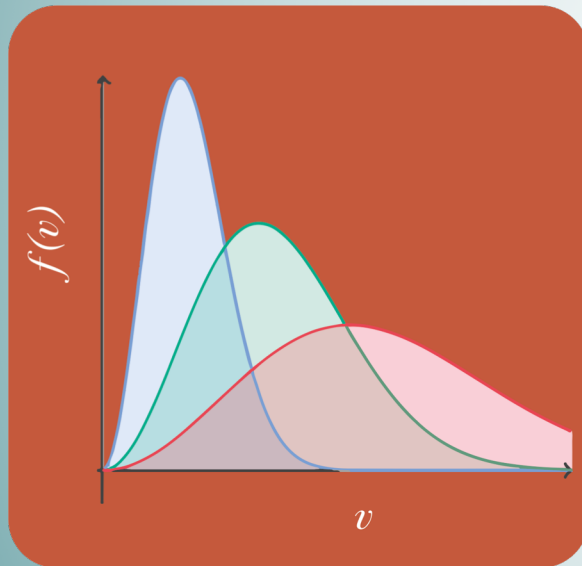


Métodos Matemáticos en la Ingeniería

BLOQUE I. ESTADÍSTICA

TEMA 1. ESTADÍSTICA DESCRIPTIVA



Vera Egorova

Sara Pérez Carabaza

DEPARTAMENTO DE MATEMÁTICA APLICADA
Y CIENCIAS DE LA COMPUTACIÓN

Este material se publica bajo la siguiente licencia:

[Creative Commons BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/)



ÍNDICE.

Introducción

Tipos de variables

Métodos de Estadística Descriptiva

- Tablas de frecuencias: tratamiento individual y de clase.
- Estadísticos: localización, dispersión, forma y posición.
- Gráficos: histograma, diagrama de barras, de caja y sectores.

Objetivos

1. Conocer el objeto de la estadística descriptiva.
2. Saber diferenciar entre población y muestra.
3. Distinguir entre tipos de datos.
4. Explorar y analizar datos de una muestra mediante tablas, estadísticos y gráficos.



INTRODUCCIÓN A LA ESTADÍSTICA DESCRIPTIVA.

La **estadística descriptiva** es una rama de la Estadística que comprende procedimientos empleados para resumir y describir las características importantes de un conjunto de mediciones o datos.

¿Por qué es necesaria o útil? La estadística descriptiva nace de la necesidad de extraer y **resumir información** relevante contenida en grandes volúmenes de datos. Esta necesidad está motivada por la incapacidad de la mente humana para comprender la información contenida en conjuntos de grandes volúmenes de datos.

¿De dónde son los alumnos de la escuela?

Santander, Trasmiera, Santander, Besaya, Santander, Santander, Santander, Pas-Miera Pisuerga , Besaya, Santander, Santander, Santander, Asón-Agüera, Costa Occidental, Liébana, Costa Oriental , Besaya, Santander, Saja-Nansa, Saja-Nansa , Besaya, Otra comunidad , Otra comunidad , Francia, Santander, Trasmiera, Liébana

n=26 alumnos



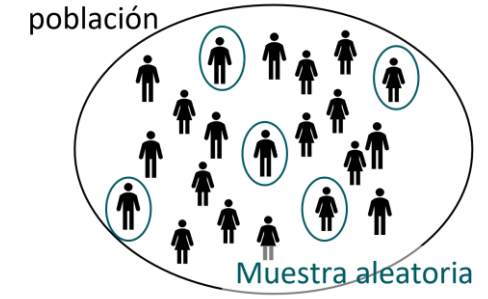
Comarca	Frecuencia absoluta	Frecuencia relativa
Santander	10	0.38
Besaya	4	0.11
Trasmiera	2	0.08
Pas-Miera Pisuerga	1	0.03
Asón-Agüera	1	0.03
Costa Occidental	1	0.03
Liébana	2	0.08
Costa Oriental	1	0.03
Saja-Nansa	2	0.08
Otro	3	0.11



INTRODUCCIÓN A LA ESTADÍSTICA DESCRIPTIVA.

La **población** de un estudio estadístico es el conjunto de elementos objeto de estudio. Cada elemento de la población se denomina **individuo**.

Cuando el número de individuos de la población es muy grande, tomamos una parte de ésta, denominada muestra. La **muestra** es un subconjunto de la población y tiene que ser representativa de la misma. El **tamaño de la muestra n** es el número de individuos que componen la muestra.



La **variable estadística** es la característica de la población que estamos interesados en estudiar.

Ejemplo. Encuesta para saber el lugar de nacimiento de los alumnos de la UC.

población: todos los estudiantes de la Universidad de Cantabria

muestra: alumnos de 1º de la Universidad de Cantabria

variable estadística: lugar de nacimiento de los alumnos de la UC

¿Es una buena muestra? ¿Sería una buena muestra para analizar la variable estadística *edad*?

La **estadística inferencial**, es una rama fundamental de la Estadística cuyo objetivo es hacer inferencias acerca de características poblacionales, a partir de información contenida en una muestra sacada de esa población.

Estadística Descriptiva
Estadística Inferencial



TIPOS DE VARIABLES ESTADÍSTICAS.

Atendiendo a su naturaleza de las variables estadísticas se pueden clasificar en:

- **Variables cualitativas** (o variables categóricas) toman valores no numéricos, miden una cualidad o característica. P. ej. colores favoritos de un grupo de personas.
- **Variables cuantitativas** toman valores numéricos, son medibles numéricamente.
 - **Variables cuantitativas discretas:** solo pueden tomar un conjunto numerable de posibles valores. P. ej. audiencia de Saber y Ganar.
 - **Variables cuantitativas continuas:** pueden tomar cualquier valor real dentro de un cierto intervalo (conjunto no numerable). P. ej. altura de los alumnos de una clase.
- **Variables semi-cuantitativas (u ordinales):** sus valores no son numéricos pero admiten una ordenación natural. P. ej. encuesta satisfacción a la salida de un servicio.

Dependiendo del tipo de variables que consideremos, pueden variar los **métodos estadísticos** apropiados para su tratamiento.

VALÓRANOS HOY



TIPOS DE VARIABLES ESTADÍSTICAS.

Variables cualitativas (o variables categóricas) toman valores no numéricos, miden una cualidad o característica. P. ej. colores favoritos de un grupo de personas.

A veces los datos categóricos se codifican con números en lugar de los nombres de las categorías, pero **siguen siendo categóricos**.

1 baloncesto
2 fútbol
3 tenis
4 natación
5 atletismo
6 otros

Nombre	Deporte
Juan	1
Ana	3
Raúl	5
Beatriz	1
Laura	1
Marta	2
Lorena	4
Ignacio	2



TIPOS DE VARIABLES ESTADÍSTICAS.

Ejercicio. Determina la naturaleza de las siguientes variables del fichero *encuestaETSIIT.txt*.

	A	B	C	D	E	F	G	H	I	J	K
1	% encuestaETSIIT - Datos de encuesta en la ETSIIT de UC (ver encuestaETSIIT_descr.txt)										
2	% Lectura desde Octave/Matlab:										
3	% [Ingenieria	Curso	Asignaturas	Sexo	Edad	Beca	Horario	Motivacion	Estudio	Transporte	Desplazamie
4	Mecanica	1	10	Hombre	18	Si	Mañ±ana	Motivado	2	Coche	15
5	Tecnologias Industriales	2	10	Hombre	19	No	Mañ±ana	Motivado	4	Andando	25
6	Quimica	1	8	Mujer	19	No	Mañ±ana	Motivado	1	Autobus	20
7	Electrica	2	7	Mujer	21	No	Mañ±ana	Motivado	4	Andando	10
8	Tecnologias Industriales	4	9	Mujer	20	No	Mañ±ana	Motivado	4	Coche	10
9	Quimica	1	10	Mujer	18	No	Mañ±ana	Muy motivac	2	Autobus	30
10	Tecnologias de Telecomunicacion	3	12	Mujer	21	No	Mañ±ana	Poco motivac	3	Autobus	20
11	Mecanica	3	8	Hombre	21	No	Mañ±ana	Motivado	2	Coche	30
12	Tecnologias de Telecomunicacion	2	10	Mujer	20	No	Mañ±ana	Motivado	4	Coche	15
13	Quimica	4	10	Mujer	21	Si	Mañ±ana	Motivado	3	Andando	5
14	Electronica Industrial y Automatica	2	9	Hombre	27	Si	Tarde	Muy motivac	3	Autobus	35
15	Quimica	3	10	Mujer	20	Si	Mañ±ana	Muy poco m	0	Moto	5
16	Tecnologias de Telecomunicacion	3	11	Mujer	21	No	Mañ±ana	Motivado	3	Andando	5
17	Quimica	3	10	Hombre	21	No	Mañ±ana	Poco motivac	4	Andando	15
18	Quimica	1	10	Mujer	17	No	Mañ±ana	Muy motivac	2	Autobus	30
19	Tecnologias Industriales	2	10	Hombre	19	No	Mañ±ana	Motivado	1	Autobus	15
20	Mecanica	1	10	Hombre	18	No	Mañ±ana	Muy motivac	2	Autobus	35
21	Tecnologias de Telecomunicacion	3	11	Hombre	22	No	Mañ±ana	Muy motivac	3	Autobus	30



TIPOS DE VARIABLES ESTADÍSTICAS.

Atendiendo al número de observaciones que se realizan del mismo individuo las variables estadísticas se pueden clasificar en:

- **Variables unidimensionales:** solamente se observa un dato de cada individuo. Por ejemplo, resistencia de una viga de hormigón.
- **Variables bidimensionales:** si se observan dos datos de cada individuo. Por ejemplo, se mide la resistencia y la deformación de una viga de hormigón.
- **Variables multidimensionales:** si se observan más de dos datos de cada individuo.

El estudio de las variables unidimensionales y bidimensionales es propio de la Estadística descriptiva básica. El estudio y tratamiento de variables multidimensionales suele requerir técnicas avanzadas para reducir la dimensionalidad de los datos como por ejemplo los métodos de componentes principales.



MÉTODOS DE ESTADÍSTICA DESCRIPTIVA BÁSICA.

La Estadística Descriptiva básica se ocupa del estudio de los datos muestrales a variables unidimensionales y bidimensionales. Para ello se utilizan los siguientes métodos:



- **Tablas de frecuencias**. Adecuado tanto para variables cualitativas como cuantitativas.
- **Estadísticos**. Son funciones de los datos, que pretenden poner de manifiesto ciertas propiedades de los mismos. Un estadístico es, por definición, cualquier función de los datos de una muestra cuantitativa. Por tanto, generalmente no tiene sentido hablar de estadísticos cuando la variable observada no es cuantitativa.
- **Gráficos**. El tipo de gráficos adecuados para representar la variable de estudio dependerá del tipo de esta.

Veremos tablas de frecuencias unidimensionales, estadísticos unidimensionales y gráficos unidimensionales.



MÉTODOS DE ESTADÍSTICA DESCRIPTIVA BÁSICA.

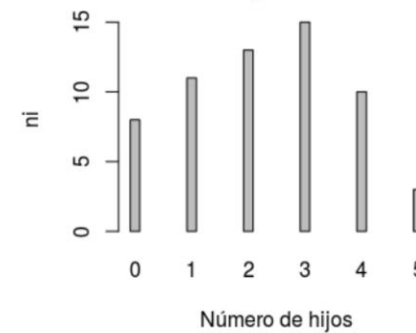
Encuesta a 60 familias de una ciudad sobre el número de hijos:

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



Hijos	n_i
0	8
1	11
2	13
3	15
4	10
5	3

Tablas



Gráficos

media=2.283
varianza=2.005
desv.stand.=1.416
moda=3
mediana=2

Estadísticos



MÉTODOS ESTADÍSTICOS: TABLAS DE FRECUENCIAS.

Una tabla de frecuencias resume la información contenida en los datos de una muestra. Según las características de la variable podemos realizar tablas de tratamiento individual o de clase.

Tratamiento individual: Este tipo de tablas es adecuado para variables cuantitativas discretas (con pocos valores) o cualitativas, aunque en el caso de las cualitativas no tiene sentido calcular las frecuencias acumuladas.

x_i : posibles valores de la variable.

n_i : **frecuencia absoluta**. Número de veces que aparece x_i en la muestra.

f_i : **frecuencia relativa**. Proporción de veces que ocurre x_i en la muestra.

N_i : **frecuencia absoluta acumulada**

F_i : **frecuencia relativa acumulada**

n : tamaño de la muestra (número de datos)

Encuesta a 60 familias de una ciudad sobre el número de hijos:

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



x_i	n_i	f_i	N_i	F_i
0	8	0.13	8	0.13
1	11	0.18	19	0.31
2	13	0.22	32	0.53
3	15	0.25	47	0.78
4	10	0.17	57	0.95
5	3	0.05	60	1

n



MÉTODOS ESTADÍSTICOS: TABLAS DE FRECUENCIAS.

f_i : frecuencia relativa. N_i : frecuencia absoluta acumulada.

F_i : frecuencia relativa acumulada.

$$f_i = \frac{n_i}{n}$$

$$N_i = \sum_{j \leq i} n_j$$

$$F_i = \sum_{j \leq i} f_j = \frac{1}{n} \sum_{j \leq i} n_j$$

Pregunta. A partir de la tabla de frecuencias responde a las siguientes preguntas:

- a) Número de familias con 2 hijos.
- b) Porcentaje de familias con 2 hijos
- c) Número de familias con 3 hijos o menos.
- d) Porcentaje de familias con menos de 5 hijos.

Encuesta a 60 familias sobre el número de hijos:

Hijos	n_i	f_i	F_i	N_i
0	8	0.13	0.13	8
1	11	0.18	0.31	19
2	13	0.22	0.53	32
3	15	0.25	0.78	47
4	10	0.17	0.95	57
5	3	0.05	1	60

A tener en cuenta... Las tablas de frecuencias facilitan mucho la comprensión de los datos respecto al listado exhaustivo de estos. Pero proporcionan una *visión estática* (se pierde información sobre el orden de las observaciones). Una tabla sobre el registro del número de hijos de familias desde 1950 no sería muy conveniente.



MÉTODOS ESTADÍSTICOS: TABLAS DE FRECUENCIAS.

Ejercicio. Construye la tabla de frecuencias en Matlab a partir del conjunto de datos.

Encuesta a 60 familias de una ciudad sobre el número de hijos:

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

```
clear all;
%Tabla de frecuencias. Diapositivas
datos=[0 2 2 4 0 3 3 2 5 2 3 2 4 3 4 ...
        3 1 4 1 1 0 4 1 1 4 2 4 2 0 3 ...
        1 3 0 5 2 2 3 0 3 0 5 1 1 4 0 ...
        3 2 3 2 3 3 1 2 4 2 3 1 3 1 4 ];
n=length(datos);%60
%xi: valores que toma la variable (0,1,2,3,4,5)
[xi, ~, pos] = unique(datos);%xi(pos) reconstruye datos
%frecuencias absolutas ni
ni = hist(pos, length(xi));
%podemos comprobar ni(1)=sum(datos==0), ni(2)=sum(datos==1) etc
%frecuencias relativas fi
fi=ni/n;
%frecuencias absolutas acumuladas
Ni=cumsum(ni);
%frecuencias relativas acumuladas
Fi=Ni/n;% o cumsum(fi)
format short g
[xi' ni' fi' Fi' Ni']
```

Hijos	n_i	f_i	F_i	N_i
0	8	0.13	0.13	8
1	11	0.18	0.31	19
2	13	0.22	0.53	32
3	15	0.25	0.78	47
4	10	0.17	0.95	57
5	3	0.05	1	60

ans =

0	8	0.13333	0.13333	8
1	11	0.18333	0.31667	19
2	13	0.21667	0.53333	32
3	15	0.25	0.78333	47
4	10	0.16667	0.95	57
5	3	0.05	1	60

[enlace datos](#)



MÉTODOS ESTADÍSTICOS: TABLAS DE FRECUENCIAS.

Tratamiento de clase (agrupamiento por intervalos): Este tipo de tablas es adecuado para variables cuantitativas continuas o cuantitativas discretas que tomen multitud de valores. Cada uno de los intervalos recibe el nombre de clase. La agrupación por intervalos conlleva pérdida de información.

$[L_{i-1}, L_i)$: límites de clase. Valor inferior y superior del intervalo que define la clase.

x_i : marcas de clase. Valor medio de los límites de clase.

$$x_i = \frac{L_{i-1} + L_i}{2}$$

n_i : frecuencia absoluta. Número de ocurrencias de la muestra en cada intervalo.

f_i : frecuencia relativa. Proporción de veces que ocurre x_i en la muestra.

N_i : frecuencia absoluta acumulada.

F_i : frecuencia relativa acumulada.

El **criterio de Sturges** nos da un número adecuado de clases: $c = 1 + \frac{\log_{10}(n)}{\log_{10}(2)}$

Estatura de 40 alumnos

1,55 1,66 1,69 1,63 1,64 1,67 1,63 1,56
1,62 1,68 1,68 1,62 1,66 1,62 1,69 1,56
1,57 1,60 1,65 1,64 1,67 1,69 1,63 1,64
1,60 1,62 1,63 1,71 1,62 1,72 1,61 1,61
1,64 1,60 1,70 **1,76** 1,65 1,65 1,68 1,66

Estatura	Marcas de clase	Frecuencia absoluta	Frecuencia relativa	Frecuencia absoluta acumulada	Frecuencia relativa acumulada
[1,55 – 1,60)	1,575	4	0,1	4	0,1
[1,60 – 1,65)	1,625	18	0,45	22	0,55
[1,65 – 1,70)	1,675	14	0,35	36	0,9
[1,70 – 1,75)	1,725	3	0,075	39	0,975
[1,75 – 1,80)	1,775	1	0,025	40	1

amplitud del intervalo=0.05m

MÉTODOS ESTADÍSTICOS: TABLAS DE FRECUENCIAS.

Ejercicio. Construye la tabla de frecuencias a partir del siguiente conjunto de datos.

Tabla 1. El tiempo de acceso al disco duro (milisegundos) medido en 30 instantes de tiempo distintos.

6.2 7.2 5.3 4.4 6.3 9.1 6.4 9.9 6.7 8.7
5.5 8.4 6.9 4.1 8.5 7.3 8.5 7.2 9.1 4.4
7.3 8.8 5.8 7.5 4.4 7.8 6.9 6.1 8.2 6.6

$$c = 1 + \frac{\log_{10}(30)}{\log_{10}(2)} = 5.9 \rightarrow 6 \text{ clases}$$

$$a \gtrsim \frac{\max(x) - \min(x)}{c} = \frac{9.9 - 4.1}{6} = 0.97 \Rightarrow a = 1$$

$(L_{i-1}, L_i]$	x_i	n_i	N_i	f_i	F_i
4-5	4.5	4	4	0.133	0.133
5-6	5.5	3	7	0.100	0.233
6-7	6.5	8	15	0.267	0.500
7-8	7.5	6	21	0.200	0.700
8-9	8.5	6	27	0.200	0.900
9-10	9.5	3	30	0.100	1.000
Total		30		1	



MÉTODOS ESTADÍSTICOS: TABLAS DE FRECUENCIAS.

Ejercicio. Construye la tabla de frecuencias con MATLAB a partir del siguiente conjunto de datos.

```
datos=[6.2 7.2 5.3 4.4 6.3 9.1 6.4 9.9 6.7 8.7...  
       5.5 8.4 6.9 4.1 8.5 7.3 8.5 7.2 9.1 4.4 ...  
       7.3 8.8 5.8 7.5 4.4 7.8 6.9 6.1 8.2 6.6];  
n=length(datos);%30  
%criterio de Sturges  
c=round(1+log10(n)/log10(2));  
%frecuencias absolutas ni  
xi=4.5:1:9.5;  
ni=hist(datos, xi);  
%frecuencias relativas fi  
fi=ni/n;  
%frecuencias absolutas acumuladas  
Ni=cumsum(ni);  
%frecuencias relativas acumuladas  
Fi=Ni/n;% o cumsum(fi)  
format short g  
[xi' ni' Ni' fi' Fi']
```

$(L_{i-1}, L_i]$	x_i	n_i	N_i	f_i	F_i
4-5	4.5	4	4	0.133	0.133
5-6	5.5	3	7	0.100	0.233
6-7	6.5	8	15	0.267	0.500
7-8	7.5	6	21	0.200	0.700
8-9	8.5	6	27	0.200	0.900
9-10	9.5	3	30	0.100	1.000
Total		30		1	

ans =

4.5	4	4	0.13333	0.13333
5.5	3	7	0.1	0.23333
6.5	8	15	0.26667	0.5
7.5	6	21	0.2	0.7
8.5	6	27	0.2	0.9
9.5	3	30	0.1	1

[enlace datos](#)

MÉTODOS ESTADÍSTICOS: ESTADÍSTICOS.

Los estadísticos son funciones de los datos de la muestra que sirven para cuantificar ciertas características relevantes de esta.

$$f(x_1, x_2, \dots, x_n)$$

Dependiendo del tipo de información que proporcionen se distinguen los siguientes tipos de estadísticos:

Tipos de
estadísticos

- estadísticos de tendencia central o de localización. (media, mediana y moda)
- estadísticos de dispersión (cuasivarianza y cuasidesviación típica muestrales)
- estadísticos de forma (coeficientes de asimetría y curtosis muestrales)
- estadísticos de posición (cuantiles muestrales)

En general, los estadísticos se definen para datos cuantitativos.



ESTADÍSTICOS DE TENDENCIA CENTRAL.

Media muestral (mean): media aritmética (promedio) de los datos:

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k$$

También se puede obtener a partir de las tablas de frecuencia. Aunque en el caso de las tablas de frecuencia de tratamiento por clases la media no tiene por qué coincidir con esta expresión.

$$\bar{x} = \frac{1}{n} \sum_{k=1}^c x_k n_k = \sum_{k=1}^c x_k f_k$$

siendo c el número de clases definido

Ejercicio 1. Obtén la media muestral del número de hijos a partir de la tabla de frecuencias.

Ejercicio 2. ¿Afectaría igual a la media muestral eliminar el dato de una familia con 3 hijos que una con 5 hijos?

Encuesta a 60 familias sobre el número de hijos:

Hijos	n_i	f_i	F_i	N_i
0	8	0.13	0.13	8
1	11	0.18	0.31	19
2	13	0.22	0.53	32
3	15	0.25	0.78	47
4	10	0.17	0.95	57
5	3	0.05	1	60



ESTADÍSTICOS DE TENDENCIA CENTRAL.

Mediana (median): es un valor numérico tal que tanto la frecuencia relativa de los datos iguales inferiores como la frecuencia relativa de los datos iguales o superiores es al menos 0.5.

Para obtener la mediana de un conjunto de datos se ordenan en orden creciente y nos quedamos con el dato central:

- n impar: la mediana será un valor único, el dato central

1, 1, 1, 1, 1, 1, **2**, 2, 2, 2, 3, 3, 4
Mitad inferior Mediana Mitad superior

- n par: puede haber dos medianas diferentes, por convenio se toma su valor medio.

1, 1, 1, 1, 1, **1, 2**, 2, 2, 3, 3, 4
Valores inferiores Valores intermedios Valores superiores

La mediana tiene la ventaja respecto a la media es que es un **estadístico robusto** (no es sensible a los valores extremos).

Ejercicio 1. Obtén la mediana de los siguientes conjuntos de datos:

- a) {3,5,6,10,15}
- b) {3,5,6,10,400}
- c) {3,5,6,10,40,50}
- d) Número de hijos (tabla)

Hijos	n_i	f_i	F_i	N_i
0	8	0.13	0.13	8
1	11	0.18	0.31	19
2	13	0.22	0.53	32
3	15	0.25	0.78	47
4	10	0.17	0.95	57
5	3	0.05	1	60



ESTADÍSTICOS DE TENDENCIA CENTRAL.

Moda (mode): es el valor que más se repite, es decir, el valor que tienen una mayor frecuencia relativa (o absoluta). Un grupo de datos puede tener una moda (unimodal), dos modas (bimodal) o más modas (multimodal).

En los datos agrupados por clases, la **clase modal** es la clase (el intervalo) que presenta la mayor frecuencia absoluta o relativa.

A diferencia de la media y mediana este tipo de estadístico sí que tiene sentido para variables cualitativas.

Ejercicio 1. Obtén la moda o la clase modal de los siguientes conjuntos de datos:

- a) {3,5,6,6,6,15,15}
- b) {1,1,1,2,2,100,100,101,101,101}
- c) El número de hijos
- d) Puntuaciones del test

Encuesta a 60 familias sobre el
número de hijos:

Hijos	n_i	f_i	F_i	N_i
0	8	0.13	0.13	8
1	11	0.18	0.31	19
2	13	0.22	0.53	32
3	15	0.25	0.78	47
4	10	0.17	0.95	57
5	3	0.05	1	60

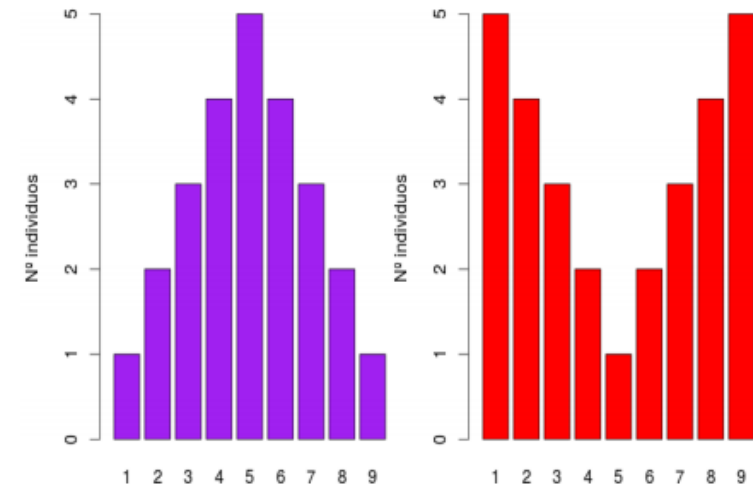
Puntuaciones del test

$[L_{i-1}, L_i)$	x_i	n_i	N_i
[5, 10)	7.5	3	3
[10, 15)	12.5	6	9
[15, 20)	17.5	13	22
[20, 25)	22.5	7	29
[25, 30)	27.5	2	31

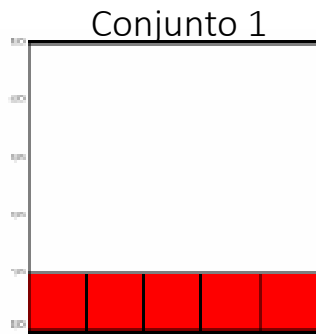


ESTADÍSTICOS DE DISPERSIÓN. MOTIVACIÓN.

Pregunta. Dos clases diferentes, de 25 y 29 alumnos respectivamente, han obtenido las calificaciones mostradas en los gráficos en una asignatura ¿Qué diferencia ves?

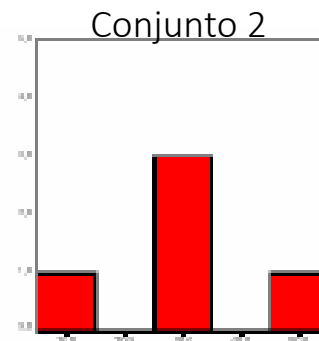


Los estadísticos de tendencia central coinciden, ¡Sin embargo los datos son totalmente distintos!



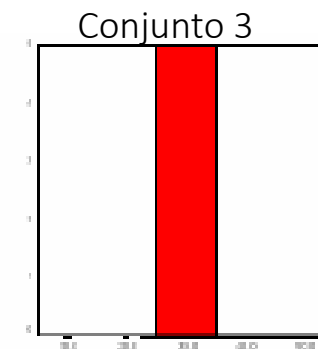
Conjunto 1: 10 20 30 40 50

media=30 ,mediana=30,moda=no tiene



Conjunto 2: 10 30 30 30 50

media=30 mediana=30 moda=30



Conjunto 3: 30 30 30 30 30

media=30 mediana=30 moda=30



ESTADÍSTICOS DE DISPERSIÓN. MOTIVACIÓN.

Los **estadísticos de dispersión** indican la mayor o menor concentración de los datos con respecto a las medidas de localización. Indican la representatividad de los estadísticos de localización.

Dispersión alta → datos alejados de la media → datos heterogéneos → media poco representativa
Dispersión baja → datos cercanos a la media → datos homogéneos → media representativa

- Rango o recorrido `range(datos)`
- Rango intercuartílico (RIC) `iqr(datos)`
- varianza, cuasivarianza `var(datos, 1)` y `var(datos)`
- desviación típica, cuasidesviación típica `std(datos, 1)` y `std(datos)`
- coeficiente de variación.



ESTADÍSTICOS DE DISPERSIÓN. MOTIVACIÓN.

Rango o recorrido (range): Diferencia entre los valores máximo y el mínimo de la muestra.

Muy sensible a valores extremos.

Mismas unidades que la variable.

El rango no refleja la variación de todos los datos.

Ejemplo

Alturas (cm): 160, 165, 172, 174, 174, 176, 179, 180, 180, 180, 180, 187

Rango = $187 - 160 = 27$ cm

MATLAB

```
datos = [160,165,172,174,174,176,179,180,180,180,180,187];  
range(datos)
```

Rango intercuartílico (iqr): Diferencia entre el tercer y el primer cuartil. (Volveremos a él cuando veamos los cuantiles). $RIC = C_{0.75} - C_{0.25}$



ESTADÍSTICOS DE DISPERSIÓN. MOTIVACIÓN.

Varianza y cuasi-varianza. La **varianza** S_n^2 se define como el promedio del cuadrado de las desviaciones de los datos muestrales respecto a su media. La **cuasi-varianza** S^2 es otro estadístico de dispersión íntimamente relacionado con la varianza, pero dividido a su media. Ambos estadísticos tienen unidades de la variable al cuadrado y son sensibles a valores extremos.

varianza

$$S_n^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2 = \overline{x^2} - \bar{x}^2$$

var(datos, 1)

cuasivarianza

$$S^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2$$

var(datos) o var(datos, 0)

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

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

Ejercicio 1. Una manera intuitiva de medir la dispersión de los datos podría ser promediar las diferencias de estos respecto a la media, ¿Por qué no se utiliza esta expresión?

Ejercicio 2. Demuestra la expresión $S_n^2 = \overline{x^2} - \bar{x}^2$



ESTADÍSTICOS DE DISPERSIÓN.

Desviación típica (o estándar) S y cuasi-desviación típica (S). La desviación típica es la raíz cuadrada positiva de la varianza y la cuasi-desviación típica de la cuasi-varianza. Ambos estadísticos tienen unidades de la variable y son sensibles a valores extremos.

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

`std(datos, 1)`

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

`std(datos) o std(datos, 0)`

Coefficiente de variación o dispersión relativa (CV). Cociente entre la cuasi-desviación típica y la media muestral. Es adimensional, y es frecuente usarlo en porcentaje.

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

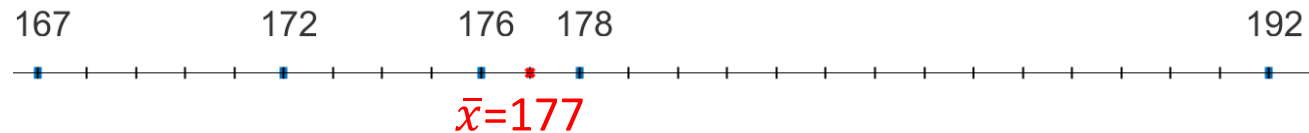
Idea clave: uso del coeficiente de variación para comparar la dispersión entre conjuntos de datos diferentes.

Ejercicio 1. En una muestra de adultos se obtiene los siguientes estadísticos respecto a su peso $\bar{x} = 73.24 \text{ kg}$, $S^2 = 169.93 \text{ kg}^2$. En otra muestra de bebés se tiene $\bar{x} = 3.12 \text{ kg}$, $S^2 = 1.80 \text{ kg}^2$. ¿Qué grupo presenta una mayor dispersión en el peso?



ESTADÍSTICOS DE DISPERSIÓN.

Ejercicio 1. Estadísticos de dispersión del siguiente conjunto de datos correspondiente a las alturas (cm) de 5 personas:



$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k = 177 \text{ cm}$$

$$\frac{1}{n} \sum_{k=1}^n (x_k - \bar{x}) = \frac{1}{5} \sum_{k=1}^5 (x_k - 177) = \frac{1}{5} (-10 - 5 - 1 + 1 + 15) = 0 \text{ cm}$$

varianza:

$$S_n^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2 = \frac{1}{5} ((167 - 177)^2 + (172 - 177)^2 + \dots + (192 - 177)^2) = \frac{1}{5} 352 = 70.4 \text{ cm}^2,$$
$$S^2 = \frac{1}{4} 352 = 88 \text{ cm}^2$$

desviación típica: $S_n = \sqrt{S_n^2} = \sqrt{70.4} = 8.3905 \text{ cm}$ $S = \sqrt{S^2} = 9.38 \text{ cm}$

desviación media: $DM = \frac{1}{n} \sum_{k=1}^n |x_k - \bar{x}| = \frac{1}{5} (10 + 5 + 1 + 1 + 15) = 6.4 \text{ cm}$



Ejercicio 2. Dados las siguientes 30 medidas de tiempos de acceso al disco duro (milisegundos) obtén tanto para los datos agrupados y sin agrupar su media, varianza y desviación típica:

4.1 , 4.4 , 4.4 , 4.4, 5.3 , 5.5 , 5.8 , 6.1 , 6.2, 6.3
6.4 , 6.6 , 6.7 , 6.9 , 6.9 , 7.2 , 7.2 , 7.3 , 7.3 , 7.5
, 7.8 , 8.2 , 8.4 , 8.5 , 8.5 , 8.7 , 8.8 , 9.1 , 9.1 9.9

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k = 6.98 \text{ ms}$$

$$S_n^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2 = 2.34 \text{ ms}^2$$

Otra opción:

$$S_n^2 = \overline{x^2} - \bar{x}^2 = \frac{1}{n} \sum_{k=1}^n x_k^2 - 6.98^2 = 2.34 \text{ ms}^2$$

$$S_n = \sqrt{S_n^2} = 1.53 \text{ ms}$$

$[L_{i-1}, L_i)$	x_i	n_i	N_i	f_i	F_i	$x_i n_i$	$x_i^2 n_i$
4-5	4.5	4	4	0.133	0.133	18	81
5-6	5.5	3	7	0.100	0.233	16.5	90.75
6-7	6.5	8	15	0.267	0.500	52	338
7-8	7.5	6	21	0.200	0.700	45	337.5
8-9	8.5	6	27	0.200	0.900	51	433.5
9-10	9.5	3	30	0.100	1.000	28.5	270.75
Total		30		1.000		211	1551.5

$$\bar{x} = \frac{1}{n} \sum_{k=1}^c x_k n_k = 7.0333 \text{ ms}$$

$$S_n^2 = \frac{1}{n} \sum_{k=1}^c (x_k - \bar{x})^2 n_k = 2.25 \text{ ms}^2$$

Otra opción:

$$S_n^2 = \overline{x^2} - \bar{x}^2 = \frac{1551.5}{30} - \left(\frac{211}{30} \right)^2 = 2.25 \text{ ms}^2$$

$$S_n = \sqrt{S_n^2} = 1.5 \text{ ms}$$

[enlace datos](#)



ESTADÍSTICOS DE FORMA. MOMENTOS DE ORDEN.

Las medidas de forma, como el coeficiente de asimetría y de curtosis, sirven para tener una idea acerca de la forma de una distribución de frecuencias con sólo un número. Comencemos definiendo los momentos de orden.

Momento de orden r . Se llama momento muestral m_r de orden r , respecto de una constante real a , a la siguiente media:

$$m_r = \frac{1}{n} \sum_{k=1}^n (x_k - a)^r$$

Si $a = 0$, se habla de **momentos respecto al origen**.

Si $a = \bar{x}$ se dice que son **momentos respecto a la media**.

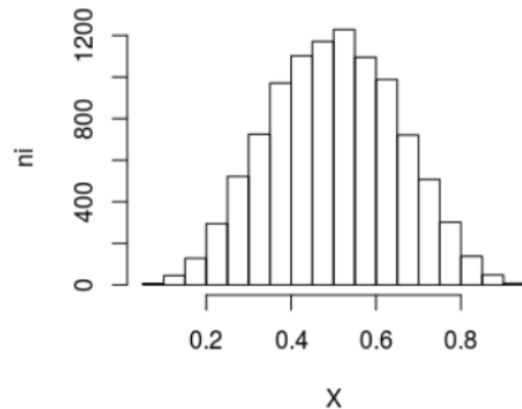
Pregunta. Hemos visto anteriormente dos estadísticos que se corresponden con momentos de orden r . ¿Sabes cuáles son? ¿Con que orden r y constante a se corresponden?



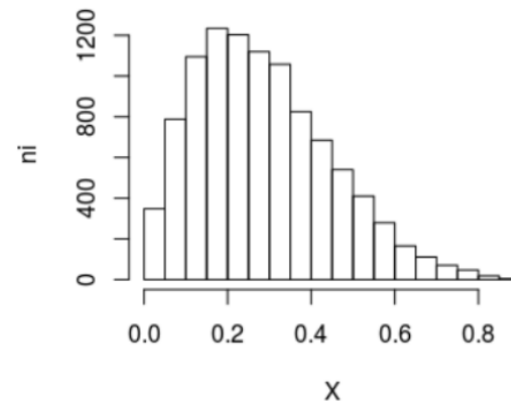
Coeficiente de asimetría (C_A), en inglés *skewness*. El **coeficiente de asimetría de Fisher** se define como el cociente entre el momento de tercer orden estandarizado y la desviación típica al cubo. Nos indica si la distribución es simétrica. Adimensional

$$C_A = \frac{m_3}{S_n^3} = \frac{\frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^3}{\left(\frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2 \right)^{3/2}}$$

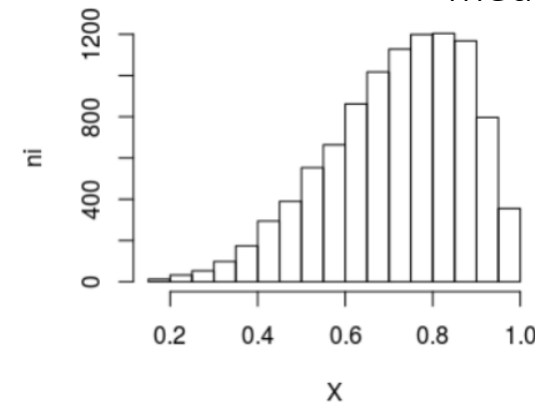
$C_A = 0$, la distribución es simétrica
(media = mediana)



$C_A > 0$, la distribución es
asimétrica por la derecha
media > mediana



$C_A < 0$, la distribución es
asimétrica por la izquierda
media < mediana



ESTADÍSTICOS DE FORMA.

Coeficiente de asimetría (C_A), en inglés *skewness*. El **coeficiente de asimetría de Fisher** se define como el cociente entre el momento de tercer orden estandarizado y la desviación típica al cubo. Nos indica si la distribución es simétrica. Adimensional.

$$C_A = \frac{m_3}{S_n^3} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2)^{3/2}}$$

Distribución simétrica: $C_A=0$, media=mediana

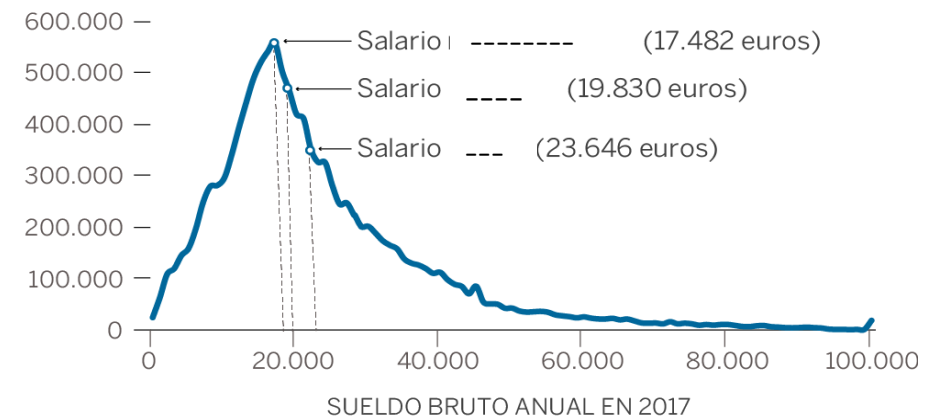
Distribución asimétrica a la derecha: $C_A>0$, media>mediana

Distribución asimétrica a la izquierda: $C_A<0$, media<mediana

La imagen de la derecha muestra la distribución de salarios de los trabajadores en *España* durante 2017.

Pregunta 1. ¿Se trata de una distribución simétrica, asimétrica a la derecha o asimétrica a la izquierda?

Pregunta 2. Identifica el salario modal, salario mediano y salario medio.



Fuente: INE. EL PAÍS



ESTADÍSTICOS DE POSICIÓN.

Los estadísticos de posición están definidos en términos de **estadísticos de orden**.

Estadísticos de orden k $x_{(k)}$: es igual al k -ésimo valor más pequeño de la muestra.

Ejemplo. Dados los datos $x_1 = 7, x_2 = 10, x_3 = 9, x_4 = 3$ los estadísticos de orden correspondientes: $x_{(1)} = 3, x_{(2)} = 7, x_{(3)} = 9, x_{(4)} = 10$

Entre los estadísticos de orden ellos destacan:

Mínimo: estadístico de orden más pequeño $x_{(1)}$

Máximo: estadístico de orden más grande $x_{(n)}$

Cuantil de orden α (C_α): valor numérico que verifica las siguientes dos desigualdades:

$$\sum_{\forall i | x_i \leq C_\alpha} f_i \geq \alpha \quad y \quad \sum_{\forall i | x_i \geq C_\alpha} f_i \geq 1 - \alpha \quad \text{con } \alpha \text{ entre } 0 \text{ y } 1.$$

Por ejemplo, $C_{0.1}$ es igual o superior al 10% o más de los datos, e inferior o igual al 90% o más de los datos de una muestra.



ESTADÍSTICOS DE POSICIÓN.

Si αn no es un número natural C_α es único e igual a $x_{(\lceil n\alpha \rceil)}$. Si por el contrario, αn es natural todos los números comprendidos en $[x_{(n\alpha)}, x_{(n\alpha+1)}]$ cumplirán la definición de cuantil. En dicho caso se seguirá el criterio de tomar el valor medio:

$$C_\alpha = \begin{cases} x_{(\lceil n\alpha \rceil)} & \text{si } n\alpha \notin \mathbb{N} \\ \frac{x_{(n\alpha)} + x_{(n\alpha+1)}}{2} & \text{si } n\alpha \in \mathbb{N} \end{cases}$$

Ejemplo. Obtener $C_{0.5}$ y $C_{0.1}$ Alturas (cm): 187, 160, 180, 172, 174, 179, 176, 174, 180, 165, 180, 180

Valores ordenados: 160, 165, 172, 174, 174, 176, 179, 180, 180, 180, 180, 187

$$\alpha n = 0.5 \cdot 12 = 6 \in \mathbb{N}$$

$$C_{0.5} = (x_{(6)} + x_{(7)})/2 = 0.5 \cdot 176 + 0.5 \cdot 179 = 177.5 \text{ cm}$$

Al menos el 50% de los individuos de la muestra tiene una altura menor o igual a 177.5 cm.

$$\alpha n = 0.1 \cdot (12) = 1.2 \notin \mathbb{N}$$

$$C_{0.1} = x_{(\lceil 1.2 \rceil)} = x_{(2)} = 165 \text{ cm}$$

tanto 176 como 179 cumplen la definición de $C_{0.5}$

$$\sum_{\forall i | x_i \leq 176} f_i = \frac{1}{12} \cdot 6 = 0.5 \geq 0.5 \quad \text{y} \quad \sum_{\forall i | x_i \geq 176} \frac{1}{12} \cdot 7 = 0.58 \geq 1 - 0.5$$

$$\sum_{\forall i | x_i \leq 179} f_i = \frac{1}{12} \cdot 7 = 0.58 \geq 0.5 \quad \text{y} \quad \sum_{\forall i | x_i \geq 179} \frac{1}{12} \cdot 6 = 0.5 \geq 1 - 0.5$$

165 cumple la definición de $C_{0.1}$

$$\sum_{\forall i | x_i \leq 165} f_i = \frac{1}{12} \cdot 2 = 0.1667 \geq 0.1 \quad \text{y} \quad \sum_{\forall i | x_i \geq 165} \frac{1}{12} \cdot 11 = 0.9167 \geq 1 - 0.1$$

`quantile(datos, 0.5); quantile(datos, 0.1)`



ESTADÍSTICOS DE POSICIÓN.

Para **datos agrupados en intervalos**, una vez identificado el intervalo $[L_{i-1}, L_i)$ donde se encuentra el cuantil C_α obtenemos su valor exacto mediante la siguiente expresión:

$$C_\alpha = L_{i-1} + a_i \frac{\alpha n - N_{i-1}}{n_i}$$

α orden del cuantil

i intervalo que contiene al cuantil

L_{i-1} límite inferior del intervalo i

a_i amplitud del intervalo i

n_i frecuencia absoluta del intervalo i

N_{i-1} frecuencia absoluta acumulada del intervalo $i-1$ (el anterior)

Ejemplo. ¿Cuál es el peso mediano?

$$C_{0.5} = 65 + 10 \frac{0.5 \cdot 30 - 9}{9} = 71.66 \text{ kg}$$

Pregunta. ¿ $C_{0.7}$?

$(L_{i-1}, L_i]$	x_i	n_i	f_i	N_i	F_i
45-55	50	2	0.067	2	0.067
55-65	60	7	0.233	9	0.300
65-75	70	9	0.300	18	0.600
75-85	80	7	0.233	25	0.833
85-95	90	3	0.100	28	0.933
95-105	100	2	0.067	30	1.000



ESTADÍSTICOS DE POSICIÓN.

Podríamos dividir los datos en la cantidad de partes que queramos. Sin embargo, tradicionalmente se dividen en 4 (cuartiles), 10 (deciles) o 100 (percentiles):

- Los **cuartiles** permiten dividir un conjunto de datos en 4 partes con frecuencias similares: primer cuartil $Q_1 = C_{0.25}$, segundo cuartil $Q_2 = C_{0.5}$, y tercer cuartil $Q_3 = C_{0.75}$.
- Los **deciles** en diez: $C_{0.1}, C_{0.2}, C_{0.3}, C_{0.4}, C_{0.5}, C_{0.6}, C_{0.7}, C_{0.8}, C_{0.9}$.
- Los **percentiles** en 100, y suele expresarse en porcentajes **100 α** . Por ejemplo, $C_{0.25}$ es el percentil 25.

Pregunta. ¿Puede ser mayor el primer decil que el primer cuartil?



MÉTODOS ESTADÍSTICOS: GRÁFICOS

Los gráficos son una herramienta de resumen de la información contenida en los datos que permiten sacar conclusiones acerca de la muestra de un solo vistazo.

El tipo de gráficos adecuados dependerá del tipo de variable.

- Diagrama de barras (variables cualitativas o discretas con un número pequeño de posibles valores)
- Histograma (variables cuantitativas)
- Diagrama de sectores (variable cualitativas o discretas con un número pequeño de posibles valores)
- Diagramas de caja (variables cuantitativas)

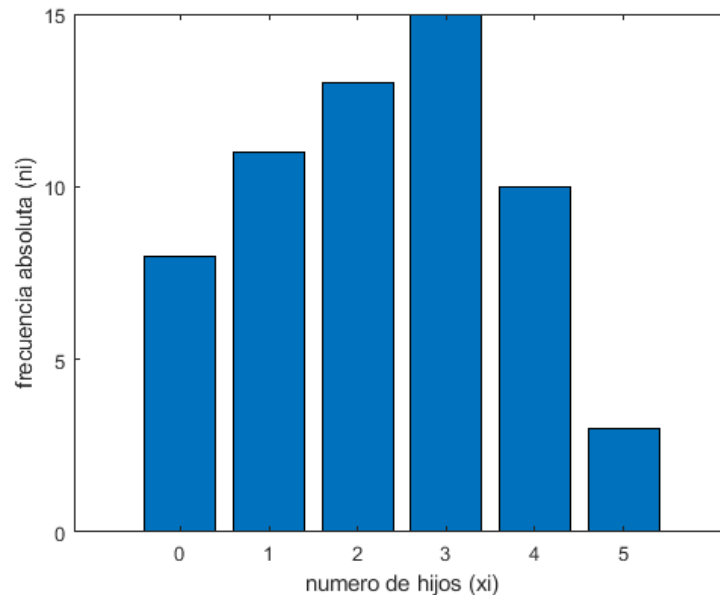


GRÁFICOS: DIAGRAMA DE BARRAS.

El **diagrama de barras** representa mediante barras la información contenida en la tabla de frecuencias. Gráfico apropiado para variables cuantitativas discretas con un número pequeño de posibles valores o variables cualitativas. Se construye a partir de la tabla de frecuencias (tratamiento individual).

Ejemplo. Diagrama de barras del número de hijos de las 60 familias.

Hijos	n_i
0	8
1	11
2	13
3	15
4	10
5	3



Matlab

```
datos = [0,2,2,4,0,3,3,2,5,2,3,2,...];  
[xi, ~, pos] = unique(datos);  
ni = hist(pos, length(xi));  
bar(xi,ni)  
xlabel('numero de hijos (xi)');  
ylabel('frecuencia absoluta (ni)');
```



GRÁFICOS: DIAGRAMA DE BARRAS E HISTOGRAMA.

Tabla de frecuencias (tratamiento individual)

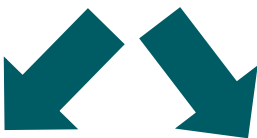


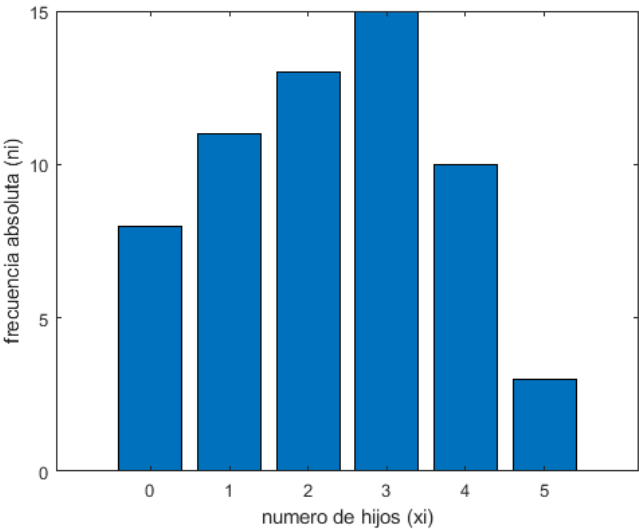
Tabla de frecuencias (tratamiento de clase)

Encuesta a 60 familias sobre el número de hijos:

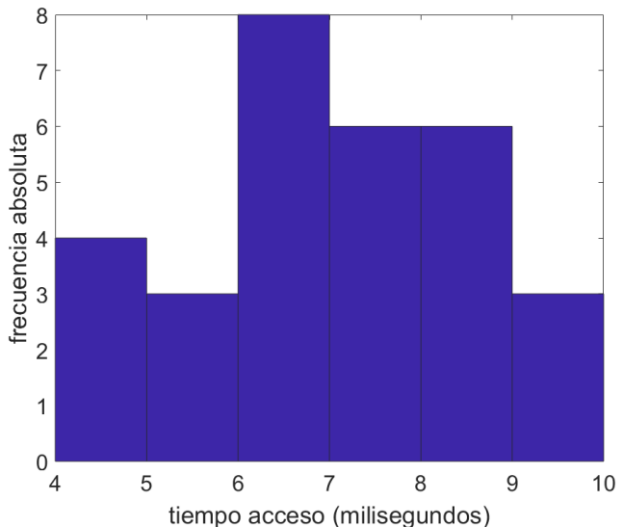
Hijos	n_i	f_i	F_i	N_i
0	8	0.13	0.13	8
1	11	0.18	0.31	19
2	13	0.22	0.53	32
3	15	0.25	0.78	47
4	10	0.17	0.95	57
5	3	0.05	1	60

$[L_{i-1}, L_i)$	x_i	n_i	N_i	f_i	F_i
4-5	4.5	4	4	0.133	0.133
5-6	5.5	3	7	0.100	0.233
6-7	6.5	8	15	0.267	0.500
7-8	7.5	6	21	0.200	0.700
8-9	8.5	6	27	0.200	0.900
9-10	9.5	3	30	0.100	1.000
Total		30		1.000	

Diagrama de barras



Histograma



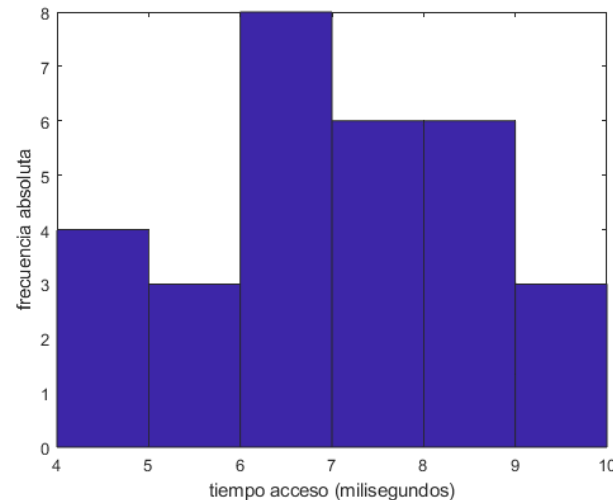
GRÁFICOS: HISTOGRAMA.

El **histograma de frecuencias** muestra la distribución de una serie de datos de variables cuantitativas agrupadas en intervalos de clase. Se construye a partir de la tabla de frecuencias (tratamiento de clase).

Se trata de un gráfico de barras verticales en el que el ancho de cada barra corresponde con la amplitud del intervalo mientras que la altura representa la frecuencia absoluta (n_i) o relativa (f_i).

Ejemplo. Histograma de frecuencias absolutas asociado 30 medidas del tiempo de acceso al disco duro.

$[L_{i-1}, L_i)$	x_i	n_i
4-5	4.5	4
5-6	5.5	3
6-7	6.5	8
7-8	7.5	6
8-9	8.5	6
9-10	9.5	3
Total		30



Matlab

```
datos = [6.2,7.2,5.3,4.4,6.3,9.1,...];  
n = length(datos);  
c = 1+log10(n)/log10(2);  
xi = 4.5:1:9.5;  
hist(datos, xi)  
xlabel("Tiempo (ms)")  
ylabel("ni")
```

[enlace
datos](#)



GRÁFICOS: DIAGRAMA DE SECTORES.

El **diagrama de sectores** (o gráfico de tarta, en inglés **pie chart**) es una representación en la que cada sector del círculo tiene un ángulo directamente proporcional a la frecuencia relativa de cada posible valor de la variable. A lado de cada sector se suele indicar el nombre de la clase correspondiente. Apropiado para variables cualitativas o cuantitativas discretas con un número pequeño de posibles valores.

Encuesta a 60 familias sobre el

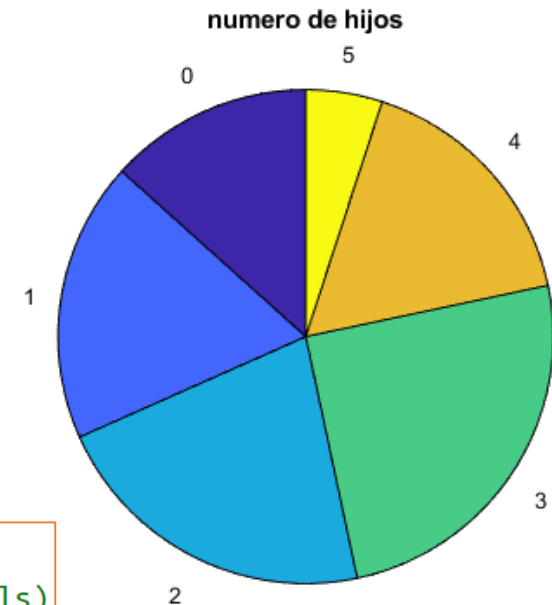
Hijos	n_i	f_i	F_i	N_i
0	8	0.13	0.13	8
1	11	0.18	0.31	19
2	13	0.22	0.53	32
3	15	0.25	0.78	47
4	10	0.17	0.95	57
5	3	0.05	1	60

$$\alpha_i = f_i \cdot 360 \text{ grados}$$



$$1 \text{ hijo: } 0.18 \times 360 = 65^\circ$$

```
figure
pie(ni,["0","1","2","3","4","5"]);%pie(ni,labels)
title('numero de hijos');
```



GRÁFICOS: DIAGRAMA DE CAJA.

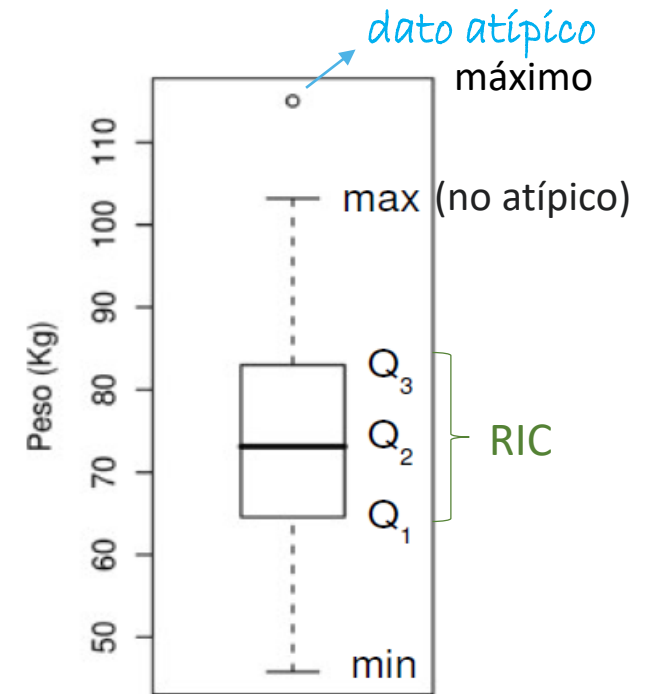
El **diagrama de caja** (o **boxplot** en inglés) resume gráficamente los datos de una variable cuantitativa a través de la representación estadísticos de posición.

El rango intercuartílico (RIC) define la anchura de la caja

$$RIC = Q_3 - Q_1 = C_{0.75} - C_{0.25}$$

Los datos que son menores que $C_{0.25} - 1.5RIC$ o mayores que $C_{0.75} + 1.5RIC$ se les denomina **datos atípicos** (en inglés *outliers*). Se representan con marcas individuales.

Los «**bigotes**» se extiende hasta los datos más extremos que no sean atípicos.

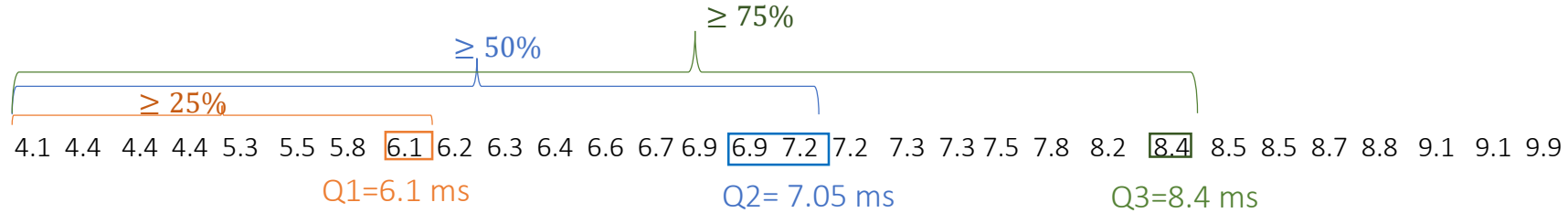


Pregunta. ¿Qué porcentaje de datos se encuentran dentro de la caja (delimitada por Q_1 y Q_3) del diagrama de caja?



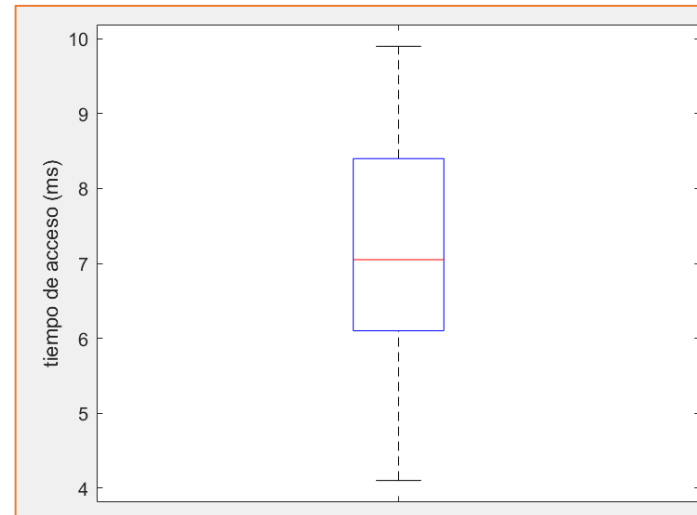
GRÁFICOS: DIAGRAMA DE CAJA.

Ejercicio. Realiza el diagrama de cajas de los datos de acceso al disco duro.

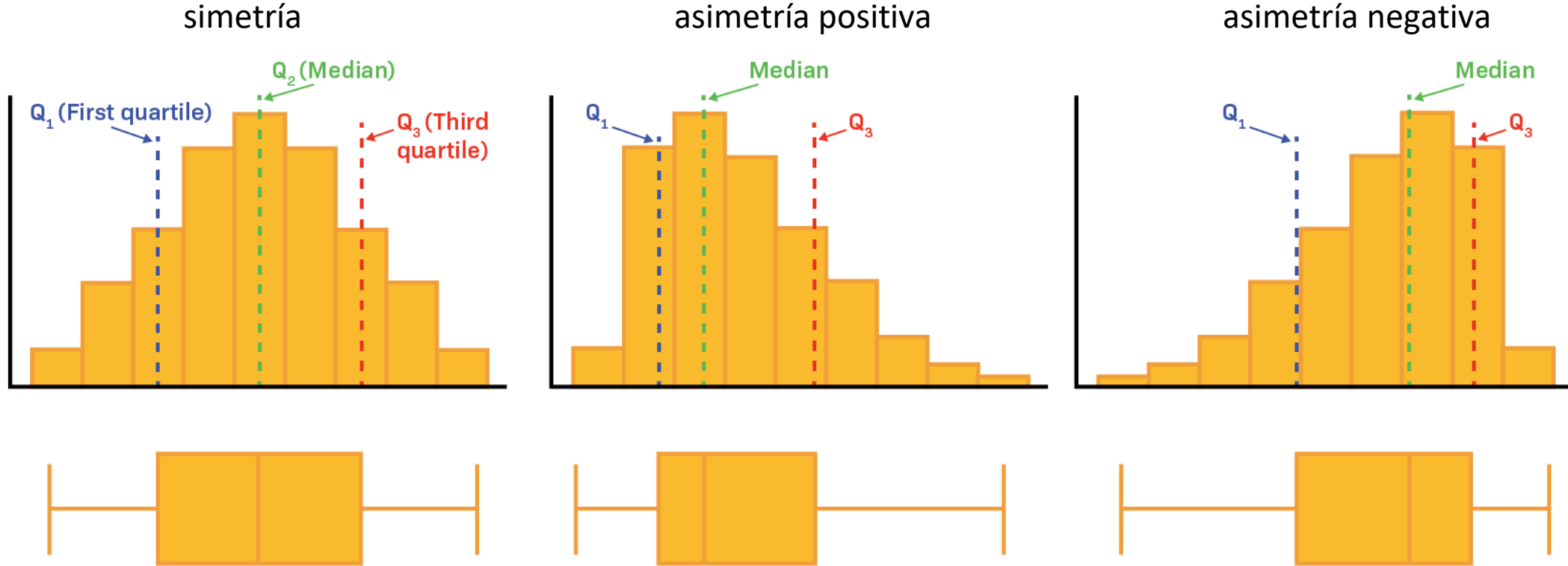


Matlab

```
%diagrama de caja  
figure;  
boxplot(datos)  
ylabel('tiempo de acceso (ms)')
```



GRÁFICOS: INTERPRETACIÓN DE LA ASIMETRÍA.



Además del coeficiente de asimetría de Fisher ya estudiado

$\left(c_A = \frac{m_3}{s_n^3}\right)$, el coeficiente de asimetría de Bowley se obtiene a partir de los cuartiles

$$C_{Bowley} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

