Universidad del País Vasco
Euskal Herriko Unibertsitatea
The University of the Basque Country

E.U.I.T.I. Bilbao

Asignatura:
MÉTODOS ESTADÍSTICOS DE LA INGENIERÍA

E.U.I.T.I. Bilbao

Asignatura:
MÉTODOS ESTADÍSTICOS DE LA INGENIERÍA

TEMA 2:
ESTADÍSTICA DESCRIPTIVA

1. RESUMEN

Métodos para resumir y describir conjuntos de datos mediante distintos tipos de tablas, gráficos y medidas estadísticas

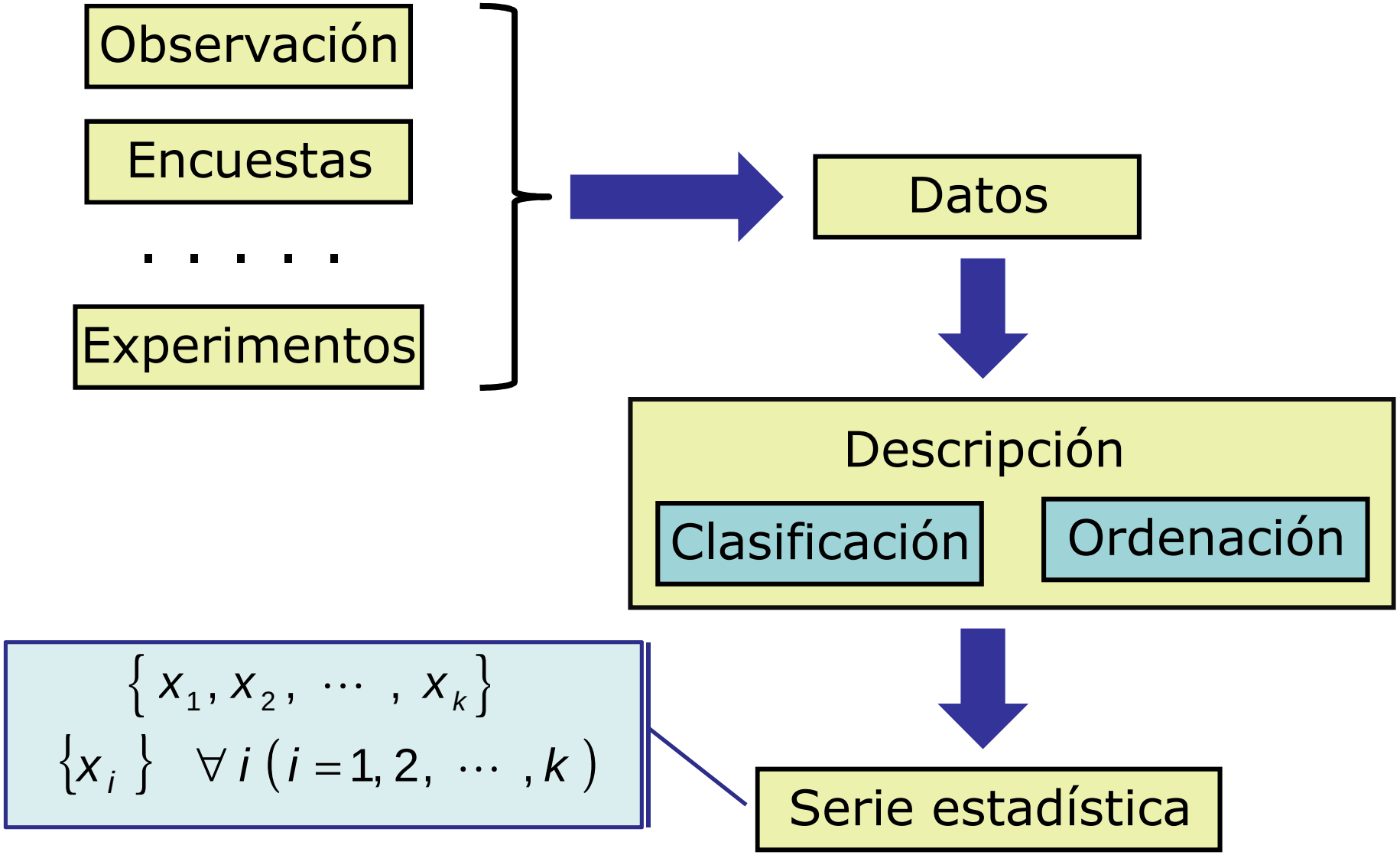
Palabras clave:

- datos cuantitativos y cualitativos
- datos discretos y continuos
- distribución de frecuencias
- diagramas de barras y de sectores, histograma
- media, mediana, moda, cuantiles
- varianza, desviación típica, asimetría, datos atípicos

2. ÍNDICE DEL TEMA

2. ÍNDICE DEL TEMA

2. INTRODUCCIÓN



2. INTRODUCCIÓN

Objetivo básico: describir de forma sencilla los datos para presentarlos de forma clara y concisa

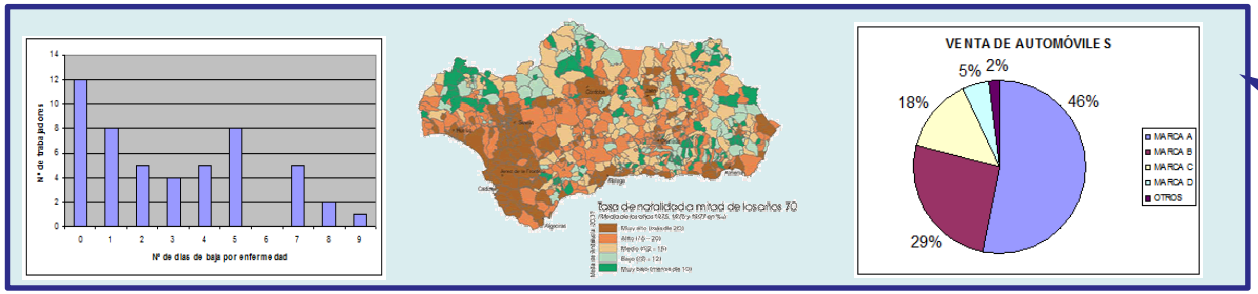
Serie estadística

Presentación

Métodos numéricos

$$\bar{x} = \frac{\sum_{i=1}^{i=n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} \qquad s_x^2 = \frac{\sum_{i=1}^{i=n} (x_i - \bar{x})^2}{n-1}$$

Métodos gráficos



Una imagen vale más que mil palabras

3. DATOS: DEFINICIÓN

Dato (ó caracter): cualidad o propiedad inherente del individuo

- la observación del individuo se describe mediante uno o más datos
- EJEMPLOS

INDIVIDUO	DATOS
LIBRO de una biblioteca	Peso, dimensiones, número de hojas, color de pastas, autor, …
OPERARIO de un taller A	Sexo, edad, grupo sanguíneo, número de hijos, número de piezas producidas por día, …
PIEZAS producidas en taller A	Tamaño, calidad, peso, color, …

4. DATOS: TIPOS

Cualitativos (ó categóricos): características de la población que no pueden asociarse a valores numéricos

- Representación gráfica: diagrama de barras y de sectores
- EJEMPLOS

INDIVIDUO	DATOS
LIBRO de una biblioteca	Color de pastas, autor, …
OPERARIO de un taller A	Sexo, grupo sanguíneo, color de ojos, …
PIEZAS producidas en taller A	Calidad, color, …

4. DATOS: TIPOS

Cuantitativos (ó numéricos): características de la población que pueden asociarse a valores numéricos

- Son medibles
- Se llaman **variables estadísticas** (X, Y, Z, …)
- Sus distintas posibilidades se llaman **valores** (x_i, y_i, z_i, …)
- Tipos: discretos y continuos

4. DATOS: TIPOS

<u>Cuantitativos (ó numéricos)</u>: características de la población que pueden asociarse a valores numéricos

- **<u>Discretos</u>**
  - conteo de una característica
  - toman valores enteros
  - <u>representación gráfica</u>: diagrama de barras
  - EJEMPLOS

Asociarlos con el verbo CONTAR

INDIVIDUO	DATOS
LIBRO de una biblioteca	Nº de páginas, nº de volúmenes, …
OPERARIO de un taller A	Nº de hijos, nº accidentes laborales, …
PIEZAS producidas en taller A	Nº piezas/día, …

4. DATOS: TIPOS

<u>Cuantitativos (ó numéricos)</u>: características de la población que pueden asociarse a valores numéricos

- **<u>Continuos</u>**
  - toman infinitos valores en un intervalo dado
  - <u>representación gráfica</u>: histograma
  - EJEMPLOS

Asociarlos con el verbo MEDIR

INDIVIDUO	DATOS
LIBRO de una biblioteca	Peso, …
OPERARIO de un taller A	Altura, peso, …
PIEZAS producidas en taller A	Peso, diámetro de una arandela, …

4. DATOS: TIPOS

- **EJEMPLO**
  - POBLACIÓN: personas residentes en una ciudad
  - MUESTRA: 50 personas (N=50)
  - DATO 1: opinión sobre una película (buena, mala, regular)
  - DATO 2: edad (para clasificar opinión por edades)

muestra cualitativa

muestra cuantitativa unidimensional

5. DATOS: DESCRIPCIÓN

Recorrido: diferencia entre el valor mayor y el valor menor de una serie estadística

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

- una variable estadística, X, puede tomar diferentes valores

$$\{ x_i \} \quad \forall i \, (i = 1, 2, \cdots, k)$$

- EJEMPLO: temperatura de una persona enferma a lo largo de una semana:

$$R = 41{,}5º - 37{,}5º = 4º$$

5. DATOS: DESCRIPCIÓN

Frecuencia (absoluta): número de veces que se repite el valor x_i en el conjunto de las observaciones

- notación: $n_i \quad \forall\, i\,(i = 1, 2, \cdots, k)$

Frecuencia relativa: cociente entre la frecuencia absoluta y el número de observaciones realizadas, N

- notación: $f_i = \dfrac{n_i}{N} \quad \forall\, i\,(i = 1, 2, \cdots, k)$

$f_i \times 100 \rightarrow \%$

- interpretación: proporción de apariciones del valor x_i respecto del total de observaciones
- propiedad: $\sum_{i=1}^{i=k} f_i = 1$

5. DATOS: DESCRIPCIÓN

Frecuencia absoluta acumulada: número de observaciones hasta la *i* inclusive; se obtiene sumando las frecuencias absolutas de los valores x_j / j≤i (i=1,…, k)

- notación: $N_i = \sum_{j=1}^{j=i} n_j$ ⇒ $N_k = N$

Frecuencia relativa acumulada: cociente entre la frecuencia absoluta acumulada y el número de observaciones realizadas, *N*

- notación: $F_i = \frac{N_i}{N} \quad \forall\, i\ (i = 1, 2, \cdots, k)$
- $F_i = \frac{N_i}{N} = \frac{1}{N} \sum_{j=1}^{j=i} n_j = \sum_{j=1}^{j=i} \frac{n_j}{N} = \sum_{j=1}^{j=i} f_j$

5. DATOS: DESCRIPCIÓN

Tabla de distribución de frecuencias : formato de ordenación y presentación de los datos

- variable estadística, X, cuantitativa o cualitativa discreta

Valores de la variable: x_i	Frecuencia absoluta: n_i	Frecuencia relativa: f_i	Frecuencia absoluta acumulada: N_i	Frecuencia relativa acumulada: F_i
x_1	n_1	$f_1 = n_1/N$	$N_1 = n_1$	$F_1 = N_1/N = f_1$
x_2	n_2	$f_2 = n_2/N$	$N_2 = N_1 + n_2$	$F_2 = N_2/N = F_1 + f_2$
... ...	... ...	... ...	... ...	... ...
x_k	n_k	$f_k = n_k/N$	$N_k = N_{k-1} + n_k = N$	$F_k = N_k/N = F_{k-1} + f_k = 1$
Total	N	1		

5. DATOS: DESCRIPCIÓN

$100 \times f_i = 12\%$: el número de veces que se repite el valor x_i supone el 12% de la muestra (ejemplo)

Valores de la variable: x_i	Frecuencia absoluta: n_i	Frecuencia relativa: f_i	Frecuencia absoluta acumulada: N_i	Frecuencia relativa acumulada: F_i
x_1	n_1	$f_1 = n_1/N$	$N_1 = n_1$	$F_1 = N_1/N = f_1$
x_2	n_2	$f_2 = n_2/N$	$N_2 = N_1 + n_2$	$F_2 = N_2/N = F_1 + f_2$
... ...	... ...	... ...	... ...	... ...
x_k	n_k	$f_k = n_k/N$	$N_k = N_{k-1} + n_k = N$	$F_k = N_k/N = F_{k-1} + f_k = 1$
Total	N	1		

$100 \times F_i = 48\%$: todos los valores menores o iguales que x_i totalizan el 48% de la muestra (ejemplo)

5. DATOS: DESCRIPCIÓN

- EJEMPLO. La siguiente tabla muestra los titulados de la U.P.V./E.H.U. en el año 2001 por tipos de titulación:

Tipo titulación	Nº titulados		
Valores variable: *X*	Frec. absoluta	Frec. relativa	Porcentaje (%)
Experimentales	791	0.085	8,5
Técnicas	2053	0.221	22,1
Salud	586	0.063	6,3
Sociales	2669	0.287	28,7
Económico-jurídicas	2577	0.277	27,7
Humanidades	622	0.067	6,7
Total promoción	9298	1	100

5. DATOS: DESCRIPCIÓN

- EJERCICIO. Los siguientes datos son los números de taras observadas en una muestra de 48 sábanas producidas por una empresa textil :

0	0	0	1	2	6	4	6	4	3	2	2	1	1	1	0
0	6	3	4	5	2	3	1	1	1	0	0	0	3	4	5
2	1	2	1	4	3	1	1	0	0	3	2	0	1	1	0

Obtener la tabla de frecuencias

5. DATOS: DESCRIPCIÓN

- EJERCICIO. Solución

Taras: X	Cantidad						
Valores	Conteo	n_i	N_i	f_i	%	F_i	
0	//// //// //	12	12	0,2500	25,00	0,2500	25,00
1	//// //// ///	13	25	0,2708	27,08	0,5208	52,08
2	//// //	7	32	0,1458	14,58	0,6666	66,66
3	//// /	6	38	0,1250	12,50	0,7916	79,16
4	////	5	43	0,1042	10,42	0,8958	89,58
5	//	2	45	0,0417	4,17	0,9375	93,75
6	///	3	48	0,0625	6,25	1	100
Total		48		1	100		

5. DATOS: DESCRIPCIÓN

Variables agrupadas en intervalos de clase : simplificación de conjuntos de datos con un número de valores elevado (ó variables continuas)

- se clasifican los valores en grupos (**intervalos de clase**)
- se sustituye cada medida por el valor del centro del intervalo (**marca de clase**)

simplificación ➡ pérdida de información

5. DATOS: DESCRIPCIÓN

Variables agrupadas en intervalos de clase

INTERVALOS DE CLASE

- acotados
- extremo inferior de un intervalo igual al extremo superior del intervalo anterior

serie: $\{x_j\}$ $\quad [L_i, L_{i+1}) = \{x_j \ / \ L_i \le x_j < L_{i+1}\}$ $\quad i = 1, 2, \cdots, k$

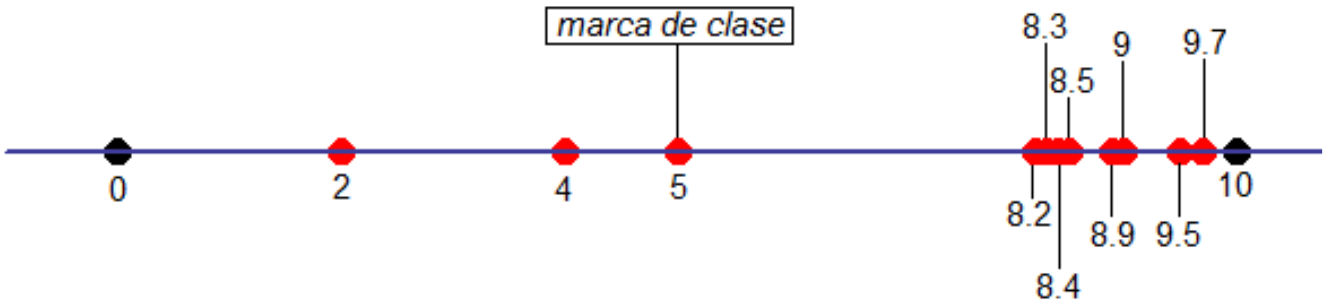
- número apropiado de intervalos: $k \approx \sqrt{N}$
- amplitud constante (siempre que el problema lo permita)

$$A_i = L_{i+1} - L_i \qquad \forall i \ (i = 1, 2, \cdots, k)$$

5. DATOS: DESCRIPCIÓN

Variables agrupadas en intervalos de clase

- **MARCA DE CLASE**
  - punto medio de cada intervalo $[L_i, L_{i+1})$ $\quad M_i = \frac{L_{i+1} + L_i}{2}$
  - valor que representa la información que contiene un intervalo
  - facilita representaciones gráficas y el cálculo de parámetros estadísticos
  - pérdida de información: sólo tiene en cuenta el número de observaciones dentro de cada intervalo y no la distribución en su interior



5. DATOS: DESCRIPCIÓN

Variables agrupadas en intervalos de clase

- **EJEMPLO** Edades en años de 26 niños

3.5	7	9.5	9	6	5	4.5	12	11	10	15	10.5	6
2	10.5	9	2	8	15	3.5	14	12	7	14	6.5	8

- número de intervalos: $\sqrt{26} \approx 5$
- recorrido: $x_{max} - x_{min} = 15 - 2 = 13$

$\Rightarrow \frac{13}{5} \approx 2{,}6$

- se amplían los intervalos inicial y final para contener todas las observaciones en intervalos de amplitud constante: A=3

5. DATOS: DESCRIPCIÓN

Variables agrupadas en intervalos de clase

- **EJEMPLO** Edades en años de 26 niños

3.5	7	9.5	9	6	5	4.5	12	11	10	15	10.5	6
2	10.5	9	2	8	15	3.5	14	12	7	14	6.5	8

Intervalos	Valores interiores	Frecuencias: n_i
$[1, 4)$	3.5, 2, 2, 3.5	4
$[4, 7)$	6, 5, 4.5, 6, 6.5	5
$[7, 10)$	7, 9.5, 9, 9, 8, 7, 8	7
$[10, 13)$	12, 11, 10, 10.5, 10.5, 12	6
$[13, 16)$	15, 15, 14, 14	4

5. DATOS: DESCRIPCIÓN

Variables agrupadas en intervalos de clase

- **TABLA DE FRECUENCIAS**

1. rango de datos (recorrido): $rango = max(x_i) - min(x_i)$
2. número de intervalos: $k \approx \sqrt{N}$
3. intervalos de igual amplitud (si es posible)

Intervalos	*Marcas de clase: x_i*	*Frecuencias absolutas*		*Frecuencias relativas*	
$[L_1, L_2)$	x_1	n_1	$N_1 = n_1$	f_1	$F_1 = f_1$
$[L_2, L_3)$	x_2	n_2	$N_2 = N_1 + n_2$	f_2	$F_2 = F_1 + f_2$
... ...	... ...	... ...	... ...	... ...	... ...
$[L_k, L_{k+1})$	x_k	n_k	$N_k = N_{k-1} + n_k$	f_k	$F_k = F_{k-1} + f_k = 1$
Total		N		1	

REPRESENTACIONES GRÁFICAS

1. INTRODUCCIÓN

- Analizar el siguiente gráfico y opinar sobre la información que suministra

El comercio, al filo de la «catástrofe»

Las ventas en Vizcaya cayeron un 10% el año pasado y la patronal alerta de que la situación irá todavía a peor por la subida del IVA

JOSU GARCÍA
josugarcia@diario-elcorreo.es

BILBAO. «Estamos al límite. No hay más que ver que la mayoría de las tiendas están totalmente vacías. Muchos dependientes no pueden

Ventas del sector comercio en Vizcaya

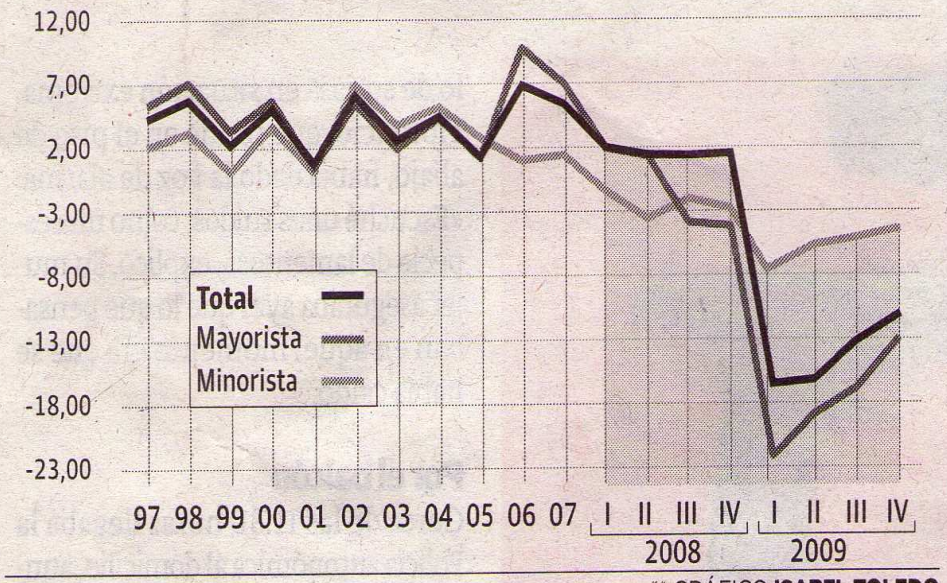


GRÁFICO ISABEL TOLEDO

LOS DATOS

-7,7%

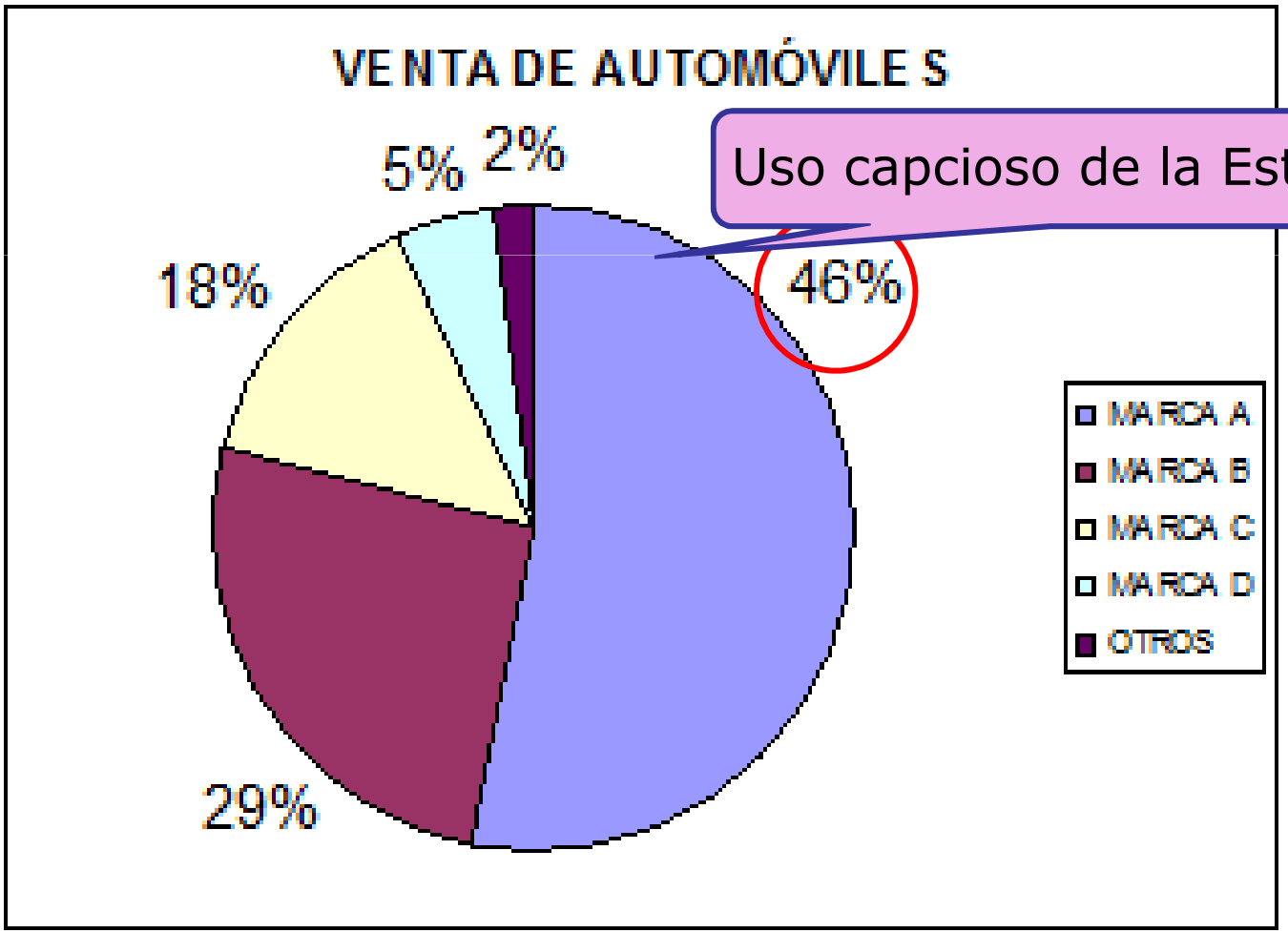
El comercio tradicional de alimentación fue uno de los más castigados.

-3,9%

La destrucción de empleo castigó también al comercio vizcaíno. Casi 4 de cada 100 operarios perdieron su puesto en 2009.

1. INTRODUCCIÓN

- Analizar el siguiente gráfico y opinar sobre la información que suministra



2. VARIABLE CUANTITATIVA DISCRETA

Diagrama de barras

- en el eje de abscisas se colocan los diferentes valores de la variable

$$\{ x_i \} \quad \forall\, i \,(i = 1, 2, \cdots, k)$$

- sobre cada una de ellos se levanta una barra (ó línea) cuya altura es la frecuencia (que se mide, por tanto, en el eje de ordenadas)
- conjunto de barras verticales cuyas alturas suman **N** (frecuencia absoluta) ó **1** (frecuencia relativa)

2. VARIABLE CUANTITATIVA DISCRETA

Polígono de frecuencias

- gráfico en el que se muestran las frecuencias de los diferentes valores de la variable y, luego, se conectan los puntos del gráfico mediante líneas rectas

- es decir, se obtiene uniendo los extremos superiores de las barras de un diagrama de barras

2. VARIABLE CUANTITATIVA DISCRETA

Diagrama de barras y polígono de frecuencias

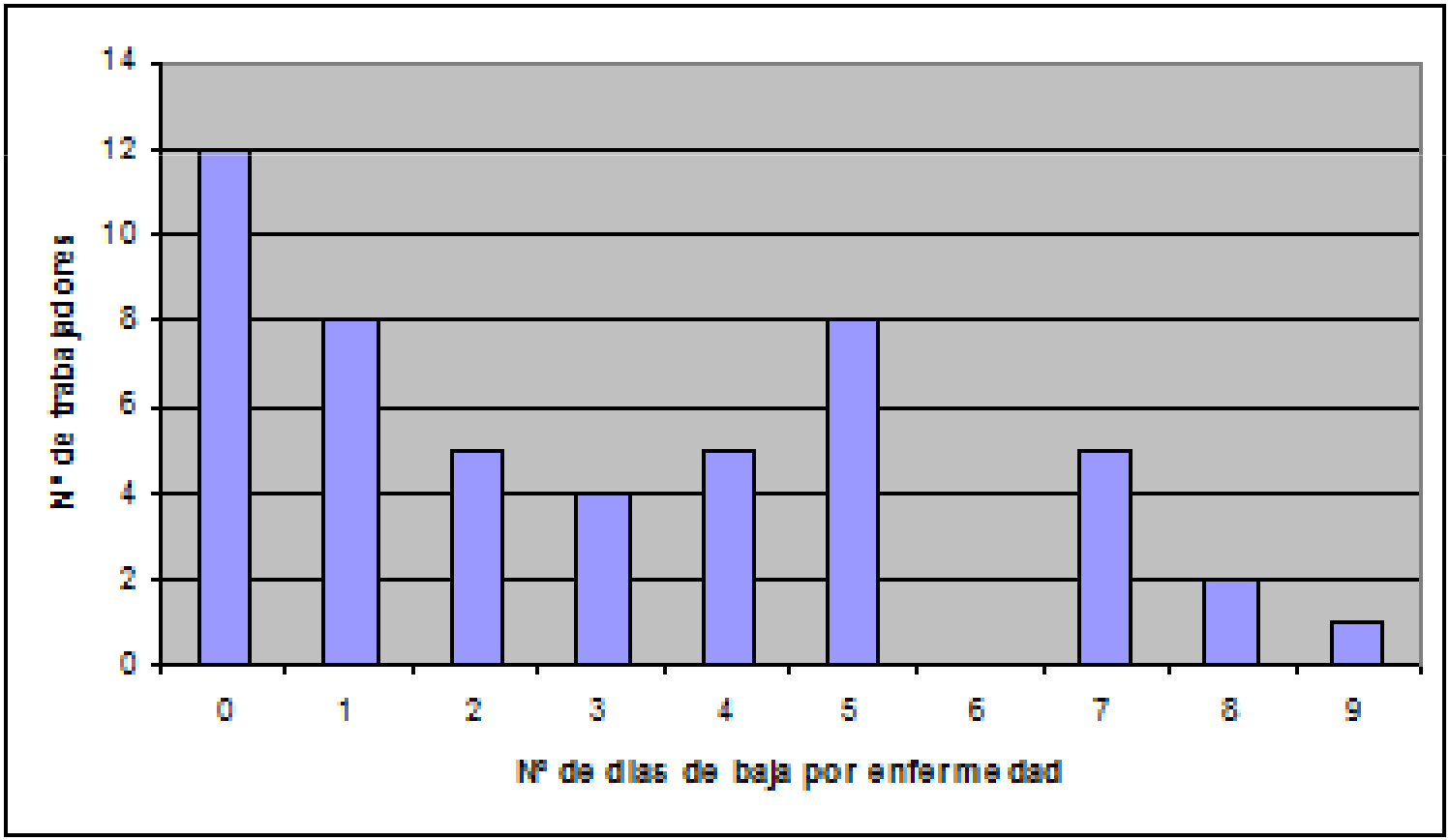
- EJEMPLO: número de días de baja de los trabajadores de una fábrica

X: días de baja	Nº trabajadores	
Valores variable: *X*	Frec. Absoluta: n_i	Frec. Relativa: f_i
0	12	$\frac{12}{50} = 0{,}24$
1	8	$\frac{8}{50} = 0{,}16$
2	5	$\frac{5}{50} = 0{,}10$
3	4	$\frac{4}{50} = 0{,}08$
4	5	$\frac{5}{50} = 0{,}10$
5	8	$\frac{8}{50} = 0{,}16$
6	0	$\frac{0}{50} = 0{,}00$
7	5	$\frac{5}{50} = 0{,}10$
8	2	$\frac{2}{50} = 0{,}04$
9	1	$\frac{1}{50} = 0{,}02$
Totales	50	1

2. VARIABLE CUANTITATIVA DISCRETA

Diagrama de barras

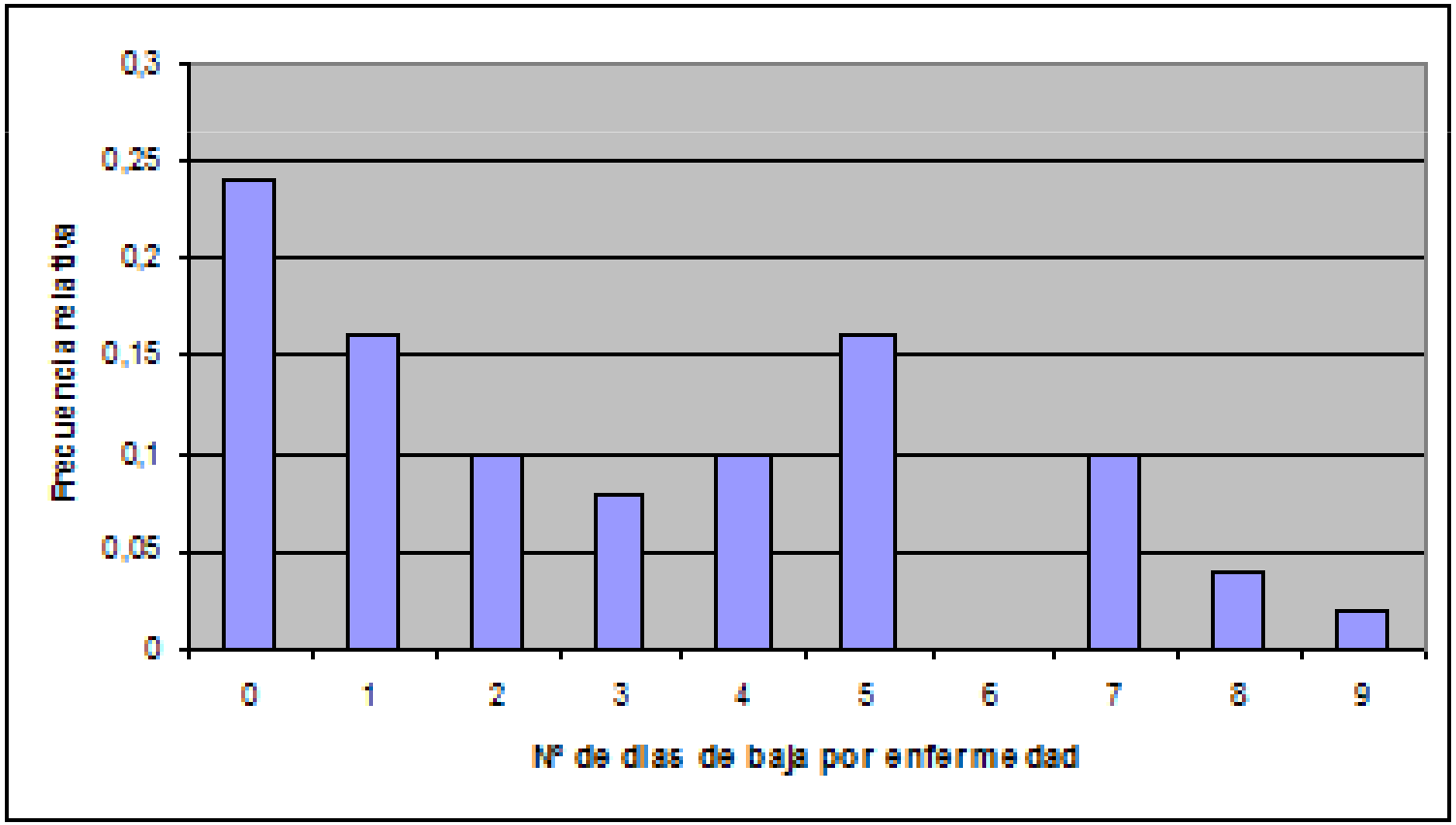
- EJEMPLO: diagrama de barras de frecuencia absoluta



2. VARIABLE CUANTITATIVA DISCRETA

Diagrama de barras

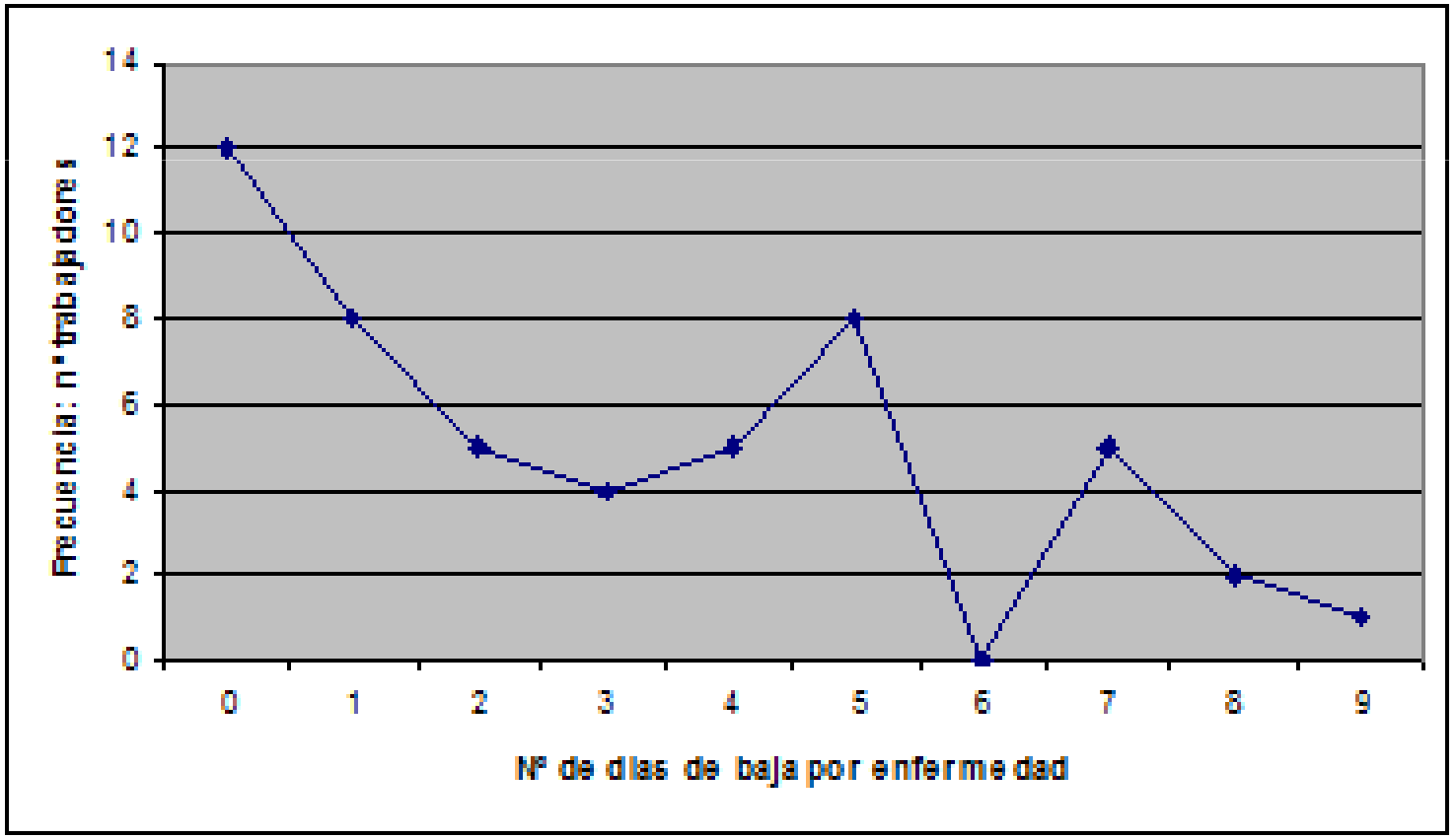
- EJEMPLO: diagrama de barras de frecuencia relativa



2. VARIABLE CUANTITATIVA DISCRETA

Polígono de frecuencias

- EJEMPLO: para la frecuencia absoluta



2. VARIABLE CUANTITATIVA DISCRETA

Diagrama de tallo y hojas

- forma eficiente de representar un conjunto de datos numéricos de tamaño pequeño o mediano
- se divide cada valor de datos en dos partes (el tallo y la hoja)
- método semigráfico

2. VARIABLE CUANTITATIVA DISCRETA

Diagrama de tallo y hojas

- **Procedimiento**
  - se redondean los datos a dos o tres cifras significativas (unidades adecuadas)
  - se disponen los datos en dos columnas separadas por una línea
    - datos de dos cifras: decena en la columna izquierda (tallo) y unidades en la columna derecha (hoja) alineadas con la decena
    - datos de tres cifras: decenas y centenas en la columna izquierda (tallo) y unidades en la columna derecha (hojas) alineadas

- ejemplo 1: número 75

Tallo		Hoja
7	|	5

- ejemplo 2: números 751 y 757

Tallo		Hojas
75	|	1 , 7

2. VARIABLE CUANTITATIVA DISCRETA

Diagrama de tallo y hojas

- **Ejemplo:** El siguiente gráfico de tallo y hojas muestra el peso en kilos de 50 deportistas masculinos de una competición de atletismo

Tallo	Hojas	Nº valores
5	2, 3, 5, 6, 7, 7, 8, 8, 8, 9, 9, 9	12
6	0, 0, 1, 1, 1, 2, 3, 3, 4, 4, 5, 6, 7, 7, 8, 9	16
7	0, 0, 1, 1, 2, 3, 3, 4, 5, 6, 6, 7, 7	13
8	1, 2, 2, 3	4
9	1, 3	2
10	2, 5	2
11	2	1

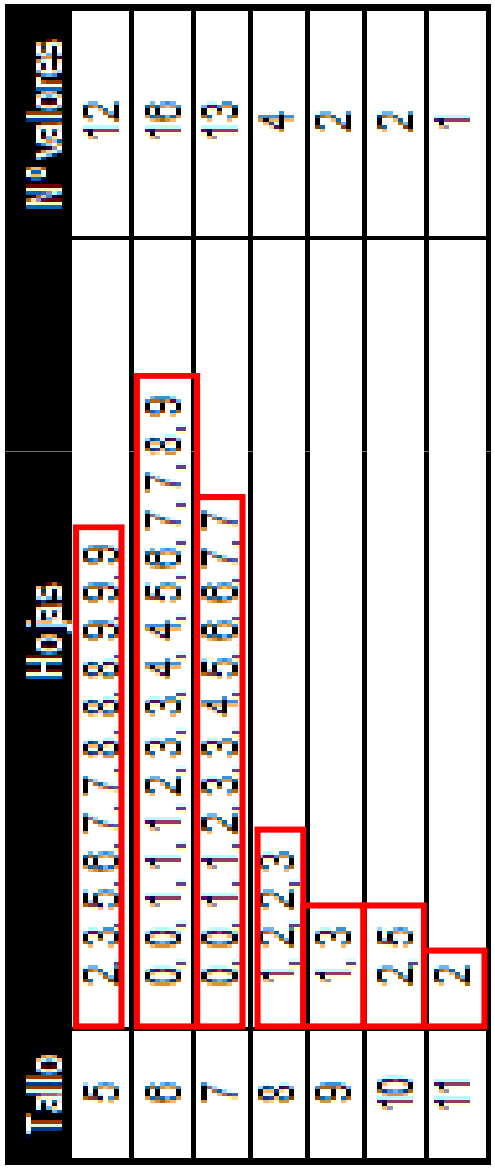
16 hojas en el tallo 6
16 individuos con pesos entre 60kg. y 69kg.

2. VARIABLE CUANTITATIVA DISCRETA

Diagrama de tallo y hojas

- **Ejemplo:**

Parece un histograma girado con el añadido de que presenta todos los valores existentes en cada clase



3. VARIABLE CUANTITATIVA CONTINUA

Histograma

- caso particular del diagrama de barras
- variables agrupadas en intervalos de clase: representación de frecuencias mediante áreas de rectángulos
- se obtiene levantando sobre cada intervalo de clase un rectángulo cuyo área sea igual a la frecuencia del mismo
- si los intervalos son correlativos los rectángulos aparecen pegados en la representación gráfica

3. VARIABLE CUANTITATIVA CONTINUA

Histograma

- **Procedimiento**
  - ordenar los datos en forma creciente
  - elegir los intervalos de clase de forma que todos aparezcan en uno de ellos
  - construir una tabla de frecuencias
  - dibujar las barras adyacentes con alturas iguales a las frecuencias obtenidas
  - si la amplitud de los intervalos no es igual para todos hay que hacer coincidir el área del rectángulo con la frecuencia del intervalo

3. VARIABLE CUANTITATIVA CONTINUA

Histograma

- **Ejemplo:** niveles de colesterol de 40 empleados de una empresa
  - observaciones

211	171	227	200	189	190	199	221	178	198
200	187	188	191	176	193	201	202	220	221
178	189	178	198	199	204	208	218	172	187
213	217	219	198	173	174	181	190	210	213

3. VARIABLE CUANTITATIVA CONTINUA

Histograma

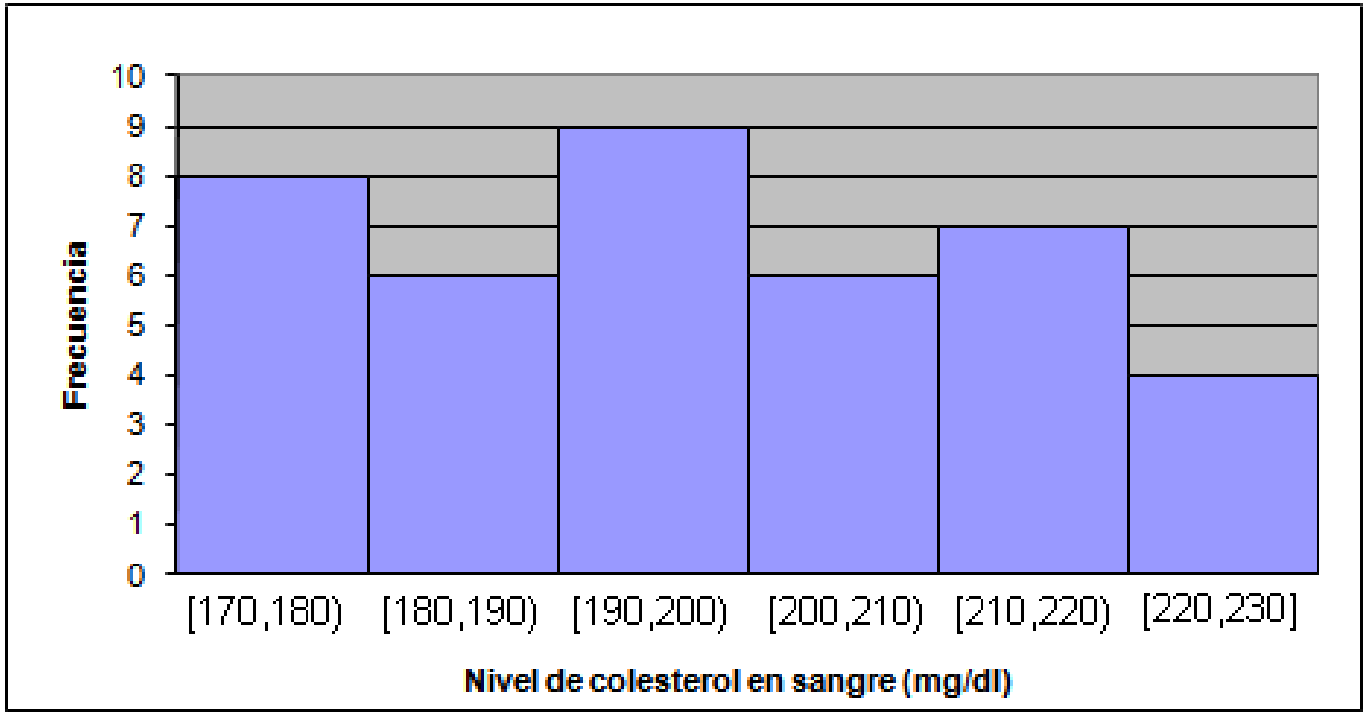
- **Ejemplo:** niveles de colesterol de 40 empleados de una empresa
  - tabla de frecuencias

Intervalos de clase	Frec. absoluta	Frec. relativa	Porcentaje (%)
[170, 180)	8	$8/40 = 0.200$	20,0
[180, 190)	6	$6/40 = 0.150$	15,0
[190, 200)	9	$9/40 = 0.225$	22,5
[200, 210)	6	$6/40 = 0.150$	15,0
[210, 220)	7	$7/40 = 0.175$	17,5
[220, 230]	4	$4/40 = 0.100$	10,0
Totales	**40**	**1**	**100**

3. VARIABLE CUANTITATIVA CONTINUA

Histograma

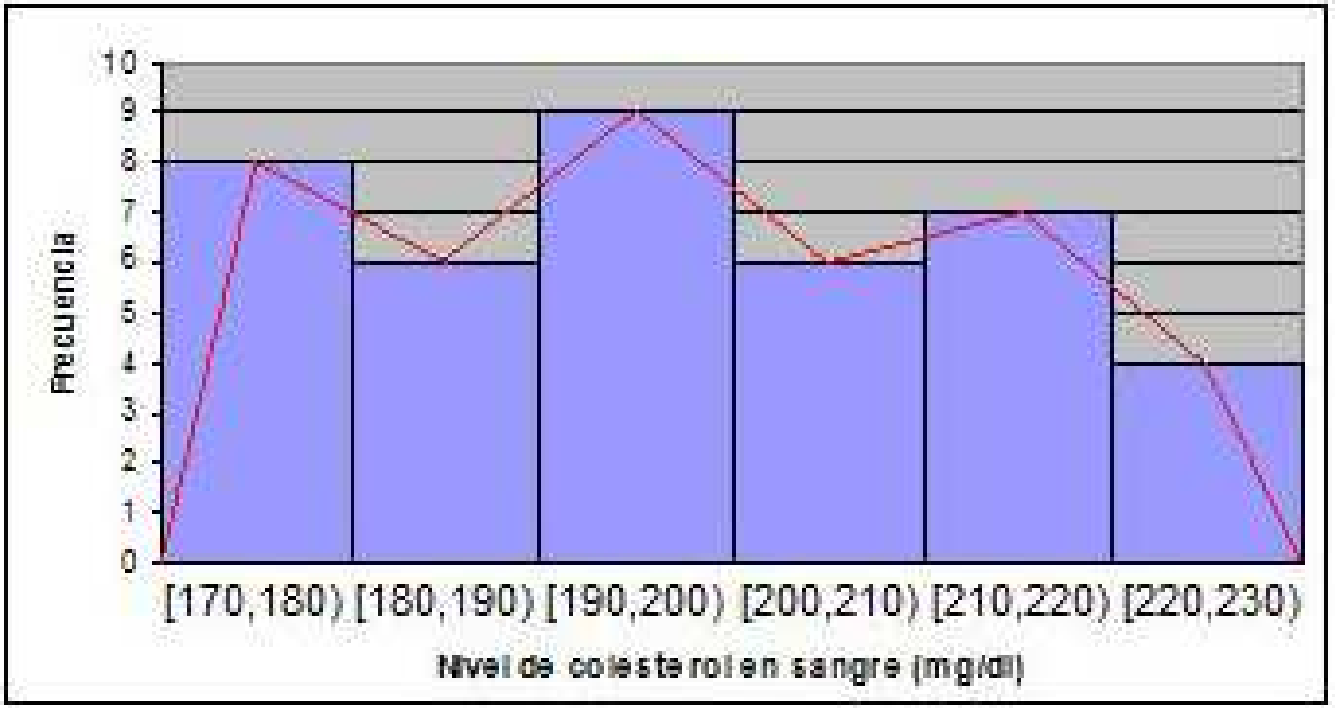
- **Ejemplo:** niveles de colesterol de 40 empleados de una empresa
  - histograma



3. VARIABLE CUANTITATIVA CONTINUA

Histograma

- **Ejemplo:** niveles de colesterol de 40 empleados de una empresa
  - polígono de frecuencias



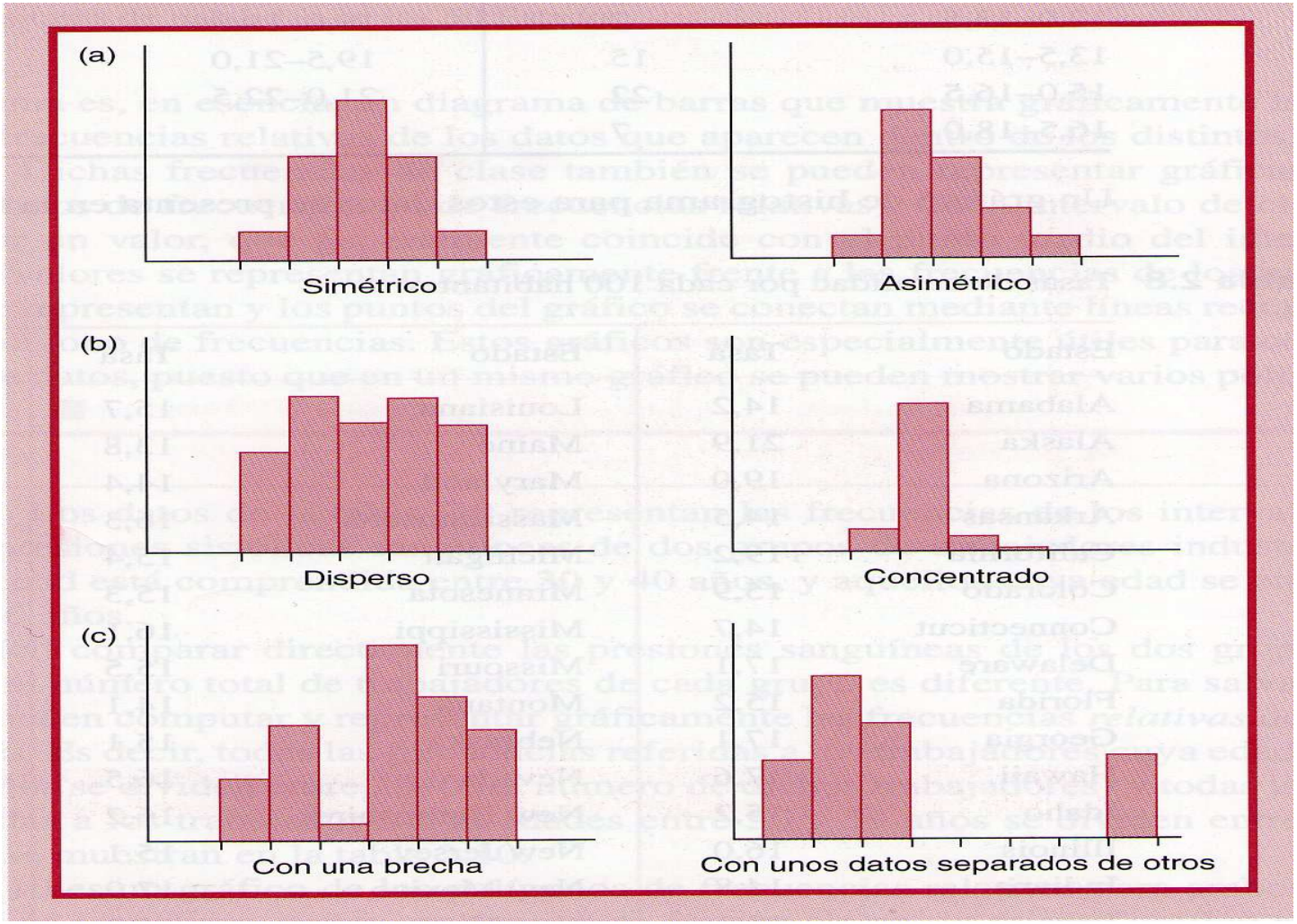
3. VARIABLE CUANTITATIVA CONTINUA

Histograma

- permite organizar y presentar los datos gráficamente con lo que puede prestarse atención a determinadas características importantes de dichos datos:
  - la simetría
  - la dispersión
  - la concentración en los diferentes intervalos
  - si existen brechas entre los datos
  - si algunos valores están muy separados de otros

3. VARIABLE CUANTITATIVA CONTINUA

Histograma



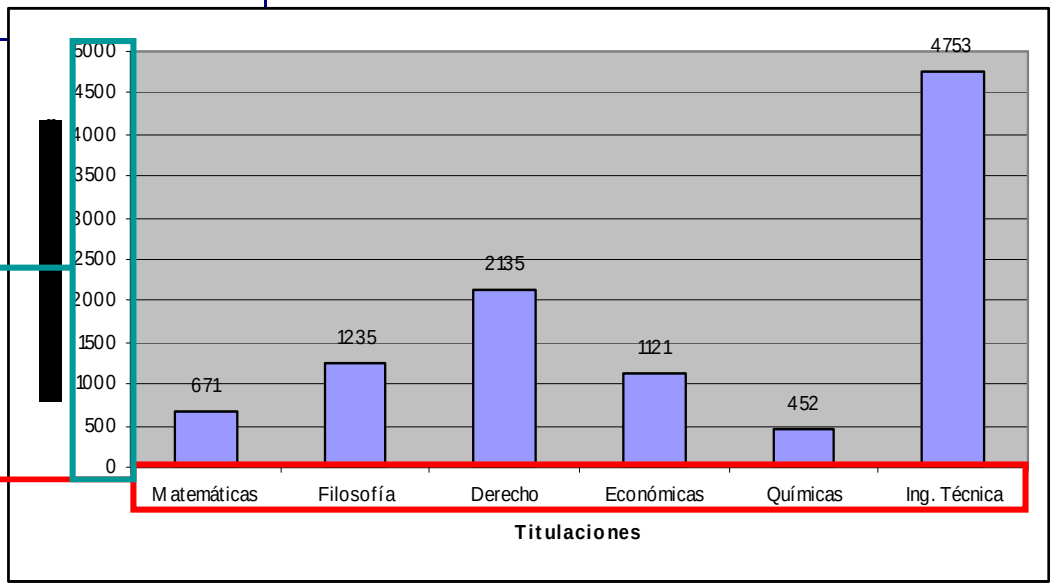
4. VARIABLE CUALITATIVA

Diagrama de barras

Titulación	Nº alumnos matriculados
Matemáticas	671
Filosofía	1235
Derecho	2135
Económicas	1121
Químicas	452
Ingeniería técnica	4753

Eje de ordenadas: frecuencias

Eje de abscisas: caracteres cualitativos



4. VARIABLE CUALITATIVA

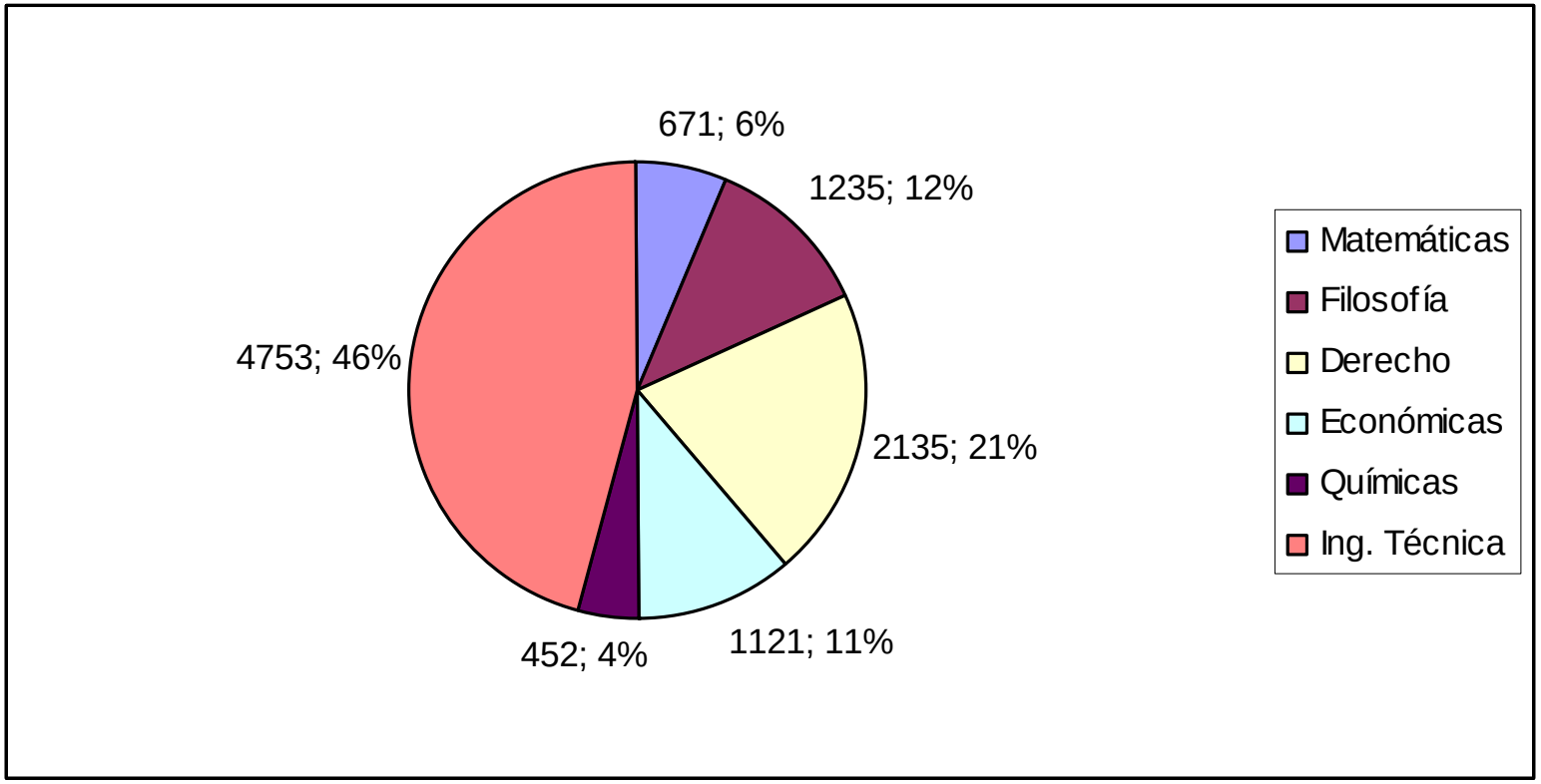
Diagrama de sectores

- también: gráficos de tarta o "*de quesitos*"
- se traza un círculo y se asigna un sector circular a cada uno de los caracteres cualitativos siendo la amplitud del sector proporcional a la frecuencia del carácter
- se consigue haciendo corresponder 360º a la suma de todas las frecuencias de los caracteres y hallando la correspondiente proporcionalidad

4. VARIABLE CUALITATIVA

Diagrama de sectores

- para el ejemplo anterior



SÍNTESIS DE CONJUNTOS DE DATOS
REPRESENTACIONES NUMÉRICAS

1. INTRODUCCIÓN

- se ha visto como describir y representar de forma gráfica los conjuntos de datos
- ahora se van a obtener medidas que permitan sintetizar los datos
- son magnitudes numéricas cuyos valores vienen determinados por dichos datos

2. DEFINICIÓN: PARÁMETROS

Parámetro: cantidad numérica calculada sobre una población

- ejemplo: altura media de los individuos de un país
- se pretende resumir toda la información de la población en unos pocos números: parámetros
- **Notación:** letras griegas
  - media poblacional: μ
  - varianza poblacional: σ^2
  - número de elementos de la población: *N*

2. DEFINICIÓN: ESTADÍSTICOS

Estadístico (ó estadístico muestral): cantidad numérica calculada sobre una muestra

- medida cuantitativa, derivada de un conjunto de datos de una muestra, cuyo objetivo es estimar o contrastar características de una población
- ejemplo: altura media de los presentes en este aula

¿es una muestra representativa de la población (país, en este caso)?

- **Notación:** letras latinas
  - media muestral: $\overline{x}$
  - varianza muestral: s^2
  - número de elementos de la muestra: *n*

2. PARÁMETROS Y ESTADÍSTICOS

Estimador: estadístico que se usa para aproximar un parámetro

- generalmente, interesa conocer un parámetro pero, por la dificultad de estudiar toda la población, se calcula un estimador en una muestra y se "confía" en que sean próximos
- se verá como elegir muestras para que el error sea pequeño con una determinada confianza

3. TIPOS DE ESTADÍSTICOS

Estadísticos de posición

- dan información sobre la *posición relativa* de una observación dentro del conjunto de datos
- para su cálculo es necesario que los datos se ordenen de menor a mayor
- **Tipos:**
  - **Centrales** (de tendencia central ó de centralización), indican valores (*valor central*) respecto a los cuales los datos parecen agruparse: media, mediana y moda
  - **No centrales** para conocer otros puntos característicos de la distribución que no son los valores centrales; los cuantiles son aquellos valores de la variable que dividen a la distribución en partes que contienen el mismo número de individuos

3. TIPOS DE ESTADÍSTICOS

Estadísticos de dispersión

- señalan la dispersión en conjunto de todos los datos de la distribución respecto de la medida o medidas de posición adoptadas
- **Tipos:**
  - **Dispersión absoluta** : recorrido, varianza, desviación típica, recorrido intercuartílico
  - **Dispersión relativa** : coeficiente de variación de Pearson
  - Diagrama de caja

3. TIPOS DE ESTADÍSTICOS

Estadísticos de forma

- Estudian la simetría y la deformación respecto de una distribución modelo denominada distribución normal

- **Tipos:**
  - simetría (sesgo): coeficiente de asimetría deformaciones respecto a la moda y a la mediana
  - curtosis (apuntamiento): coeficiente de curtosis

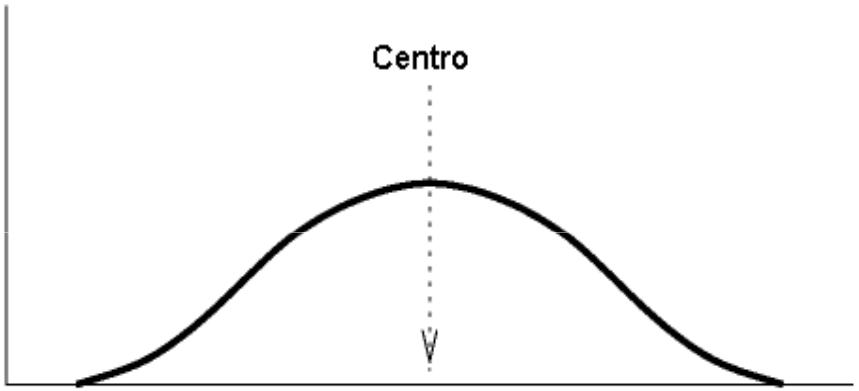
3. TIPOS DE ESTADÍSTICOS

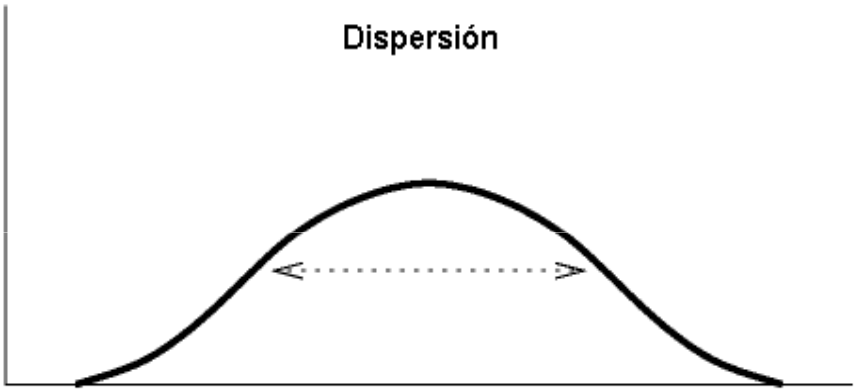
Estadísticos de concentración

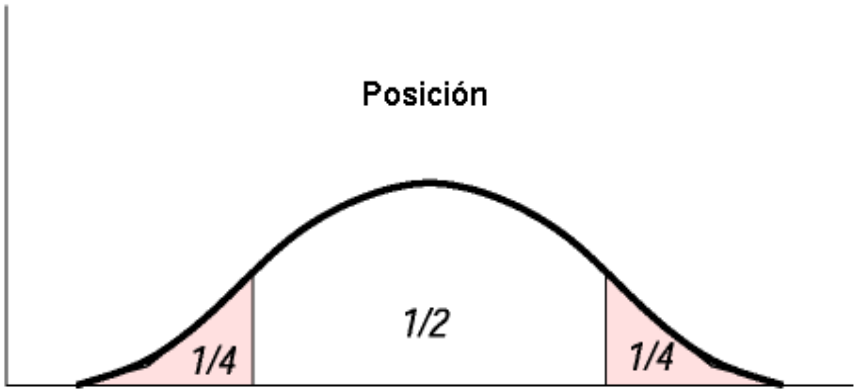
- concentración de una distribución frente a la uniformidad
- **Tipos:**
  - curva de Lorenz
  - índice de Gini

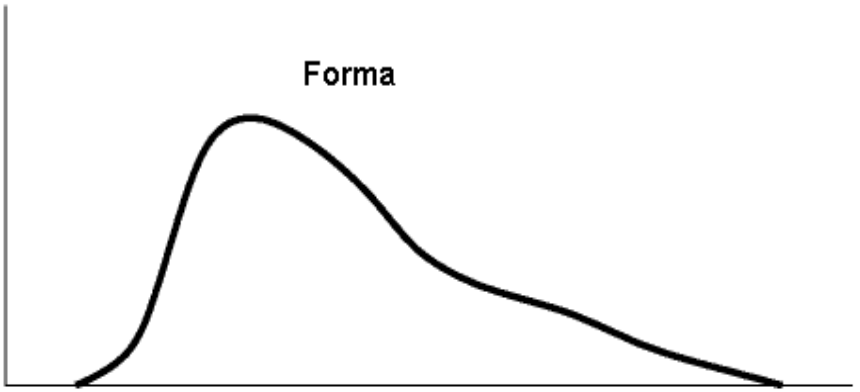
3. TIPOS DE ESTADÍSTICOS

Ejemplos gráficos









3. TIPOS DE ESTADÍSTICOS

Notación

- en las siguientes definiciones se van a denotar las *n* observaciones del conjunto de datos estudiado como sigue:

$$x_1, x_2, \cdots, x_n$$

4. MEDIDAS DE CENTRALIZACIÓN

Media muestral: estadístico que se usa para indicar el centro de un conjunto de datos

- definida como la **media aritmética** de los valores de los datos (notación, $\bar{x}$):

$$\bar{x} = \frac{\sum_{i=1}^{i=n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- ejemplo: media de 7, 11, 11, 8, 12, 7, 6, 6

$$\bar{x} = \frac{\sum_{i=1}^{i=8} x_i}{n} = \frac{7+11+11+8+12+7+6+6}{8} = \frac{68}{8} = 8{,}5$$

4. MEDIDAS DE CENTRALIZACIÓN

Media muestral: cálculo

- si la muestra de n datos se organiza en una tabla de frecuencias, los valores x_i se suceden con sus frecuencias n_i, entonces:

$$\bar{x} = \frac{\sum_{i=1}^{i=n} n_i \cdot x_i}{n}$$

- ejemplo:

X: peso	Nº mujeres	
Valores: x_i	F. absoluta: n_i	$x_i \cdot n_i$
51	2	102
53	3	159
60	4	240
64	1	64
Total	10	565

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{i=4} n_i \cdot x_i = \frac{565}{10} = 56{,}5\ kg.$$

4. MEDIDAS DE CENTRALIZACIÓN

Media muestral: cálculo

- si la variable esta agrupada en *k* intervalos de clase se asignan las frecuencias a las marcas de clase y se procede como si la variable fuera discreta :

$$\bar{x} = \frac{\sum_{i=1}^{i=k} n_i \cdot x_i}{\sum_{i=1}^{i=k} n_i} = \frac{1}{n} \sum_{i=1}^{i=k} n_i \cdot x_i = \sum_{i=1}^{i=k} f_i \cdot x_i$$

- ejemplo:

Intervalos de clase	Marcas clase: $c_i=x_i$	F. absoluta: n_i	$x_i \cdot n_i$
$[30, 40)$	35	4	140
$[40, 50)$	45	3	135
$[50, 60)$	55	3	165
Totales		10	440

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{i=3} n_i \cdot x_i = \frac{440}{10} = 44 \text{ años}$$

4. MEDIDAS DE CENTRALIZACIÓN

Media muestral: debilidad

- debilidad como indicador del centro de un conjunto de datos ya que resulta muy afectada por los valores extremos

- ejemplo: $X = \{1, 6, 211, 2, 8, 12, 7, 9, 10, 4\}$

$$\bar{x} = \frac{270}{10} = 27$$

Un solo valor de la muestra es mayor que la media

4. MEDIDAS DE CENTRALIZACIÓN

Mediana: valor que divide el conjunto de datos en dos partes iguales; es decir, es el *valor medio* cuando los datos están ordenados de menor a mayor

- la mitad de los valores son menores que la mediana y la otra mitad son mayores que la mediana
- notación: M_e ó $\tilde{x}$

$$\tilde{x} = M_e = \begin{cases} \text{tér min o}\left[(n+1)/2\right] & \text{cuando n es impar} \\ \dfrac{\text{tér min o}\,(n/2) + \text{tér min o}\left[(n/2)+1\right]}{2} & \text{cuando n es par} \end{cases}$$

4. MEDIDAS DE CENTRALIZACIÓN

Mediana: valor que divide el conjunto de datos en dos partes iguales; es decir, es el *valor medio* cuando los datos están ordenados de menor a mayor

- ejemplo: $X = \{1, 6, 8, 12, 17\} \Rightarrow M_e = 8 \quad \left(n = 5 \neq \dot{2} \right)$

- ejemplo: $X = \{1, 6, 8, 12, 17, 21\} \Rightarrow M_e = \frac{8+12}{2} = 10 \quad \left(n = 6 = \dot{2} \right)$

- indica **orden** dentro de la muestra

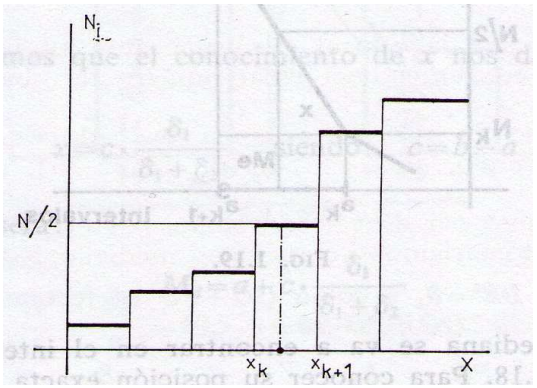
- sólo usa un único valor central ó un par de valores centrales: no le afectan los valores extremos

4. MEDIDAS DE CENTRALIZACIÓN

Mediana: cálculo con **tabla de frecuencias**

- se divide el número de observaciones entre 2: $\frac{n}{2}$

  - $\frac{n}{2} = N_k$: $M_e = \frac{x_k + x_{k+1}}{2}$



Valores: x_i	F. absoluta: n_i	F. abs. acumulada: N_i
x_1=1	2	2
x_2=2	3	5
x_3=5	4	9
x_4=7	1	10
x_5=9	8	18
Total	n=18	

$$N_3 = 9 = \frac{n}{2}$$

$$M_e = \frac{x_3 + x_4}{2} = \frac{5+7}{2} = 6$$

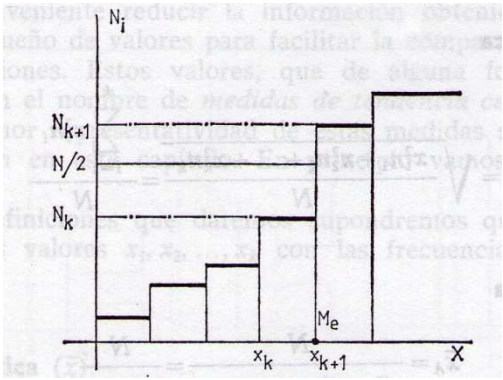
$$\bar{x} = 5{,}944$$

4. MEDIDAS DE CENTRALIZACIÓN

Mediana: cálculo con **tabla de frecuencias**

- se divide el número de observaciones entre 2: $\frac{n}{2}$

  - $N_k < \frac{n}{2} < N_{k+1}$: $M_e = x_{k+1}$



Valores: x_i	F. absoluta: n_i	F. abs. acumulada: N_i
x_1=1	2	2
x_2=2	3	5
x_3=5	4	9
x_4=7	1	10
x_5=9	1	11
x_6=11	4	15
Total	n=15	

$$N_2 = 5 < \frac{n}{2} = 7{,}5 < 9 = N_3$$

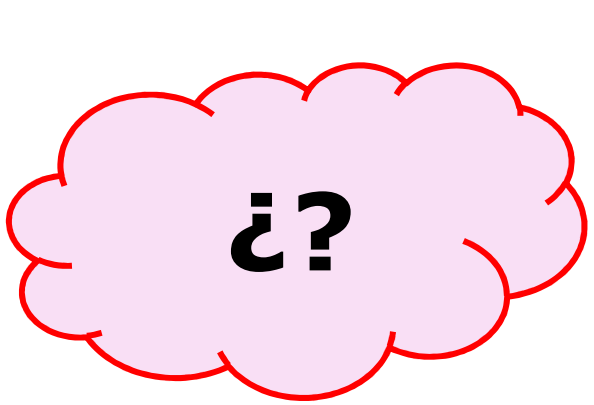
$$M_e = x_3 = 5$$

$$\bar{x} = 5{,}867$$

4. MEDIDAS DE CENTRALIZACIÓN

Mediana: cálculo con intervalos de clase

- fundamentalmente se habla de clase mediana que es aquélla cuya frecuencia acumulada sobrepasa o es igual n/2
- la mediana para datos agrupados se calcula hallando el valor de la variable que divide al histograma en dos partes iguales



4. MEDIDAS DE CENTRALIZACIÓN

Moda: valor de la variable que más veces se repite, es decir, el valor que tiene mayor frecuencia absoluta en un conjunto de datos

- en un conjunto de datos puede que:
  - no exista moda (todas las frecuencias son iguales)
  - haya más de una moda (multimodal)
- notación: Mo
- con datos cualitativos no se puede calcular la media ni la mediana pero sí la moda
- con datos agrupados en intervalos de clase se habla de clase modal: aquélla(s) con mayor frecuencia

4. MEDIDAS DE CENTRALIZACIÓN

¿Cuál es la más apropiada?

- depende del tipo de datos
- la moda puede calcularse para datos cualitativos y cuantitativos
- si el número de observaciones es muy pequeño la media se ve muy afectada por los valores extremos, sin embargo, la mediana no
- si el número de observaciones es grande la mediana y la media tienden a ser iguales
- la media es más fácil de calcular que la mediana

5. MEDIDAS DE DISPERSIÓN

Introducción

- Las medidas de centralización proporcionan un valor para representar cierta información de los datos
  - hay información importante que no se tiene en cuenta
  - se necesita conocer lo "dispersos" que están los datos

- ejemplo

xi	ni
2	1
4	2
6	3
8	2
10	1

yi	ni
5,5	2
6	5
6,5	2

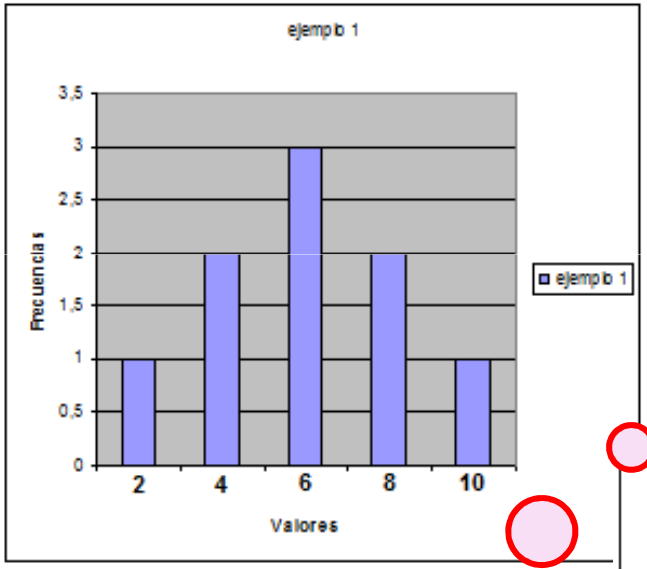
5. MEDIDAS DE DISPERSIÓN

Introducción

- ejemplo

xi	ni
2	1
4	2
6	3
8	2
10	1

yi	ni
5,5	2
6	5
6,5	2

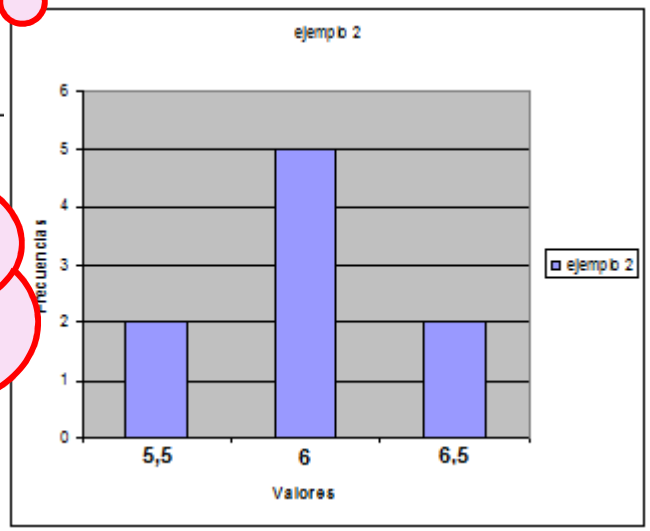


$$\bar{x} = \bar{y} = 6$$

$$\tilde{x} = \tilde{y} = 6$$

$$(M_O)_x = (M_O)_y = 6$$

Conjuntos de datos diferentes aunque las medidas centrales coinciden



5. MEDIDAS DE DISPERSIÓN

Introducción: concepto de dispersión

- concepto de dispersión o variabilidad
- **ejemplo**: los estudiantes de una asignatura reciben diferentes calificaciones; es decir, hay variabilidad en las notas
  - causa principal: diferencias individuales en el conocimiento de la materia
  - aunque todos los alumnos tengan exactamente los mismos conocimientos lo más probable es que las calificaciones no sean iguales para todos por diferentes motivos
    - ◇ diferencias individuales en la habilidad ante un examen
    - ◇ variabilidad por error de medida: un examen escrito no es una herramienta perfecta para medir el conocimiento
    - ◇ aleatoriedad o variabilidad por azar

5. MEDIDAS DE DISPERSIÓN

Recorrido: diferencia entre la mayor observación y la menor observación del conjunto de datos

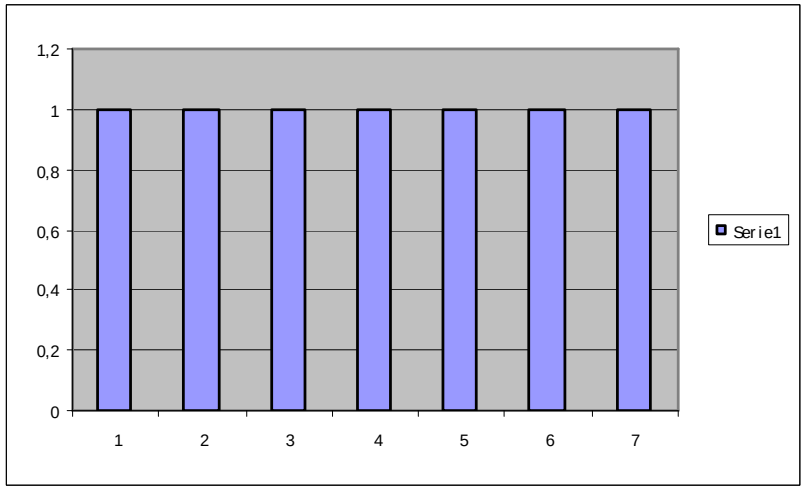
- es la medida de dispersión más simple
  - da una idea sencilla sobre la dispersión de los datos
  - se basa en los datos extremos y no informa sobre como se distribuyen

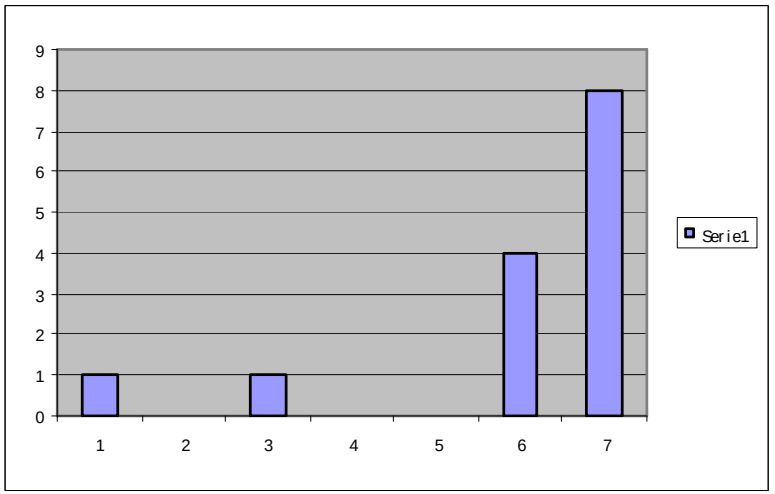
$$Re = max\left(x_k\right) - min\left(x_k\right) = H - L$$

5. MEDIDAS DE DISPERSIÓN

Recorrido: diferencia entre la mayor observación y La menor observación del conjunto de datos

- se usa en control de calidad porque se trabaja con muestras de tamaño pequeño
- muy sensible a los datos extremos (*outliers*) por lo que no es la mejor medida de dispersión





5. MEDIDAS DE DISPERSIÓN

Varianza: estadístico que indica el mayor o menor grado de dispersión de los valores de la muestra respecto de su media aritmética

- media de la suma de las desviaciones a la media al cuadrado (notación, s^2):

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

- para datos agrupados en intervalos de clase:

$$s^2 = \frac{\sum_{i=1}^{i=r} (x_i - \bar{x})^2 \cdot n_i}{n}$$

- ***r***: número de clases
- ***xi***: marca de clase
- ***ni***: frecuencia de la clase

5. MEDIDAS DE DISPERSIÓN

Varianza

$$s^2 = \frac{\sum_{i=1}^{i=r} (x_i - \bar{x})^2 \cdot n_i}{n} \quad \Rightarrow \quad s^2 = \sum_{i=1}^{i=r} (x_i - \bar{x})^2 \cdot f_i$$

$$s^2 = \sum_{i=1}^{i=r} x_i^{\,2} \cdot f_i - 2 \cdot \bar{x} \underbrace{\sum_{i=1}^{i=r} x_i \cdot f_i}_{\bar{x}} + \bar{x}^2 \underbrace{\sum_{i=1}^{i=r} f_i}_{1}$$

$$s^2 = \sum_{i=1}^{i=r} x_i^{\,2} \cdot f_i - 2 \cdot \bar{x}^2 + \bar{x}^2 = \sum_{i=1}^{i=r} x_i^{\,2} \cdot f_i - \bar{x}^2$$

$$s^2 = \frac{\sum_{i=1}^{i=r} x_i^{\,2} \cdot n_i}{n} - \bar{x}^2$$

5. MEDIDAS DE DISPERSIÓN

Varianza: inconveniente

- sus unidades no coinciden con las unidades de la variable
- sus unidades son el cuadrado de las unidades de la variable
  - si las unidades son minutos → ¿minutos al cuadrado?
  - si se miden alturas → metros al cuadrado
- para obtener una idea aproximada, nunca exacta, de la dispersión hay que obtener la raíz cuadrada que es la desviación típica

5. MEDIDAS DE DISPERSIÓN

Desviación típica: raíz cuadrada de la varianza

- notación: s

$$s = \sqrt{\frac{\sum_{i=1}^{i=n} (x_i - \bar{x})^2}{n}}$$

- para datos agrupados en intervalos de clase:

$$s = \sqrt{\frac{\sum_{i=1}^{i=r} (x_i - \bar{x})^2 \cdot n_i}{n}}$$

- r: número de clases
- x_i: marca de clase
- n_i: frecuencia de la clase

5. MEDIDAS DE DISPERSIÓN

Ejemplo

- tabla de frecuencias con los pesos, medidos en kilos, de 10 recién nacidos en una maternidad

X: peso	Nº nacidos		
Valores: x_i	F. absoluta: n_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
2,550	1	2,550	6,503
2,890	1	2,890	8,352
3,050	2	6,100	18,605
3,250	3	9,750	31,688
3,500	2	7,500	24,500
4,100	1	4,100	16,810
Total	10	32,390	106,458

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{i=6} n_i \cdot x_i = \frac{32,39}{10} = 3,239\ kg.$$

$$s_x^2 = 0,1547 kg^2.$$

$$s_x = \sqrt{s_x^2} = 0,3933\ kg.$$

6. CAMBIOS DE VARIABLE

Medidas de centralización y dispersión

- a veces, es necesario realizar un cambio de escala dado que los valores considerados son grandes
- en estos casos se realiza un cambio de variable de tipo lineal:

$$y = a + b \cdot x$$

- los estadísticos estudiados se ven afectados de la siguiente manera:
  - media: $\bar{y} = a + b \cdot \bar{x}$
  - varianza: $s_y^2 = b^2 \cdot s_x^2$
  - desviación típica: $s_y = |b| \cdot s_x$

7. MEDIDAS DE POSICIÓN NO CENTRAL

Introducción

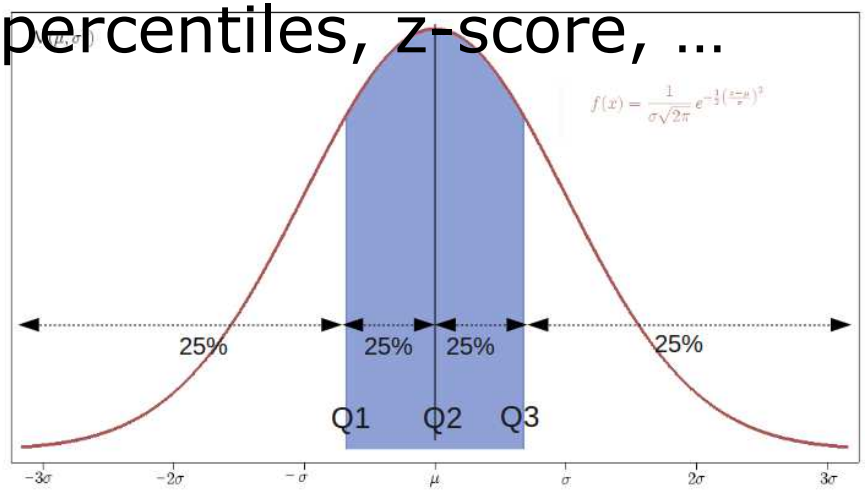
- Constituyen una medida única y se calcula cuando los datos son numéricos para situar la distribución y dar una idea de su posición

- Indican
  - la posición de un dato en un conjunto **ordenado** de datos ó
  - la posición relativa respecto a la media

7. MEDIDAS DE POSICIÓN NO CENTRAL

<u>Cuantil de orden *p*</u>: valor de la variable por debajo del cual se encuentra una frecuencia relativa acumulada *p*.

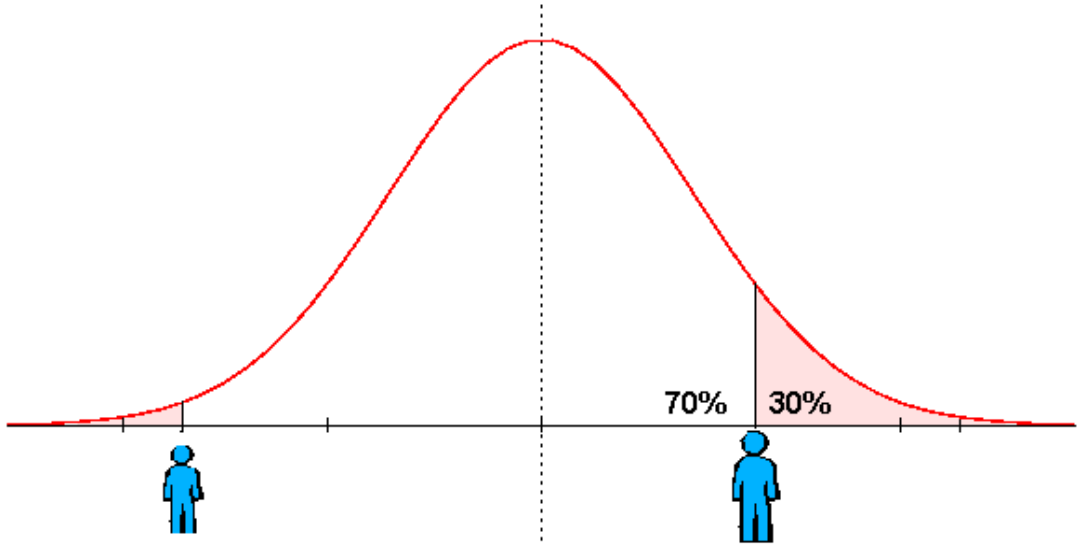
- medidas de posición que dividen a la distribución en un cierto número de partes de manera que en cada una de ellas hay el mismo porcentaje, *p*, de valores de la variable
- <u>tipos</u>: cuartiles, deciles, percentiles, z-score, …
- <u>notación</u>: $C_p \quad (0 < p < 1)$



7. MEDIDAS DE POSICIÓN NO CENTRAL

Cuantil de orden *p*: valor de la variable por debajo del cual se encuentra una frecuencia relativa acumulada *p*.

- informan del valor de la variable que ocupa la posición (en tanto por ciento) que nos interese respecto del conjunto de variables



7. MEDIDAS DE POSICIÓN NO CENTRAL

Cuartiles: tres valores de la variable que dividen el conjunto de datos en cuatro partes iguales (tres divisiones)

- notación: $Q_1, Q_2 = M_e, Q_3$

- **primer cuartil** (ó cuartil inferior): mediana de la primera mitad de los valores; es decir, valor de la variable que, aproximadamente, deja el 25% de las observaciones menores ó iguales que él

  - notación: Q_1
  - cálculo: como el de la mediana pero considerando $\frac{n}{4}$

7. MEDIDAS DE POSICIÓN NO CENTRAL

► **segundo cuartil** (ó cuartil medio): se trata de la *mediana* ya que es el valor de la variable que deja la mitad (50%) de las observaciones menores ó iguales que él

- notación: $Q_2 = M_e$

► **tercer cuartil** (ó cuartil superior): es la mediana de la segunda mitad de las observaciones; es decir, el valor de la variable que, aproximadamente, deja el 75% de las observaciones menores ó iguales que él

- notación: Q_3
- cálculo: como el de la mediana pero considerando $\frac{3n}{4}$

7. MEDIDAS DE POSICIÓN NO CENTRAL

Ejemplo: cálculo de cuartiles

Valores: x_i	F. absoluta: n_i	F. abs. acumulada: N_i
$x_1=1$	3	3
$x_2=5$	2	5
$x_3=6$	5	10
$x_4=8$	3	13
$x_5=10$	3	16
$x_6=11$	8	24
$x_7=12$	11	35
$x_8=14$	7	42
Total	n=42	

7. MEDIDAS DE POSICIÓN NO CENTRAL

Ejemplo: cálculo de cuartiles

- primer cuartil:

$$N_3 = 10 < \frac{n}{4} = \frac{42}{4} = 10{,}5 < N_4 = 13 \Rightarrow \boxed{Q_1 = 8}$$

x_2=5	2	5
x_3=8	5	10
x_4=8	3	13
x_5=10	3	16

7. MEDIDAS DE POSICIÓN NO CENTRAL

Ejemplo: cálculo de cuartiles

- tercer cuartil:

$$N_6 = 24 < \frac{3n}{4} = 31{,}5 < N_7 = 35 \Rightarrow \quad Q_3 = 12$$

x_5=10	3	16
x_6=11	8	24
x_7=12	11	35
x_8=14	7	42

7. MEDIDAS DE POSICIÓN NO CENTRAL

Percentil (ó centil *k*–ésimo): valor de la variable que deja inferiores o iguales a él, aproximadamente, el *k*% de los datos

- notación: $P_k \quad (k = 1, 2, 3, \cdots, 99)$
- cálculo: análogo al de los cuartiles y mediana
- ejemplo anterior: P_{90}

$$N_7 = 35 < \frac{90n}{100} = 37{,}8 < N_8 = 42 \Rightarrow \quad P_{90} = x_8 = 14$$

x_7=12	11	35
x_8=14	7	42

7. MEDIDAS DE POSICIÓN NO CENTRAL

***z*−score (unidades tipificadas):** indica el número de desviaciones típicas en que un valor dado, x_i, se sitúa por encima o por debajo de la media de su muestra ó población

- cálculo: $z = \frac{x_i - \bar{x}}{s}$ $z = \frac{x_i - \mu}{\sigma}$
- describe la posición de un dato respecto a la media medida en términos de la desviación típica
  - z<0: la observación está a la izquierda de la media
  - z>0: la observación está a la derecha de la media
- se puede usar, también, para comparar valores de diferentes muestras o poblaciones

8. DATOS ATÍPICOS

Datos atípicos (*outliers*): observaciones que son numéricamente distantes del resto de los datos y, por tanto, son heterogéneas con los demás datos

- son observaciones extremadamente pequeñas ó grandes en comparación con los otros datos
- pueden ser debidas a:
  - error de medida, por ejemplo, la observación procede de una población diferente
  - error en la introducción o la transcripción de los datos
  - observación correcta pero corresponde a un suceso poco común (atípico)

8. DATOS ATÍPICOS

Detección

- *z-score*: $z = \dfrac{x_i - \bar{x}}{s}$

  - dato atípico es aquel cuyo $z > |3|$

  - de otra forma, son atípicos aquellos x_i tales que:

$$x_i \notin \left[\bar{x} - 3 \cdot s \, , \, \bar{x} + 3 \cdot s \right]$$

  - método adecuado cuando el histograma tiene forma de campana

8. DATOS ATÍPICOS

Detección

- mediante los percentiles:
  - se obtienen los percentiles 25 y 75 (cuartiles 1 y 3):

$$P_{25} = Q_1 \, , P_{75} = Q_3$$

  - se calcula el **rango intercuartílico** (mide la dispersión de la mitad central del conjunto de datos), ***IQR=RIC***:

$$IQR = Q_3 - Q_1 = P_{75} - P_{25}$$

8. DATOS ATÍPICOS

Detección

- mediante los percentiles:

  - **datos atípicos**:

$$\begin{cases} x_i < P_{25} - 1.5 \cdot IQR \\ x_i > P_{75} + 1.5 \cdot IQR \end{cases}$$

  - **datos atípicos extremos**:

$$\begin{cases} x_i < P_{25} - 3 \cdot IQR \\ x_i > P_{75} + 3 \cdot IQR \end{cases}$$

8. DATOS ATÍPICOS

Detección: ejemplo

- los datos siguientes corresponden al tiempo necesario para procesar 25 trabajos en una CPU

1.17	1.61	1.16	1.38	3.53
1.23	3.76	1.94	0.96	4.75
0.15	2.41	0.71	0.02	1.59
0.19	0.82	0.47	2.16	2.01
0.92	0.75	2.59	3.07	1.40

  - tabla de frecuencias con datos en intervalos de clase
  - histograma
  - detección de datos atípicos

8. DATOS ATÍPICOS

Detección: ejemplo

- intervalos de clase
  - número: $\sqrt{25} = 5$
  - valor mínimo: $x_{min} = 0.02$
  - valor máximo: $x_{max} = 4.75$

  $\Rightarrow$ recorrido: 4.73
  - rango del histograma: 4.80
  - amplitud de los intervalos: $\frac{4.80}{5} = 0.96$

8. DATOS ATÍPICOS

Detección: ejemplo

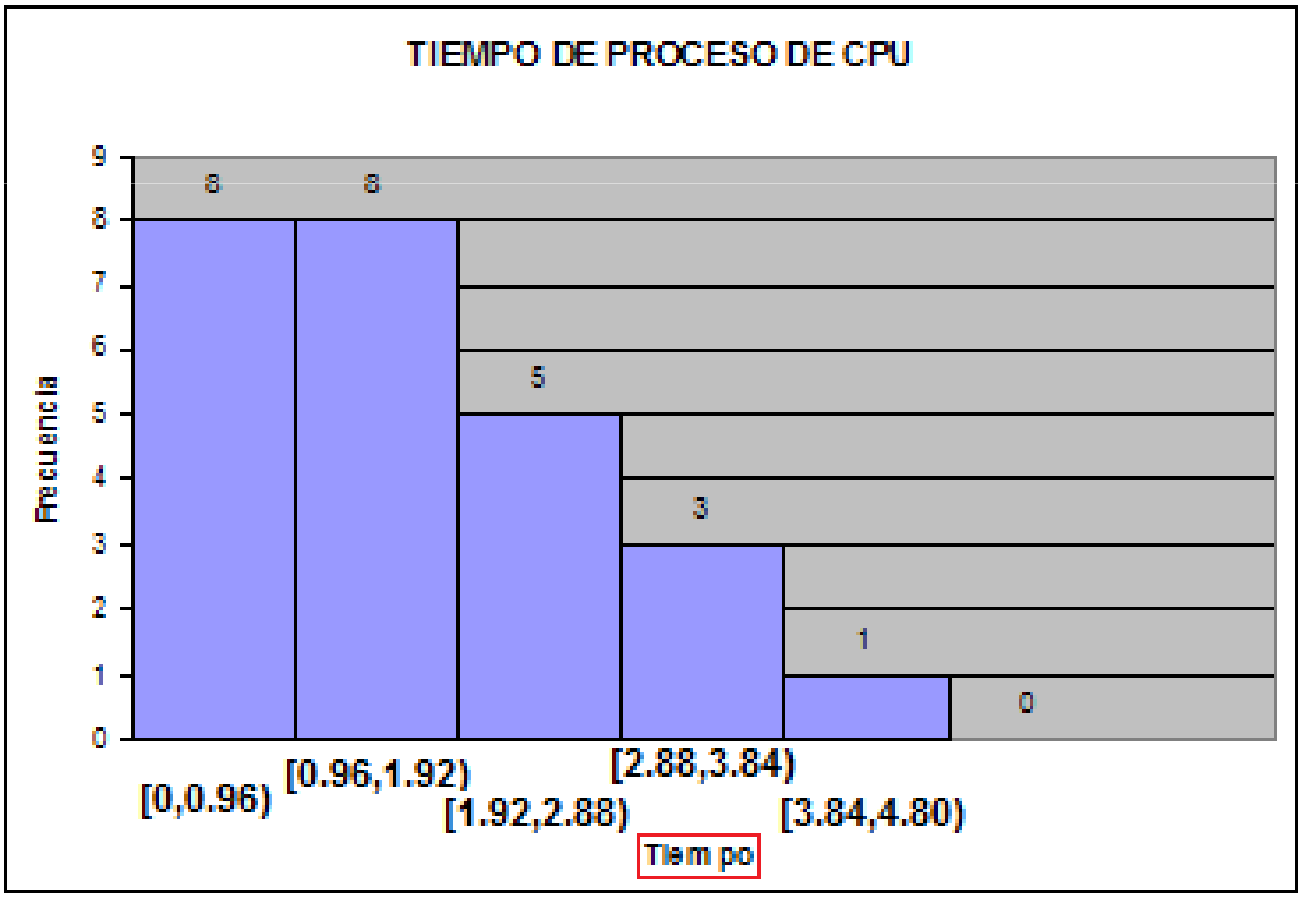
- tabla de frecuencias y cálculo de estadísticos

Datos	z-score	Clases	Extremo derecho	Marcas de clase:xi	ni	Ni
0,02	-1,51	[0,0.96)	0,96	0,48	8	8
0,15	-1,39	[0.96,1.92)	1,92	1,44	8	16
0,19	-1,35	[1.92,2.88)	2,88	2,40	5	21
0,47	-1,10	[2.88,3.84)	3,84	3,36	3	24
0,71	-0,88	[3.84,4.80)	4,80	4,32	1	25
0,75	-0,84	Amplitud	0,96		0	
0,82	-0,78					
0,92	-0,68			nº observaciones	25	
0,96	-0,65			N	L	H
1,16	-0,47			25	4,75	0,02
1,17	-0,46					
1,23	-0,40			Mínimo		0,00
1,38	-0,26			Máximo		4,80
1,40	-0,25			Nº Intervalos		5
1,59	-0,07			Amplitud intervalos		0,96
1,61	-0,06					
1,94	0,25	MEDIA	MEDIANA	MODA	DESV. TÍPICA	VARIANZA
2,01	0,31	1,670318	1,3800	#N/A	1,095613083	1,20036803
2,16	0,45					
2,41	0,68	CUARTIL 1	CUARTIL 2	CUARTIL 3	IQR	
2,59	0,84	0,8200	1,3800	2,1600	1,3400	
3,07	1,28					
3,53	1,70	Cota interior 1	-1,19	Datos atípicos	0	
3,76	1,91	Cota interior 2	4,17			
4,75	2,81	Cota exterior 1	-3,2	D. atípicos extremos	0	
		Cota exterior 2	6,18			

8. DATOS ATÍPICOS

Detección: ejemplo

- histograma



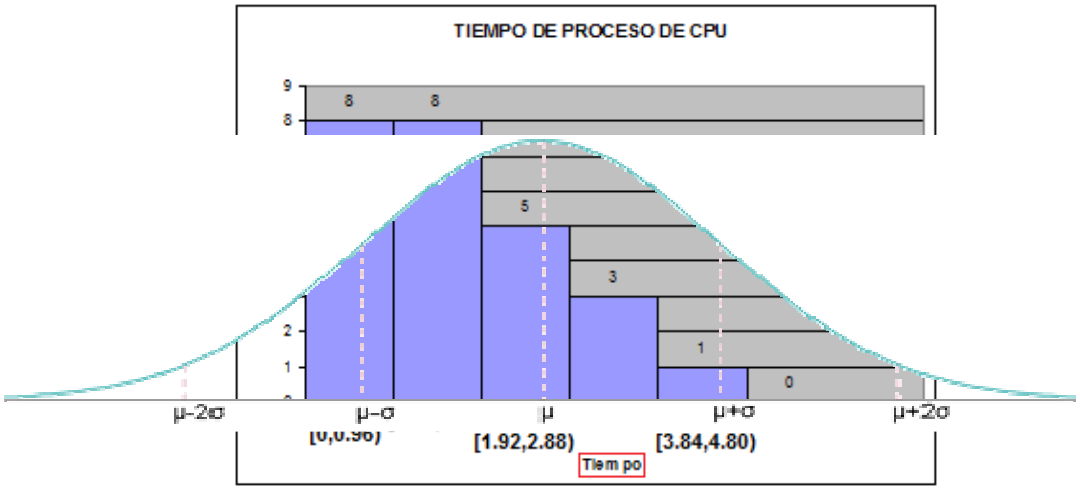
8. DATOS ATÍPICOS

Detección: ejemplo

- datos atípicos: *z-score*

$$\forall\, x_i :\ -3 < z_i < 3$$

  - por tanto, no hay datos atípicos en la muestra
  - sin embargo, como el histograma no tiene forma de campana este método no es el más adecuado para detectar datos atípicos



8. DATOS ATÍPICOS

Detección: ejemplo

- datos atípicos: percentiles
  - $IQR = Q_3 - Q_1 = P_{75} - P_{25} = 2.16 - 0.82 = 1.34$
  - $$\begin{cases} P_{25} - 1.5 \cdot IQR = 0.82 - 1.5 \times 1.34 = -1.19 \\ x_i > P_{75} + 1.5 \cdot IQR = 2.16 + 1.5 \times 1.34 = 4.17 \end{cases}$$
  - dato atípico: $4.75 \notin [-1.19, 4.17]$
  - no hay datos atípicos extremos: $x_i \in [-3.2, 6.18] \quad \forall i$

8. DATOS ATÍPICOS

Detección: diagrama de caja (*boxplot*)

- gráfico representativo de las distribuciones de un conjunto de datos
- se construye usando cinco medidas descriptivas (**números resumen**) de los datos
  - menor valor: ***L***
  - mayor valor: ***H***
  - los tres cuartiles: Q_1, Q_2, Q_3

L	H
0,02	4,75

ejemplo anterior

CUARTIL 1	CUARTIL 2	CUARTIL 3
0,8200	1,3800	2,1600

8. DATOS ATÍPICOS

Detección: diagrama de caja (*boxplot*)

- los cinco números resumen dividen la muestra en cuatro intervalos que contienen el 25% de los datos, aproximadamente

$$\lfloor L, Q_1 \rfloor \quad \lfloor Q_1, Q_2 \rfloor \quad \lfloor Q_2, Q_3 \rfloor \quad \lfloor Q_3, H \rfloor$$

- información que suministra:
  - sobre la tendencia central, la dispersión y la simetría de los datos de estudio
  - permite identificar, con claridad, lo datos atípicos

8. DATOS ATÍPICOS

Detección: diagrama de caja (*boxplot*)

- etimología
  - ideado por John Tukey en 1977
  - nombre original: *box and whisker plot* (diagrama de caja y bigotes)

Origen de las palabras, razón de su existencia de su significación y de su forma

8. DATOS ATÍPICOS

Detección: diagrama de caja (*boxplot*)

- diseño
  - en el eje horizontal se dibuja un segmento entre los valores menor (*L*) y mayor (*H*) de los datos
  - superpuesta a este segmento se coloca una caja que comienza en el primer cuartil y termina en el tercer cuartil, con lo que contiene el 50% central de las observaciones (rango intercuartílico)
  - dentro de la caja se indica el valor del segundo cuartil mediante una línea vertical

8. DATOS ATÍPICOS

Detección: diagrama de caja (*boxplot*)

- diseño
  - se denominan "bigotes" a las partes del segmento rectilíneo que quedan a ambos lados de la caja
  - generalmente, los "bigotes" se trazan de forma que no lleguen hasta los extremos (valores máximo y mínimo) sino que comprendan los valores del intervalo

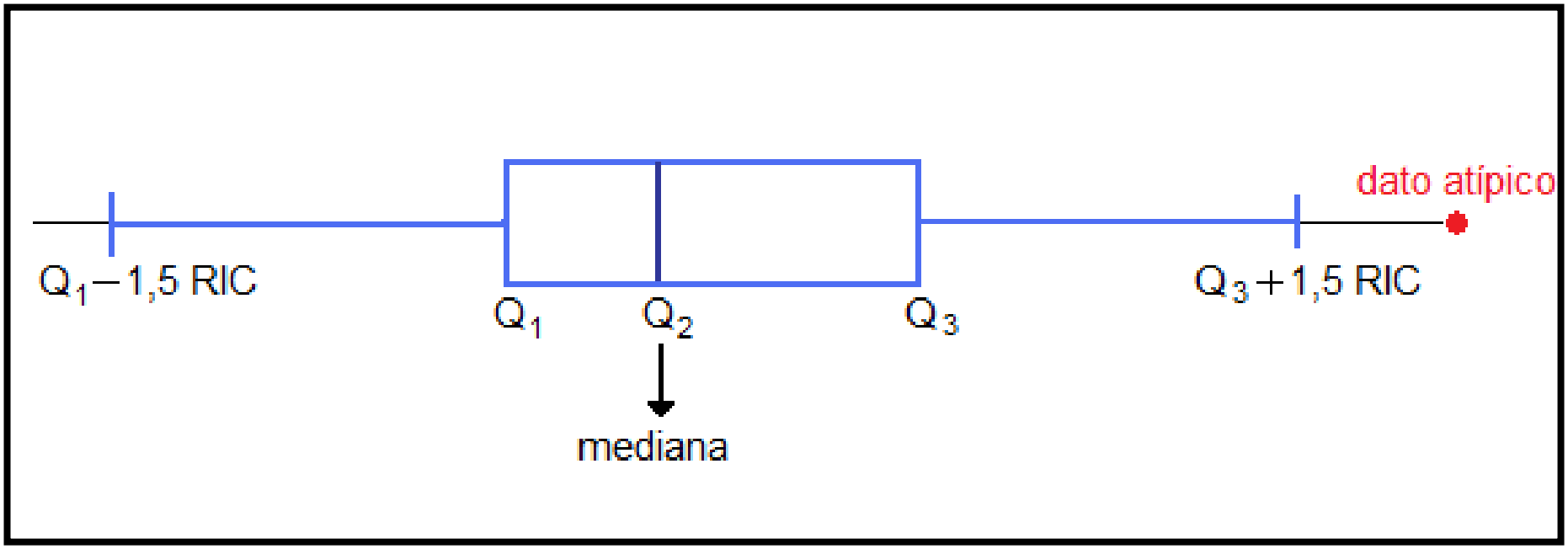
$$\left[Q_1 - 1{,}5 \cdot IQR \, , Q_3 + 1{,}5 \cdot IQR \right]$$

  - fuera de ese intervalo los valores se consideran atípicos

8. DATOS ATÍPICOS

Detección: diagrama de caja (*boxplot*)

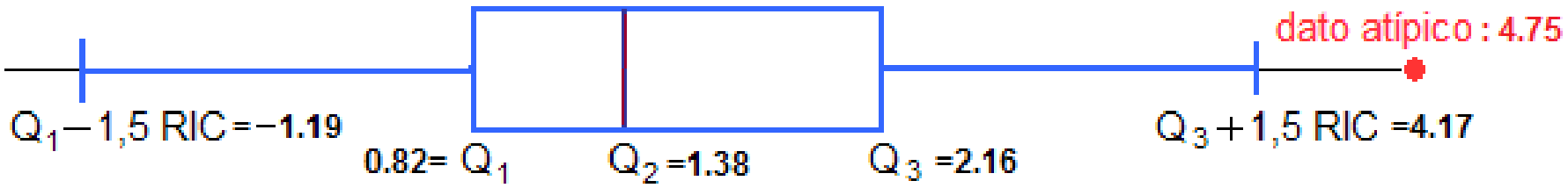
- diseño



8. DATOS ATÍPICOS

Detección: diagrama de caja (*boxplot*)

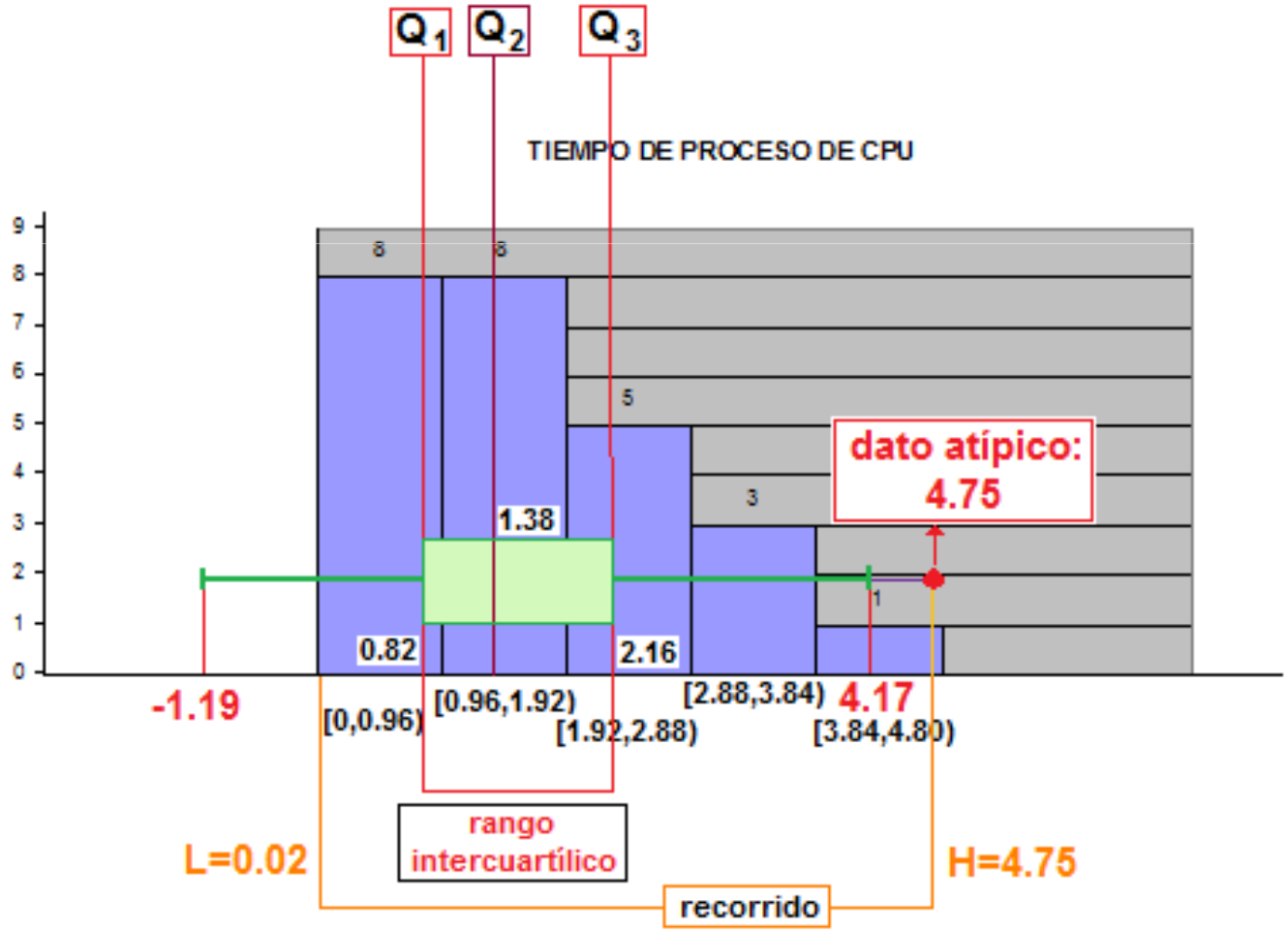
- volviendo al ejemplo anterior:



8. DATOS ATÍPICOS

Detección: diagrama de caja (*boxplot*)

- ejemplo:



8. DATOS ATÍPICOS

Detección: desigualdad de Chebycheff

- sea una muestra de *n* valores:
  - media: $\overline{x}$
  - desviación típica: *s*
- sea el intervalo: $(\overline{x} - k \cdot s, \overline{x} + k \cdot s)$
  - en ese intervalo hay un porcentaje del total de datos que es superior o igual a :

$$\left(1 - \frac{1}{k^2}\right) 100$$

8. DATOS ATÍPICOS

Detección: desigualdad de Chebycheff

- es decir, si n_k es el número de datos contenidos en el intervalo se cumple :

$$\frac{n_k}{N} \geq 1 - \frac{1}{k^2}$$

- este resultado permite establecer otro criterio que determina si un dato es atípico
  - suelen considerarse atípicos los datos que se encuentran a más de tres desviaciones típicas de la media (ya usado en la detección de datos atípicos con el *z-score*
  - la frecuencia relativa de estos datos atípicos es menor de $\frac{1}{9}$

8. DATOS ATÍPICOS

Detección: desigualdad de Chebycheff

- la proporción de datos cuyo valor dista de la media menos de 3 veces la desviación típica será:

$$1-\frac{1}{3^2}=\frac{8}{9}\approx 0.\widehat{8} \quad \Rightarrow \quad 88.\widehat{8}\%$$

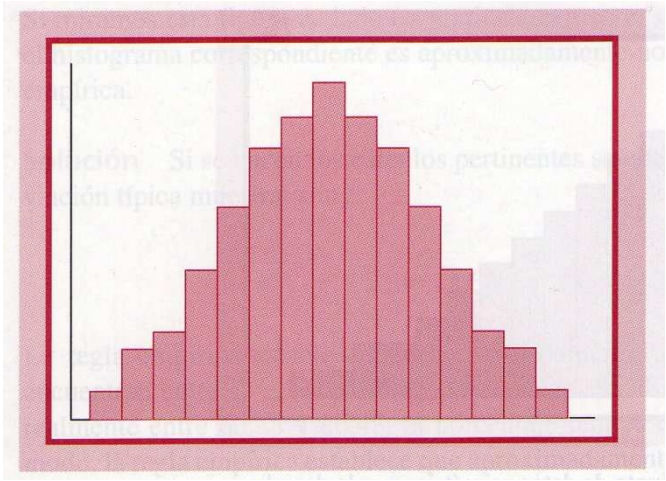
- la proporción de datos cuyo valor dista de la media menos de 4 veces la desviación típica será:

$$1-\frac{1}{4^2}=\frac{15}{16}\approx 0.9375 \quad \Rightarrow \quad 93.75\%$$

9. REGLA EMPÍRICA

<u>Conjunto de datos normal</u>: es aquel conjunto cuyo histograma tiene las propiedades siguientes:

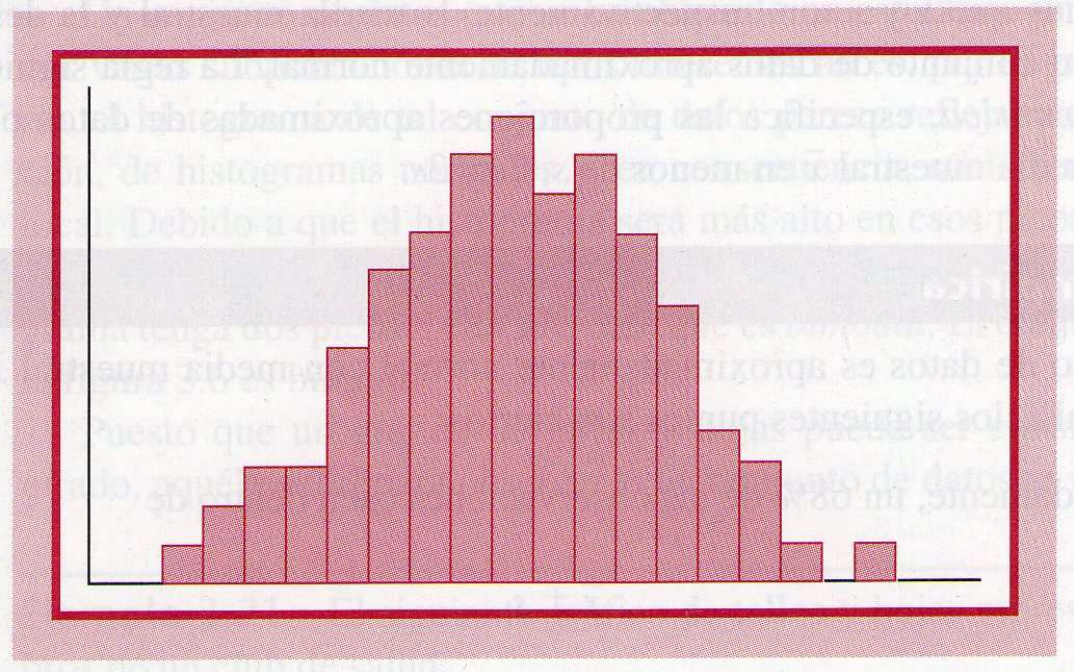
- la máxima altura se alcanza en el intervalo central
- al desplazarse desde el intervalo central en cualquier sentido la altura decrece de tal modo que el histograma completo tiene una forma acampanada
- es simétrico con respecto al intervalo central



Ross, M.S.; Introducción a la Estadística; Ed. Reverté S.A. (2005)

9. REGLA EMPÍRICA

<u>Conjunto de datos aproximadamente normal</u>:
si su histograma se aproxima al de un histograma normal



Ross, M.S.; Introducción a la Estadística;
Ed. Reverté S.A. (2005)

9. REGLA EMPÍRICA

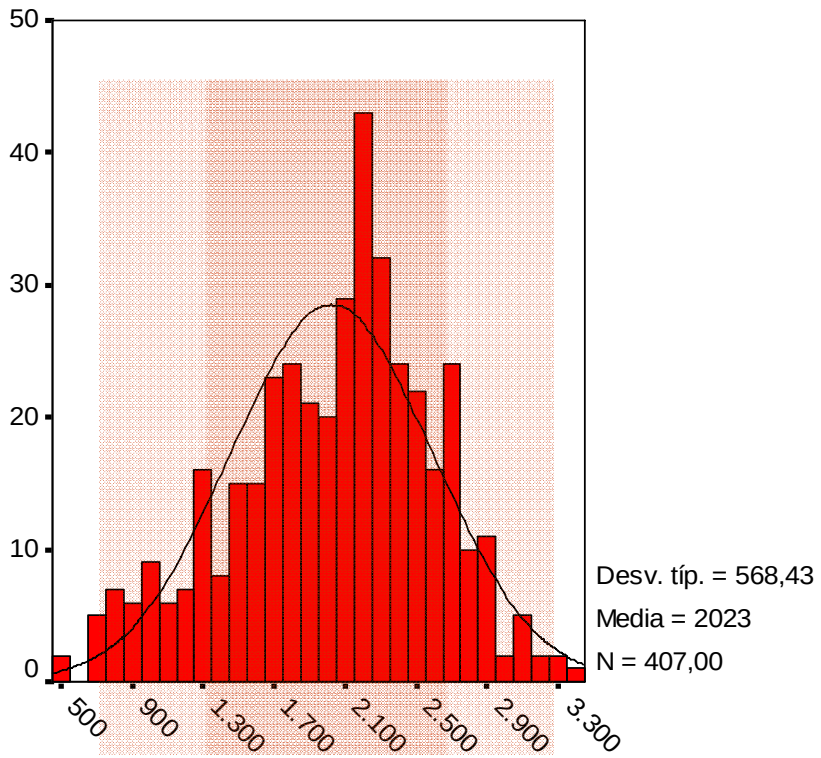
Regla empírica:

- sea un conjunto de datos aproximadamente normal
  - media muestral: $\bar{x}$
  - desviación típica muestral: s
- los siguientes asertos son ciertos:
  - aproximadamente, un 68% de las observaciones caen dentro del intervalo: $(\bar{x}-s, \bar{x}+s)$
  - aproximadamente, un 95% de las observaciones caen dentro del intervalo: $(\bar{x}-2s, \bar{x}+2s)$
  - aproximadamente, un 99.7% de las observaciones caen dentro del intervalo: $(\bar{x}-3s, \bar{x}+3s)$

9. REGLA EMPÍRICA

Regla empírica:

- sólo es válida para datos con distribuciones más o menos de campana

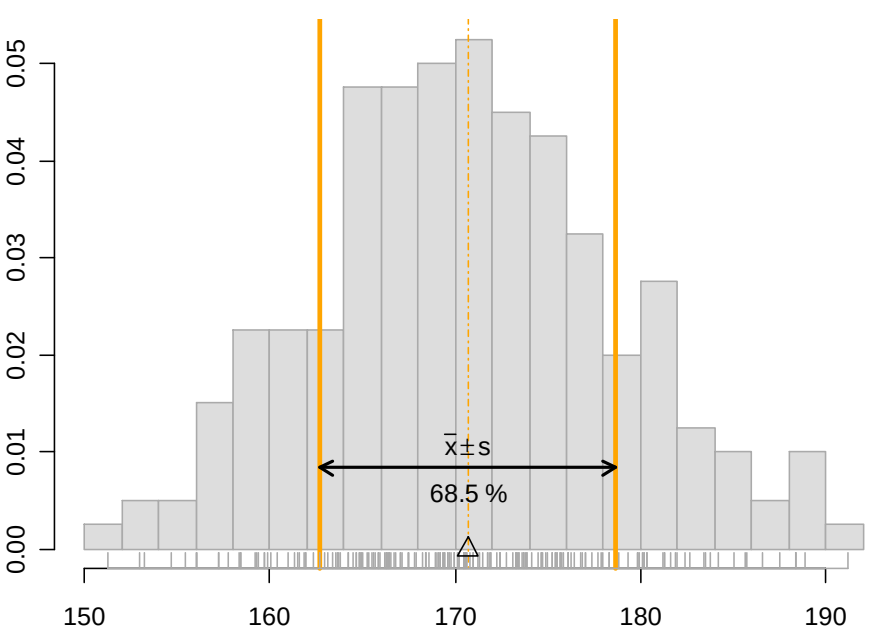


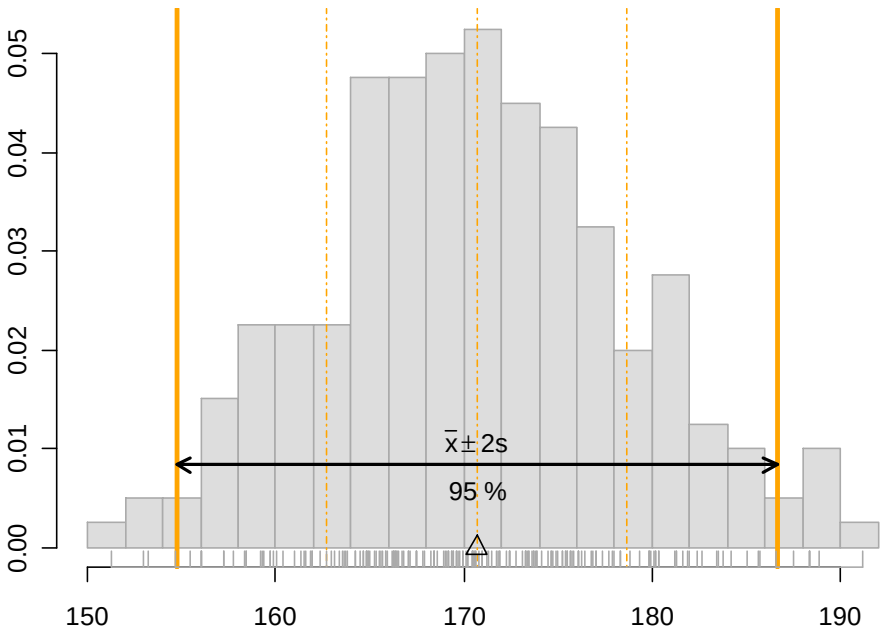
Peso recién nacidos en partos gemelares

9. REGLA EMPÍRICA

Regla empírica:

- se llama "empírica" porque deriva de la observación de lo que "suele suceder" en la práctica
- por eso se formula como "aproximadamente"





10. COEFICIENTE DE PEARSON

Coeficiente de variación de Pearson: es el cociente entre la desviación típica y la media aritmética e indica el número de veces que la desviación típica contiene a la media aritmética

- es el más usado de los estadísticos de dispersión relativa
- evidentemente no puede hallarse cuando $\overline{x} = 0$
- también, llamado *índice de dispersión de Pearson*
- notación:

$$CV = v_1 = \frac{s_x}{\overline{x}}$$

10. COEFICIENTE DE PEARSON

- también, se le denomina *variabilidad relativa* y, en ocasiones, se multiplica por 100 para trabajar con porcentajes

- interesante para comparar la dispersión de variables diferentes

- <u>ejemplo</u>: en una muestra de individuos se mide su peso y su altura
  - $CV_{peso} = 30\%$
  - $CV_{altura} = 10\%$

  → el peso de los individuos está más disperso que su altura

11. ANEXO: FÓRMULAS

Cuantiles ($C_{p/r}$): cálculo con intervalos de clase

- $0 < p/r < 1$
- se multiplica el número de observaciones por p/r :

$$n \cdot \frac{p}{r}$$

- se comprueba si el número obtenido se encuentra en la columna de frecuencias absolutas acumuladas, N_i, de la tabla de distribución de frecuencias
  - si el valor se encuentra en esa columna es que es la frecuencia absoluta acumulada de un cierto intervalo de clase $[L_k, L_{k+1})$ y, por tanto, el cuantil pedido es el extremo superior de dicho intervalo

11. ANEXO: FÓRMULAS

Cuantiles ($C_{p/r}$): cálculo con intervalos de clase

- si no se encuentra en dicha columna el valor estará comprendido entre dos valores N_{k-1} y N_k que corresponden con las frecuencias absolutas acumuladas de dos intervalos de clase $\lfloor L_{k-1}, L_k \rangle$ y $\lfloor L_k, L_{k+1} \rangle$, respectivamente; por lo tanto, el cuantil se encuentra en el intervalo $\lfloor L_k, L_{k+1} \rangle$, calculándose su posición exacta con la siguiente fórmula:

$$C_{p/r} = L_k + \frac{\frac{p}{r} \cdot n - N_{k-1}}{n_k} \cdot A_k$$

11. ANEXO: FÓRMULAS

Cuantiles ($C_{p/r}$): cálculo con intervalos de clase

- ejemplo: cálculo del primer cuartil:

Intervalos: x_i	F. absoluta: n_i	F. abs. acumulada: N_i	$d_i = n_i / A_i$
$[0,100)$	90	90	0,9
$[100,200)$	140	230	1,4
$[200,300)$	150	380	1,5
$[300,800)$	120	500	0,24
Total	**n=500**		

11. ANEXO: FÓRMULAS

Cuantiles ($C_{p/r}$): cálculo con intervalos de clase

- ejemplo: cálculo del primer cuartil:

Intervalos: x_i	F. absoluta: n_i	F. abs. acumulada: N_i	$d_i = n_i / A_i$
$[0,100)$	90	90	0,9
$[100,200)$	140	230	1,4
$[200,300)$	150	380	1,5
$[300,800)$	120	500	0,24
Total	**n=500**		

$$N_1 = 90 < \frac{n}{4} = 125 < N_2 = 230 \Rightarrow \quad C_1 \in [100, 200)$$

$$C_1 = L_2 + \frac{\frac{n}{4} - N_1}{n_2} \cdot A_2 = 100 + \frac{125 - 90}{140} \cdot 100 = 125$$