8	6,1		3,8	1	25
		15	3,7		3		2	50
		16	1,6	6,4		2,1	1	25
		17	0,5	9,2	3,3	2,8	0	0
		18	2,8	5,2		2,7	1	25
		19	2,2	6,7	2,6	2,9	0	0
		20	1,8	9	2,2	3	0	0
Ausentes por variable	Número	2	2	6	2		Total ausentes	
							12	Número
	%	10	10	30	10		15,0%	%

Ejemplo[1]

[1]Hair, J., Anderson, R., Tatham, R., & Black, W. (1999). Análisis Multivariante

Muestra original SIN V3, ni casos 3, 13, 15

Variables	4
Casos	17
Observaciones	68
N	63
Casos completos	12
	71%

Eliminar
- **Variable V3**
- **Casos 3, 13, 15**

Caso Id	V1	V2	V4	V5	Ausentes por caso	
					Número	%
1	1,3	9,9	3	2,6	0	0
2	4,1	5,7		2,9	1	25
4	0,9	8,6	2,1	1,8	0	0
5	0,4	8,3	1,2	1,7	0	0
6	1,5	6,7		2,5	1	25
7	0,2	8,8	3	2,4	0	0
8	2,1	8	3,8	1,4	0	0
9	1,8	7,6	3,2	2,5	0	0
10	4,5	8	3,3	2,2	0	0
11	2,5	9,2	3,3	3,9	0	0
12	4,5	6,4	3	2,5	0	0
14	2,8	6,1		3,8	1	25
16	1,6	6,4		2,1	1	25
17	0,5	9,2	3,3	2,8	0	0
18	2,8	5,2		2,7	1	25
19	2,2	6,7	2,6	2,9	0	0
20	1,8	9	2,2	3	0	0

Ausentes por variable						Total ausentes	
	Número	0	0	5	0	5	Número
	%	0	0	29,41	0	7,4%	%

3) Tratamiento

- **Si** los **procesos de datos ausentes** son **MAR o MNAR**, sólo se debe aplicar un **método diseñado específicamente para este proceso**. Cualquier otro método introduce sesgos en los resultados.
    4. **Modelos que construyen explícitamente el proceso de datos ausentes**
- **Si** los **procesos de ausencia de datos** son **MCAR** puede utilizarse alguna de las **4** siguientes **soluciones**, en función del método empleado **para estimar los datos ausentes o los parámetros muestrales**:
    1. **Solo casos completos**
    2. **Eliminación de casos y/o variables**
    3. **Imputación**
        1. **Solo información disponible**
        2. **Sustitución de datos ausentes**
    4. **Modelos que construyen explícitamente el proceso de datos ausentes**

1 - Aproximación de casos completos *(listwise)*

Consiste en **incluir sólo aquellos casos con datos completos**.

Ventajas

- La aproximación **más simple y directa**
- Este método se encuentra en todos los programas estadísticos y suele ser uno de los métodos por defecto.

Inconvenientes

- Los resultados pueden no son generalizables para la población.
- El tamaño de la muestra resultante puede quedar reducida a una muestra inapropiada para el propósito del análisis.

Uso con MCAR

- Situaciones en las que la extensión de la ausencia de datos es pequeña
- La muestra es suficientemente grande
- Las relaciones entre los datos son tan fuertes que no pueden verse afectadas por cualquier proceso de datos ausentes.

Ejemplo[1]

[1]Hair, J., Anderson, R., Tatham, R., & Black, W. (1999). Análisis Multivariante

Muestra con casos completos

Variables	5
Casos	5
Observaciones	25
N	25
Casos completos	5
	100%

		Caso Id	V1	V2	V3	V4	V5	Ausentes por caso Número	%
		1	1,3	9,9	6,7	3	2,6	0	0
		7	0,2	8,8	4,5	3	2,4	0	0
		8	2,1	8	3	3,8	1,4	0	0
		12	4,5	6,4	5,3	3	2,5	0	0
		20	1,8	9	5	2,2	3	0	0
Asuntes por variable	Número		0	0	0	0	0	Total ausentes 0Número	
	%		0	0	0	0	0	0%%	

N = 5

Podría ser una muestra inaceptable

2 - Eliminación de casos y/o variables

Consiste en **suprimir el caso(s) y/o variable(s) que peor se comporta(n) respecto a los datos ausentes**.

Ventajas

- Aproximación **simple** y también abordable por cualquier programa estadístico

Inconvenientes

- El tamaño de la muestra queda reducido.
- Se puede perder información al no contar con una determinada variable en el análisis multivariante.

- Cualquier decisión deberá basarse en consideraciones empíricas y teóricas.

Ejemplo[1]

[1]Hair, J., Anderson, R., Tatham, R., & Black, W. (1999). Análisis Multivariante

Muestra original

Variables	5
Casos	20
Observaciones	100
N	77
Caos completos	5
	25%

¿Eliminar alguna variable y/o caso?

Caso Id	V1	V2	V3	V4	V5	Ausentes por caso Número	%
1	1,3	9,9	6,7	3	2,6	0	0
2	4,1	5,7			2,9	2	40
3		9,9		3		3	60
4	0,9	8,6		2,1	1,8	1	20
5	0,4	8,3		1,2	1,7	1	20
6	1,5	6,7	4,8		2,5	1	20
7	0,2	8,8	4,5	3	2,4	0	0
8	2,1	8	3	3,8	1,4	0	0
9	1,8	7,6		3,2	2,5	1	20
10	4,5	8		3,3	2,2	1	20
11	2,5	9,2		3,3	3,9	1	20
12	4,5	6,4	5,3	3	2,5	0	0
13					2,7	4	80
14	2,8	6,1	6,4		3,8	1	20
15	3,7			3		3	60
16	1,6	6,4	5		2,1	1	20
17	0,5	9,2		3,3	2,8	1	20
18	2,8	5,2	5		2,7	1	20
19	2,2	6,7		2,6	2,9	1	20
20	1,8	9	5	2,2	3	0	0
Asuntes por variable Número	2	2	11	6	2	Total ausentes 23	Número
%	10	10	55	30	10	23,0%	%

Ejemplo[1]

[1]Hair, J., Anderson, R., Tatham, R., & Black, W. (1999). Análisis Multivariante

Muestra original SIN V3

Variables	4
Casos	20
Observaciones	80
N	68
Casos completos	12
	60%

Eliminar
- Variable V3

Caso Id	V1	V2	V4	V5	Ausentes por caso Número	%
1	1,3	9,9	3	2,6	0	0
2	4,1	5,7		2,9	1	25
3		9,9	3		2	50
4	0,9	8,6	2,1	1,8	0	0
5	0,4	8,3	1,2	1,7	0	0
6	1,5	6,7		2,5	1	25
7	0,2	8,8	3	2,4	0	0
8	2,1	8	3,8	1,4	0	0
9	1,8	7,6	3,2	2,5	0	0
10	4,5	8	3,3	2,2	0	0
11	2,5	9,2	3,3	3,9	0	0
12	4,5	6,4	3	2,5	0	0
13				2,7	3	75
14	2,8	6,1		3,8	1	25
15	3,7		3		2	50
16	1,6	6,4		2,1	1	25
17	0,5	9,2	3,3	2,8	0	0
18	2,8	5,2		2,7	1	25
19	2,2	6,7	2,6	2,9	0	0
20	1,8	9	2,2	3	0	0
Asuntes por variable Número	2	2	6	2	Total ausentes 12	Número
%	10	10	30	10	15,0%	%

Ejemplo[1]

[1]Hair, J., Anderson, R., Tatham, R., & Black, W. (1999). Análisis Multivariante

Muestra original SIN V3, ni casos 3, 13, 15

Variables	4
Casos	17
Observaciones	68
N	63
Casos completos	12
	71%

Eliminar
- **Variable V3**
- **Casos 3, 13, 15**

Caso Id	V1	V2	V4	V5	Ausentes por caso	
					Número	%
1	1,3	9,9	3	2,6	0	0
2	4,1	5,7		2,9	1	25
4	0,9	8,6	2,1	1,8	0	0
5	0,4	8,3	1,2	1,7	0	0
6	1,5	6,7		2,5	1	25
7	0,2	8,8	3	2,4	0	0
8	2,1	8	3,8	1,4	0	0
9	1,8	7,6	3,2	2,5	0	0
10	4,5	8	3,3	2,2	0	0
11	2,5	9,2	3,3	3,9	0	0
12	4,5	6,4	3	2,5	0	0
14	2,8	6,1		3,8	1	25
16	1,6	6,4		2,1	1	25
17	0,5	9,2	3,3	2,8	0	0
18	2,8	5,2		2,7	1	25
19	2,2	6,7	2,6	2,9	0	0
20	1,8	9	2,2	3	0	0
Asuntes por variable: Número	0	0	5	0	Total ausentes: 5	Número
Asuntes por variable: %	0	0	29,41	0	7,4%	%

3 - Imputación

La **imputación** es el proceso de estimación de valores ausentes y/o de los parámetros muestrales basado en los valores válidos de la variable evaluada y/o otras variables y/o casos de la muestra.

- **Para estimar** los valores ausentes (o parámetros):
  - Emplear **parámetros muestrales de los valores válidos** que caracterizan a la variable
  - **Estimar los valores ausentes** usando **relaciones** conocidas que puedan identificarse en los valores válidos de la muestra.
  - Otros procedimientos (sustitución de los valores ausentes por constantes, valores al azar,…)

- **Uso en variables cuantitativas!!!!!**

  Se debe considerar cuidadosamente el uso de la imputación en cada situación particular por sus potenciales impactos sobre el análisis:
  - **Subestimar la varianza, ya que disminuye la variabilidad**
  - **Distorsionar la distribución de la v.a.**

3.1 - Métodos de disponibilidad completa

Utilizan toda la información disponible a partir de un subconjunto de casos para inferir sobre la muestra entera *(**pairwise**)*.

No **se sustituyen** los valores ausentes, pero **sí algunos** de sus **parámetros muestrales** como **estimaciones calculadas a partir de los valores presentes**.

Uso

- Para estimar medias, varianzas y correlaciones

Ventajas

- Maximiza la información disponible en la muestra
- Evita el problema de tener que eliminar un caso entero por una sola o pocas variables con uno o pocos datos ausentes

Inconvenientes

- Las correlaciones pueden calcularse «fuera de rango» y de forma inconsistente con otras correlaciones de la matriz de correlación.
- Los autovalores de la matriz de correlación pueden tomar valores negativos.

Ejemplo[1]

[1]Hair, J., Anderson, R., Tatham, R., & Black, W. (1999). Análisis Multivariante

Muestra original

Variables	5
Casos	20
Observaciones	100
N	77
Caos completos	5
	25%

¿Eliminar alguna variable y/o caso?

Caso Id	V1	V2	V3	V4	V5	Ausentes por caso Número	%
1	1,3	9,9	6,7	3	2,6	0	0
2	4,1	5,7			2,9	2	40
3		9,9		3		3	60
4	0,9	8,6		2,1	1,8	1	20
5	0,4	8,3		1,2	1,7	1	20
6	1,5	6,7	4,8		2,5	1	20
7	0,2	8,8	4,5	3	2,4	0	0
8	2,1	8	3	3,8	1,4	0	0
9	1,8	7,6		3,2	2,5	1	20
10	4,5	8		3,3	2,2	1	20
11	2,5	9,2		3,3	3,9	1	20
12	4,5	6,4	5,3	3	2,5	0	0
13					2,7	4	80
14	2,8	6,1	6,4		3,8	1	20
15	3,7			3		3	60
16	1,6	6,4	5		2,1	1	20
17	0,5	9,2		3,3	2,8	1	20
18	2,8	5,2	5		2,7	1	20
19	2,2	6,7		2,6	2,9	1	20
20	1,8	9	5	2,2	3	0	0
Asuntes por variable Número	2	2	11	6	2	Total ausentes 23	Número
%	10	10	55	30	10	23,0%	%

3.2 - Métodos de sustitución

Estiman los valores de reemplazo para los datos ausentes, sobre la base de otra información existente en la muestra.

Tipos

- **Sustitución de caso**:
  - Un caso se sustituye por otro que esté disponible
- **Sustitución por parámetros muestrales**:
  - Media
  - Mediana
- **Sustitución por observaciones no muestrales**:
  - una constante
  - un valor aleatorio de entre los valores posibles de la v.a
- **Sustitución por regresión**: a partir de estimaciones realizadas con otras variables muy relacionadas

Sustitución de caso

- Las observaciones (casos) con datos ausentes se sustituyen con otras observaciones (casos) que no estaban inicialmente en la muestra, si es posible.
- Este método es el que más se utiliza para sustituir las observaciones con datos ausentes completos, aunque también puede emplearse para reemplazar observaciones con menores cantidades de datos ausentes.

Ejemplo:

En una encuesta sobre hogares, reemplazar los datos de un individuo (hogar) que está en la muestra, pero con el que no se puede contactar o que tiene gran cantidad de datos ausentes con otro hogar que no está en la muestra, preferiblemente muy similar al de la observación original.

Sustitución por parámetros muestrales

Consiste en **sustituir los valores ausentes de una v.a. por la media calculada sobre todos los valores válidos de dicha v.a.**.

La lógica de esta aproximación es que la media es el mejor valor de sustitución.

Uso

- Es uno de los métodos más empleados

Ventajas

- Se puede implementar fácilmente y proporciona información completa para todos los casos.

Inconvenientes

- Subestima la varianza y, por tanto, invalida las estimaciones a partir de ésta.
- Distorsiona la distribución real de los valores.
- Modifica la correlación observada porque todos los datos ausentes tendrán un valor único constante.

NOTA: en v.a asimétricas puede usarse la **mediana** en ves de la media

Sustitución por observaciones no muestrales

Consiste en **sustituir los datos ausentes de una v.a. por un valor constante derivado de fuentes externas o investigación previa**.

La lógica de esta aproximación es que la constante utilizada es el mejor valor de sustitución.

Uso

- El valor de sustitución de una fuente externa es más válido que el valor generado internamente por la media

Ventajas (como el anterior)

- Se puede implementar fácilmente y proporciona información completa para todos los casos.

Inconvenientes (como el anterior)

- Invalida las estimaciones de la varianza.
- Distorsiona la distribución real de los valores.
- Modifica la correlación observada porque todos los datos ausentes tendrán un valor único constante.

Sustitución por regresión

Consiste en **sustituir los datos ausentes de una v.a. por las predicciones realizadas mediante modelos de regresión**.

Uso

- El valor de sustitución estimado es más válido que la media, …

Inconvenientes (como el anterior)

- Sobrestima las relaciones ya existentes en los datos.
- Conforme aumente el uso de este método, los datos resultantes son más característicos de la muestra y menos generalizables.
- Se subestima la varianza de la distribución
- Se asume la existencia de relaciones significativas entre la v.a. a imputar y el resto de v.a.
- Los valores predichos puede que no correspondan a los rangos válidos de las variables

Imputación múltiple

Es en realidad una combinación de varios métodos.

La lógica de esta aproximación es que el uso de la aproximación múltiple minimiza los problemas específicos con cualquier método simple siendo su composición la mejor estimación.

La elección de esta aproximación se basa fundamentalmente en la concesión mutua entre la percepción del investigador de los potenciales beneficios ponderada por el esfuerzo sustancialmente superior que requiere realizar y combinar las múltiples estimaciones.

DATATHON 2022
#Oddatathon
- Producción y consumo responsable
- Aspectos medioambientales
- Cultura

Iniciación a la estadística

Elena Vázquez Barrachina
evazquez@eio.upv.es

Ángeles Calduch Losa
mcalduch@eio.upv.es