UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Departamento de Estadística e Investigación Operativa Aplicadas y Calidad

www.upv.es bigdata.inf.upv.es

Herramientas estadísticas para Big Data

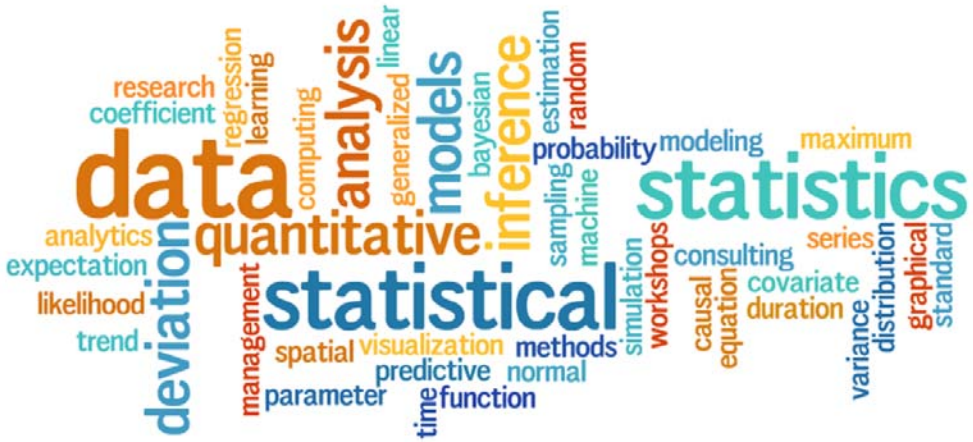
Introducción a la Inferencia Estadística, Muestreo y Preproceso de datos

Máster Big Data Analytics

Valencia, Octubre 2017

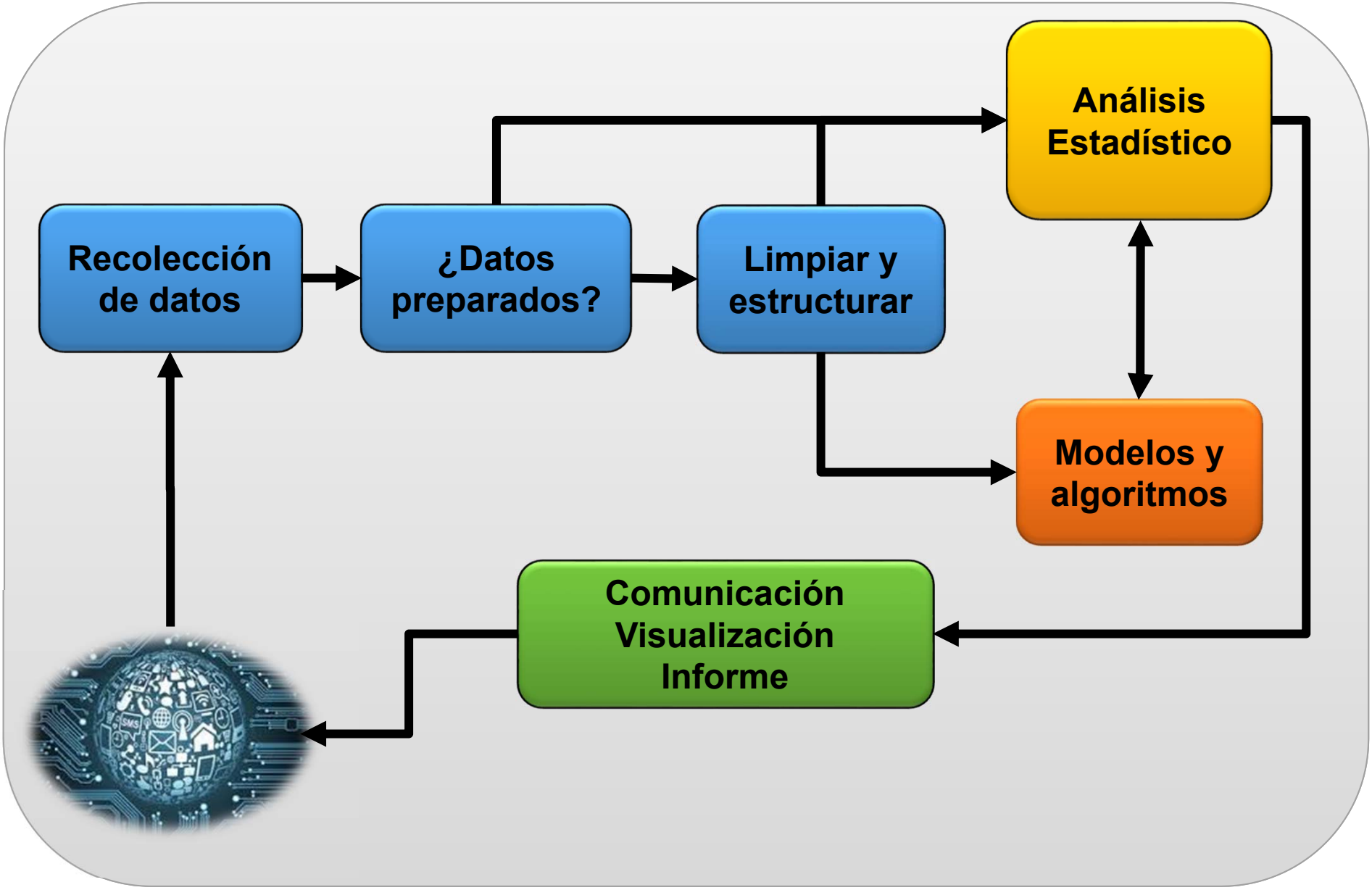
Elena Vázquez

Contenidos



1 Conceptos básicos

Introducción



El proceso de investigación estadística

La pregunta

- **Observación inicial**
  - **Encontramos algo que necesita explicación**
    - Observando el mundo real
    - Leyendo alguna investigación,…
  - **Justificar la posible explicación**
    - Recolectar **datos** para ver si nuestra "intuición" es correcta

- **Teorías e hipótesis**

Qué medir

Datos vs información

- Lo más importante en *data science* es la pregunta
- La segunda cosa más importante son los datos
- Los datos determinan el tipo de preguntas que podemos hacer, pero son inútiles si no se formula la pregunta adecuada
- Los datos por si mismos no contienen la respuesta, necesitamos (Big o Small) los datos y procedimientos adecuados

Variables, observaciones y casos

```
## Cargar data frame
> load("EjemploBD.RData")
## Visualizar dataset
> View(datos1)
```

Editor de datos

	SEXO	EDAD	MES	ESTATURA	PESO	PROVINCIA
1	varon	20	1	183	76	Castellon
2	varon	21	6	185	72	Alicante
3	varon	22	10	165	75	Teruel
4	varon	22	4	174	70	Teruel
5	varon	22	7	175	70	Teruel
6	varon	20	11	175	70	Valencia
7	varon	22	7	174	65	Teruel
8	mujer	23	10	159	54	Alicante
9	varon	21	3	177	72	Teruel
10	varon	21	6	180	73	Castellon
11	varon	21	3	175	68	Castellon
12	mujer	22	6	162	48	Castellon
13	mujer	20	4	160	65	Teruel
14	varon	20	12	185	74	Teruel
15	varon	20	9	182	72	Teruel
16	varon	20	10	198	88	Valencia
17	mujer	20	6	164	54	
18	mujer	20	5	162	55	Valencia
19	mujer	20	7	164	53	

Hoja de datos o *dataframe*

En el contexto de Data Science, el objeto que se utiliza para recoger la información sobre las características a estudiar es la **hoja de datos** o ***data frame***.

SEXO	EDAD	PROVINCIA	ESTATURA
varón	20	Castellón	183,2
varón	21	Castellón	185,4
varón	22	Alicante	165,7
varón	22	Valencia	174,5
varón	22	Valencia	175,8
mujer	20	Valencia	175,9
varón	22	Castellón	174,3
mujer	23	Alicante	159,1

Dato estadístico u observación

Unidad de análisis o caso

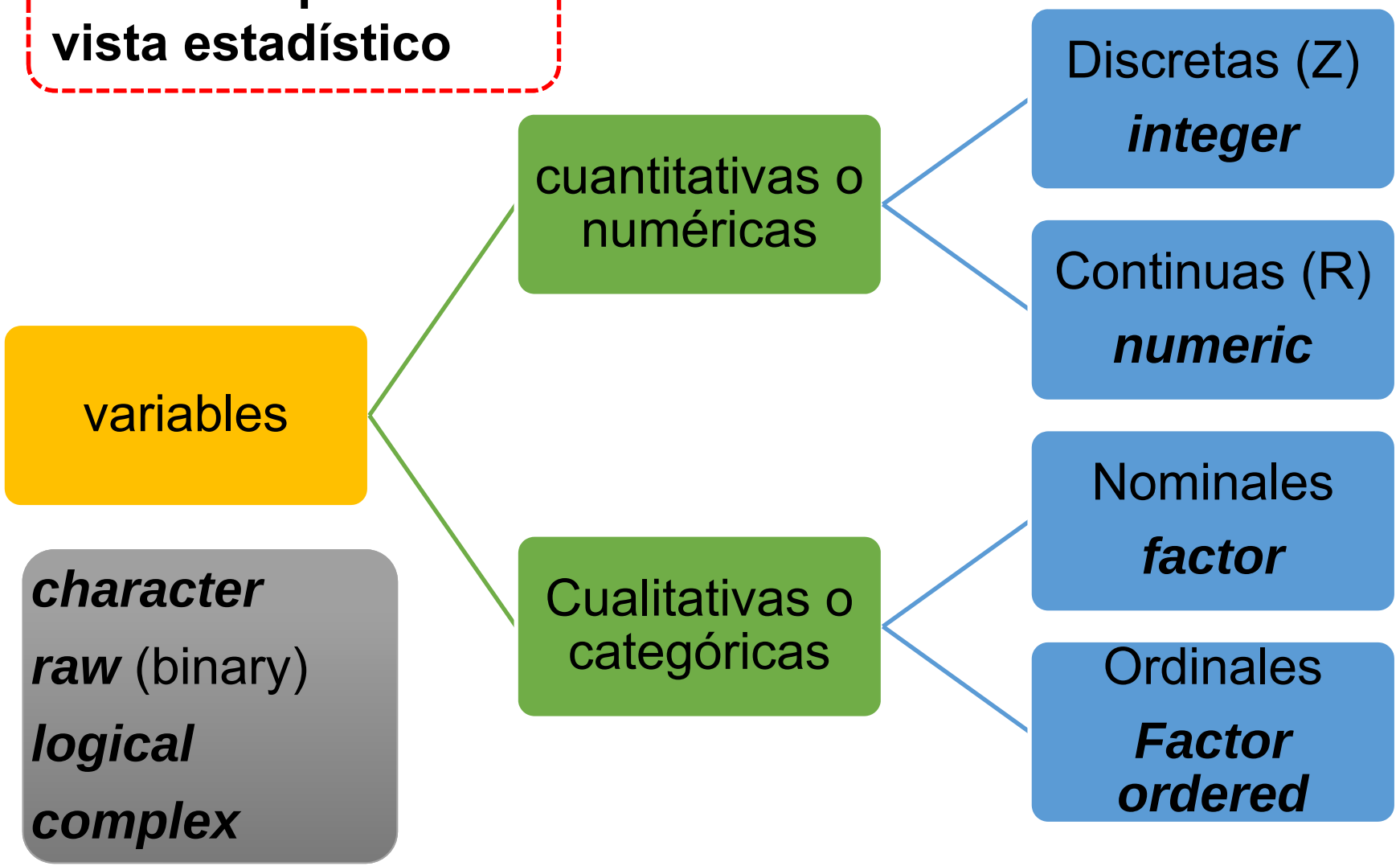
Variable

Variable aleatoria

- Características de los individuos estudiados susceptibles de adquirir diferentes valores o modalidades.
- Comprenden los valores de las características sobre las que podemos o necesitamos disponer de información
- Las herramientas o modelos estadísticos a utilizar dependen del tipo de variable aleatoria

Tipos de variables por su naturaleza

Desde un punto de vista estadístico



Tipos de variables

	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

SEXO: categórica nominal

EDAD: cuantitativa discreta

MES: categórica ordinal

ESTATURA: cuantitativa continua

Cadena, fecha, etc…

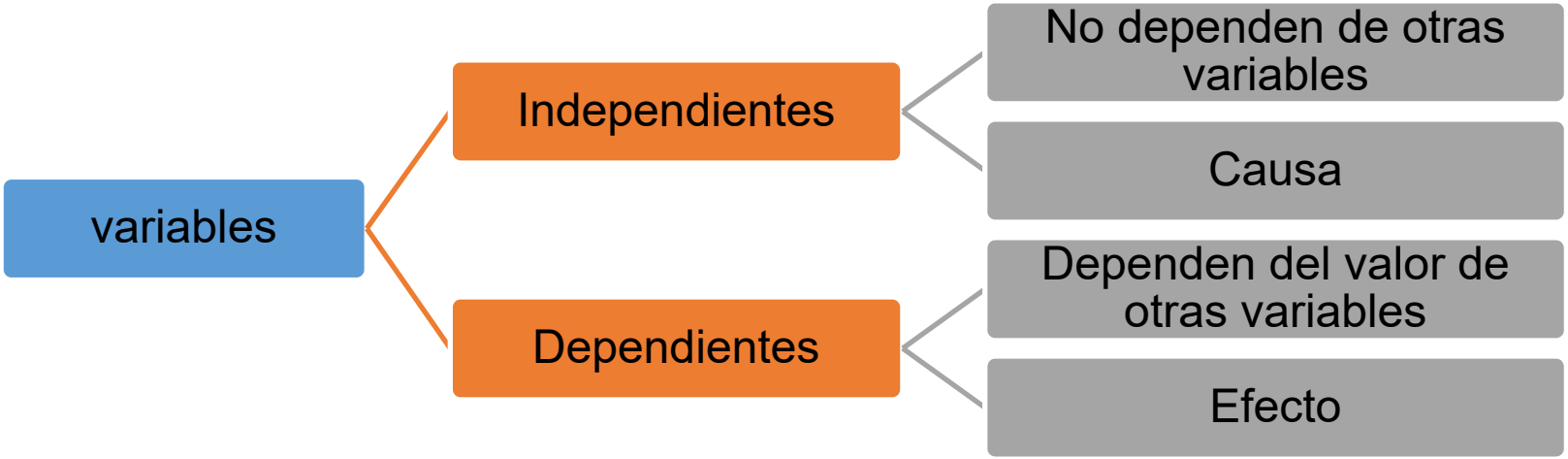
Tipos de variables

```
## Ver características de las variables del data.frame
> str(datos1)
'data.frame':    131 obs. of  12 variables:
 $ SEXO      : Factor w/ 2 levels "mujer           ",..: 2 2 2 2 2 2 2 1 2 2 ...
 $ EDAD      : int  20 21 22 22 22 20 22 23 21 21 ...
 $ MES       : int  1 6 10 4 7 11 7 10 3 6 ...
 $ ESTATURA  : int  183 185 165 174 175 175 174 159 177 180 ...
 $ PESO      : int  76 72 75 70 70 70 65 54 72 73 ...
 $ PROVINCIA : Factor w/ 5 levels "               ",..: 3 2 4 4 4 5 4 2 4 3 ...
 $ ACCESOS   : int  5 8 8 6 7 3 7 5 8 5 ...
 $ RESIDENCIA: Factor w/ 5 levels "familia         ",..: 4 2 1 5 4 1 2 1 4 4 ...
 $ TRANSPORTE: Factor w/ 6 levels "                ",..: 2 3 6 2 2 4 5 3 2 2 ...
 $ TIEMPO    : int  15 10 30 15 20 60 5 65 15 15 ...
 $ DEPORTE   : Factor w/ 6 levels "                ",..: 4 2 4 3 4 5 4 6 6 3 ...
 $ GASTO     : num  8.35 8.01 7.79 11.07 10.57 ...
```

Tipos de variables por su relación

Desde un punto de vista estadístico

- Si los valores que tome una variable dependen del que tome otra, las dos variables son dependientes.
- La dependencia no implica causalidad



Conceptos básicos

- Error de medida
- Validez y Fiabilidad
- Tipos de estudios estadísticos
- Variabilidad
- Aleatorización
- Objetivos estadística

Error de medida

- **Discrepancia entre el valor real y el observado** de la característica medida.
  - Discrepancia entre el valor real y el número que usamos para representar dicho valor.

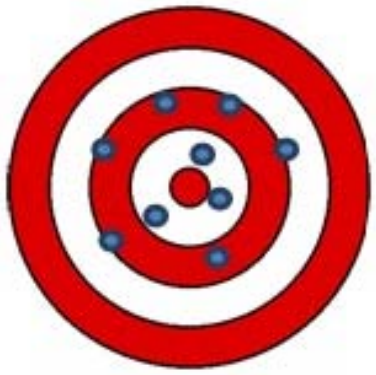
- **Ejemplo**:
  - Yo peso 60 Kg (en realidad)
  - En mi peso pone 63 kg
  - El error de medida es 3 Kg

Validez y Fiabilidad

- **Minimizar el error de medida**
  - **Validez**:
    - Capacidad de que un instrumento mida correctamente aquello para lo que ha sido diseñado.
  - **Fiabilidad**:
    - Capacidad de que un instrumento produzca los mismos resultados bajo las mismas condiciones

Validez y Fiabilidad

Fiable pero no válido



Válido pero no fiable

Ni fiable ni válido

Fiable y válido

Tipos de estudios

- **Tipos de estudios**
  - Estudios observacionales o correlacionales
    - Estudios longitudinales
    - *Cross-sectional*
  - Estudios experimentales

- Correlación vs Causa
- Causa y efecto

Experimentos y recolección de datos

- Medidas entre grupos, entre individuos o independientes (*between-groups*) :
  - Diferentes entes en las distintas condiciones experimentales
- Medidas repetidas (*within-subject*)
  - Los mismos entes participan en todas las condiciones experimentales.

Tipos de variabilidad

- Variabilidad sistemática y planificada. (“DESEABLE”)

- Variabilidad típica de la naturaleza del problema y del experimento. (“TOLERABLE”)(ruido aleatorio)
  - Azar del muestreo
  - Error de medida
- Variabilidad sistemática y no planificada.(“AMENAZA CON EL DESASTRE”)

- **Aleatorización**

Tipos de variabilidad

- **Variabilidad sistemática**: variabilidad que puede explicarse mediante el modelo que queremos ajustar a los datos (no debida al azar del muestreo).

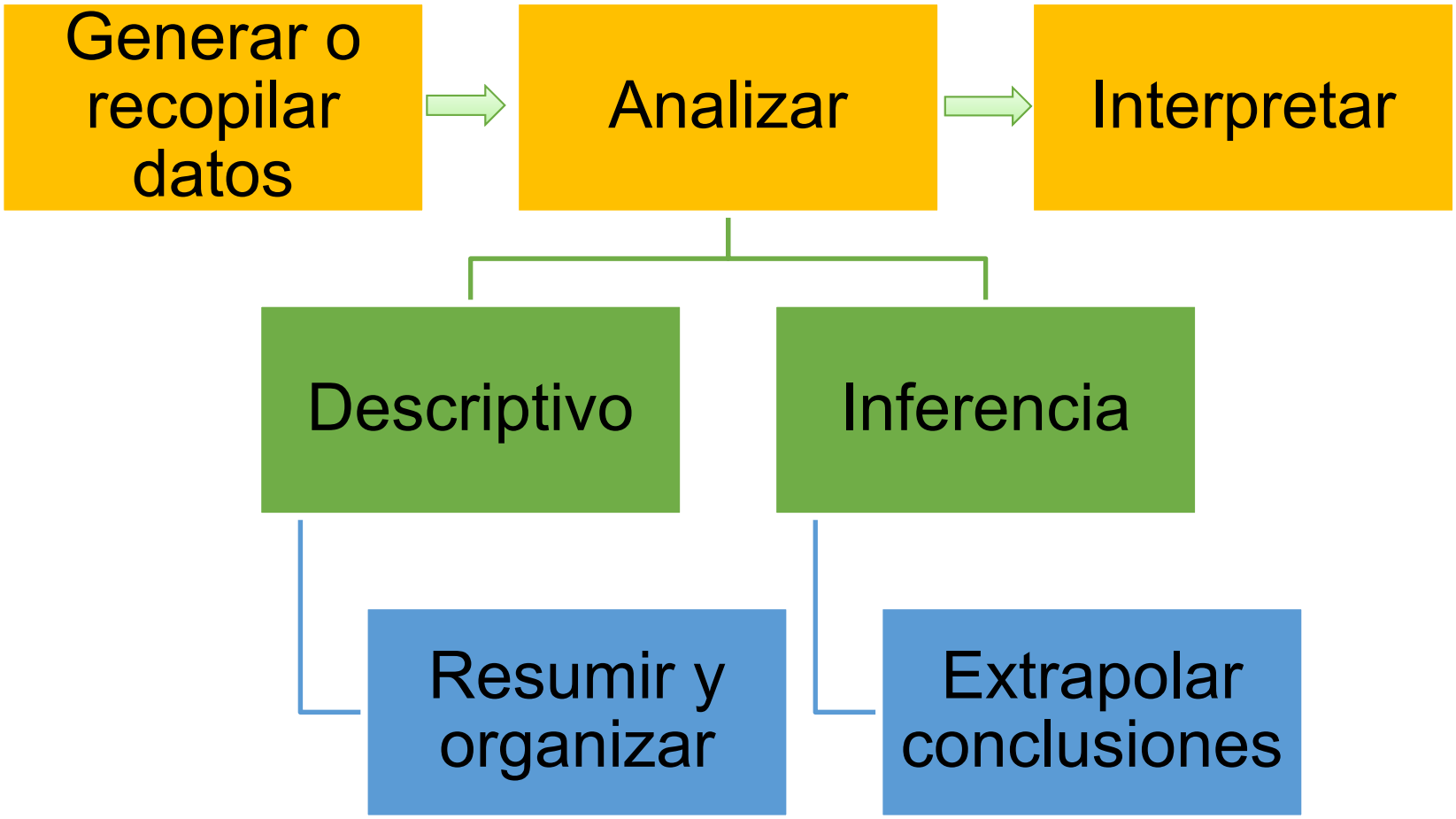
- **Variabilidad no sistemática**: variabilidad típica de la naturaleza del problema y del experimento, error o variación no atribuible al efecto que se está investigando (debida al azar del muestreo).

Aleatorización

- Es el procedimiento que permite que cada unidad experimental tenga iguales condiciones para recibir cualquier tratamiento.
- ¿Puede un análisis estadístico sofisticado resolver el problema planteado por el mal diseño utilizado? → NO

- Los **efectos** que un **mal diseño** ha "**confundido**" completamente, **no** puede "**separarlos**" **ningún** tipo de **análisis**.
- La aleatorización es la única garantía para evitar que el experimento conduzca a conclusiones sesgadas o erróneas.

Objetivos Estadística



Análisis de datos: técnicas estadísticas

Variables independientes				
Nivel de medida	Nominal u Ordinal		Intervalo o Razón	
Variables dependientes	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 nominal u ordinal.	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 nominal u ordinal.	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 de intervalo o razón.	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 de intervalo o razón.	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.

Análisis descriptivo de datos

Primer paso de cualquier análisis

Tratamiento descriptivo o exploratorio:

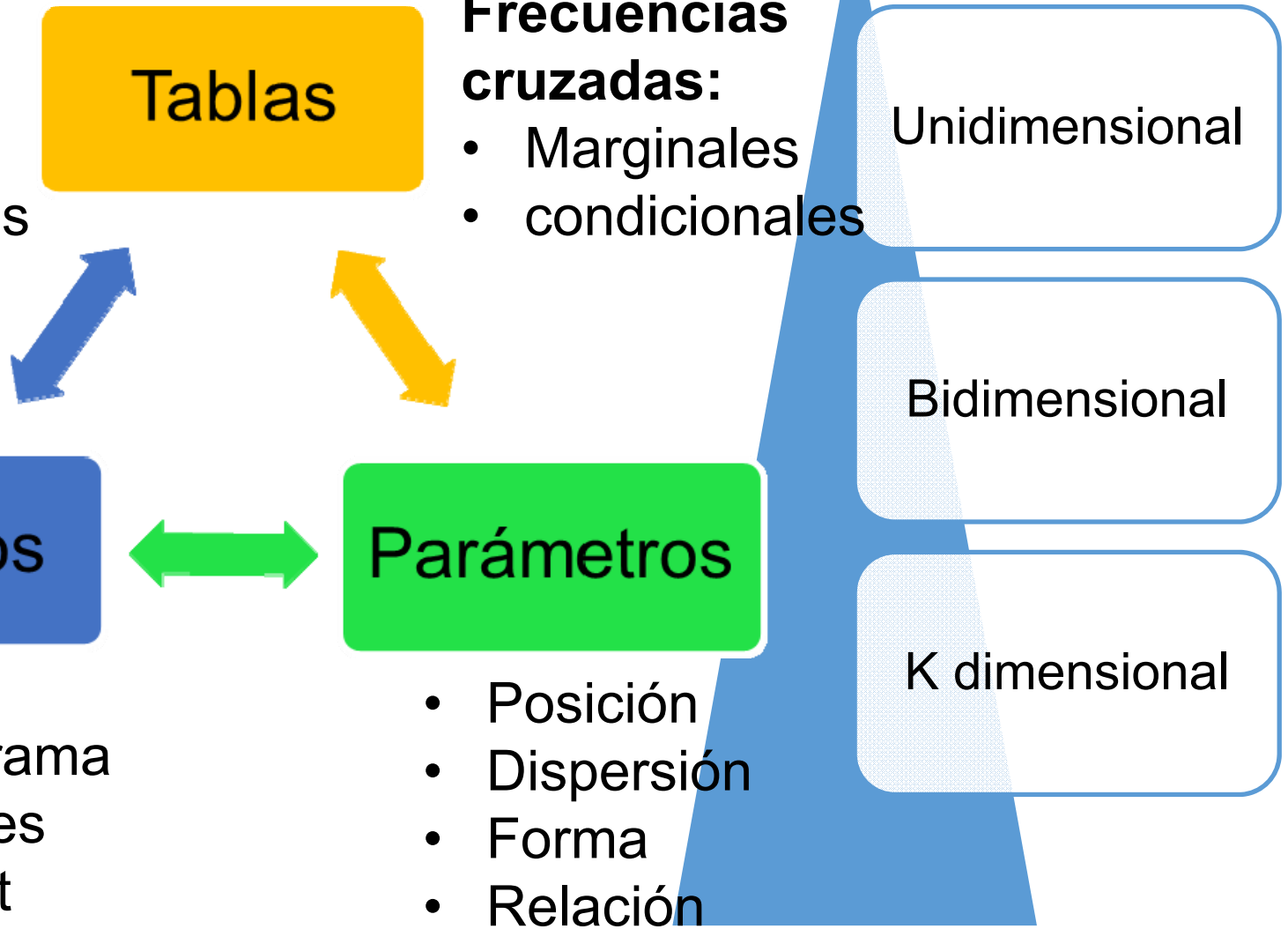
- ✓ Sintetizar y simplificar la presentación de los datos →
- ✓ Disponer los valores observados de forma clara y útil para su interpretación:
  - – Poner de manifiesto características y regularidades de los datos
  - – Detectar errores e inconsistencias

¡¡No se extrapolan todavía conclusiones de los datos a la población!!

Los detalles sobre este análisis no pretenden ser exhaustivos, dado el tiempo disponible y la accesibilidad a ellos en cualquier texto básico de estadística.
Algunas cuestiones adicionales se abordarán en la parte de preparación de datos.

Análisis descriptivo

Frecuencias:
- Absolutas
- Relativas
- Acumuladas

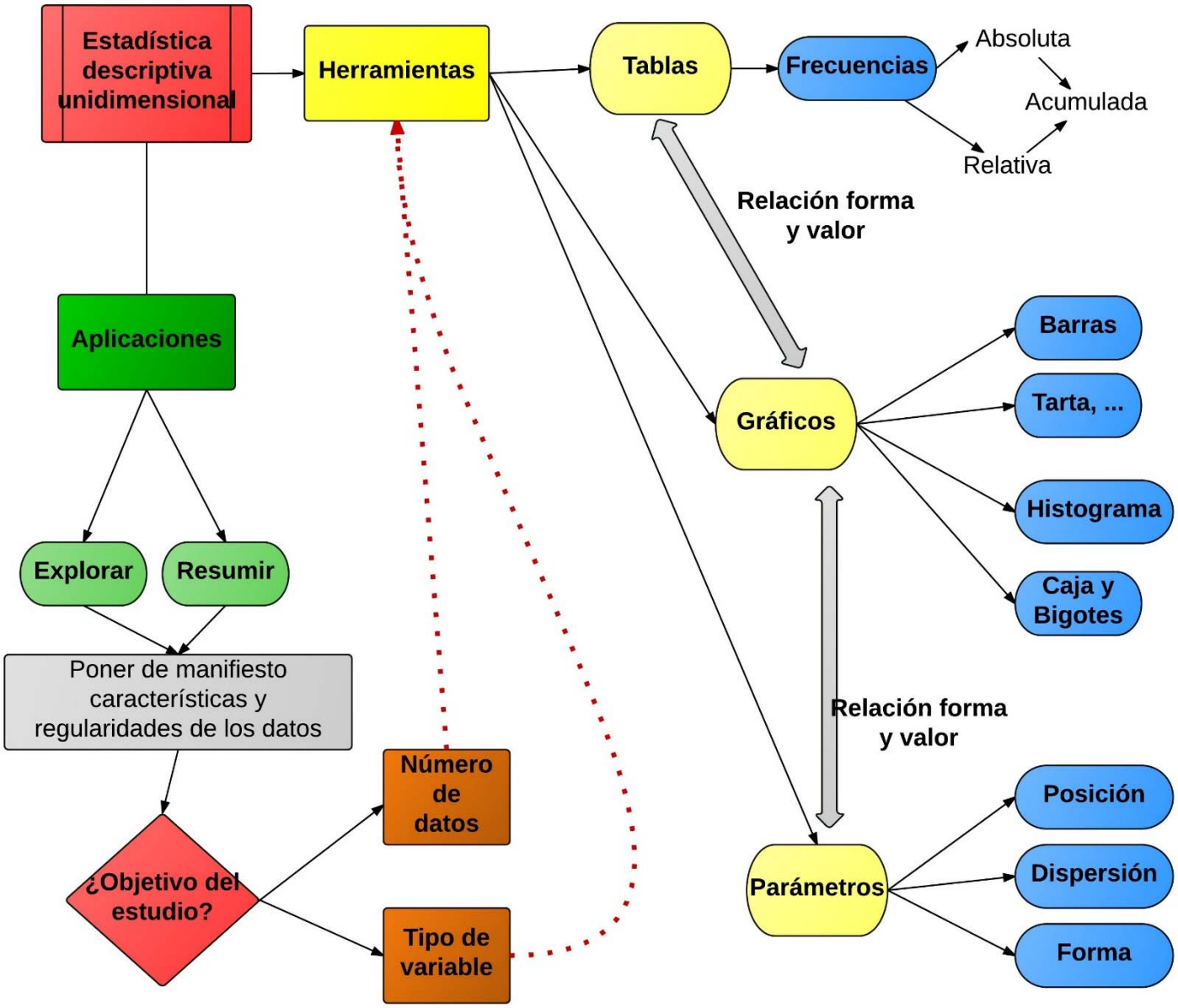


Frecuencias cruzadas:
- Marginales
- condicionales

- Barras
- Histograma
- Sectores
- Boxplot
- …

- Posición
- Dispersión
- Forma
- Relación

Análisis descriptivo unidimensional



Técnicas básicas de estadística descriptiva unidimensional

Variable		Estadísticos	Gráficos
Tipo	Subtipo		
Cualitativa/Discreta/Categórica	Nominal	Frecuencias	Diagrama de barras Diagrama de sectores
Cualitativa/Discreta/Categórica	Ordinal	Frecuencias Moda, mediana Rango, cuartiles	Diagrama de barras Box-plot
Cuantitativa/Continua		Media, mediana, moda Varianza, desviación típica Coeficiente de variación Percentiles …	Histograma Box-plot

Tablas de frecuencias

- v.a. Cualitativas
- V.a. Cuantitativas discretas con pocos valores distintos
- V.a. Cuantitativas continuas o discretas con muchos valores distintos discretizada

Tabla de frecuencias

- Variables cualitativas
- Variables cuantitativas discretas con pocos valores
- Variables cuantitativas continuas: recodifiicar y agrupar

Valores variable	**Frecuencia absoluta**	**Frecuencia relativa %**	**Frecuencia absoluta acumulada**	**Frecuencia relativa acumulada %**
x_1	f_1	$f_{r1} = f_1/N*100$	$F_1 = f_1$	$F_{r1} = f_{r1}$
X_2	f_2	$f_{r2} = f_2/N*100$	$F_2 = F_1 + f_2$	$F_{r2} = F_{r1} + f_{r2}$
X_3	f_3	$F_{r2} = f_3/N*100$	$F_3 = F_2 + f_3$	$F_{r3} = F_{r2} + f_{r3}$
...	...	...	N	100%
Total	N	100%		

Tabla de frecuencias: ejemplo

Nº de procesadores	Nº de robots	% robots	Nº de robots acumulado	% robots acumulado
0	10	6,25%	10	6,25%
1	35	21,88%	45	28,13%
2	60	37,50%	105	65,63%
3	55	34,37%	160	100%
Total	160	100%		

Diagramas de Barras Y Sectores

- v.a. Cualitativas
- V.a. Cuantitativas discretas con pocos valores distintos
- Forma gráfica de la Tabla de frecuencias

Diagrama de Barras

- A cada alternativa de la **característica cualitativa** se le hace corresponder un barra (horizontal o vertical) cuya altura es proporcional a la **frecuencia** observada (**absoluta, relativa, acumulada**) en la muestra para ese "valor" de la característica estudi

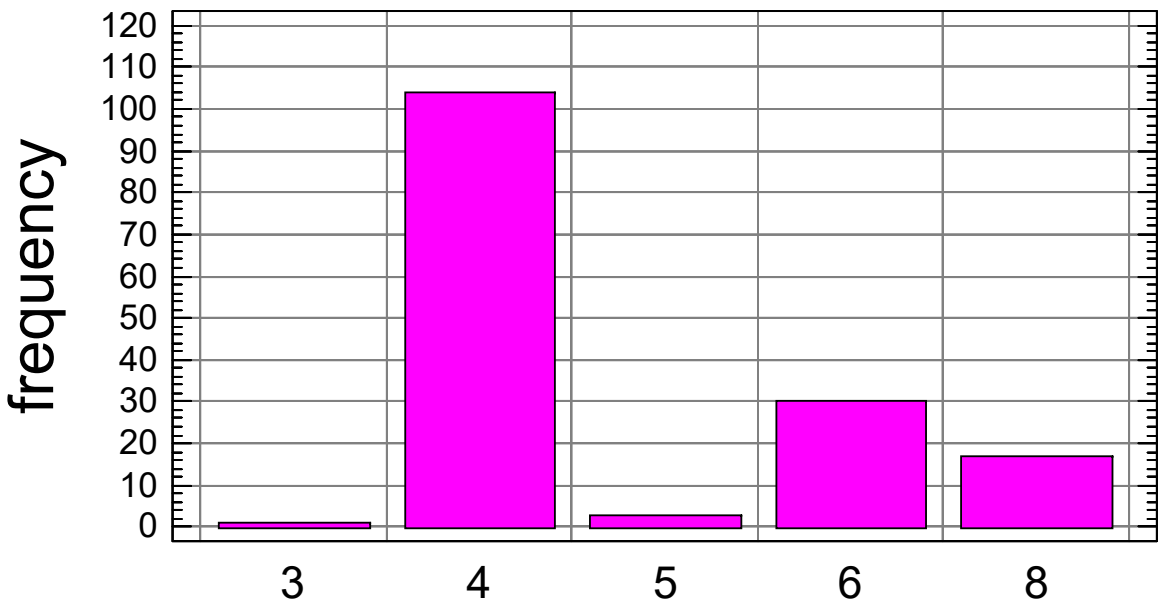


Diagrama de Barras

Equipamiento de las viviendas productos TIC (% hogares)

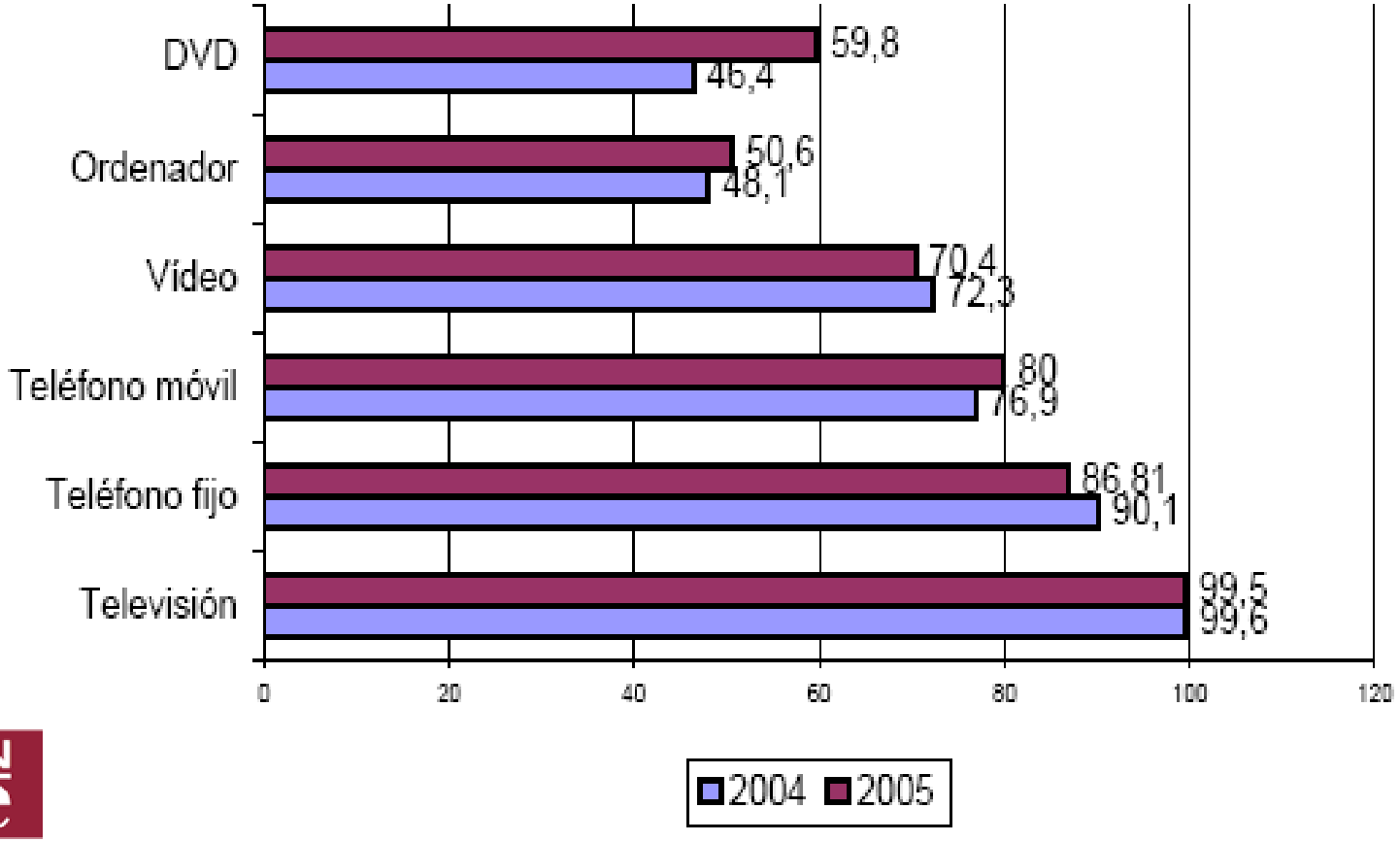


Diagrama de sectores o tarta

- La superficie total de un círculo se reparte en **sectores** cuyas áreas son proporcionales a las frecuencias observadas en la muestra para cada “valor” de la característica estudiada.
- Frecuencias absolutas o relativas

Histogramas

v.a. Cuantitativas:

- Continuas
- Discretas con muchos valores distintos

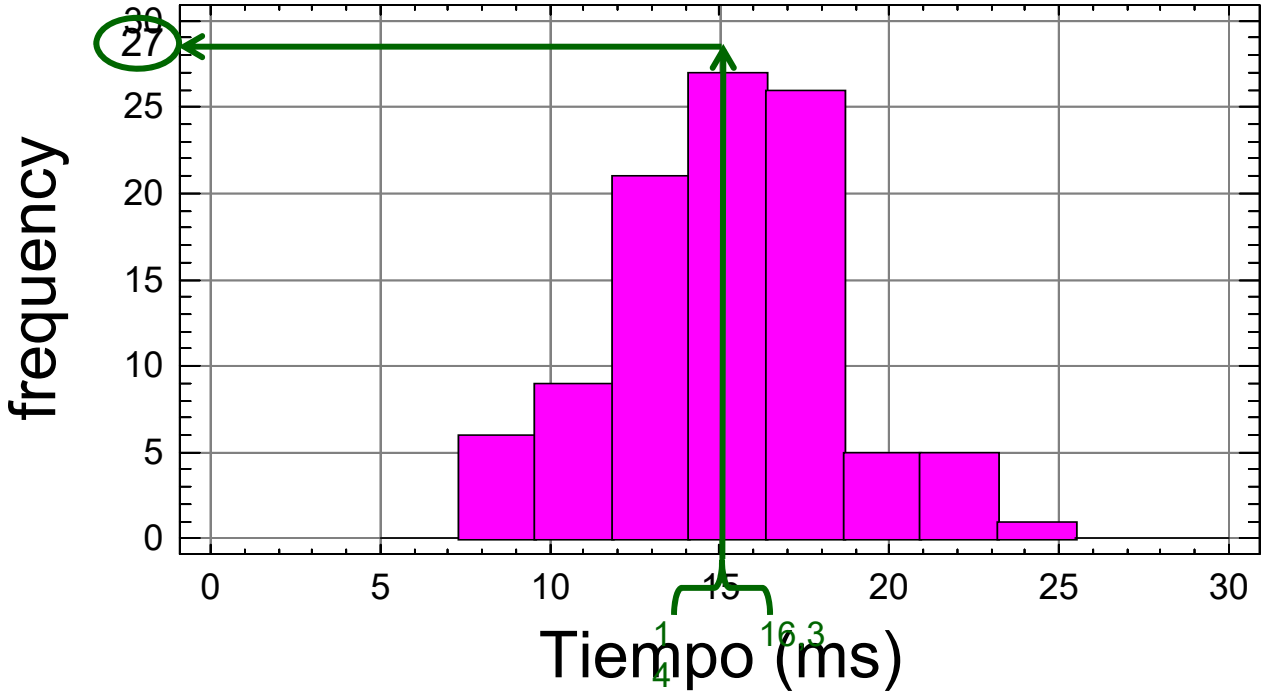
Histograma ¿Qué es?

- Es un diagrama de barras para variables cuantitativas continuas o discretas con muchos valores
- Es una representación gráfica de un conjunto de datos (mínimo 40-50 datos)
- Para cada valor o intervalo de valores de la variable (eje de abscisas) se levanta una barra de altura proporcional a la frecuencia con que aparece dicha variable los valores del intervalo (absoluta o relativa)
- Nº de intervalos
  - regla empírica: entero cercano a $\sqrt{n}$
  - en general entre 15-20 intervalos

Ejemplo

v.a.: Tiempo de ejecución (ms) de 100 programas

Frecuencias absolutas (número de programas)

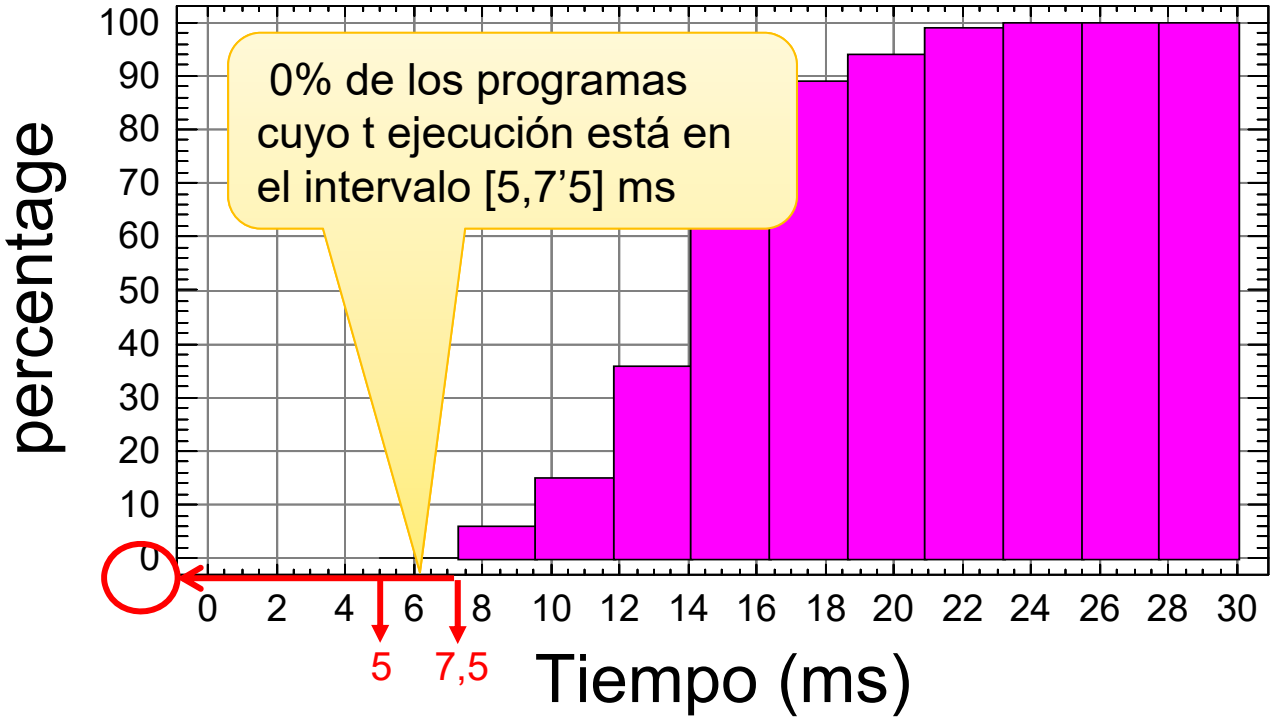


27 programas cuyo tiempo de ejecución ha estado entre 14 y 16,3 ms. (aprox.)

Ejemplo

v.a.:Tiempo de ejecución (ms) de 100 programas

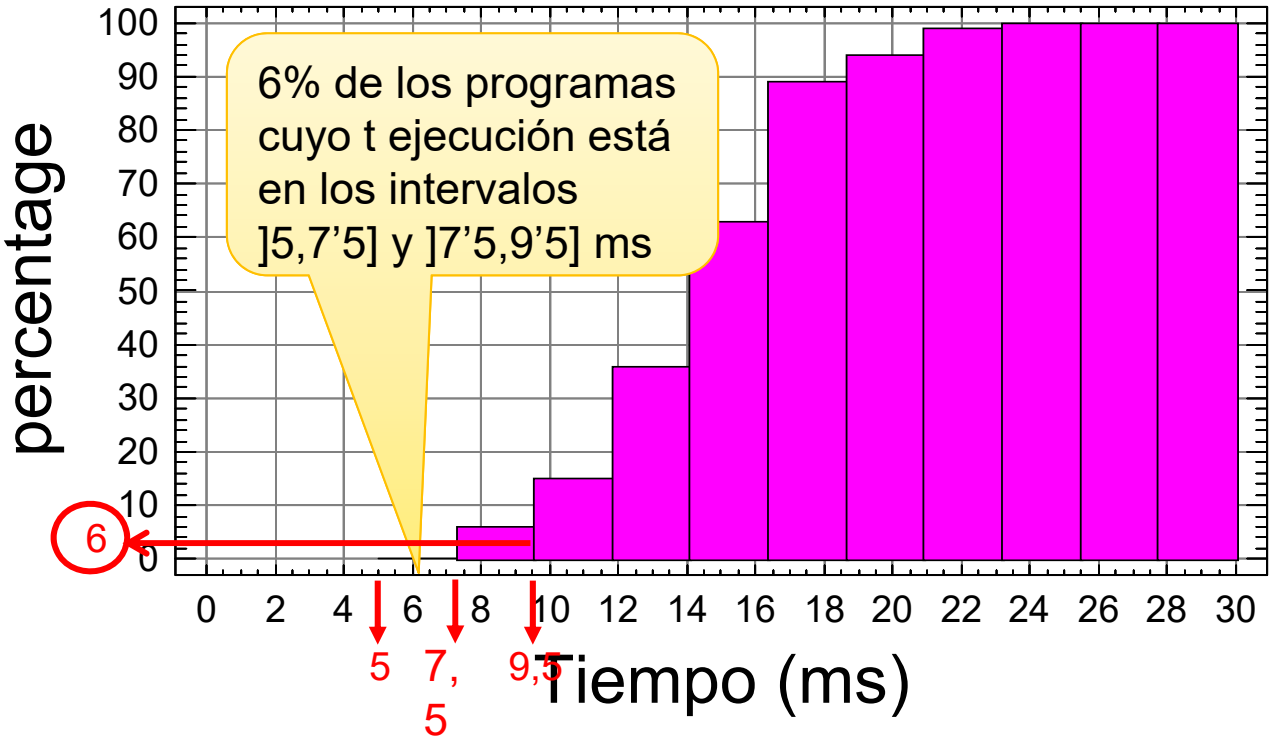
Frecuencias relativas acumuladas (% de programas)



Ejemplo

v.a.:Tiempo de ejecución (ms) de 100 programas

Frecuencias relativas acumuladas (% de programas)



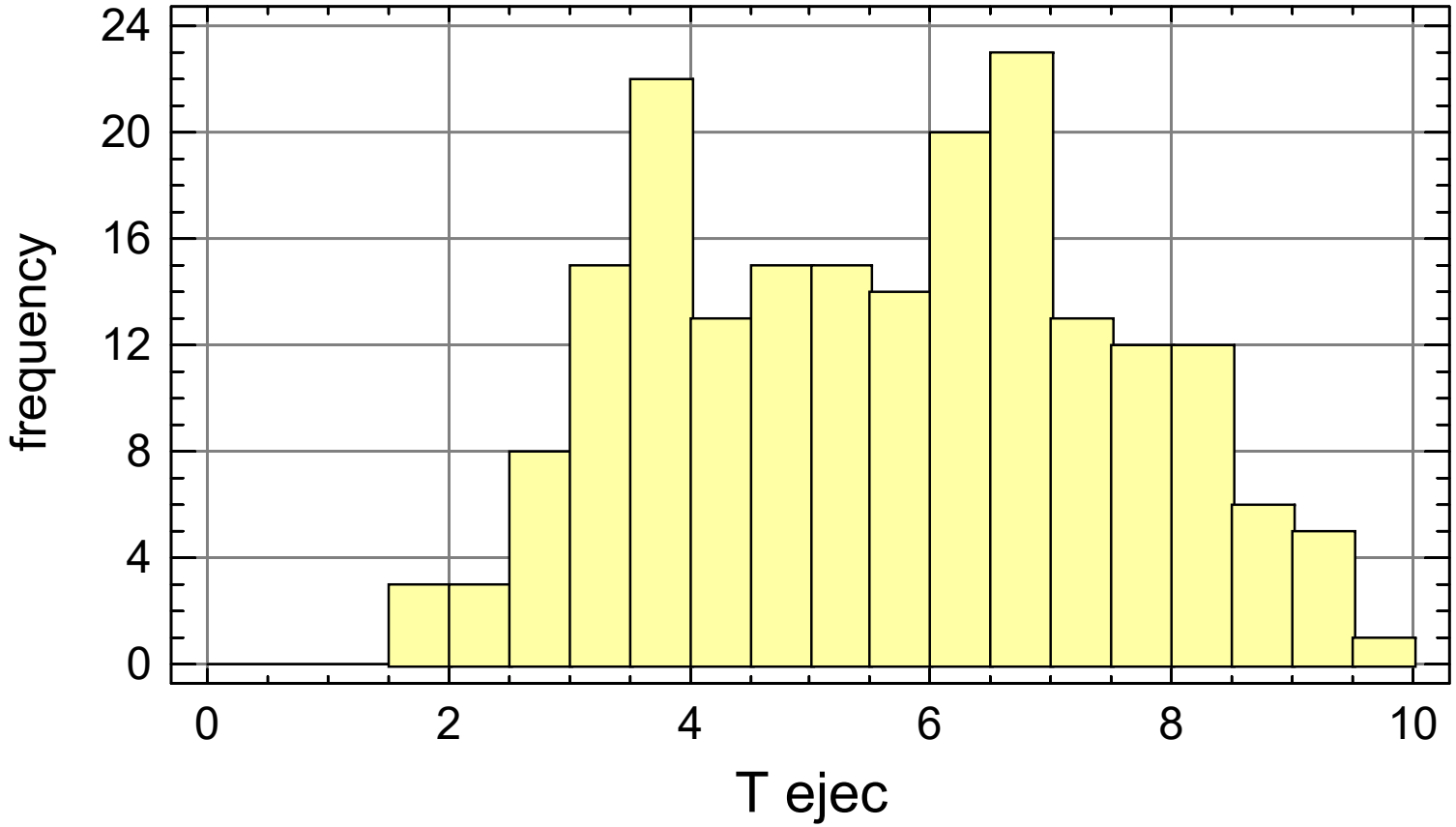
El 6% de los programas tienen tiempo de ejecución menor o igual a 9,5 ms. (aprox.)

Utilidad

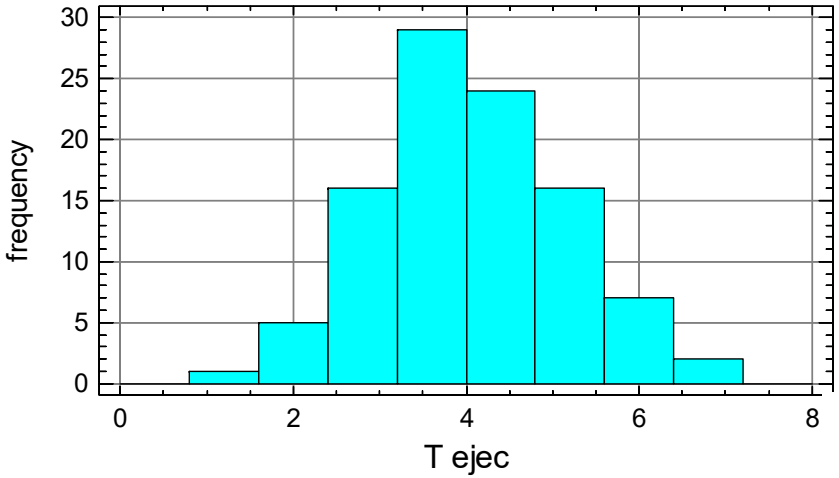
- Podemos detectar rápidamente:
  - Existencia de datos anómalos
  - Mezclas de poblaciones distintas
  - Datos artificialmente modificados
  - No normalidad de los datos

Histogramas tipo: mezcla de poblaciones

Tiempo de ejecución (ms) de 200 programas

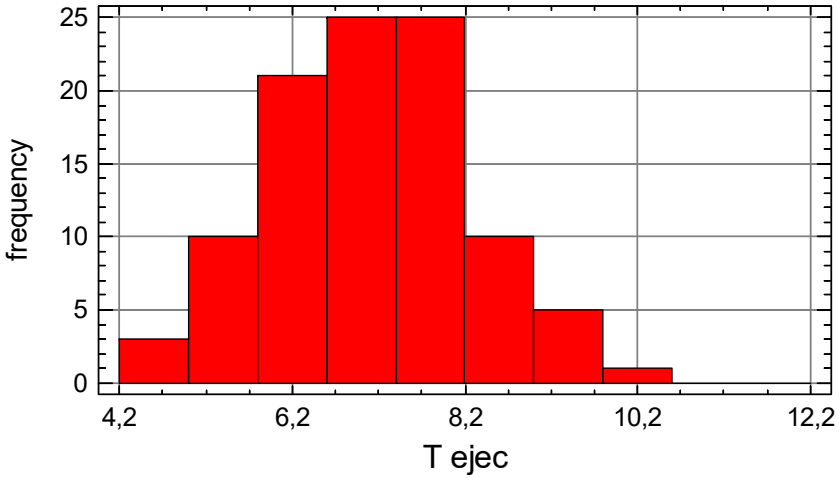


Histogramas tipo : mezcla de poblaciones

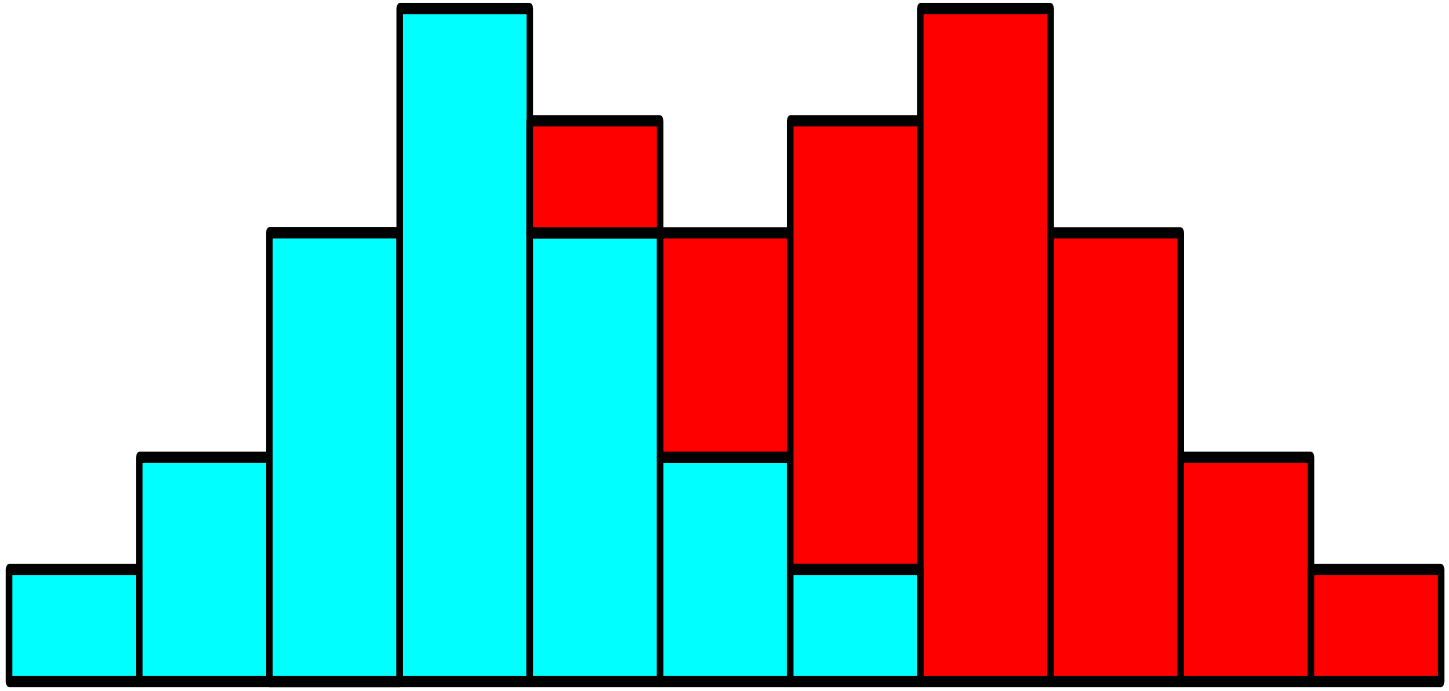


El histograma anterior es la superposición de estos dos

Mezcla de dos poblaciones próximas

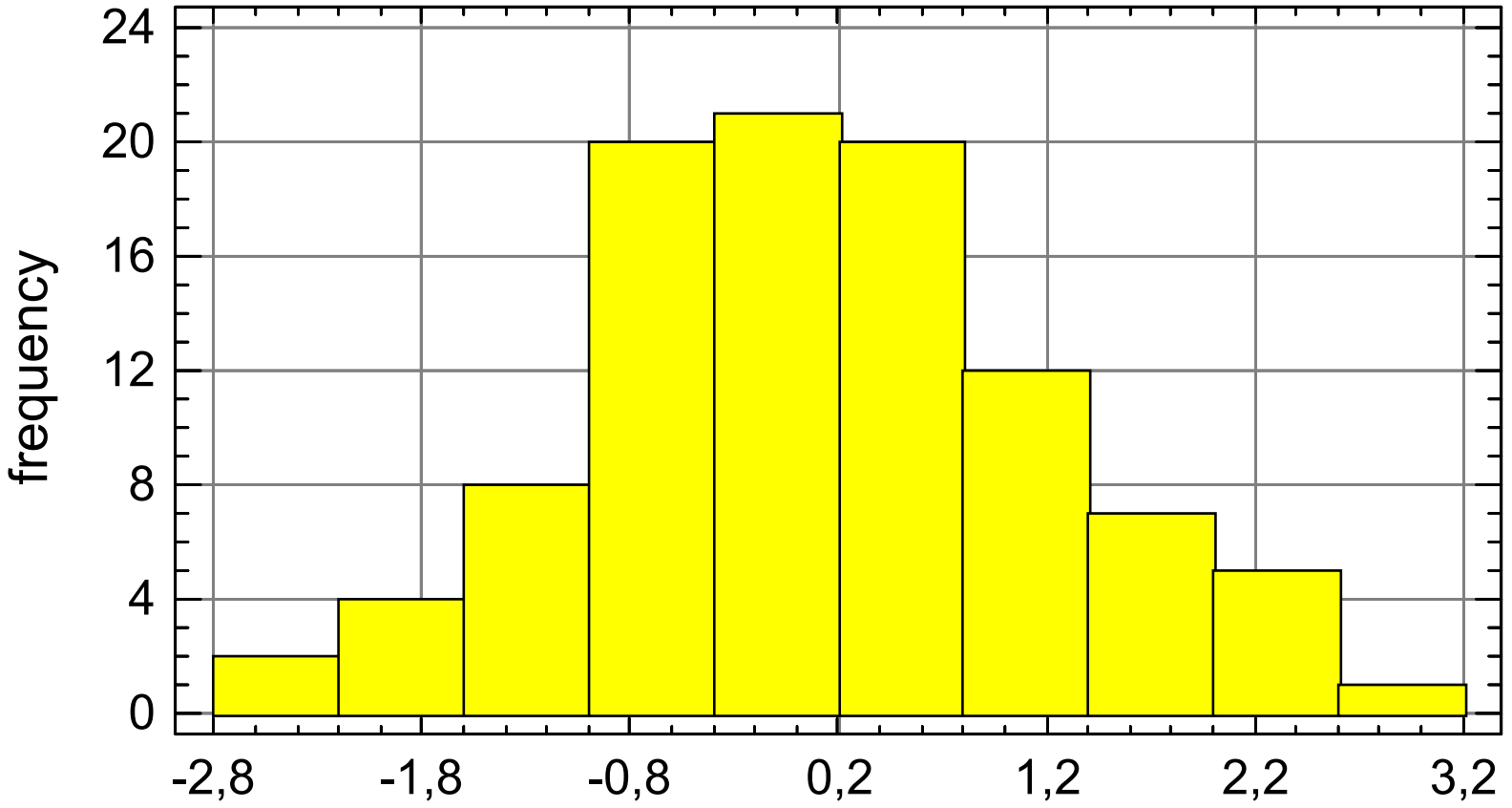


Histogramas tipo: mezcla de poblaciones

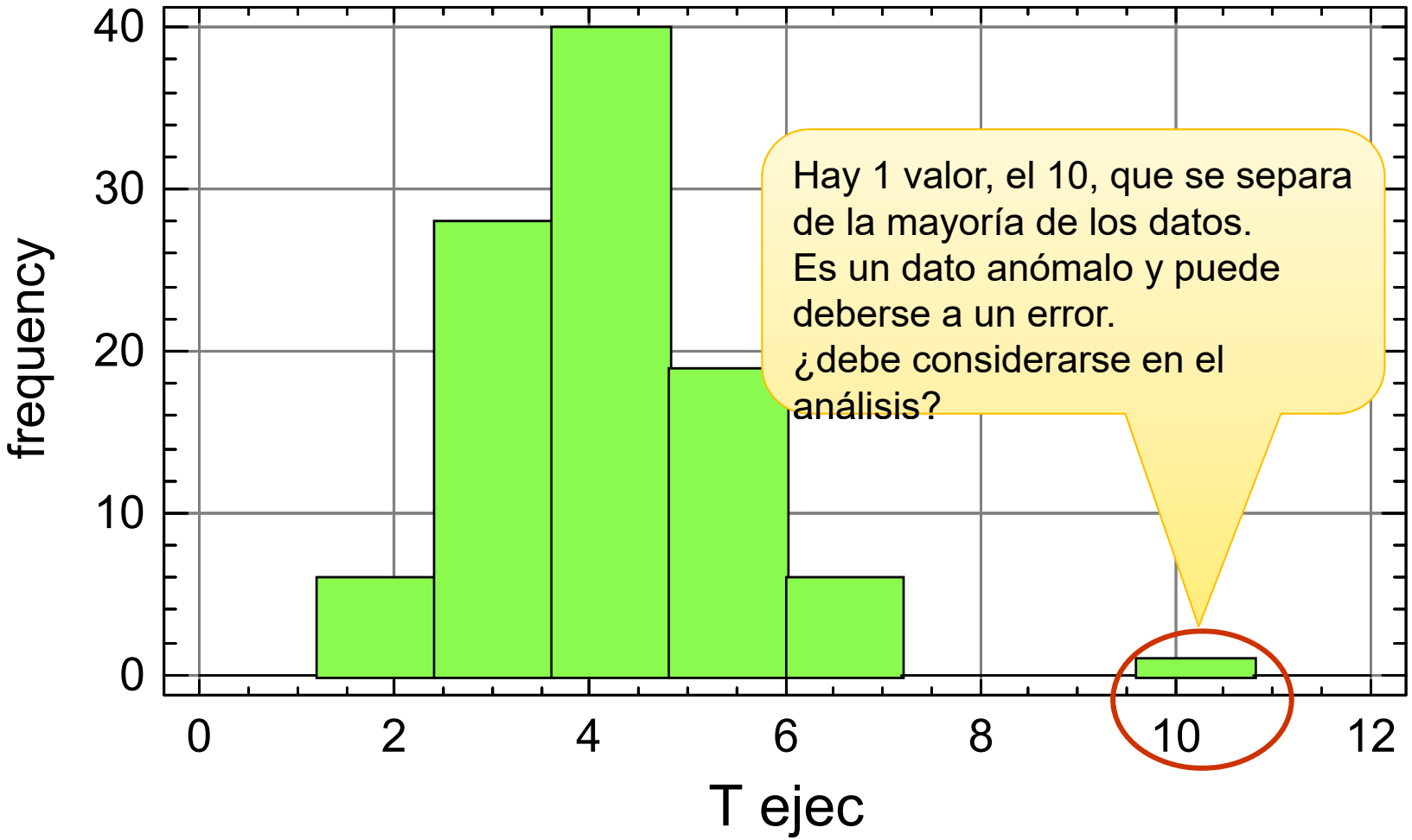


Histogramas tipo: “Normal”

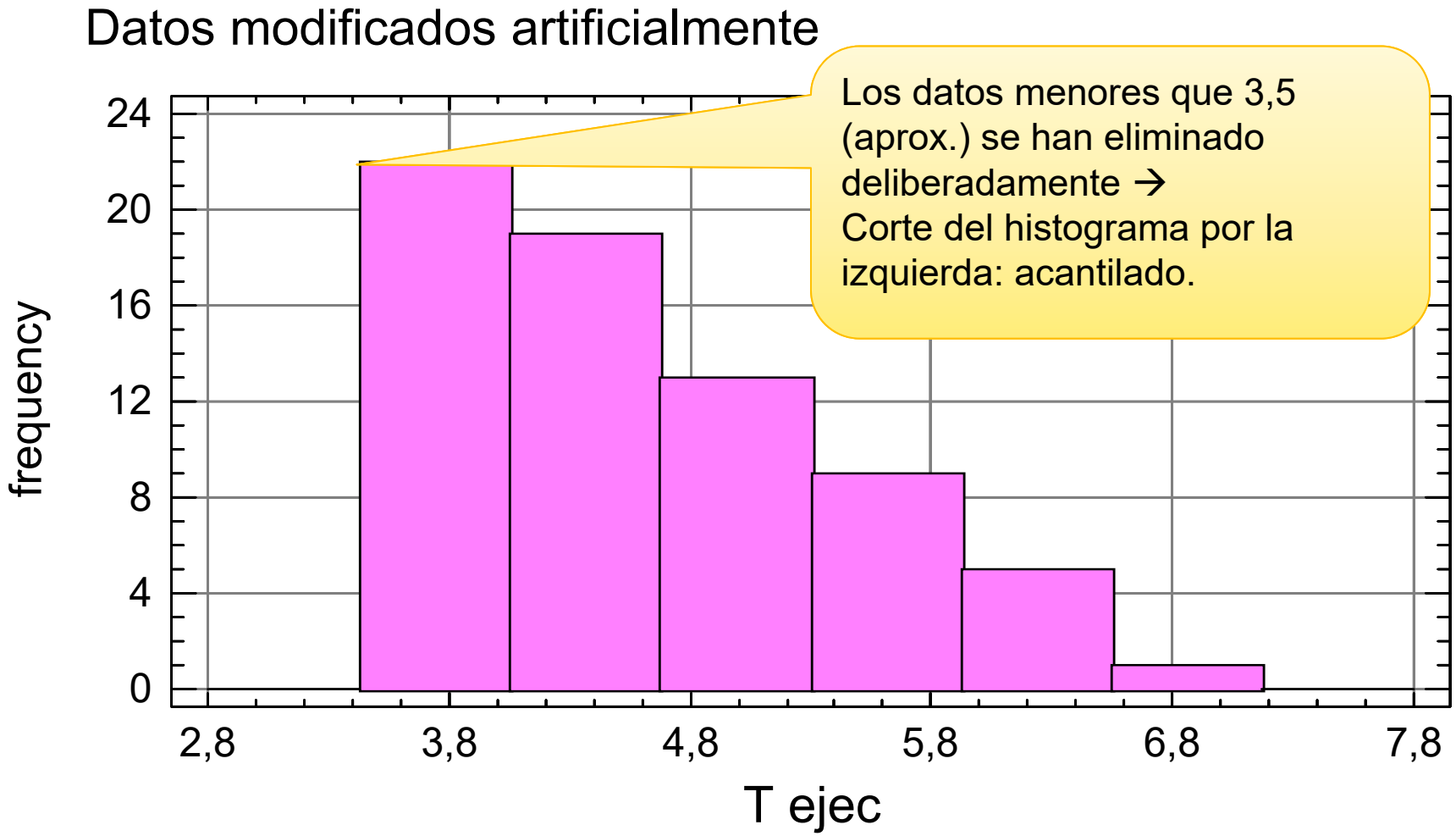
v.a. : desviaciones (mm) sobre el nominal del Φ de una pieza



Histogramas tipo: datos anómalos

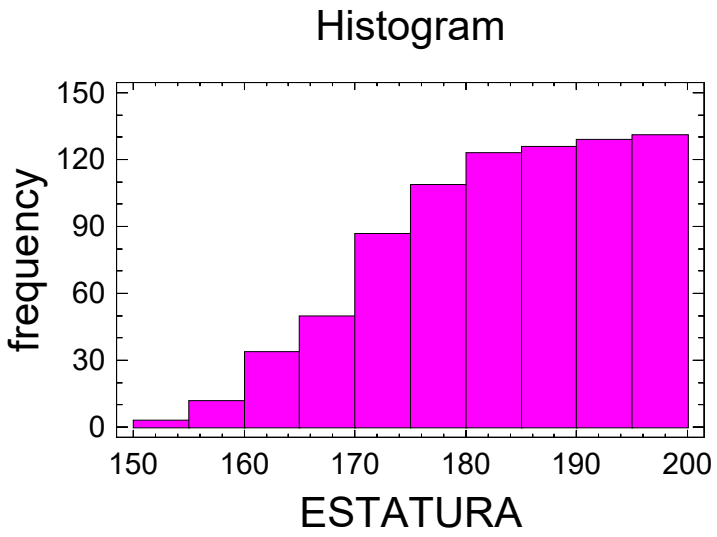


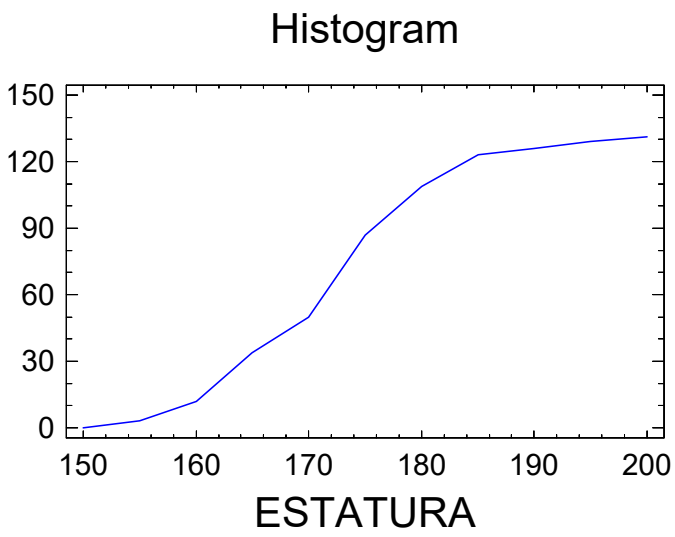
Histogramas tipo “acantilado”



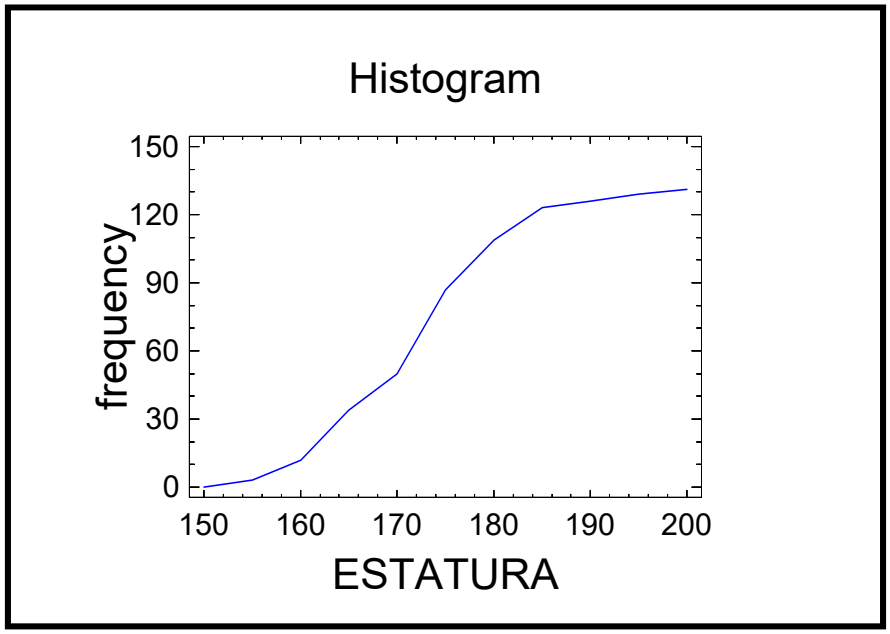
Histogramas de frecuencias acumuladas

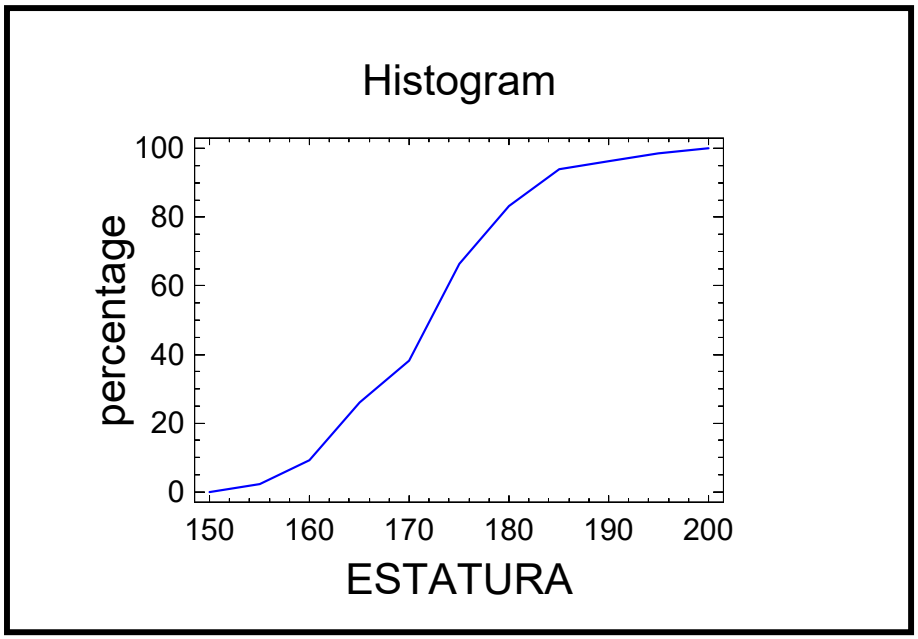
Se puede representar un histograma a partir de las **frecuencias acumuladas** para los diferentes tramos de la variable. Si las abscisas se levantan sobre el límite de cada tramo y se unen los puntos, se obtiene una gráfica con una





Polígono de frecuencias



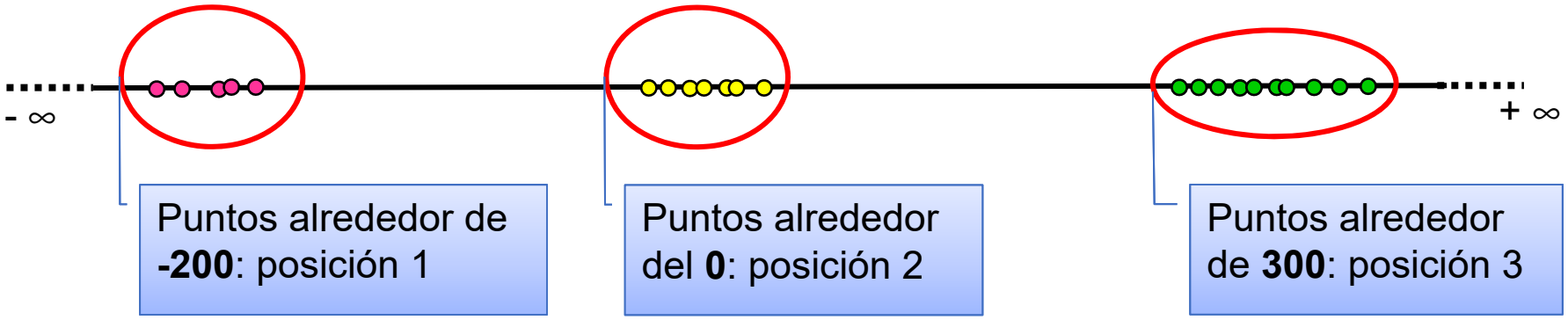


- ¿Qué % de los alumnos miden más de 170 cm?
- ¿Qué estatura es superada por un 5% de los alumnos?
- ¿Cuántos alumnos tienen una estatura menor que 165 cm?

Parámetros muestrales

- Posición
- Dispersión
- Forma

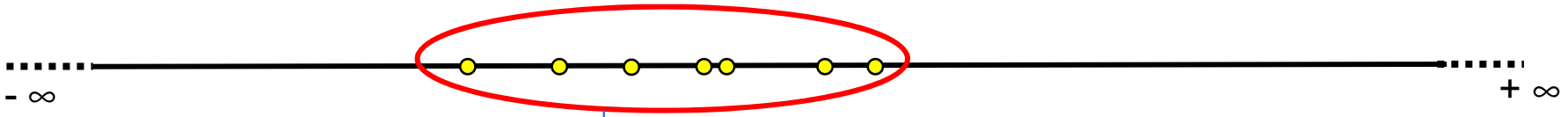
Parámetros: Posición, Dispersión y Forma



Puntos alrededor de **-200**: posición 1

Puntos alrededor del **0**: posición 2

Puntos alrededor de **300**: posición 3

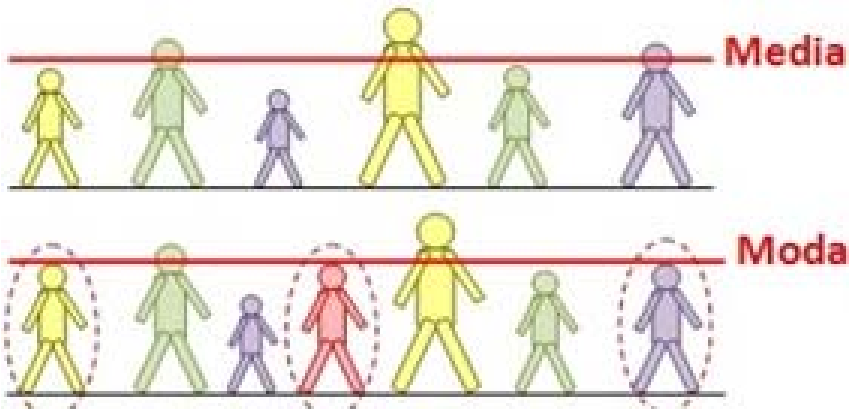


Puntos alrededor del **0**: posición 2, pero los puntos (valores) están más alejados unos de otros: distinta dispersión

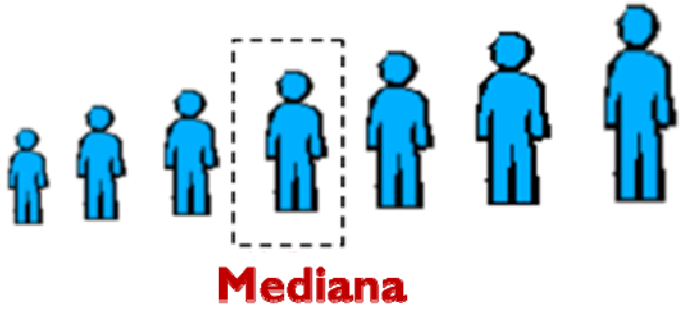
Posición o tendencia central

- Permiten cuantificar, mediante un número, la **posición** de las observaciones
- con un número nos indican “alrededor” de qué valor están las observaciones.
- Parámetros más relevantes:
  - **Media**
  - **Mediana**
  - **Moda**
  - **Cuartiles**
  - **Cuantiles o Percentiles**

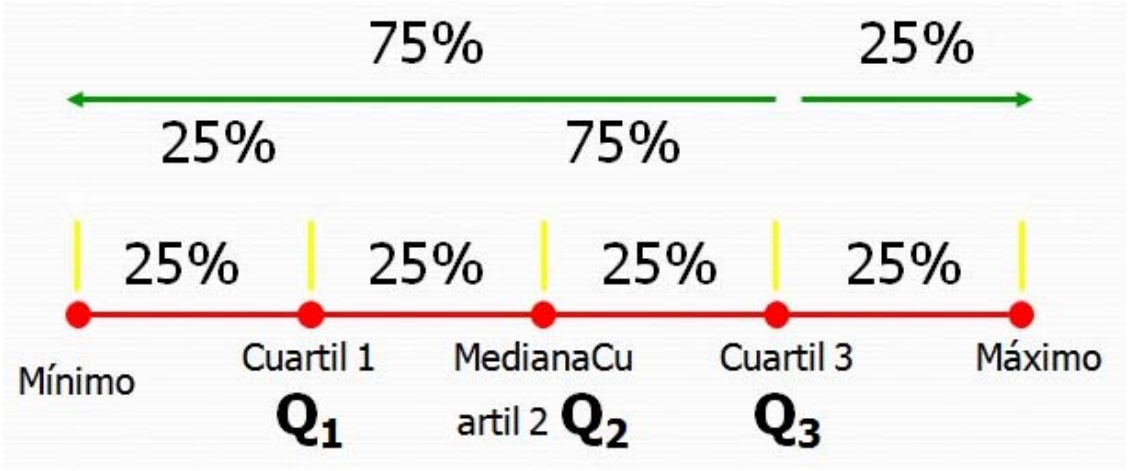
Posición o tendencia central



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



Cuartiles



Posición o tendencia central

1. Calcular N/2.
2. Buscar el intervalo de referencia, es decir, a la columna N_i (frecuencia absoluta acumulada) se busca N/2. Si no observamos el valor exacto, entonces nos quedamos con el inmediatamente superior.

$$Me = L_{i-1} + \frac{N/2 - N_{i-1}}{n_i} \times c_i$$

Donde Li-1 es el extremo inferior del intervalo de referencia, Ni-1 es la frecuencia absoluta acumulada del intervalo inmediatamente inferior, ni es la frecuencia absoluta del intervalo de referencia y ci es la amplitud del intervalo.

Dispersión

- Permiten cuantificar, mediante un número, la **dispersión** de las observaciones, su homogeneidad.
- Con un número nos indican lo separados que están unas observaciones de otras.
- Parámetros más relevantes:
  - **Rango**
  - **Varianza y Desviación típica**
  - **Coeficiente de Variación**
  - **Recorrido o Rango Intercuartílico**

Dispersión

- **Rango** Max - Min
- **Varianza y Desviación típica**

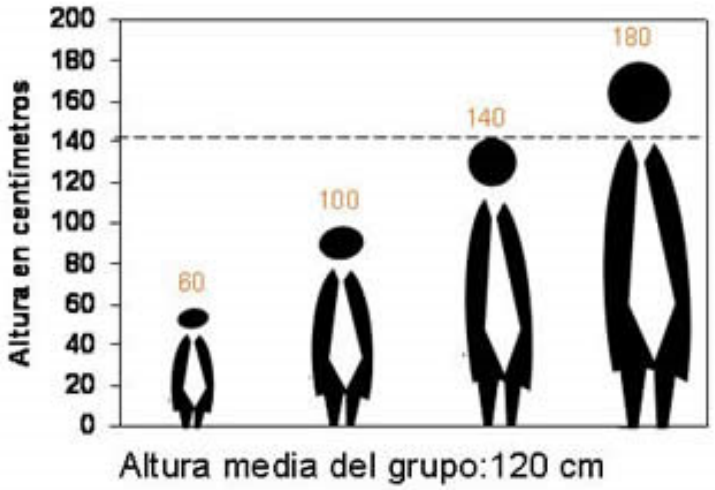
$$\text{Varianza} = S^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N-1} \qquad \text{Desv Típica} = S = \sqrt{S^2}$$

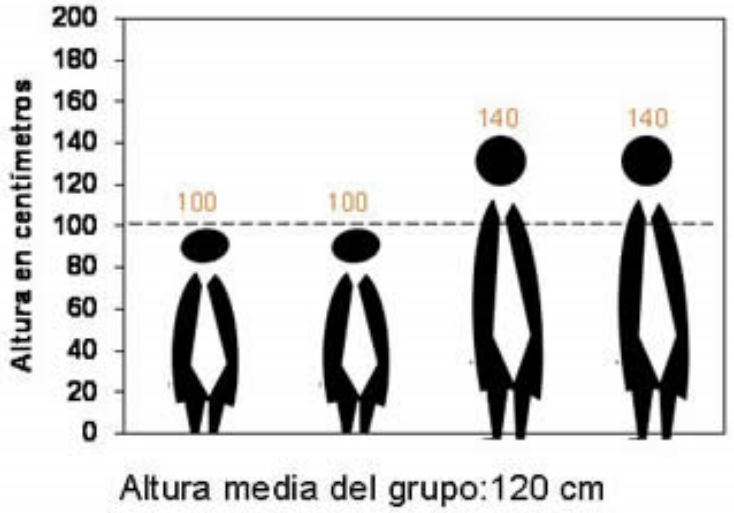
- **Coeficiente de Variación**

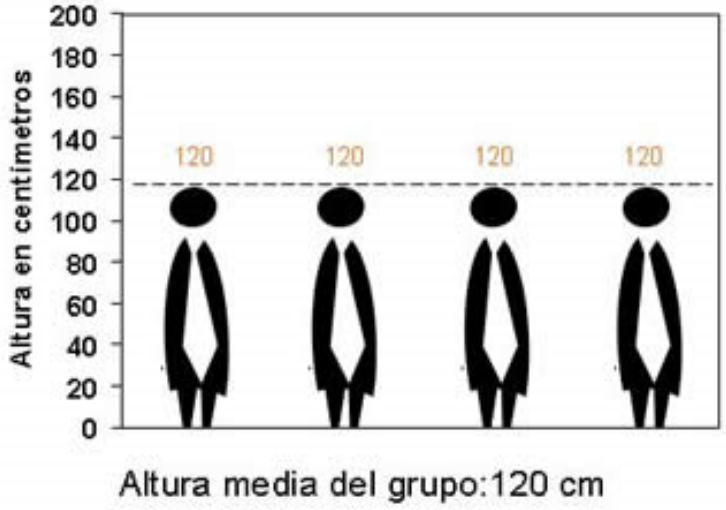
$$CV = \frac{S}{\bar{X}}$$

- **Recorrido o Rango Intercuartílico** Q3 - Q1

Desviación típica



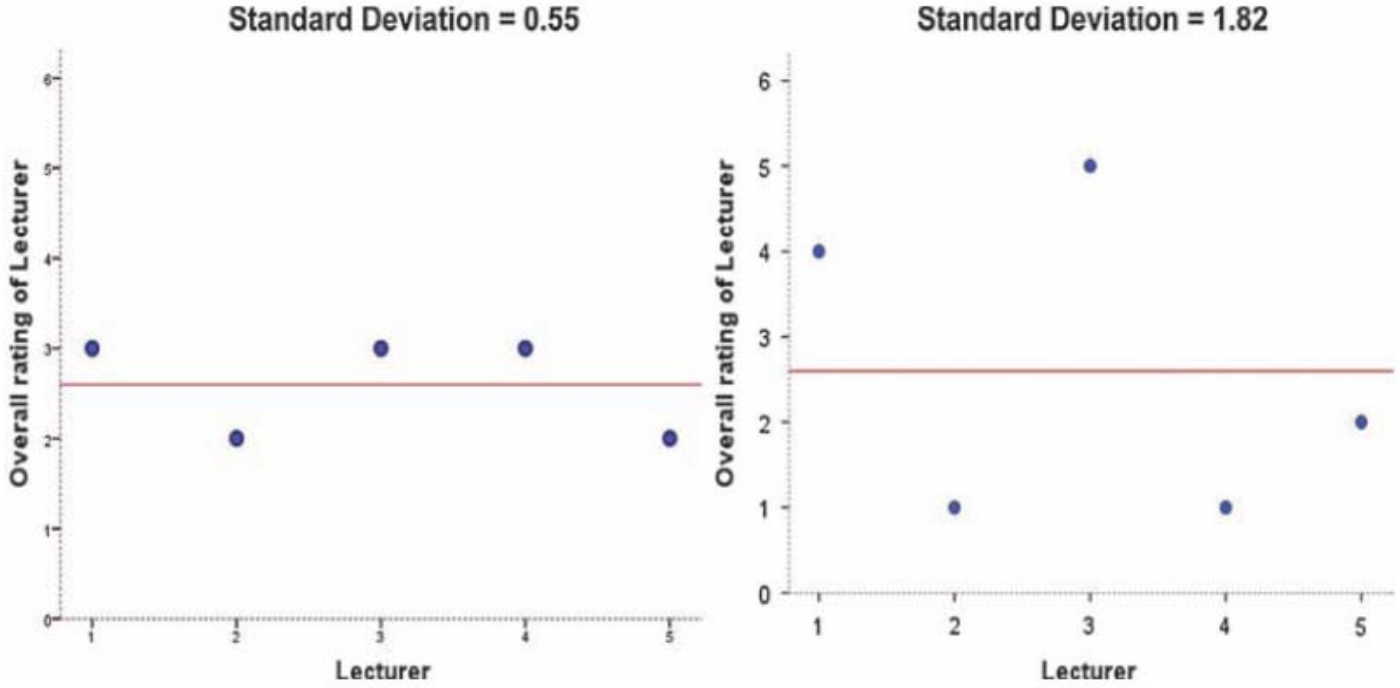




En estos tres grupos de personas, la altura media es la misma: 120 cm. ¿En qué se diferencian?

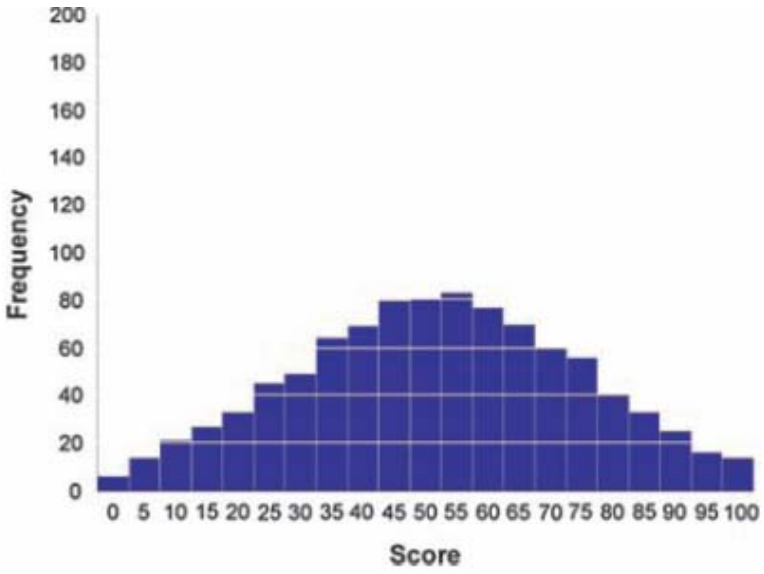
- En el primer grupo, la desviación de los individuos con respecto a la media es mayor que en el resto. Es el grupo más heterogéneo.
- Los individuos del segundo grupo se parecen más entre sí. La variabilidad de los individuos en torno a la media es menor que en el primero.
- En el tercer grupo, todos los individuos tienen la misma altura. No hay dispersión con respecto a la media. El grupo es homogéneo.

Misma media, diferente S

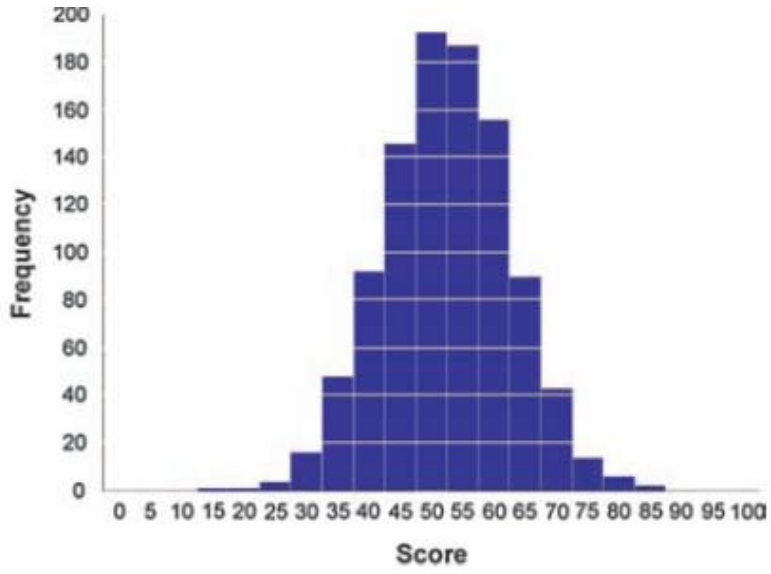


La desviación típica y la forma de la distribución

Misma media y diferentes desviaciones típicas



Desviación típica grande



Desviación típica pequeña

Forma de la distribución

• Los coeficientes de **Asimetría** y **Curtosis** son **parámetros de forma**.

• Los dos permiten comprobar si nuestros datos se asemejan suficientemente a una “campana de Gauss” (distribución Normal)

Pautas de comportamiento que se alejan sensiblemente de la Normal exigen:

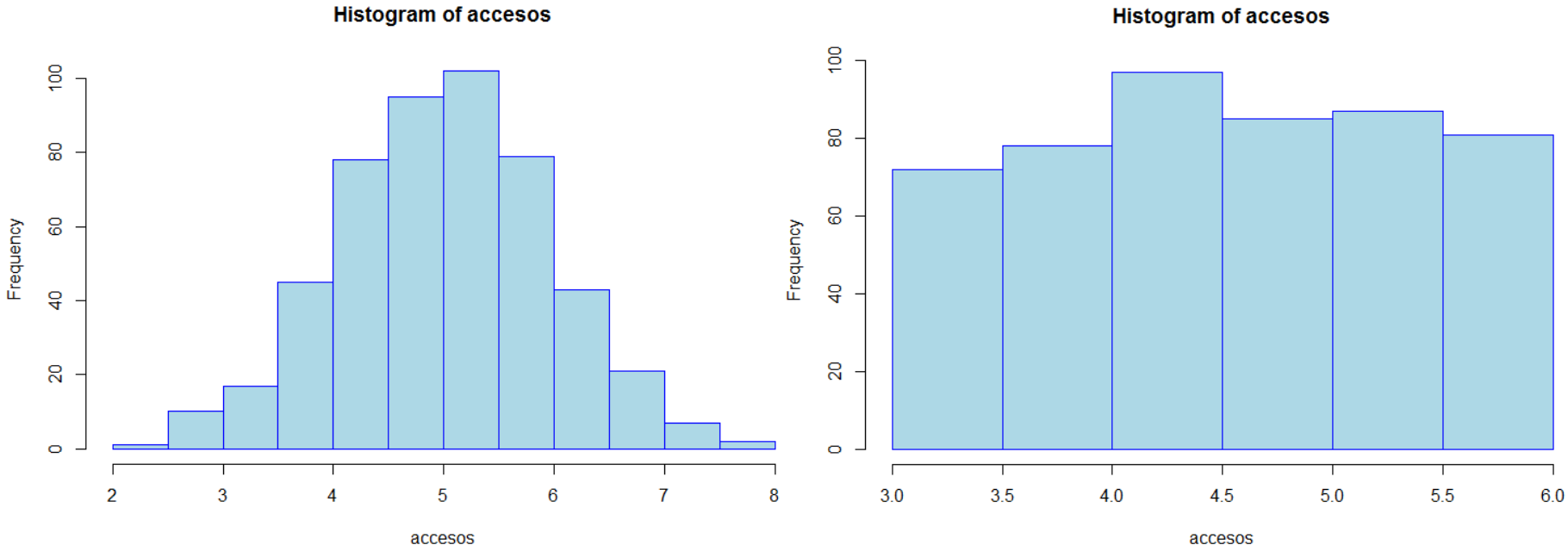
- la revisión y corrección de datos anómalos, si procede
- recurrir a modelos o tratamientos estadísticos especiales.

Mediante los parámetros de Asimetría y Curtosis podemos detectar la “no normalidad” de los datos y obrar en consecuencia

Asimetría (Skewness)

- **Asimetría**: grado de simetría, con respecto a la media, de la distribución de frecuencias

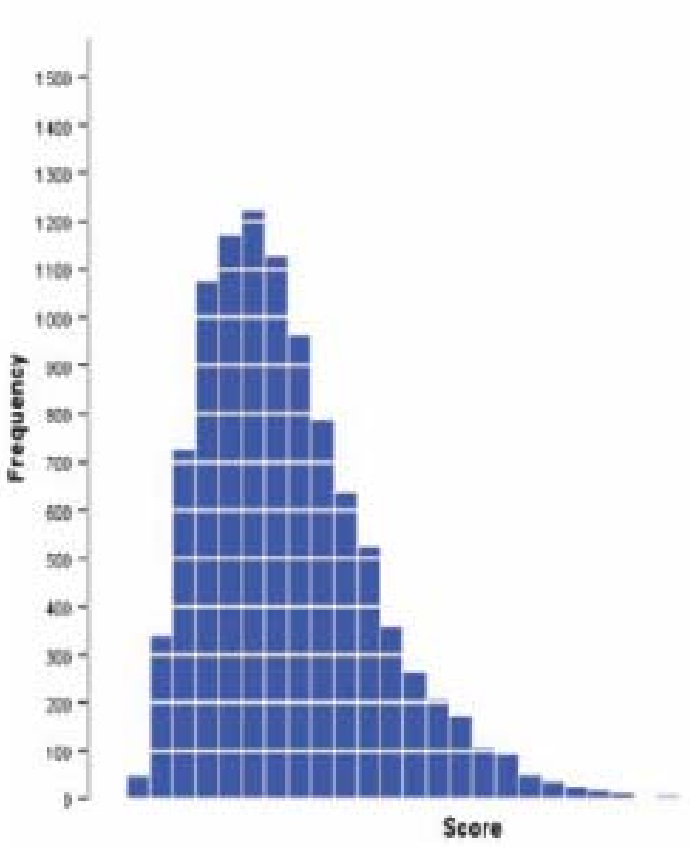
$$AS = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{n \cdot s^3}$$



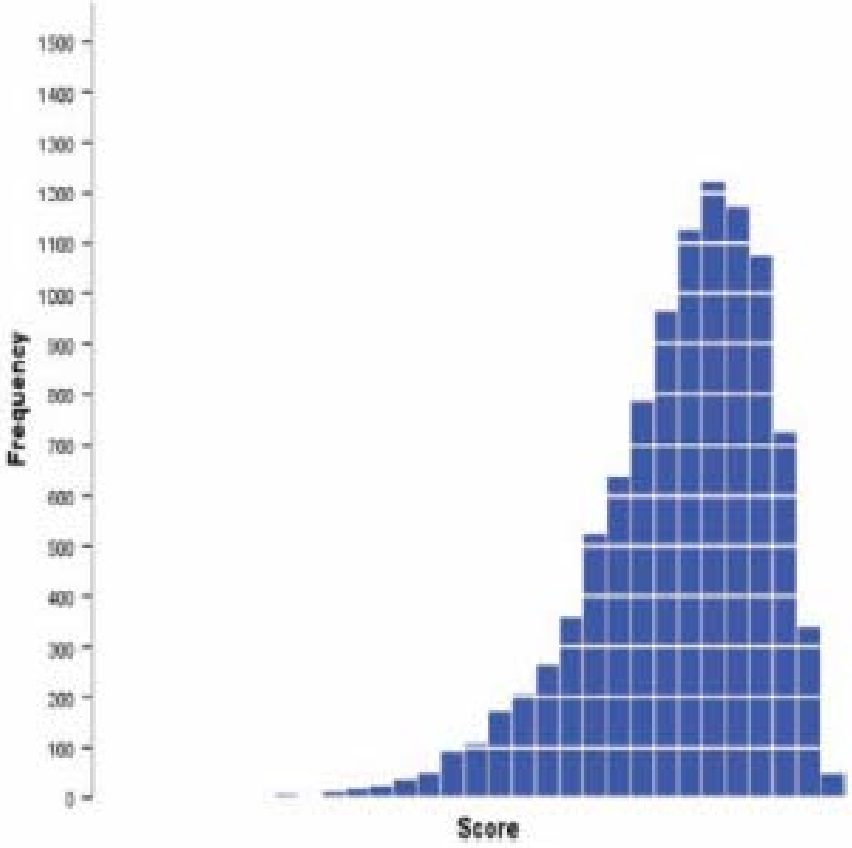
Simetría

Asimetría

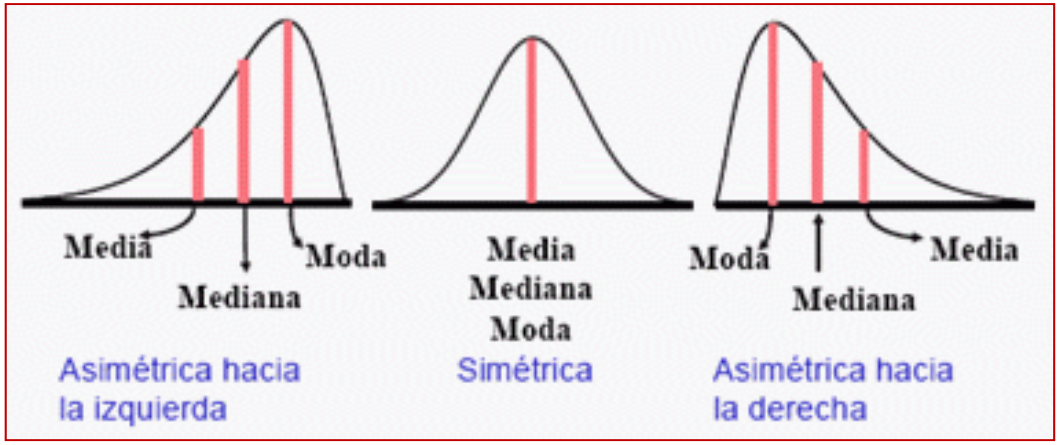
Asimetría positiva

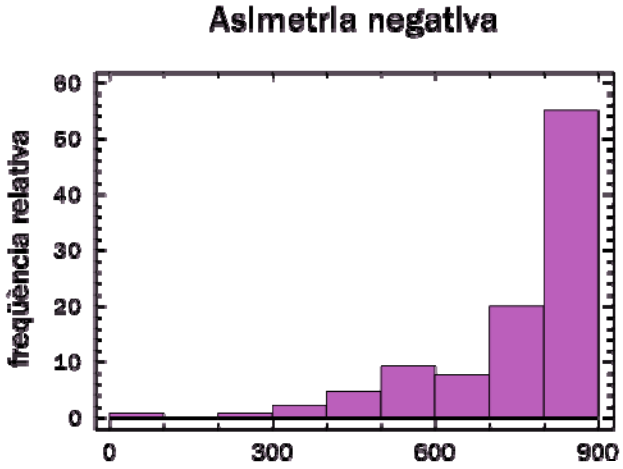


Asimetría negativa

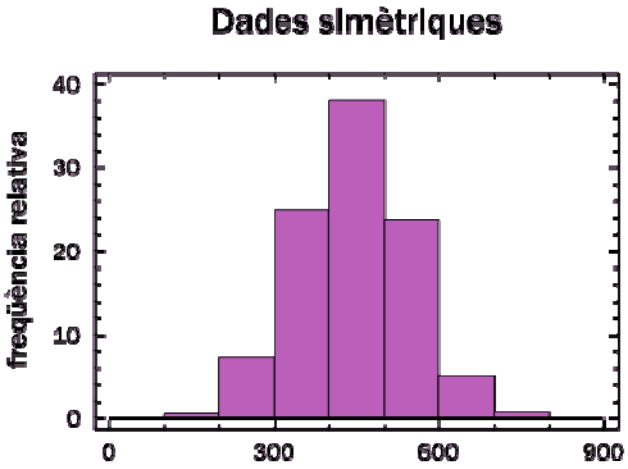


Asimetría (Skewness)

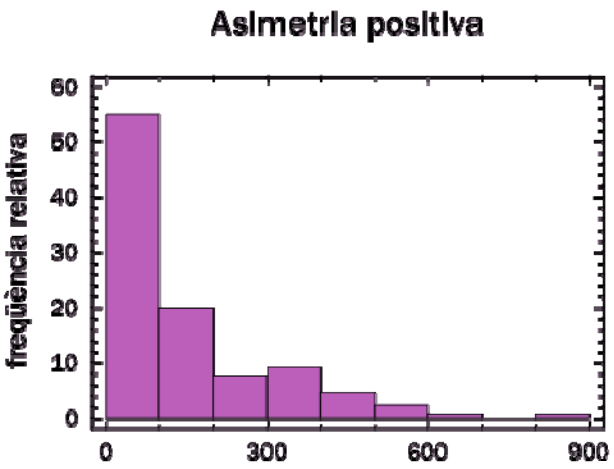




CAE < -2



CAE ∈ [-2, +2]



CAE > +2

Forma o apuntamiento

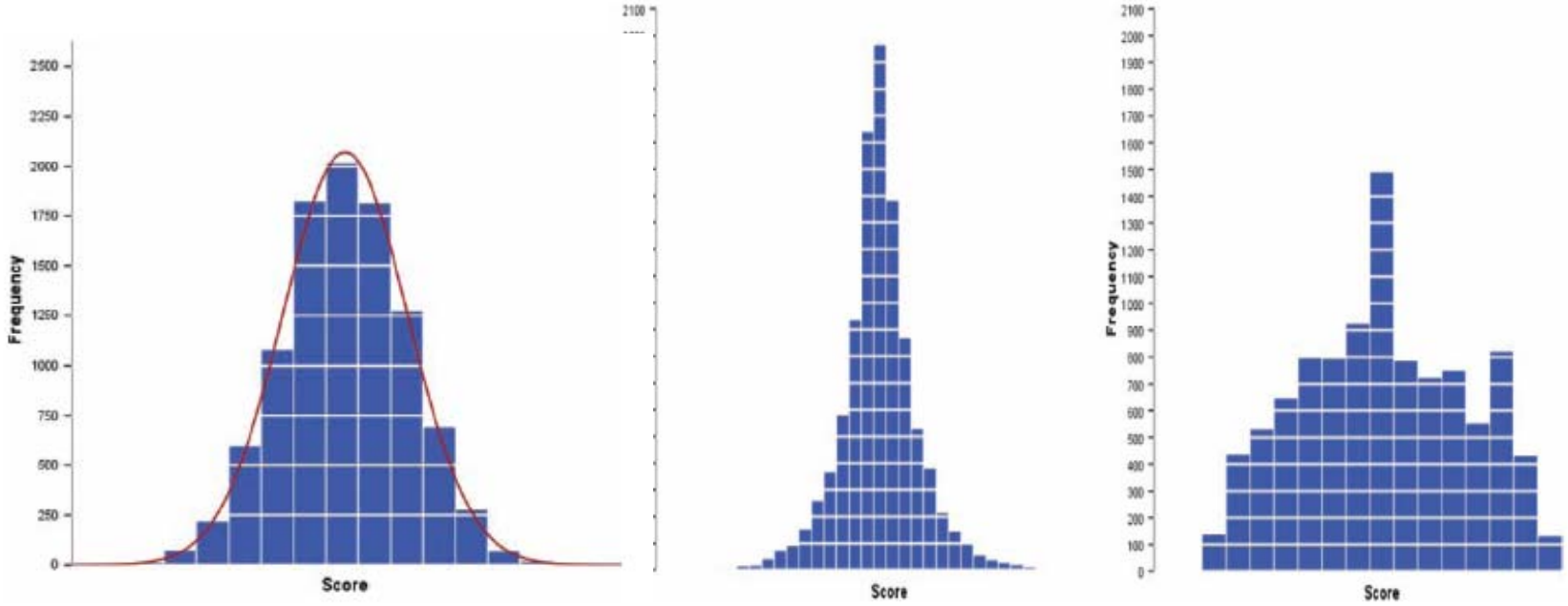
$$Cur = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{n \cdot s^4}$$

- **Curtosis**: grado de apuntamiento, comparada con la *normal*, de la distribución de frecuencias

Normales o mesocúrtidos

Leptocúrtidos

Planicúrtidos



Parámetros más adecuados

	Datos simétricos y sin valores anómalos	Datos asimétricos o con valores anómalos
Posición	Media	Mediana
Dispersión	Desviación típica	Recorrido intercuartílico

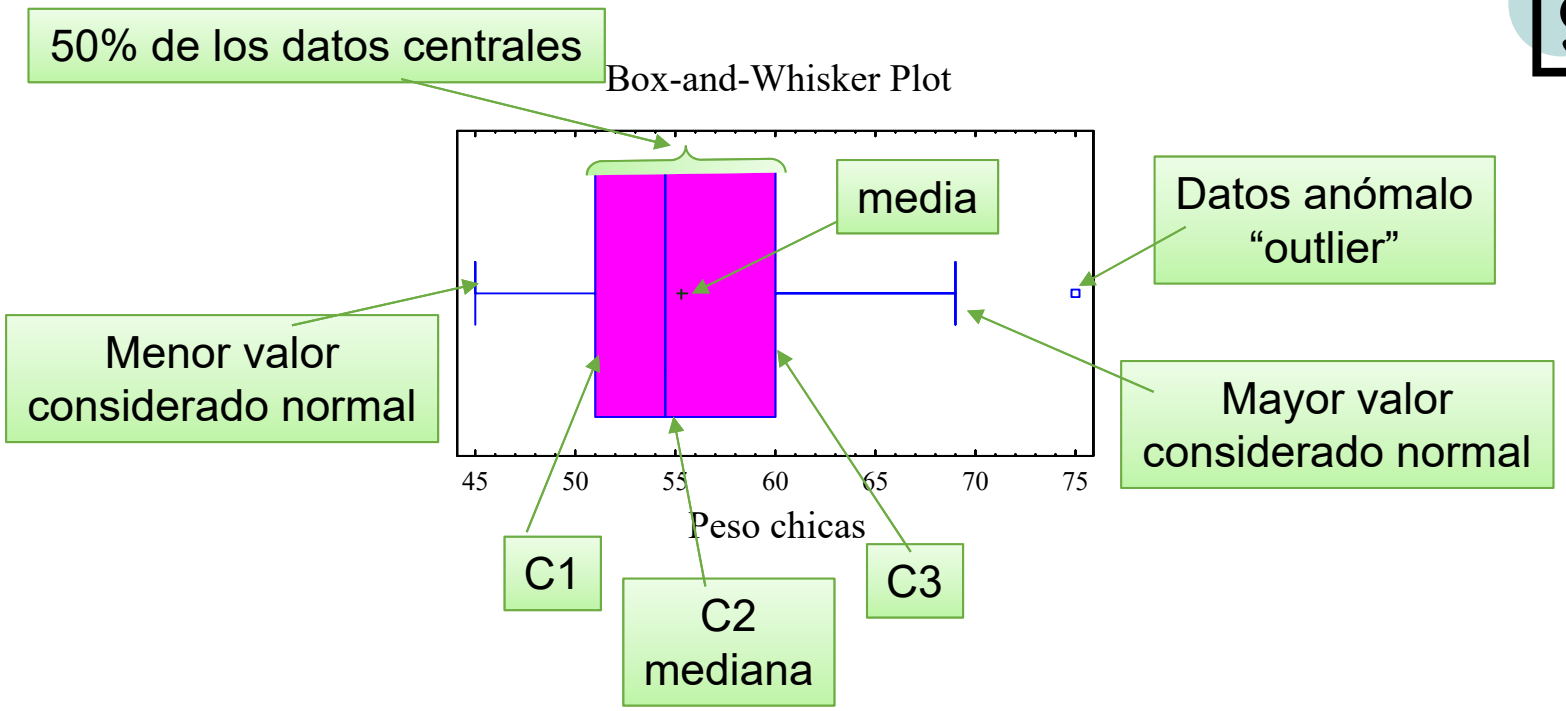
Diagrama
Box & Whisker

O Diagrama de *Caja y Bigotes*

9.- Diagramas Box-Whisker

- No exige un número elevado de datos para su construcción como el Histograma
- Muy útil para comparar 2 grupos de datos y observar de forma gráfica si hay o no diferencias entre ellos.
- La “caja” comprende el 50% de los valores centrales de los datos y se extiende entre el 1º y 3º cuartil

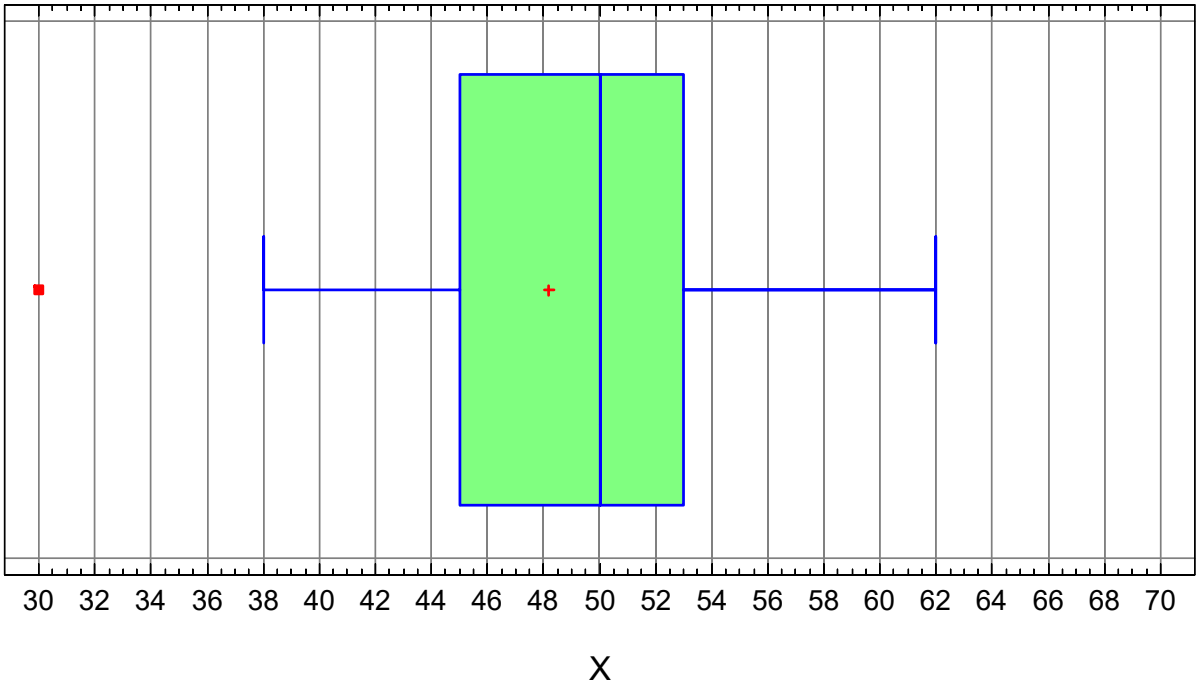
Interpretación



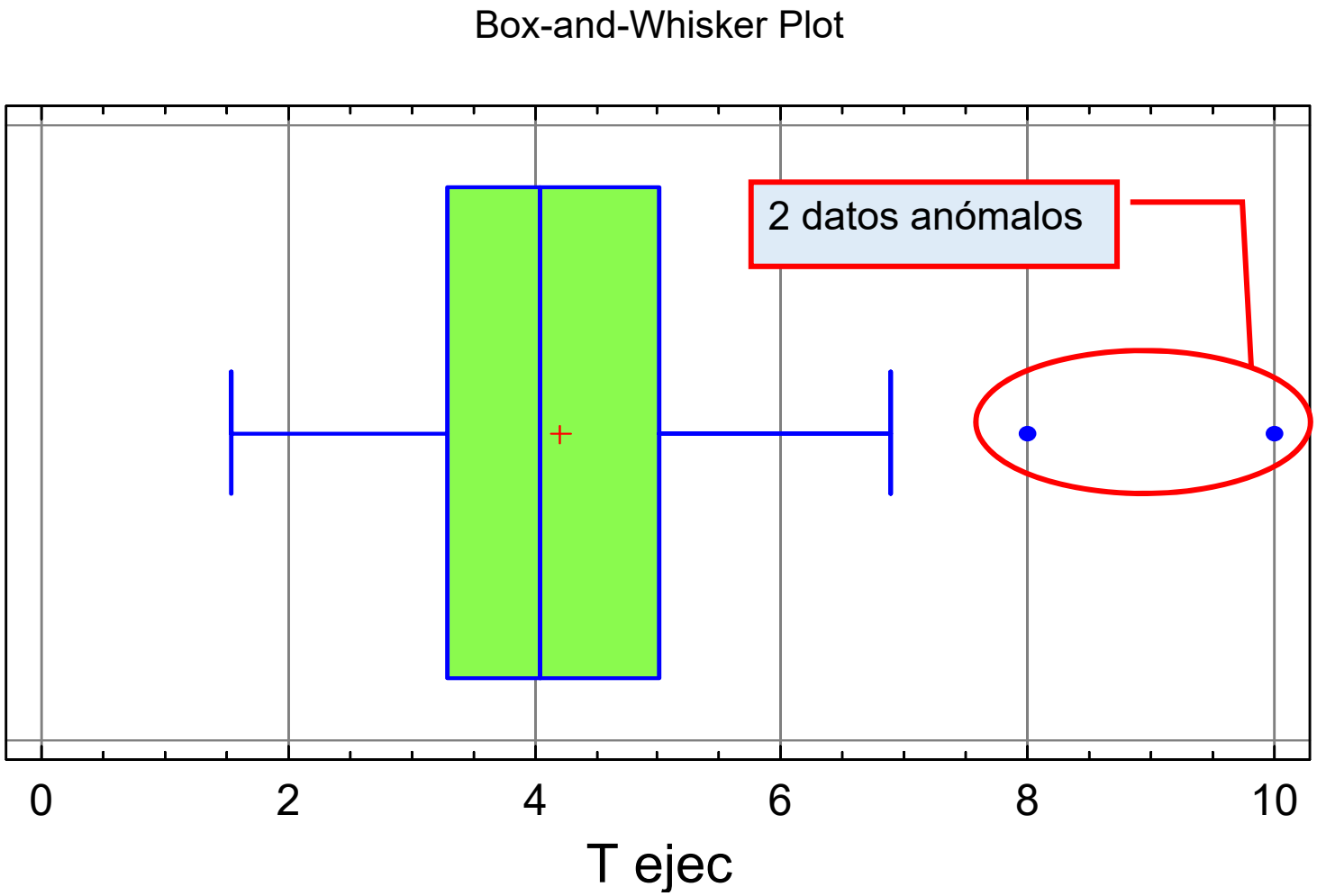
Dato anómalo ("outlier"): Valores extremos que difieren del cuartil más próximo en más de 1,5 veces el intervalo intercuartílico (C3-C1)

Ejercicio (UD2)

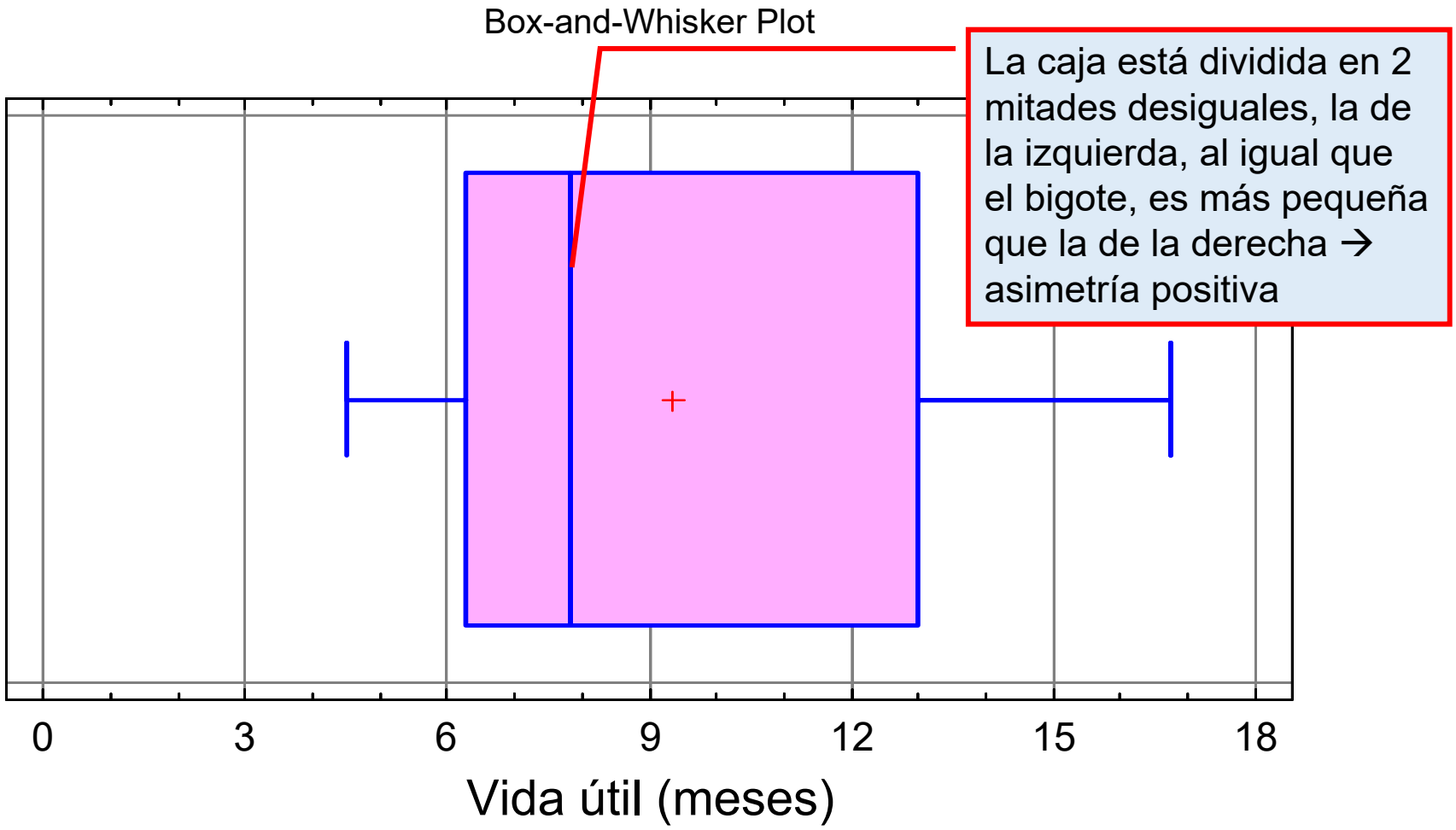
- Siguiendo con el ejercicio del tiempo de funcionamiento sin averías: X = {50, 38, 45, 30, 47, 50, 48, 62, 55, 53, 52}



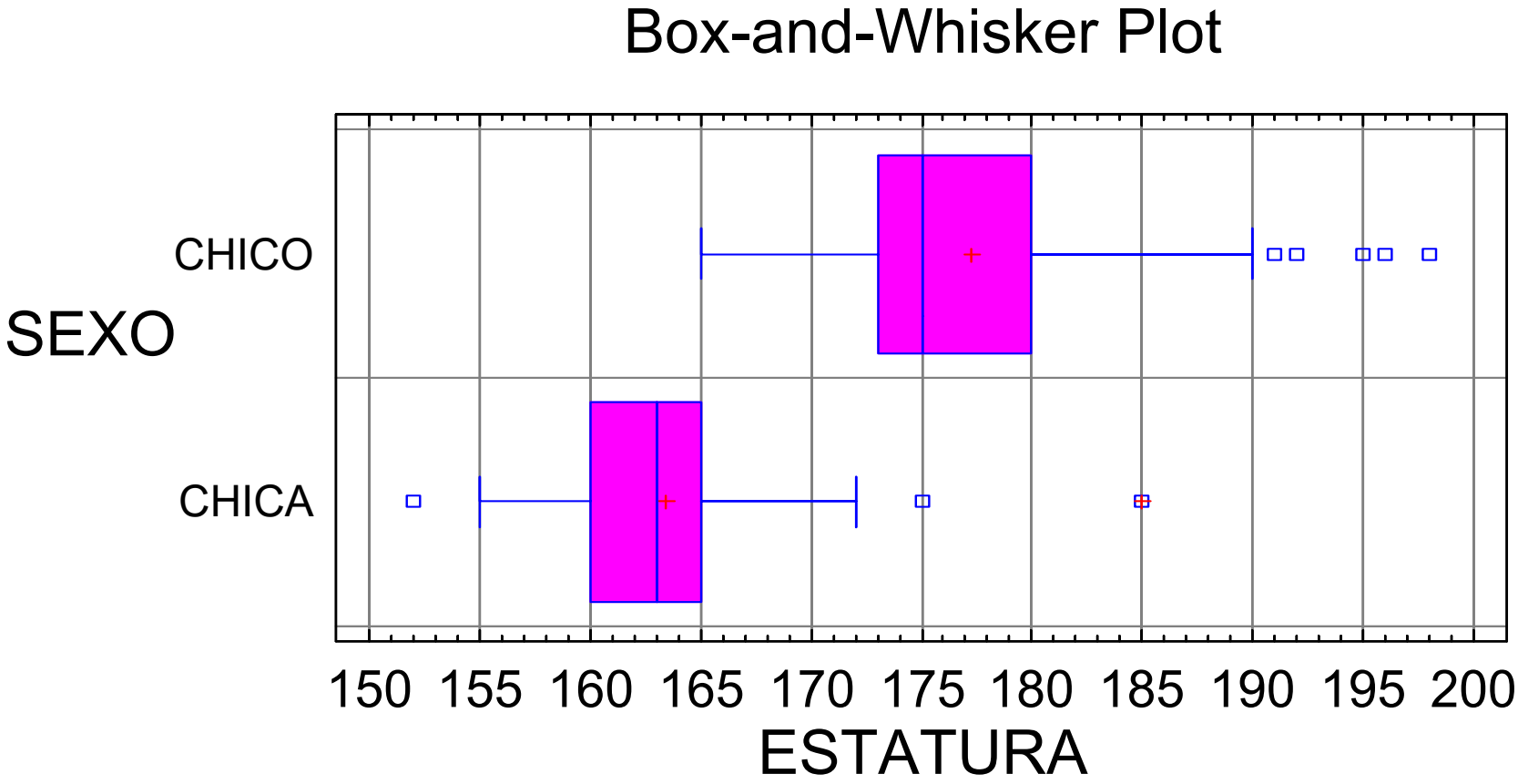
Otro ejemplo: datos anómalos



Otro ejemplo: asimetría positiva

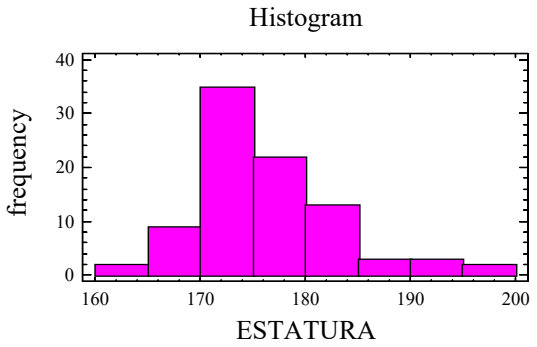


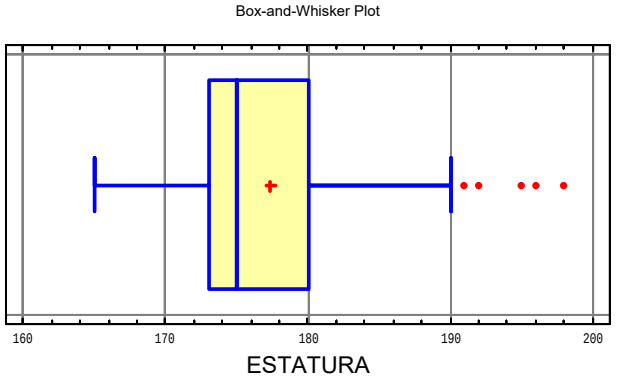
Ejercicio 20 UD2



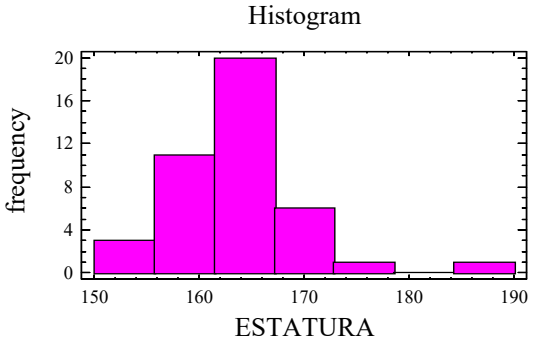
Box&Whisker e Hitogramas

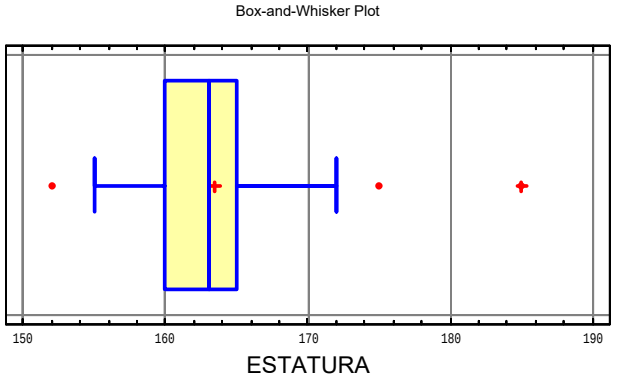
chicos



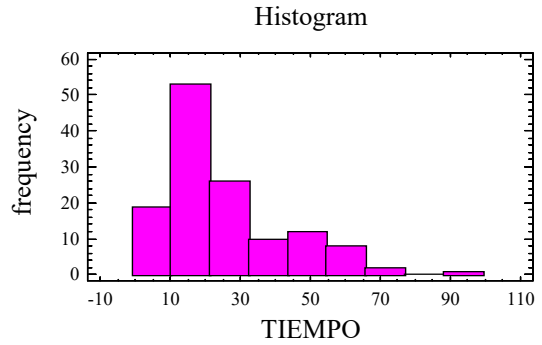


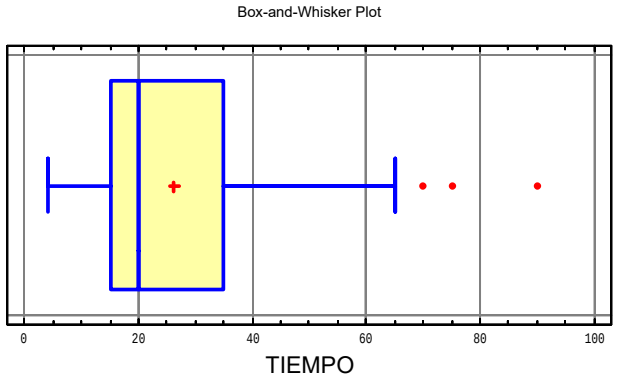
chicas

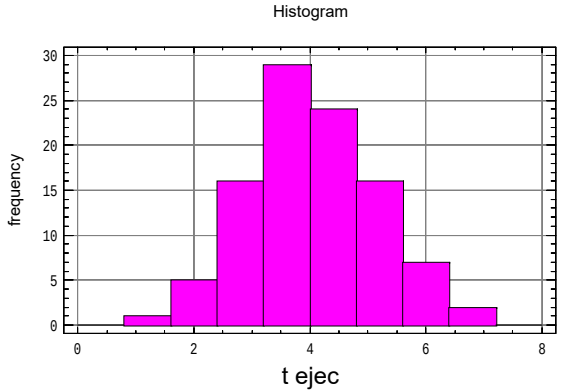


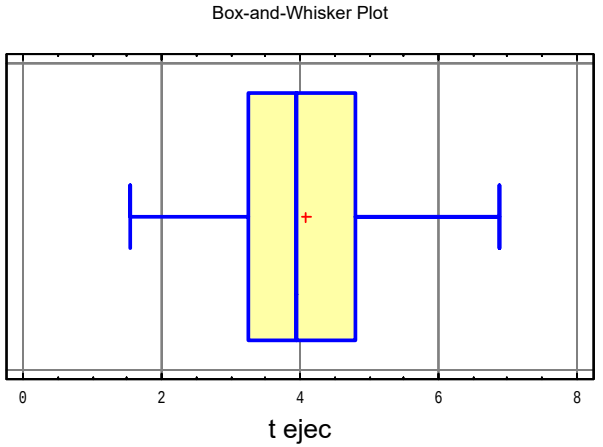


Box&Whisker e Hitogramas









Estadística descriptiva bidimensional

- En la mayoría de los análisis se estudia más de una característica sobre cada individuo de la población.
- Cuando se observan los valores de 2 características

Variable aleatoria bidimensional

- En el estudio de v.a. bidimensionales, no basta con conocer las características y regularidades de los datos para cada componente de la v.a. …
- Determinar si hay alguna relación entre las variables que se miden.

Var. Cualitativas: Tablas de contingencia

Una tabla de contingencia es una tabla de doble entrada que muestra las coincidencias y discrepancias entre los valores de dos variables categóricas distintas.

Además de la frecuencia de casos en cada celda, se suelen mostrar los porcentajes de casos por fila y/o columna.

- **Objetivo del análisis**: describir la relación existente entre las dos componentes de la v.a. bidimensional

- **Herramienta**: **Tabla de Contingencia**
  - variables **discretas**:
    - naturaleza cualitativa (codificadas)
    - cuantitativas con pocos valores
    - Cuantitativa continua discretizada en tramos

Frec. Relativas condicionales de CATEGORÍA en función de SEXO

	Administrativo	Seguridad	Directivo	Frecuencias Marginales De SEXO
Hombre	157 60,9%	27 10,5%	74 28,7%	258 54,4%
Mujer	206 95,4%	0 0%	10 4,6%	216 45,6%
Frecuencias Marginales De CATEGORÍA	363 76,6%	27 5,7%	84 17,7%	474

5 - % de las mujeres que son directivas
(10/ 216*100)

Frecuencias relativas a los totales de las filas (SEXO)

Frecuencias conjuntas

	Administrativo	Seguridad	Directivo
Hombre	157	27	74
Mujer	206	0	10

2 - Nº de mujeres que ocupan cargo de directivo

1 - Nº de empleados encuestados
Tamaño de la muestra (N)

474

Frecuencias marginales (absolutas y relativas)

	Administrativo	Seguridad	Directivo	Frecuencias Marginales De SEXO
Hombre	157		74	258 54,4%
Mujer	206	0	10	216 45,6%
Frecuencias Marginales De CATEGORÍA	363 76,6%	5,7%	17,7%	474

4 - Nº de mujeres encuestadas (206 + 0 + 10)

3 - % de empleados administrativos (363 /474*100)

Frec. Relativas condicionales de CATEGORÍA en función de SEXO

	Administrativo	Seguridad	Directivo	Frecuencias Marginales De SEXO
Hombre	153 60,9%	26 10,5%	74 28,7%	258 54,4%
Mujer	206 95,4%	0 0%	10 4,6%	216 45,6%
Frecuencias Marginales De CATEGORÍA	363 76,6%	27 5,7%	84 17,7%	474

5 - % de las mujeres que son directivas (10/ 216*100)

Frecuencias relativas a los totales de las filas (SEXO)

Frec. Relativas condicionales de SEXO en función de CATEGORÍA

	Administrativo	Seguridad	Directivo	Frecuencias Marginales De SEXO
Hombre	157 43,3%	27 100%	74 88,1%	258 54,4%
Mujer	206 56,7%			216 45,6%
Frecuencias Marginales De CATEGORÍA	363 76,6%	27 5,7%	84 17,7%	474

6 - % de las administrativos que son hombres (157/ 363 *100)

Frec. relativas a los totales de las columnas (CATEGORÍA)

Frecuencias relativas condicionales

- La frecuencia relativa condicional a calcular depende del objetivo de nuestro estudio.

- **EJEMPLO**: Si queremos estudiar si la proporción de mujeres (u hombres) es igual o diferente para los diferentes cargos:

  - ¿Frecuencia relativa condicional de CATEGORÍA en función de SEXO?

  ó

  - ¿Frecuencia relativa condicional de SEXO en función de CATEGORÍA?

Frec. Relativas condicionales de SEXO en función de CATEGORÍA

	Administrativo	Seguridad	Directivo	Frecuencias Marginales De SEXO
Hombre	157 43,3%	27 100%	74 88,1%	258 54,4%
Mujer	206 56,7%	0 0%	10 11,9%	216 45,6%
Frecuencias Marginales De CATEGORÍA	363 76,6%	27 5,7%	84 17,7%	474

NO se puede deducir que en el grupo de **directivos** hay más hombres que mujeres, ya que en el total de la muestra hay más hombres (258 hombres frente a 216 mujeres)

Frec. Relativas condicionales de CATEGORÍA en función de SEXO

	Administrativo	Seguridad	Directivo	Frecuencias Marginales De SEXO
Hombre	157 60,9%	27 10,5%	74 28,7%	258 54,4%
Mujer	206 95,4%	0 0%	10 4,6%	216 45,6%
Frecuencias Marginales De CATEGORÍA	363 76,6%	27 5,7%	84 17,7%	474

Ahora **SÍ** se puede deducir que en el grupo de **mujeres** hay más o menos directivos, ya que las frecuencias están relativizadas con respecto a los totales de mujeres y hombres para que éstos no influyan. De las mujeres un 4,6% son directivos frente al 28,7% de los hombres que son directivos.

Var. Cuantitativas: Tablas de frecuencias cruzadas

- Todo lo dicho respecto al cálculo e interpretación de las frecuencias de la Tabla de Contingencia es aplicable al caso de las v.a. cuantitativas.

- La única diferencia, como en el caso de las v.a. unidimensionales, es que previamente a la representación de la Tabla es necesario agrupar los valores de las variables en intervalos.

Tablas de frecuencias cruzadas

ESTATURA PESO	145 155	155 165	165 175	175 185	185 195	Row Total
40 - 55	9	17	0	0	0	
55 - 70	3	18	31	5	0	
70 - 85	0	3	24	12	3	
85 - 99	0	0	3	0	2	
Column Total						

Frecuencia conjunta: peso y estatura

Frecuencia marginal absoluta: peso y estatura

Frecuencia marginal relativa: peso y estatura

Frecuencia relativa de peso condicionada a estatura

Variables cuantitativas: tablas de frecuencias cruzadas

ESTATURA PESO	145 155	155 165	165 175	175 185	185 195	Row Total
40 55	9 75.0	17 44.7	0 .0	0 .0	0 .0	26 20.0
55 70	3 25.0	18 47.4	31 53.4	5 29.4	0 .0	57 43.8
70 85	0 .0	3 7.9	24 41.4	12 70.6	3 60.0	42 32.3
85 99	0 .0	0 .0	3 5.2	0 .0	2 40.0	5 3.8
Column Total	12 9.2	38 29.2	58 44.6	17 13.1	5 3.8	130 100

Frecuencia conjunta: peso y estatura

Frecuencia marginal absoluta: peso y estatura

Frecuencia marginal relativa: peso y estatura

Frecuencia relativa de peso condicionada a estatura

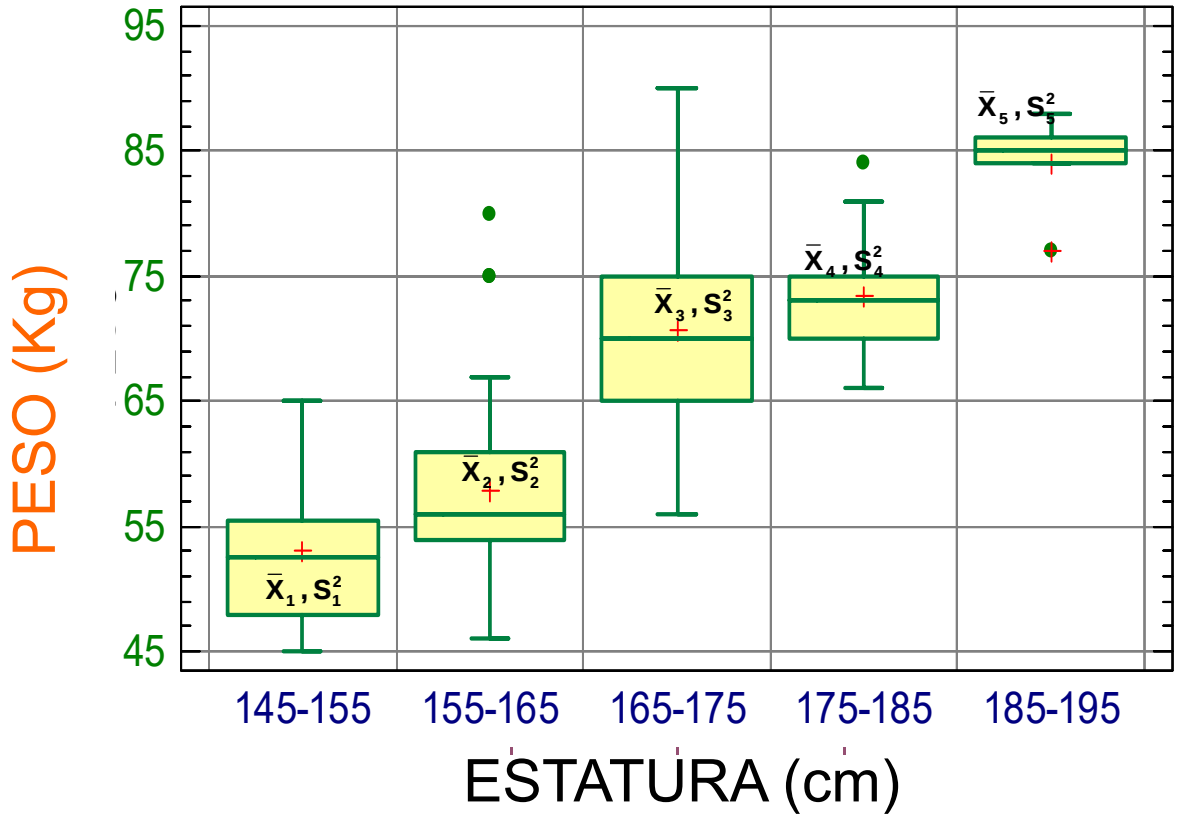
Distribuciones marginales y condicionales

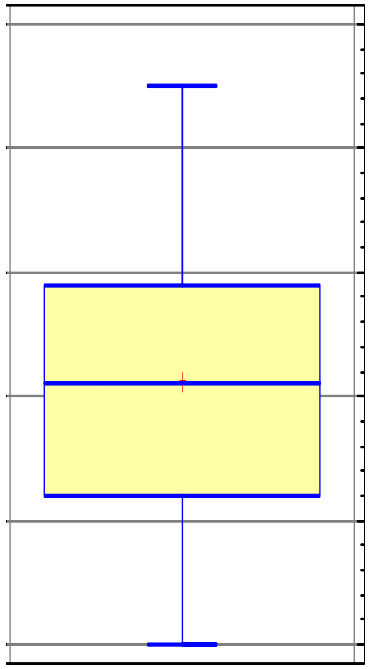
Si dispusiéramos de todos los valores de PESO y ESTATURA de todos los individuos de la población:

- Distribuciones marginales:
  - PESO
  - ESTATURA
- Distribución condicional:
  - PESO/ESTATURA
  - ESTATURA/PESO

Valores del PESO condicionados a los valores de ESTATURA

Los valores de la media de PESO en cada tramo (+) van aumentando (relación)



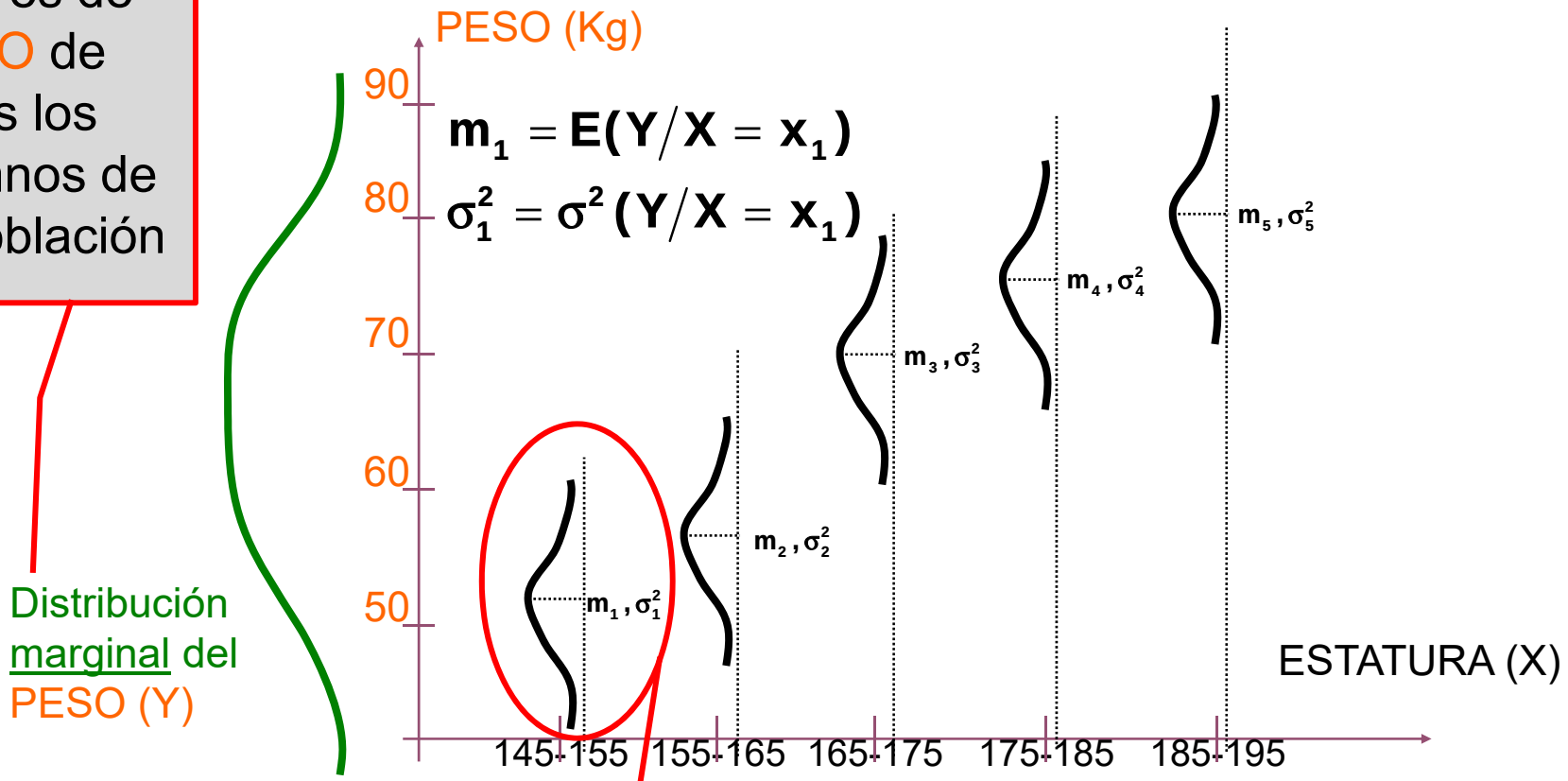


Todos los valores de PESO, al margen de los tramos de ESTATURA

Muestra

Distribuciones condicionales
del PESO respecto a ESTATURA (Y/X)

Valores de PESO de todos los alumnos de la población



Distribución marginal del PESO (Y)

Valores de PESO de todos los alumnos de la "subpoblación" ESTATURA $\in$ [145,155] cm

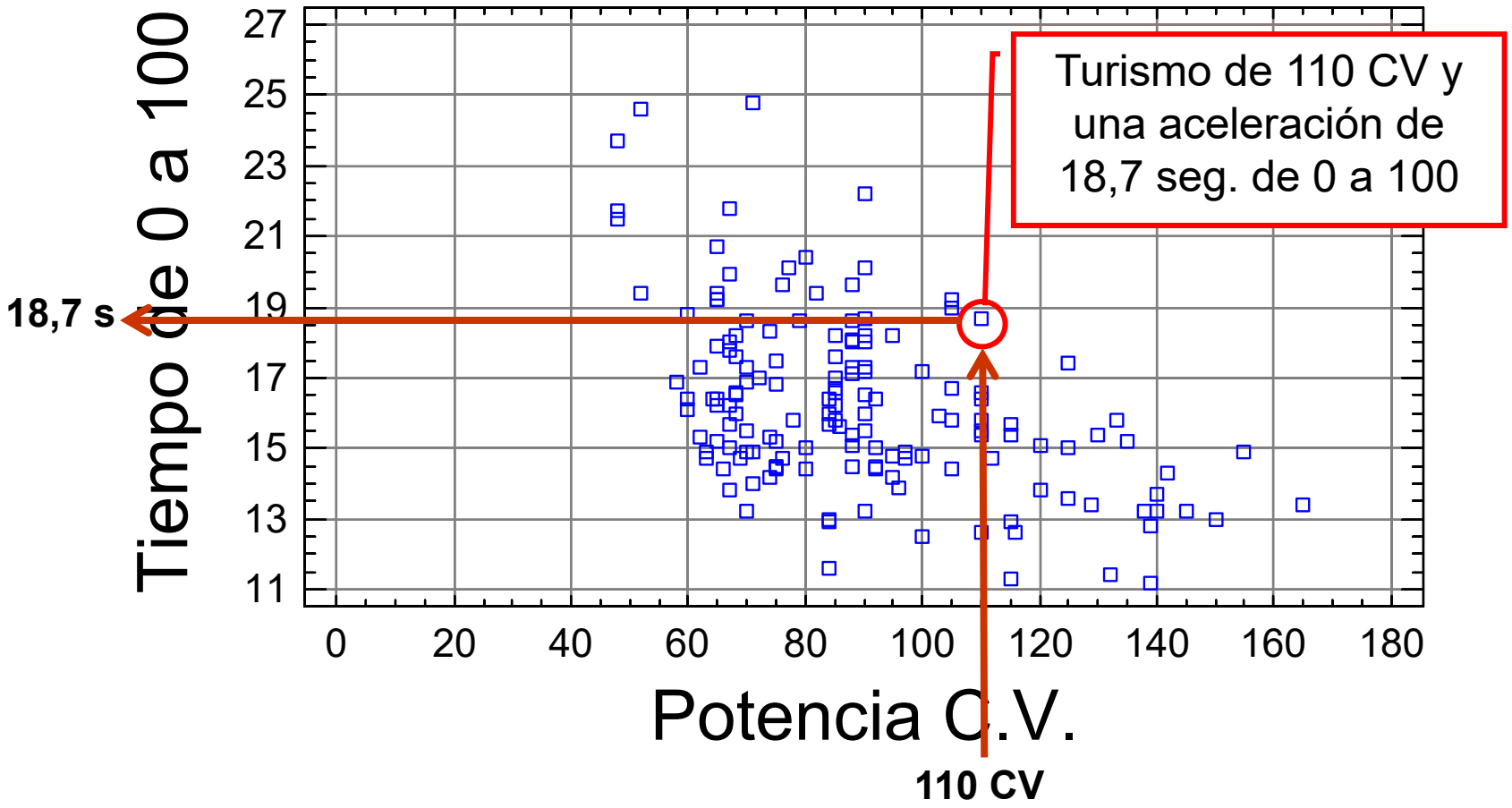
Población

Diagramas de Dispersión

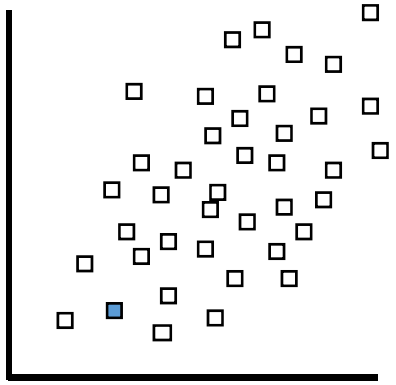
Objeto

- Se trata de una herramienta especialmente útil para estudiar e identificar las posibles relaciones entre los cambios observados en dos conjuntos diferentes de variables.
- Proporciona un medio visual para:
  - Suministrar datos para plantear hipótesis acerca de si dos variables están relacionadas.
    - Probar la fuerza de una posible relación.
  - Detectar datos anómalos.

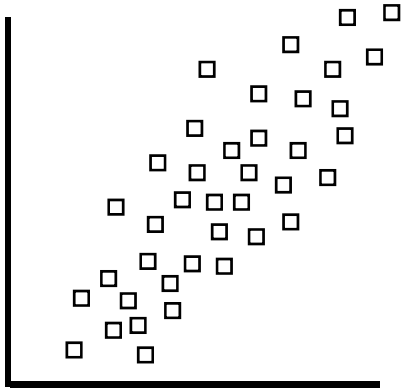
Ejemplo 1


Turismo de 110 CV y una aceleración de 18,7 seg. de 0 a 100
18,7 s
110 CV
Tiempo de 0 a 100
Potencia C.V.
27
25
23
21
19
17
15
13
11
0
20
40
60
80
100
120
140
160
180

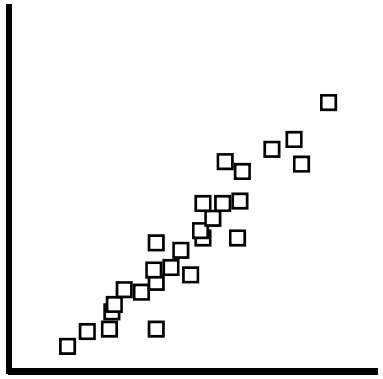
Interpretación de los Diagramas de Dispersión



Relación lineal Débil



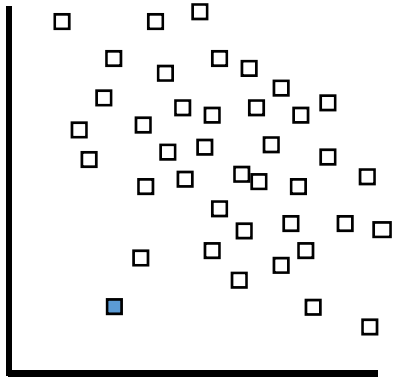
Relación lineal Intermedia



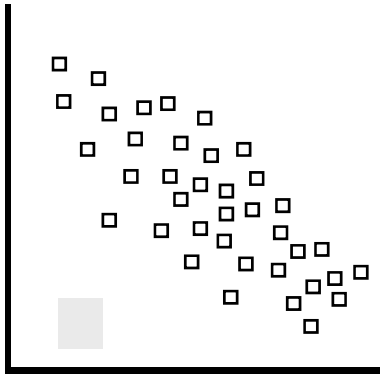
Relación lineal Fuerte

Relación positiva

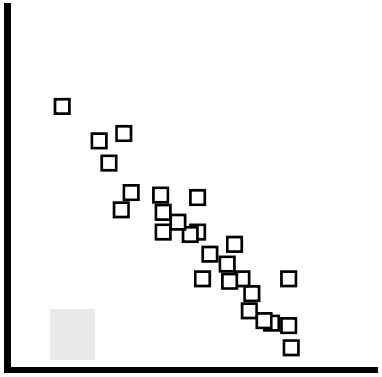
Interpretación de los Diagramas de Dispersión



Relación lineal Débil



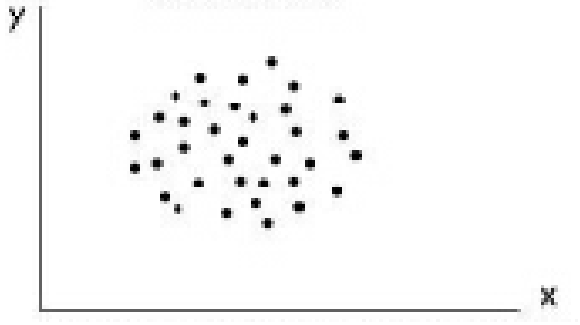
Relación lineal Intermedia



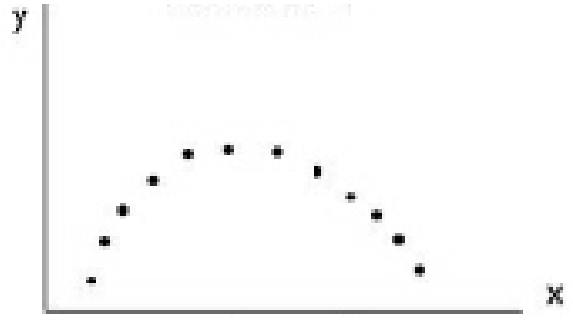
Relación lineal Fuerte

Relación negativa

Interpretación de los Diagramas de Dispersión



No Relación



Sí Relación, pero NO lineal

http://www.ruf.rice.edu/~lane/stat_sim/reg_by_eye/index.html

Covarianza

Objetivo: **cuantificar** el grado de **relación lineal** existente entre dos **componentes** de una **v.a. bidimensional** mediante en un **índice numérico**.

Cálculo:

$$COV_{(X,Y)} = S_{X,Y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{N - 1}$$

Promedio de los productos de las desviaciones de cada componente de la v.a. respecto a su media

Interpretación

$Cov_{xy} > 0$

- Es probable que exista una relación lineal positiva entre las dos componentes de la v.a. →
- A mayores valores de X, mayores valores de Y

$Cov_{xy} < 0$

- Es probable que exista una relación lineal negativa entre las dos componentes de la v.a. →
- A mayores valores de X, menores valores de Y

$Cov_{xy} \approx 0$

- No existe relación lineal entre las componentes de la v.a.
- ¡ Puede existir una relación de otro tipo (cuadrática, …) !

Coeficiente de correlación lineal

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. bidemensional

$$r_{XY} = \frac{Cov_{XY}}{S_X S_Y} = \frac{S_{XY}}{S_X S_Y}$$

Ejemplo: $$r_{A,P} = \frac{Cov_{AP}}{S_A S_P} = \frac{S_{AP}}{S_A S_P} =$$

$$= \frac{-30,41}{24,42 \times 2,52} = -0,49$$

Interpretación de los valores de r_{XY}

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

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

¡ puede haberla de otro tipo !

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

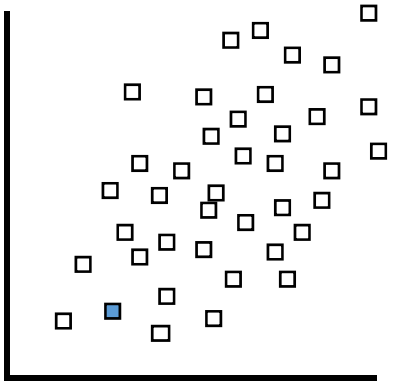
Si $r_{XY} \approx -1 \Leftrightarrow$ relación LINEAL inversa

Valores de r_{xy} que se usan frecuentemente en la práctica como referencia aproximada:

Diagramas de Dispersión y r_{XY}

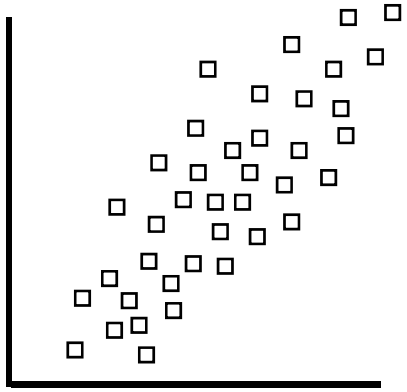
Relación positiva

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



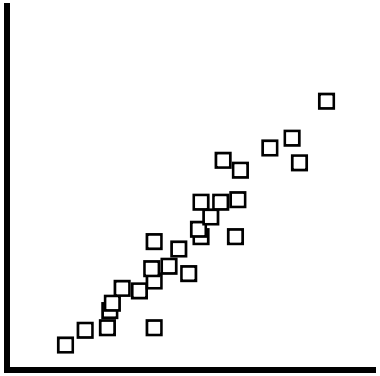
Relación lineal Débil o
No relación

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



Relación lineal
Intermedia

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

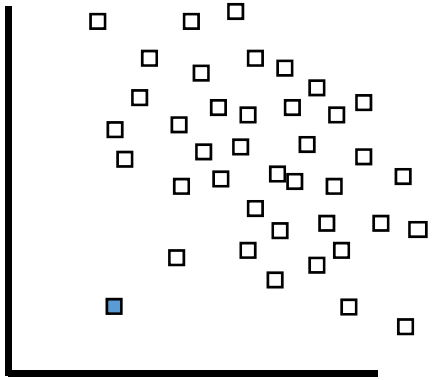


Relación lineal
Fuerte

Diagramas de Dispersión y r_{XY}

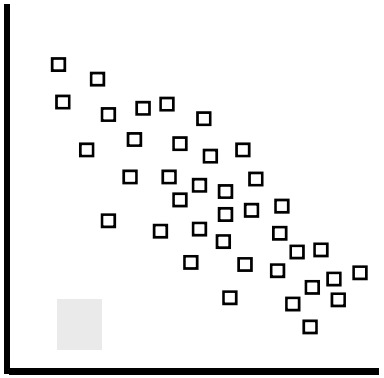
Relación negativa

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



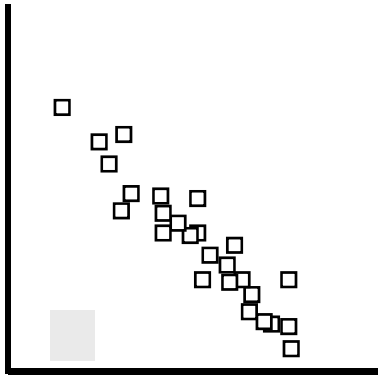
Relación lineal Débil
o No relación

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



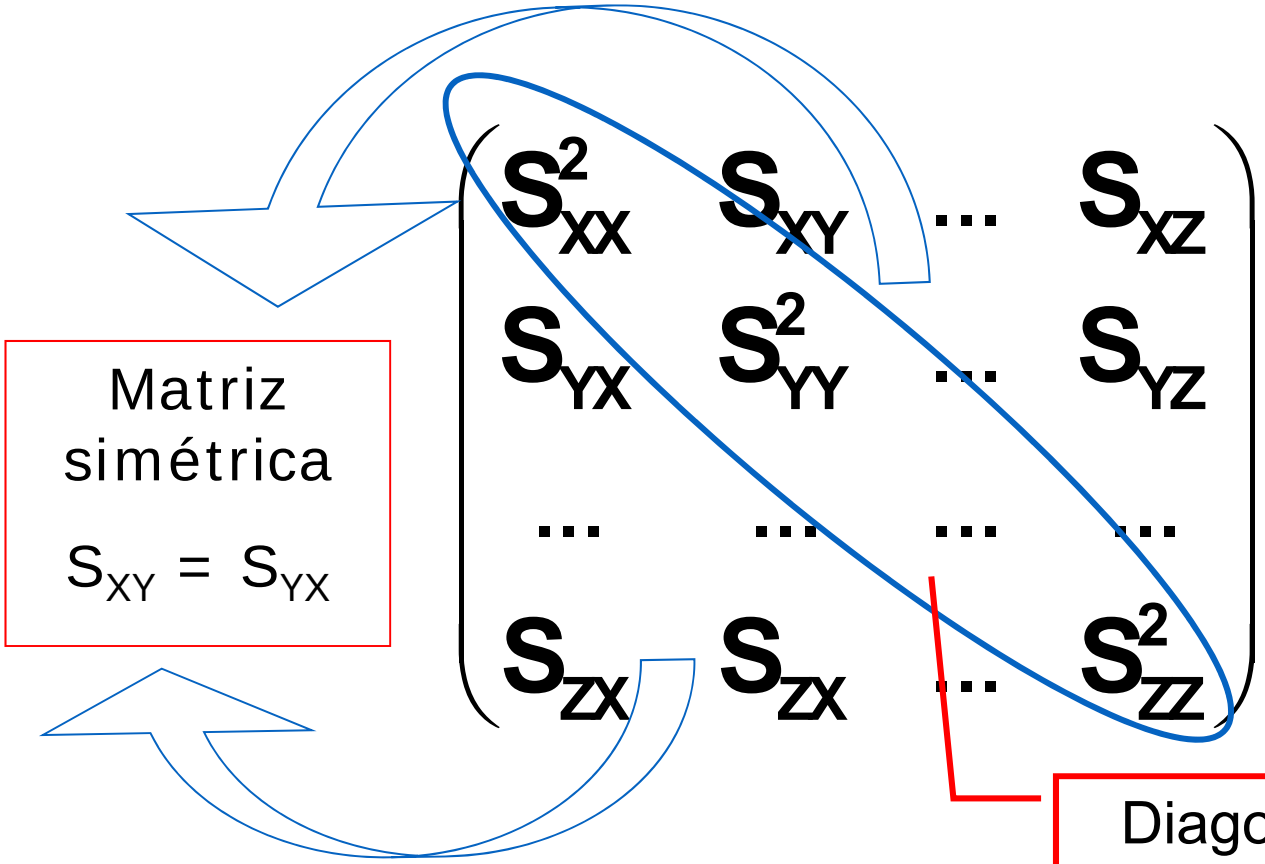
Relación lineal
Intermedia

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



Relación lineal
Fuerte

Matriz de Varianzas-Covarianzas (K-dim)



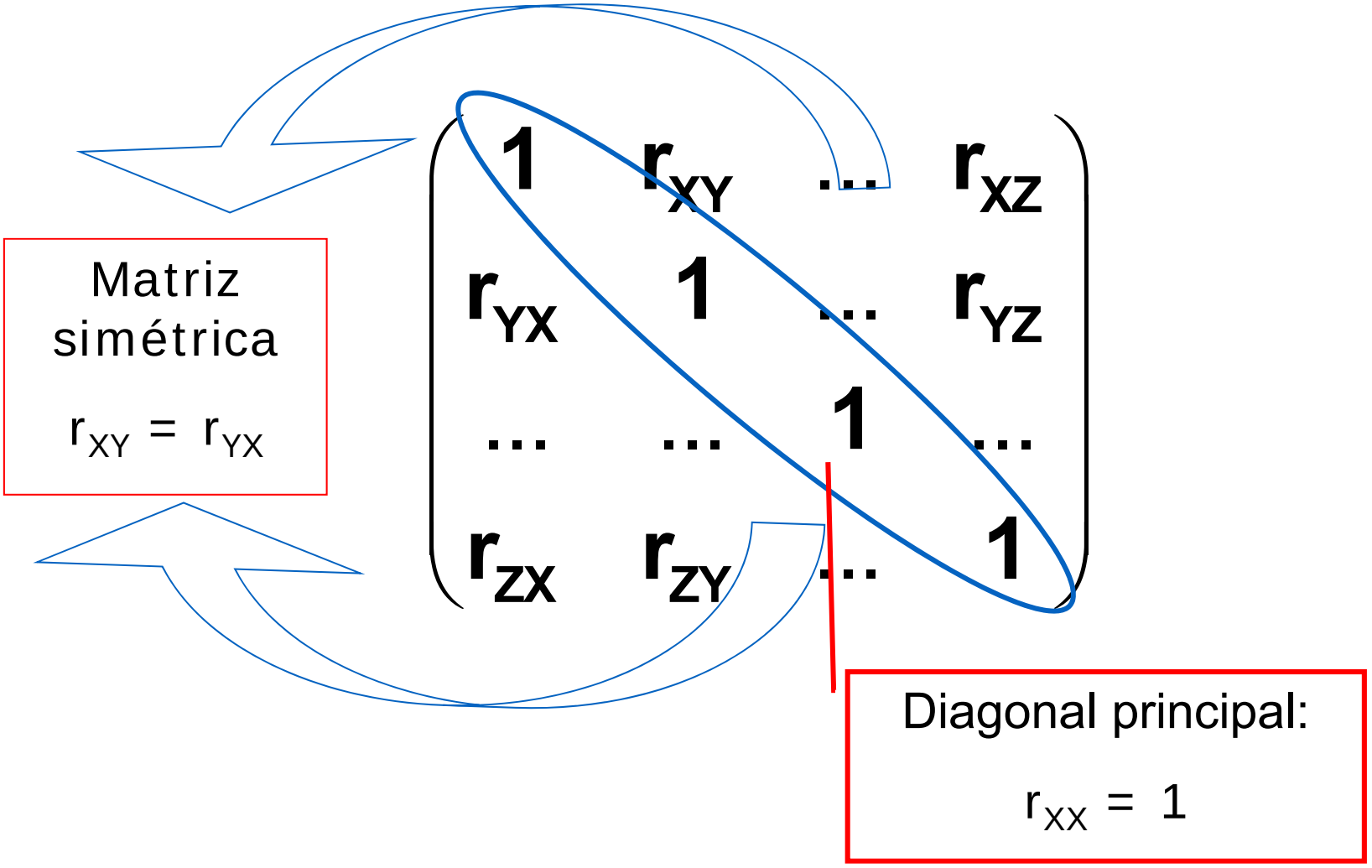
Matriz simétrica

$S_{XY} = S_{YX}$

Diagonal principal: VARIANZAS

$S_{xx} = Cov_{XX} = S^2_X$

Matriz de Correlación $\bar{\bar{R}}$



Estadística descriptiva con R

- Se van a usar funciones básicas (paquetes: **base**)
- Hay resultados que se pueden obtener de diferentes modos y usando distintos paquetes.

Hoja de Datos y descripción

```
### Cargar datos

load("JaenIndicadores.RData")
```

El fichero *JaenIndicadores.RData*[1] contiene datos sobre indicadores importantes de los municipios de la provincia de Jaén en el año 2001, e incluye las siguientes variables:

- Código INE del municipio.
- Nombre del municipio.
- Consumo de energía eléctrica en megavatios por hora.
- Consumo medio de agua en invierno, en metros cúbicos por día.
- Consumo medio de agua en verano, en metros cúbicos por día.
- Destino de los residuos sólidos urbanos: las posibilidades son vertedero controlado, vertedero incontrolado, compostaje.
- Cantidad de residuos sólidos urbanos, en toneladas.
- Tipo de municipio (Grande, Mediano o Pequeño)
- Otros calculados a partir de las variables anteriores

[1] Sáez Castillo, A.J., 2010. *Métodos Estadísticos con R y R Commander,* Jaén: Universidad de Jaén.

Análisis descriptivo

Para la variable tamaño de municipio ***(Tipo)***, obtener:

- Las **frecuencias absolutas**
- Las **frecuencias relativas** (en tanto por cien y como proporción)
- Un resumen de los parámetros de posición

Para la variable consumo medio de agua por habitante ***(agua.hab),*** obtener:

- Un resumen de los parámetros de posición con *summary()* y *fivnum()*

[1] Sáez Castillo, A.J., 2010. *Métodos Estadísticos con R y R Commander*, Jaén: Universidad de Jaén.

Tablas de frecuencias con R

```
### FRECUENCIAS ABSOLUTAS
> Tabla <- table(Datos$Tipo)
> Tabla

 Grande Mediano Pequeño
     33       30       33


### FRECUENCIAS RELATIVAS
> Tabla.rel<-prop.table(Tabla)
> Tabla.rel

 Grande Mediano Pequeño
0.34375 0.31250 0.34375
```

Tablas de frecuencias con R

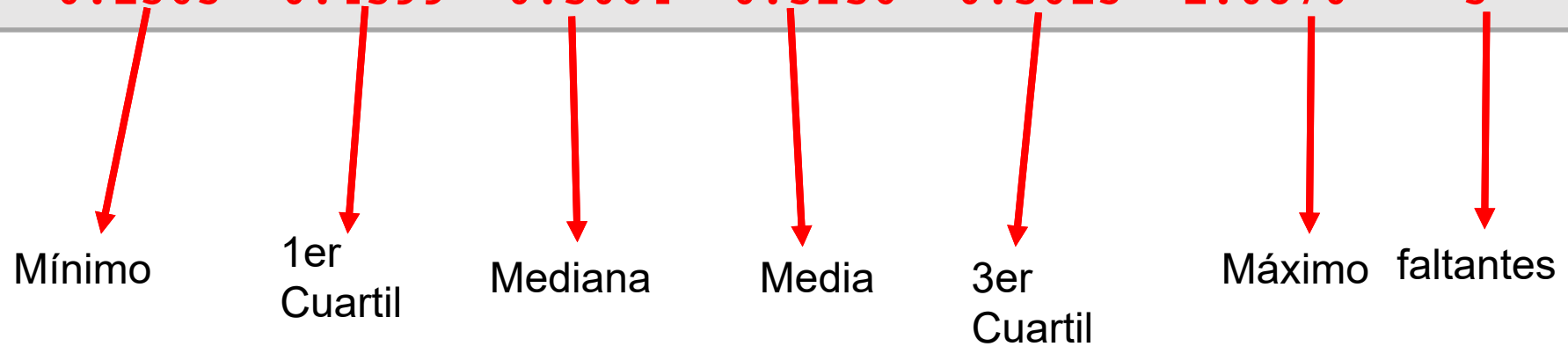
```
### FRECUENCIAS RELATIVAS %
# Frecuencias relativas %
> Tabla.rel <- round(Tabla.rel*100, 2)
> Tabla.rel

 Grande Mediano Pequeño %
  34.38   31.25   34.38
```

Análisis descriptivo

- Para las variables consumo medio de agua por habitante ***(agua.hab),*** obtener:
  - Resumen con *summary()* y *fivnum()*
  - *Mínimo y Máximo*
  - *Media, Mediana y Moda*
  - *Percentiles 5% y 95%*
  - *Varianza y Desviación típica*
  - *Recorrido*
  - *Recorrido intercuartílico*
  - *Coeficiente de variación*
  - *Coeficientes de asimetría y curtosis*

Parámetros de posición con R

```
## resumen principales parámetros
> summary(Datos$agua.hab)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.1363  0.4599  0.5064  0.5256  0.5613  1.0870       3
```



Los cuartiles están calculados como percentiles teóricos (25% y 75%)

Parámetros de posición con R

```
## resumen principales parámetros
> fivnum(Datos$agua.hab)
 0.1362745  0.4586279  0.5063754  0.5616029  1.0872600
```

- Los cuartiles están calculados como mediana de cada mitad de datos.
- No calcula la media
- No indica los datos faltantes
- Sólo admite una variable

Parámetros de posición con R

```
# Media o promedio

mean(Datos$agua.hab)

> [1] NA
```

Observamos que el resultado es `NA` porque hay datos que faltan. Hay que decirle a la función que los elimine del cálculo con el parámetro `na.rm`:

```
mean(Datos$agua.hab, na.rm=TRUE)

> [1] 0.5256292
```

Parámetros de posición con R

```
# Mediana

median(Datos$agua.hab, na.rm=T)

> [1] 0.5063754


# Percentil 5% y percentil 95%

quantile(Datos$agua.hab,probs=c(0.05,0.95),na.rm=
TRUE)

>         5%        95%
> 0.3934972 0.7649731
```

Parámetros de posición con R

```
## Moda
# Obtención de frecuencias absolutas
> frec.agua.hab<-table(Datos$agua.hab)


# Obtención del valor (y posición) de máxima
frecuencia absoluta
> moda<-frec.agua.hab[which(frec.agua.hab ==
max(frec.agua.hab))[1]]
```

Parámetros de posición con R

```
## Mínimo
min.Tasa<-min(Datos$Tasa.actividad.2001, na.rm=T)


## Máximo
max.Tasa<-max(Datos$Tasa.actividad.2001, na.rm=T)


## Primer cuartil
q1.Tasa<-quantile(Datos$Tasa.actividad.2001, probs=0.25,
na.rm=T)


## Tercer cuartil
q3.Tasa<-quantile(Datos$Tasa.actividad.2001, probs=0.75,
na.rm=T)
```

Parámetros de dispersión con R

```
# Desviación típica
sd(Datos$agua.hab,na.rm=TRUE)

# Varianza
var(Datos$agua.hab,na.rm=TRUE)

# Coeficientes de variación.
sd(Datos[,c("agua.hab","elec.hab","res.hab")],na.rm=TRUE)
/mean(Datos[,c("agua.hab","elec.hab","res.hab")],na.rm=TRUE)


# Recorrido intercuartílico
ri.Tasa<-q3.Tasa-q3.Tasa

# Rango
rango.Tasa<-max.Tasa-min.Tasa
```

Parámetros de forma con R

```
library(e1071)


# Coeficiente de asimetría
skewness(Datos$agua.hab,na.rm=TRUE)
skewness(Datos$elec.hab,na.rm=TRUE)


# Coeficiente de curtosis
kurtosis(Datos$agua.hab,na.rm=TRUE)
```

Parámetros por grupos con R

```
# Análisis por grupos dados por la variable Tipo
levels(Datos$Tipo)#Es la variable relativa al tamaño del municipio
#A la función tapply tenemos que decirle:
#1. ?Qué? datos manejamos?
#2. ?Qué? factor es el que determina los grupos?
#3. ?Qué? función queremos aplicar?
#4. Información adicional necesaria para la función
tapply(Datos$agua.hab,Datos$Tipo,mean, na.rm=TRUE)
tapply(Datos$agua.hab,Datos$Tipo,sd, na.rm=TRUE)
tapply(Datos$agua.hab,Datos$Tipo,quantile,probs=c(0.05,0.95),na.rm=TRUE)

tapply(Datos$elec.hab,Datos$Tipo,mean, na.rm=TRUE)
tapply(Datos$elec.hab,Datos$Tipo,sd, na.rm=TRUE)
tapply(Datos$elec.hab,Datos$Tipo,quantile,probs=c(0.05,0.95),na.rm=TRUE)

tapply(Datos$res.hab,Datos$Tipo,mean, na.rm=TRUE)
tapply(Datos$res.hab,Datos$Tipo,sd, na.rm=TRUE)
tapply(Datos$res.hab,Datos$Tipo,quantile,probs=c(0.05,0.95),na.rm=TRUE)
```

Parámetros relación con R

COVARIANZA

```
> cov(datos1$ESTATURA, datos1$PESO,
use="pairwise.complete.obs")
[1] 71.61585
```

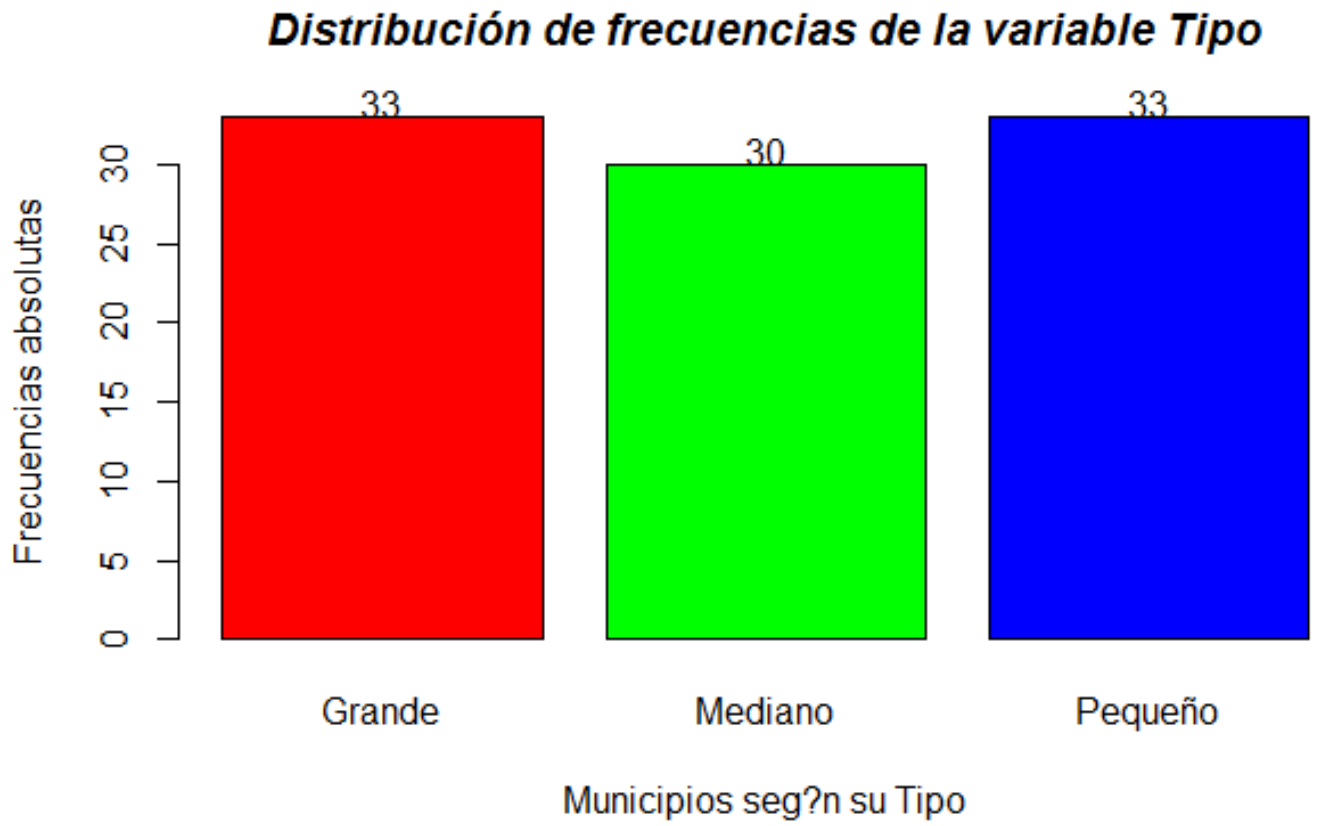
CORRELACIÓN r

```
> cor(datos1$ESTATURA, datos1$PESO,
use="pairwise.complete.obs")
[1] 0.7404422 ## Alta correlación
> cor(datos1$ESTATURA, datos1$EDAD,
use="pairwise.complete.obs")
[1] 0.08681962 ## Correlación inexistente o no lineal
```

Algunos gráficos con R

```
### DIAGRAMA DE BARRAS para la variable Tipo


# Con las frecuencias absolutas
> barras.tipo<-
barplot(Tabla,col=rainbow(3),xlab="Municipios según su
Tipo",ylab="Frecuencias absolutas")


#Ahora quiero añadir las frecuencias al gráfico:
> text(barras.tipo,Tabla + 1,labels=Tabla, xpd = TRUE)
> title(main = "Distribución de frecuencias de la
variable Tipo", font.main = 4)
```

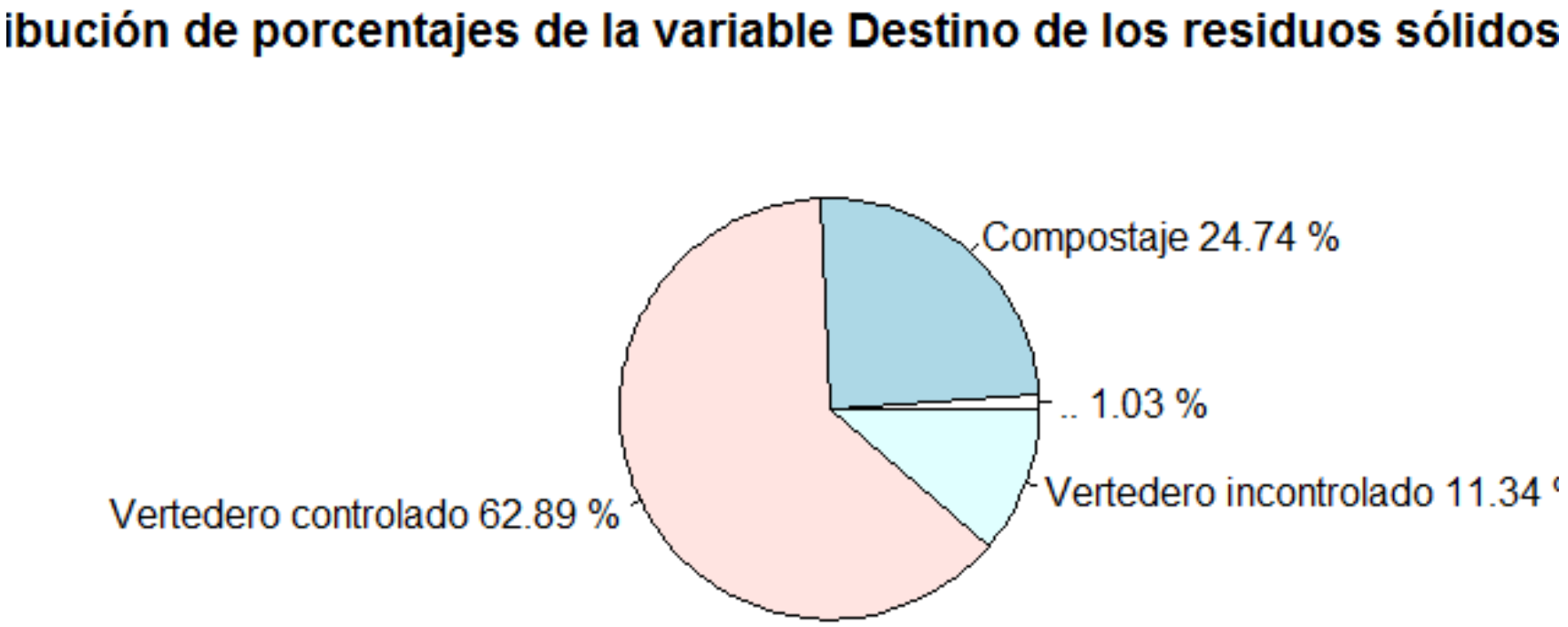
Algunos gráficos con R



Algunos gráficos con R

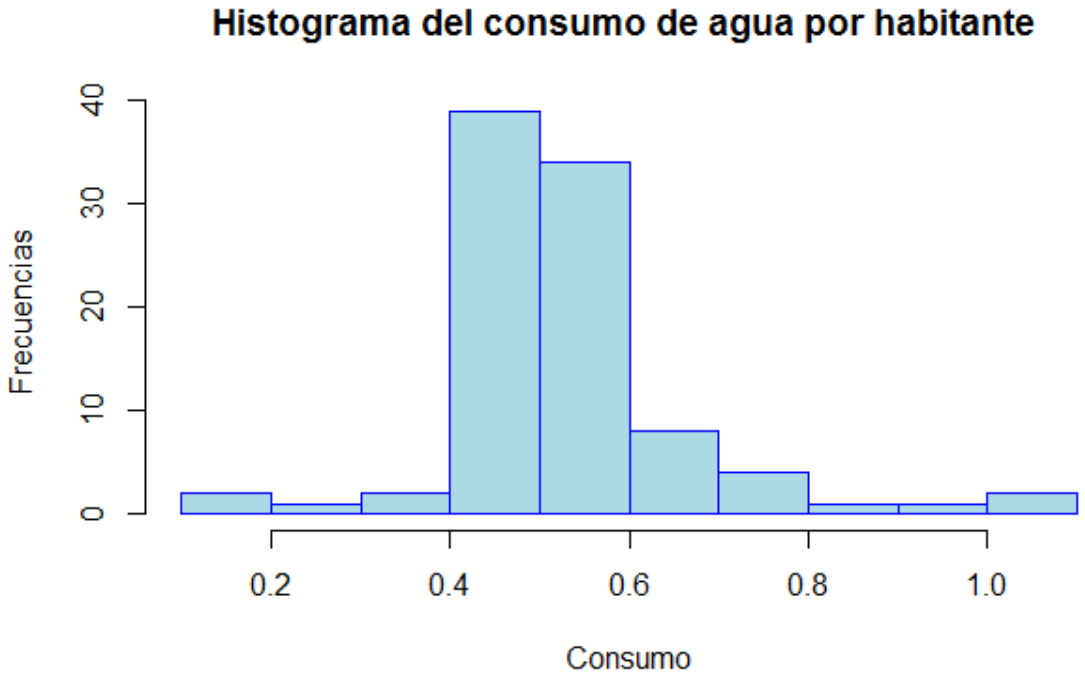
```
### DIAGRAMA DE SECTORES para la variable
Residuos.sólidos.urbanos..Destino

# Con las frecuencias relativas expresadas en
porcentajes

tabla.destino<-
prop.table(table(Datos$Residuos.sólidos.urbanos..Destino))

tabla.destino<-round(100*tabla.destino,2)#En
porcentaje y redondeando


sectores.destino<-
pie(tabla.destino,labels=paste(names(tabla.destino),tabla.destino,"%"),main="Distribuci?n de porcentajes de la variable Destino de los residuos s?lidos urbanos")
```

Algunos gráficos con R



Algunos gráficos con R

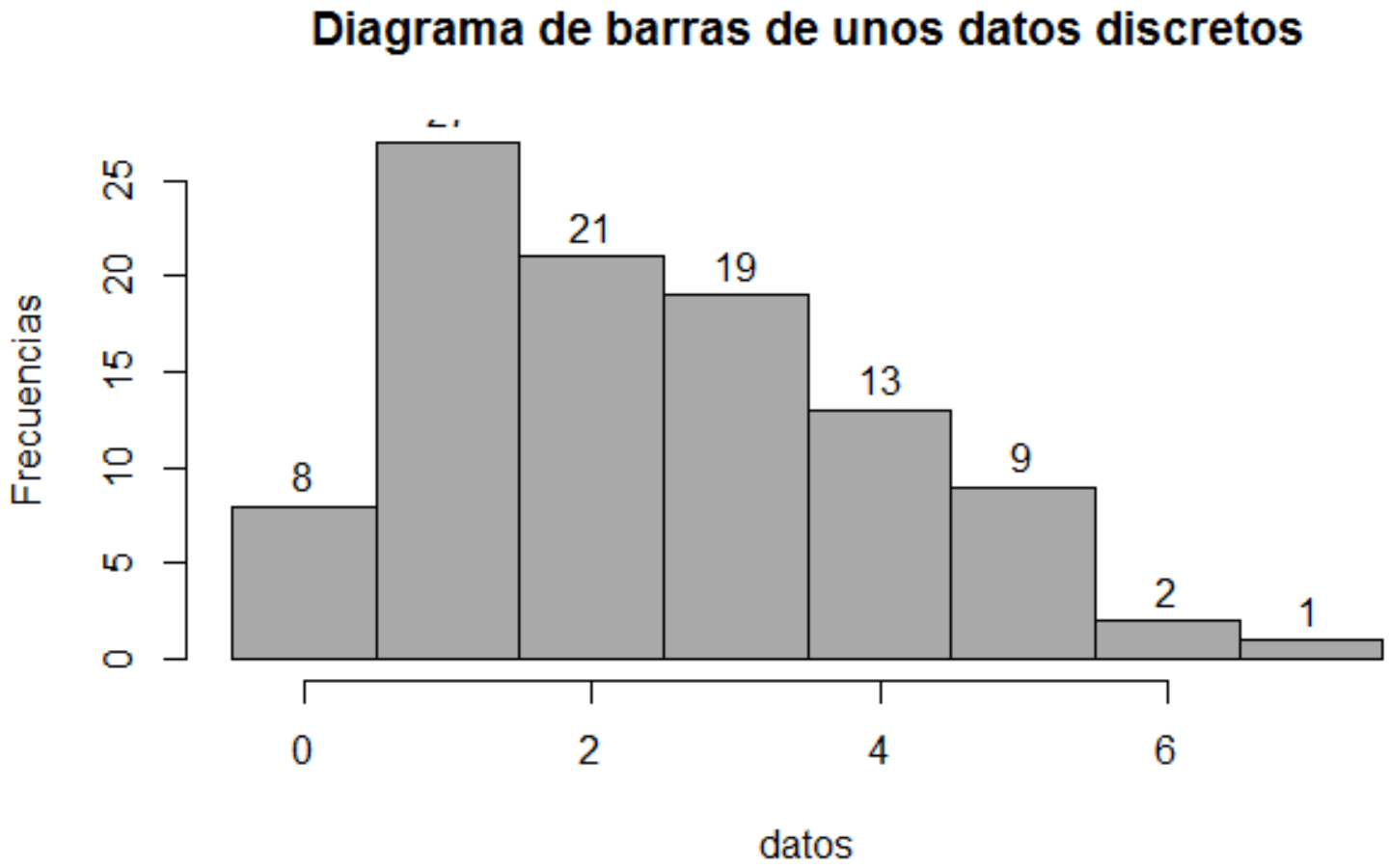
```
###  HISTOGRAMA PARA VARIABLE CONTINUA

hist(Datos$agua.hab, breaks = 10, freq = TRUE, main =
"Histograma del consumo de agua por habitante
",xlab="Consumo",ylab="Frecuencias", col="lightblue",
border="blue")
```



Algunos gráficos con R

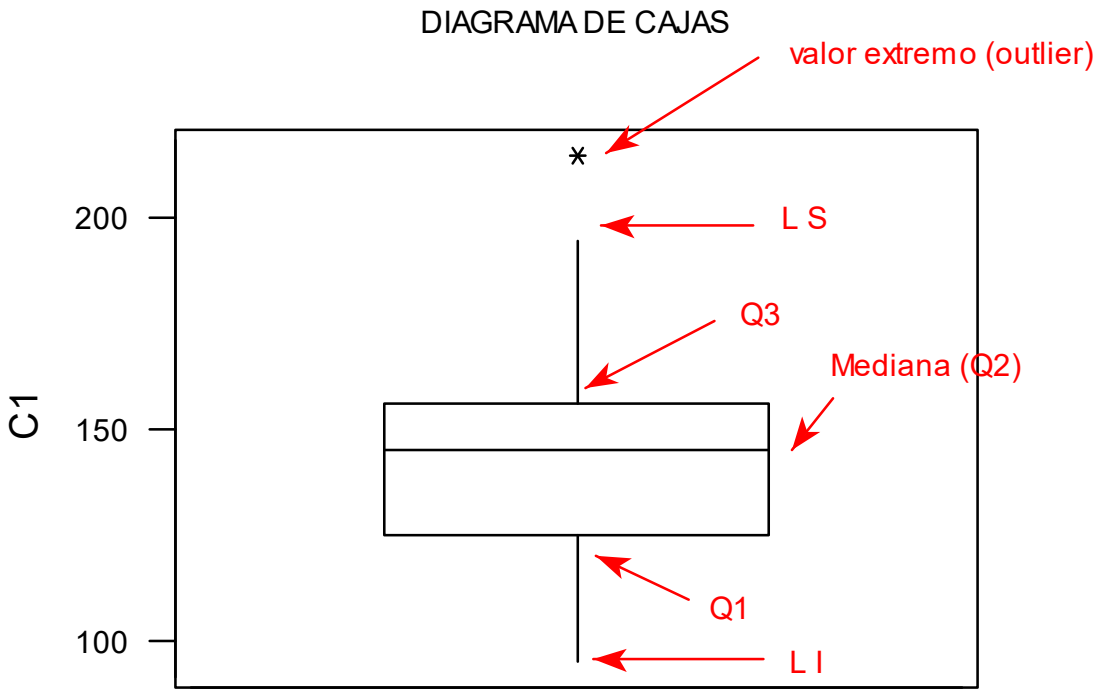
```
### HISTOGRAMA PARA VARIABLE DISCRETA

#Diagrama de barras de una variable discreta mediante
un histograma

#Tomemos estos datos

datos<-rpois(100,2.5)

#Los puntos de corte del histograma fuerzan a que
contemos los 0, los 1, ...

cortes<-(min(datos)-0.5):(max(datos)+0.5)

#El histograma:

hist(datos,breaks=cortes,freq=TRUE,labels=TRUE,ylab="F
recuencias",main="Diagrama de barras de unos datos
discretos", col="darkgray", border="black")
```

Algunos gráficos con R



Algunos gráficos con R

El **diagrama de caja y bigotes** (Box-Whisker) es una representación gráfica que permite apreciar las principales características de un conjunto de datos, señalando los datos anómalos.

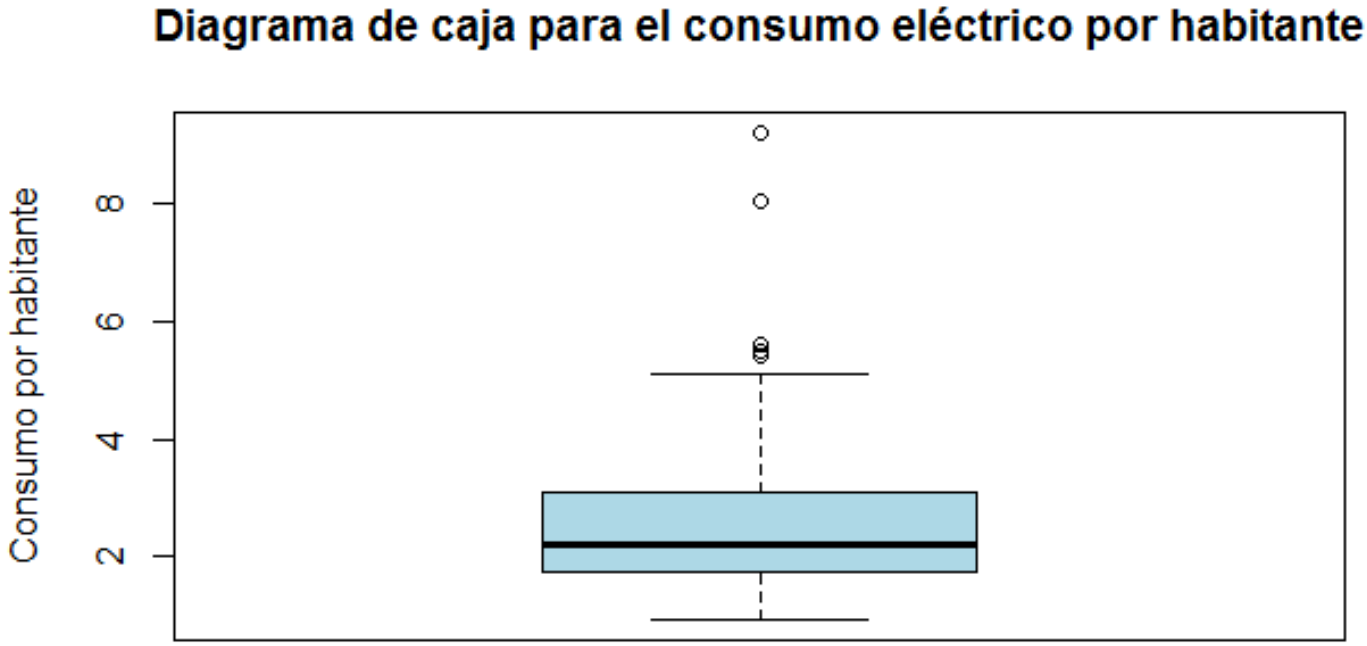


Límite inferior : L I = Q1 – 1,5*(Q3 – Q1)

Límite superior : L S = Q3 + 1,5*(Q3 – Q1)

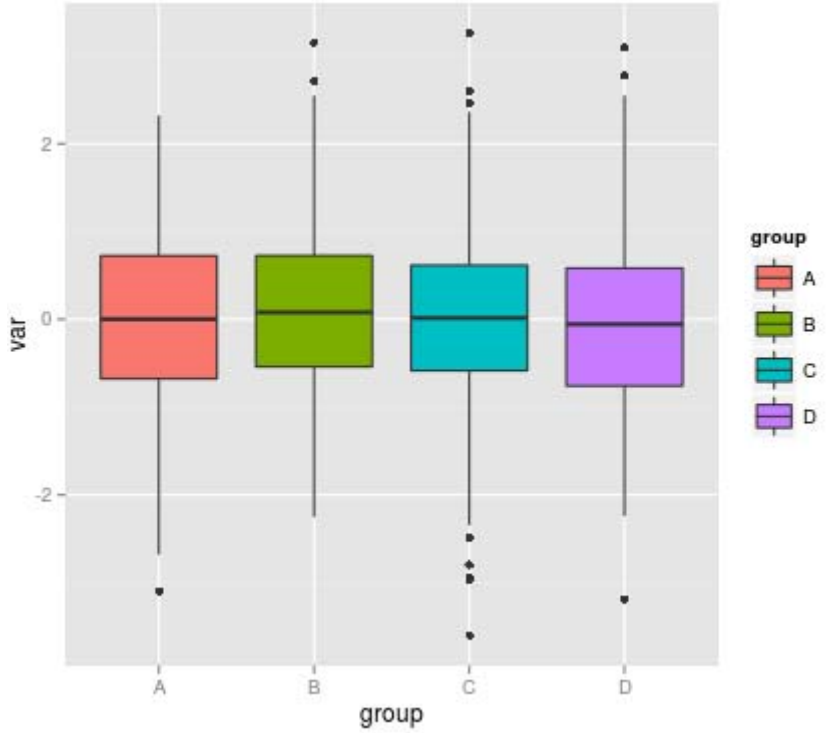
Algunos gráficos con R

```
### DIAGRAMA DE CAJA de elec.hab
> boxplot(Datos$elec.hab,main="Diagrama de caja para
el consumo eléctrico por habitante",ylab="Consumo por
habitante", col="lightblue")
```



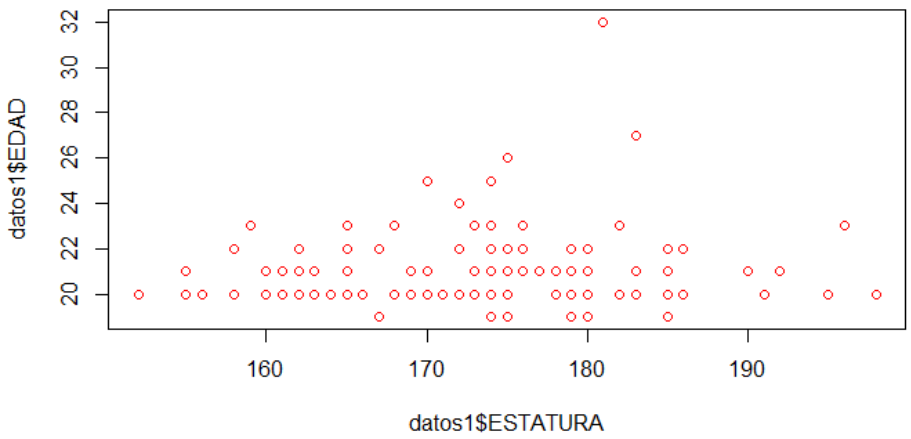
Algunos gráficos con R

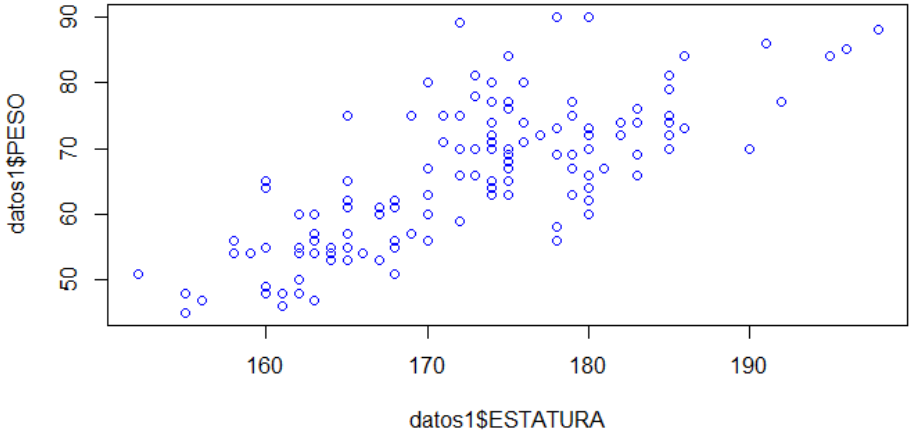
```
> library(gplot2) Esto lo veréis en el taller de Visualización de datos médicos


> gplot(group, var, geom="boxplot", data=datos, fill=group)
```



Diagramas de dispersión con R

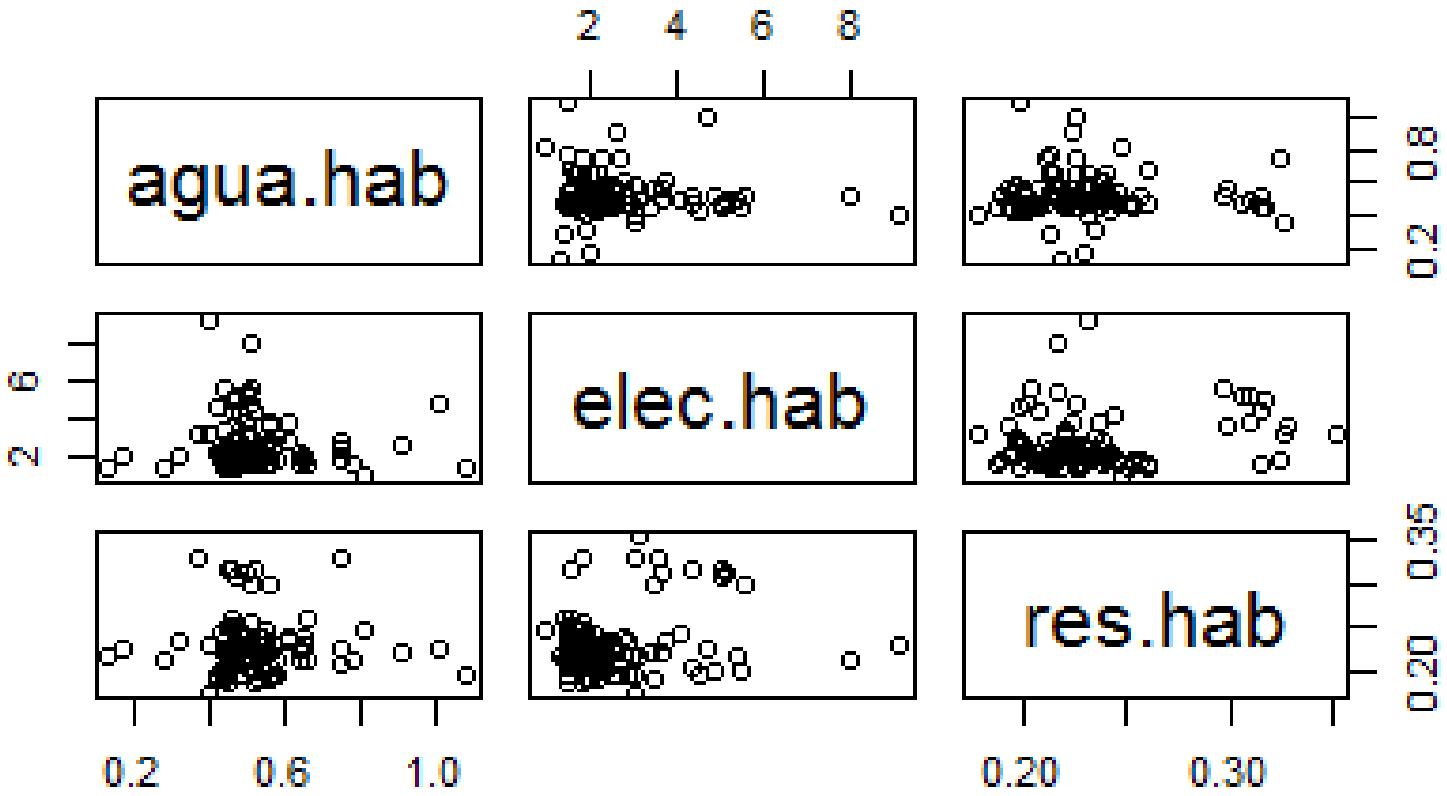
```
> plot(datos1$ESTATURA, datos1$EDAD, type="p", col="red")
> plot(datos1$ESTATURA, datos1$PESO, type="p",
col="blue")
```





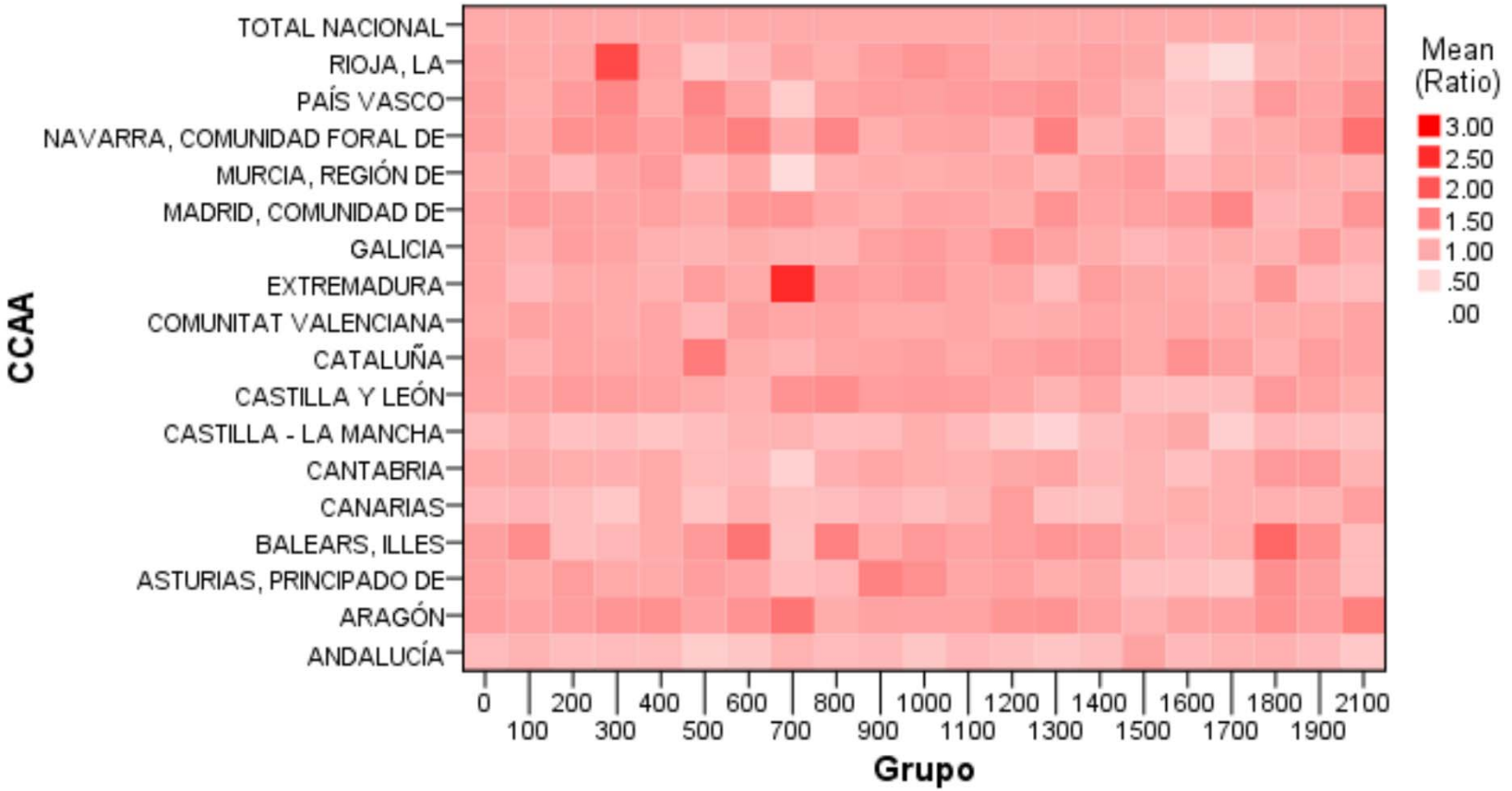
Parámetros relación con R

```
> pairs(Datos[,9:11])
```



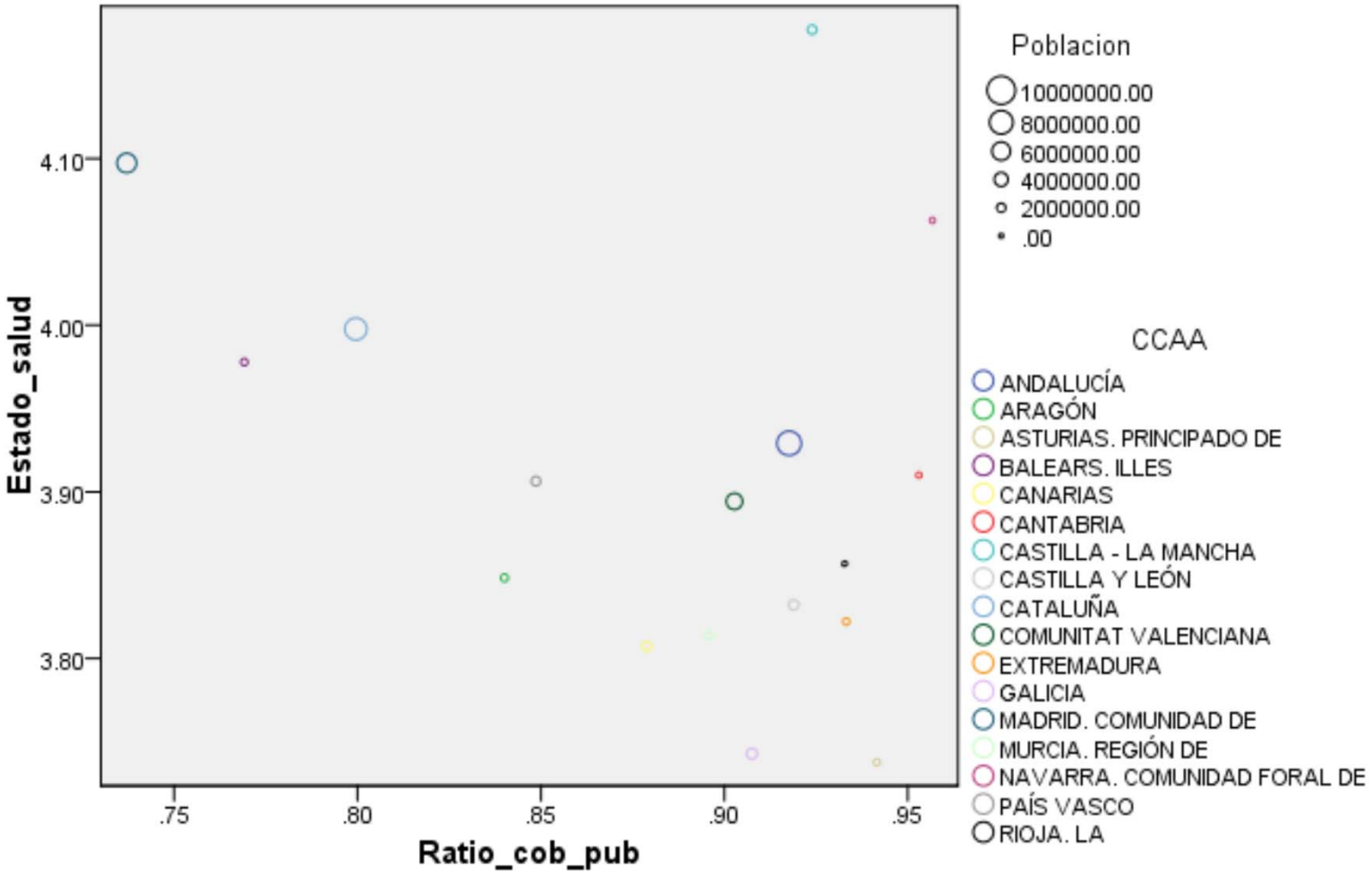
Nuevas técnicas de visualización

Mapa de calor



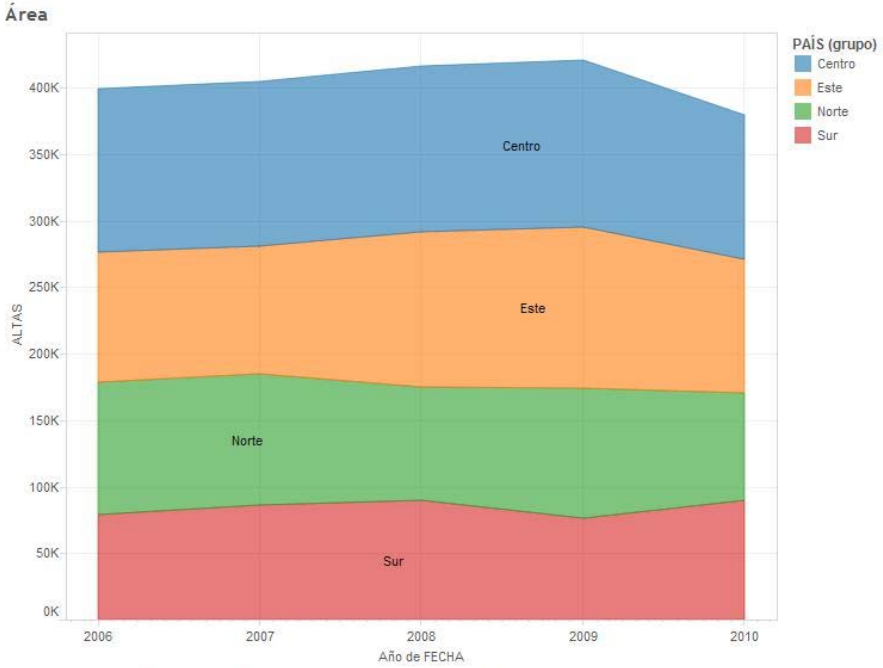
Nuevas técnicas de visualización

Gráfico de burbujas (dispersión)



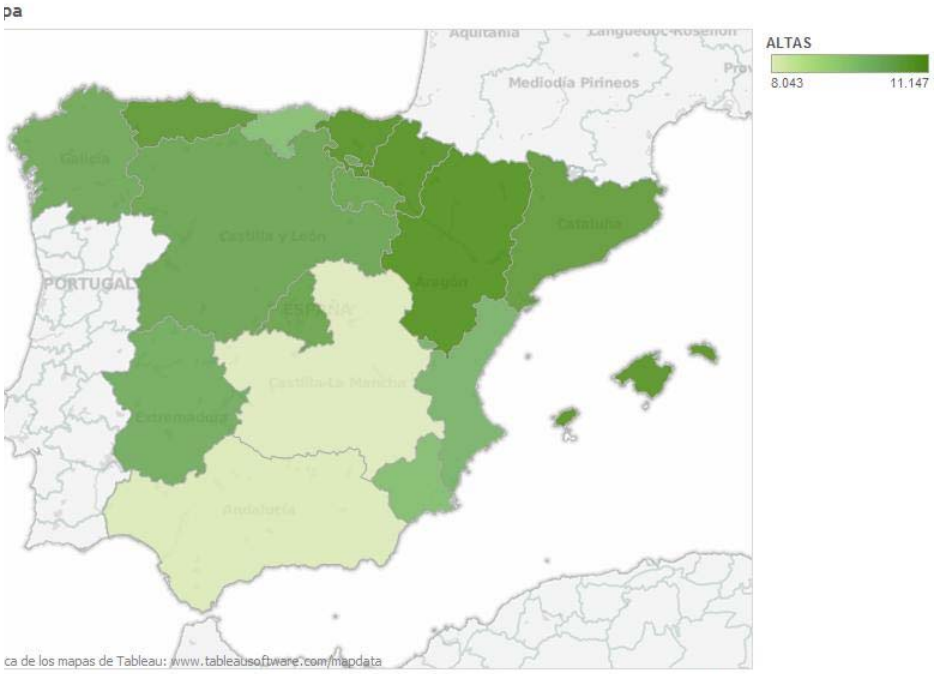
Nuevas técnicas de visualización

Gráfico de áreas



El diagrama de suma de ALTAS para FECHA año. El color muestra detalles acerca de PAÍS (grupo). Las marcas se etiquetan por PAÍS (grupo).

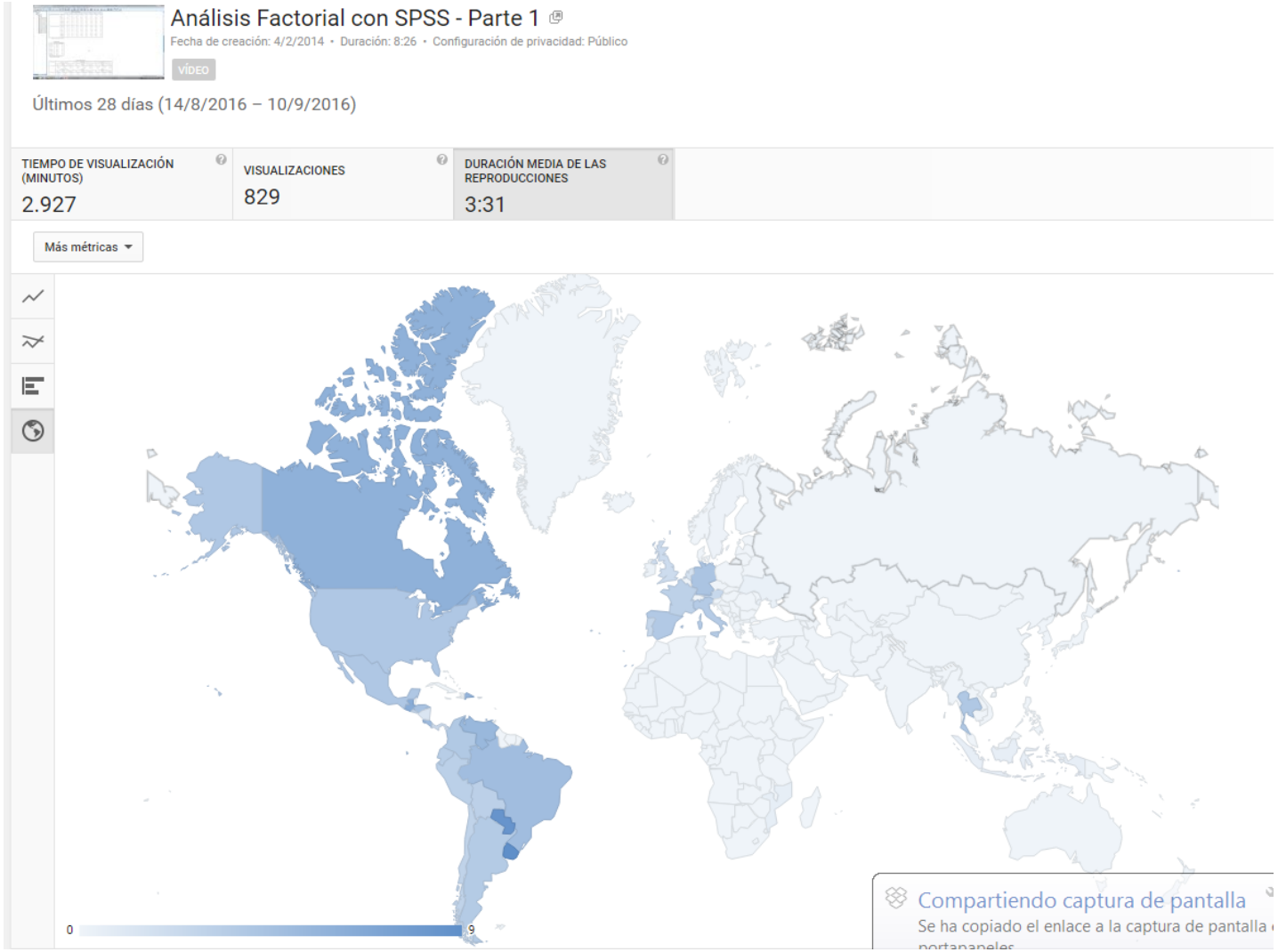
Mapa geográfico



a basado en Longitud (generado) y Latitud (generado). El color muestra suma de ALTAS. Se muestran detalles para CCAA. Los datos se filtran en GRUPO, lo que conserva Todas. La vista se filtra en CCAA, lo que excluye CANARIAS y TOTAL NACIONAL.

Nuevas técnicas de visualización
Infogramas o pictogramas

YouTube Analytics



Glosario

Apuntamiento	Frecuencia acumulada	Parámetro muestral	Posición
Box & Whisker	Frecuencia relativa	Individuo	Recorrido o Rango
Caja y Bigotes	Histograma	Inferencia	Recorrido o Rango Intercuartílico
Campana de Gauss	Desviación típica	Media	Simetría
Característica aleatoria	Diagrama de barras	Mediana	Tabla de frecuencias
Coeficiente de asimetría	Diagrama de sectores o tarta	Moda	Tabla de frecuencias cruzadas
Coeficiente de curtosis	Dispersión	Muestra	v.a. Continua
Coeficiente de Variación	Distribución Normal	Muestreo	v.a. Cualitativa
Cuartil	Estadística	Normalidad	v.a. Cuantitativa
Dato anómalo	Experimento aleatorio	Parámetro de dispersión	v.a. Discreta
Dato o valor extremo	Frecuencia absoluta	Parámetro poblacional	Variabilidad
Datos estadísticos	Parámetro de forma	Percentil	Variable aleatoria (v.a.)
Descriptiva, Estadística	Parámetro de posición	Población	Varianza

Funciones

table()	tapply()
pop.table()	sapply()
summary()	vapply()
fivnum()	cov()
mean()	cor()
quantile()	barplot()
median()	pie()
min()	hist
max()	boxplot()
sd()	plot()
var()	pairs()
skewness()	
kurtosis()	

Herramientas Estadísticas para Big Data
Introducción a la Inferencia Estadística, Muestreo y Preproceso de datos

1- Conceptos básicos

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

www.upv.es

E. Vázquez
Dto. De Estadística e Investigación Operativa, Aplicadas y Calidad