

GRADO EN DESARROLLO DE VIDEOJUEGOS

Asignatura: Probabilidad y Estadística

ECTS 6

Profesor: Óscar Vadillo Muñoz

Despacho: 321

Email: ovadillo@ucm.es

Campus virtual

HORARIO

Aula 5

Hora	Lunes	Martes	Miércoles	Jueves	Viernes
17:00-18:00					CLASE
18:00-19:00			CLASE	CLASE	CLASE

Las tutorías se realizarán a través del campus virtual, foro de dudas o presencialmente en el despacho previa solicitud de cita

Las transparencias NO sustituyen en ningún caso la información dada en clase, son material de apoyo para la realización de la clase y se facilitarán al final de cada tema.

TEMARIO

1. Estadística descriptiva unidimensional

(Software estadísticos. R)

1. E.D. Bidimensional. Regresión y correlación

2. Probabilidad

3. Variables aleatorias y Modelos de probabilidad

4. Muestreo y Teorema Central del Límite

5. Inferencia. Estimación, Intervalos de confianza y contrastes de hipótesis.

Bibliografía

- 1.- Álvarez Cáceres, Rafael: “Estadística Aplicada a Ciencias de la Salud”, Díaz de Santos, 2007.
- 2.- García, A. y otros. Estadística I. UNED 1995.
- 3.- Rodríguez L. y Tomeo V. Métodos Estadísticos para Ingeniería. Garceta Grupo Editorial. 2011
- 4.- Spiegel, M.R., Schiler, J. Srinivasan, R.A. Probabilidad y Estadística. Mc Graw Hill. 2001.
- 5.-Devore, J.L. Probabilidad y Estadística para Ingeniería y Ciencias. Thompson Learning. 2001

SISTEMA DE EVALUACIÓN

La nota final se compone de un **80% un examen final que es necesario aprobar y otras actividades 20%. (Realización de trabajos, entregas de ejercicios y/o exámenes parciales).**

El examen final es común para todos los grupos y consistirá en la resolución de un **máximo de 5 problemas.**

Para el examen final los alumnos podrán traer las tablas y la hoja de contrastes e intervalos de confianza disponibles en el campus virtual y un formulario que se les facilitará.

La duración del examen será de un **máximo 150 minutos***

Resultados del aprendizaje previstos

- Conocimiento de las técnicas de tratamiento y análisis de datos mediante cálculos estadísticos.
- Aplicación de los modelos básicos de regresión.
- Conocimiento y aplicación de las distribuciones de probabilidad más usuales.
- Aplicación de métodos para el contraste de hipótesis estadísticas.
- Interpretar resultados de análisis realizados con paquetes estadísticos.

*“Algún día será tan necesario conocer el
razonamiento estadístico como leer y escribir”*

H.G. Wells

¿Qué es la estadística?

1. *Es la ciencia que se encarga de recoger, organizar e interpretar datos.*
2. *Rama de la matemática que utiliza grandes conjuntos de datos numéricos para obtener inferencias basadas en el cálculo de probabilidades*

objetivo: obtener conclusiones de la investigación empírica usando modelos matemáticos.

Para qué sirve la Estadística

- Para ayudar a entender el resto de asignaturas
- Herramienta de futuro en la profesión
- Acceso a literatura científica
- Herramienta básica para cuantificar la incertidumbre
→ ciencias no exactas

La estadística constituye un instrumento esencial para la adquisición de conocimiento mediante la investigación científica

Para qué sirve la Estadística

Análisis de muestras

Descripción de datos

Contrastes de hipótesis

Medición de relaciones entre variables

Predicción

En general:

Dar respuesta a preguntas concretas sobre el comportamiento de conjuntos muy amplios o inaccesibles de individuos

algunos ejemplos de estas preguntas

¿Qué proporción de ciudadanos votaría a un determinado partido político, si hubiese hoy elecciones?

¿Qué porcentaje de españoles gasta más del 50% del presupuesto familiar en la adquisición de su vivienda?

¿Cumple una clase de cemento las especificaciones de una norma ISO?

¿Qué tipo de procesador es más eficiente?

¿Es cierto que las bombillas de larga duración duran hasta 8 veces más?

¿Cómo es la satisfacción de un usuario con un videojuego?

MÉTODO CIENTÍFICO

¿Qué es el método científico?

-
- El método científico es un proceso que tiene como finalidad establecer relaciones entre hechos para enunciar leyes y teorías que expliquen y fundamenten el funcionamiento del mundo.
 - Es un sistema riguroso que cuenta con una serie de pasos y cuyo fin es generar conocimiento científico a través de la comprobación empírica de fenómenos y hechos. En el método científico se utiliza la observación para proponer una hipótesis que luego se intenta comprobar a través de la experimentación.

CARACTERÍSTICAS DEL MÉTODO CIENTÍFICO

1. **Empirismo:** todo conocimiento adquirido a través del método científico debe ser **empírico, debe ser real y objetivo, adquirido mediante la observación** directa del hecho o a través de sus efectos relativamente constantes.
2. **Publicidad:** cualquier científico bajo las condiciones oportunas ha de poder realizar estas observaciones.
3. **Replicabilidad:** estas observaciones empíricas **deben ser susceptibles de volver a ser obtenidas** en cualquier momento, repitiendo una experiencia similar en la que se mantengan las condiciones bajo las que fueron obtenidas.

LAS FASES DE LA INVESTIGACIÓN CIENTÍFICA

Investigación científica: secuencia de actividades

- 1.- Definición del problema
- 2.- Deducción de hipótesis contrastables
- 3.- Establecimiento de un procedimiento de recogida de datos
- 4.- Análisis de los resultados y búsqueda de conclusiones
- 5.- Elaboración de un informe

CRITERIOS DE CALIDAD

FIABILIDAD EN LAS MEDIDAS: La aplicación de los instrumentos de recogida de datos nos aportan siempre la misma información.

VALIDEZ INTERNA: Grado en el que los cambios observados en la VD pueden ser atribuidos únicamente a los cambios introducidos en la VI.

VALIDEZ EXTERNA: Grado en el que los resultados de un experimento son generalizables a otros sujetos o contextos.

VALIDEZ DE CONSTRUCTO: Grado en el que lo que medimos es lo que realmente queremos medir.

VALIDEZ DE LA CONCLUSIÓN ESTADÍSTICA: Grado de confianza en que la conclusión estadística es o no correcta.

Tema 1

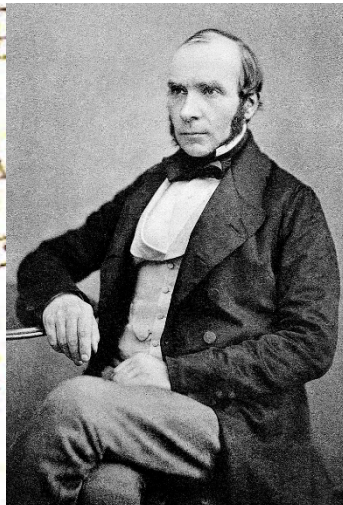
Estadística Descriptiva

“No confíes en lo que la estadística te dice hasta haber considerado con cuidado qué es lo que no dice.”

- *William W. Watt*

“Es un error capital teorizar antes de poseer datos. Insensiblemente uno comienza a alterar los hechos para encajarlos en las teorías, en lugar de encajar las teorías en los hechos”

- *Sherlock Holmes (A.C. Doyle)*



John Snow



Serie	Placa	$\overset{\circ}{A}$								
1	1	90.00	92.20	94.90	92.70	91.6	88.20	92.00	98.20	96.00
1	2	91.80	94.50	93.90	77.30	92.0	89.90	87.90	92.80	93.30
1	3	90.30	91.10	93.30	93.50	87.2	88.10	90.10	91.90	94.50
1	4	92.60	90.30	92.80	91.60	92.7	91.70	89.30	95.50	93.60
1	5	91.10	89.80	91.50	91.50	90.6	93.10	88.90	92.50	92.40
1	6	76.10	90.20	96.80	84.60	93.3	95.70	90.90	100.30	95.20
1	7	92.40	91.70	91.60	91.10	88.0	92.40	88.70	92.90	92.60
1	8	91.30	90.10	95.40	89.60	90.7	95.80	91.70	97.90	95.70
1	9	96.70	93.70	93.90	87.90	90.4	92.00	90.50	95.20	94.30
1	10	92.00	94.60	93.70	94.00	89.3	90.10	91.30	92.70	94.50
1	11	94.10	91.50	95.30	92.80	93.4	92.20	89.40	94.50	95.40
1	12	91.70	97.40	95.10	96.70	77.5	91.40	90.50	95.20	93.10
2	1	93.00	89.90	93.60	89.00	93.6	90.90	89.80	92.40	93.00
2	2	91.40	90.60	92.20	91.90	92.4	87.60	88.90	90.90	92.80
2	3	91.90	91.80	92.80	96.40	93.8	86.50	92.70	90.90	92.80
2	4	90.60	91.30	94.90	88.30	87.9	92.20	90.70	91.30	93.60
2	5	93.10	91.80	94.60	88.90	90.0	97.90	92.10	91.60	98.40
2	6	90.80	91.50	91.50	91.50	94.0	91.00	92.10	91.80	94.00
2	7	88.00	91.80	90.50	90.40	90.3	91.50	89.40	93.20	93.90
2	8	88.30	96.00	92.80	93.70	89.6	89.60	90.20	95.30	93.00
2	9	94.20	92.20	95.80	92.50	91.0	91.40	92.80	93.60	91.00
2	10	101.50	103.10	103.20	103.50	96.1	102.50	102.00	106.70	105.40
2	11	92.80	90.80	92.20	91.70	89.0	88.50	87.50	93.80	91.40
2	12	92.10	93.40	94.00	94.70	90.8	92.10	91.20	92.30	91.10

fecha informe	hospitalizados	uci	fallecidos	curados
30/03/2020	14917	1460	3392	8301
31/03/2020	15140	1514	3603	9330
01/04/2020	15227	1514	3865	10827
02/04/2020	15227	1528	4175	12400
03/04/2020	15050	1506	4483	13850
04/04/2020	14741	1498	4723	15362
05/04/2020	14551	1499	4941	16543
06/04/2020	14501	1510	5136	17322
07/04/2020	13950	1494	5371	18410
08/04/2020	13289	1450	5586	19836
09/04/2020	12853	1433	5800	21121
10/04/2020	12432	1399	5972	22412
11/04/2020	11894	1376	6084	23636
12/04/2020	11424	1332	6278	24683
13/04/2020	11233	1327	6423	25385
14/04/2020	10753	1299	6568	26247
15/04/2020	10116	1244	6724	27433
16/04/2020	9653	1206	6877	28491
17/04/2020	9141	1154	7007	29436
18/04/2020	8597	1094	7132	30475
19/04/2020	8291	1123	7239	31313
20/04/2020	8191	1111	7351	31762
21/04/2020	7930	1076	7460	32277
22/04/2020	7464	1024	7577	33032
23/04/2020	7077	981	7684	33645
24/04/2020	6601	949	7765	34212
25/04/2020	6183	892	7848	34902
26/04/2020	5892	873	7922	35367
27/04/2020	5820	855	7986	35565
28/04/2020	5484	815	8048	35841
29/04/2020	5171	773	8105	36314
30/04/2020	4945	727	8176	36707
01/05/2020	4546	706	8222	37154
02/05/2020	4270	681	8292	37530
03/05/2020	4221	674	8332	37704
04/05/2020	4228	660	8376	37808
05/05/2020	4103	640	8420	38002

Objetivos de la Estadística descriptiva

Las técnicas de la estadística descriptiva y del análisis exploratorio de datos tienen como objetivo **ordenar** los datos, en base a obtener el **máximo** de información, y a orientar la investigación. para ello se usan herramientas tales como:

1. Tablas
2. Gráficos: Diagramas de barras, histogramas, diagramas de cajas,...
3. Medidas numéricas:
 - ❖ De centralización: Media, mediana, moda . . .
 - ❖ De dispersión: Rango, varianza, desviación típica, . . .
 - ❖ Otros índices: Percentiles, asimetría, curtosis, . .

VARIABLES ESTADISTICAS

- **Población**: conjunto completo de elementos, con alguna característica común, objeto del estudio estadístico.

Ejemplo: Estudiantes de primero de segundo de farmacia

Puede ser finita

Ejemplo: Población española

Puede ser infinita (también cuando es muy grande)

Ejemplo: estrellas, medidas de la velocidad de la luz, medidas de la resistividad de un material

- **Individuo**: Cada uno de los elementos de la población

- **Muestra**: subconjunto de elementos de la población. Conjunto de individuos seleccionados para la obtención de datos.

Ejemplo: Estudiantes de 1ºC)

- **Tamaño**: número de elementos de la muestra

Tipología de los conjuntos de datos

Variables o atributos: Son las características que se pueden estudiar u observar en los individuos de la población. Dato o carácter objeto de estudio de los elementos de la muestra y que puede tomar un conjunto de valores.

El tipo de datos, así como el problema que originó su recogida, condiciona la clase de análisis estadístico que conviene realizar. Los conjuntos de datos, de manera general, se clasifican como:

- ❖ **Datos cualitativos:** cada dato es una cualidad, como por ejemplo un color, un estado civil, una posición, . . .
- ❖ **Datos numéricos o cuantitativos :** cada dato es un número.

Variable cuantitativa:

- **Datos discretos:** sólo pueden tomar valores en un conjunto asimilable a un subconjunto de los números enteros. Por ejemplo:
Número de hijos de una persona, número de veces que alguien va al cine al cabo de un año, . . .
- **Datos continuos:** pueden tomar cualquier valor en un rango.
Por ejemplo:
Resistencia de un material, duración de un aparato, altura h

Variable cualitativa:

- **Datos ordinales:** admiten ordenación lógica
Por ejemplo: *gravedad, estado, etc.*
- **Datos nominales:** No admiten ninguna ordenación lógica
Por ejemplo: *Sexo, grupo sanguíneo, etc.*

Cuantitativa: las diferencias entre los números, reflejan diferencias equivalentes en las magnitudes de la característica

```
graph TD; A[Cuantitativa: las diferencias entre los números, reflejan diferencias equivalentes en las magnitudes de la característica] --> B[DISCRETAS: entre dos valores consecutivos no es posible encontrar ningún otro valor]; A --> C[CONTINUAS: entre dos valores consecutivos, por muy próximos que estén, siempre es posible encontrar infinitos valores];
```

DISCRETAS: entre dos valores consecutivos no es posible encontrar ningún otro valor

CONTINUAS: entre dos valores consecutivos, por muy próximos que estén, siempre es posible encontrar infinitos valores

Ejercicio de ejemplo

Clasifica las siguientes variables:

Gravedad de un infarto (leve, moderado, fuerte), Número de ataques al servidor, Sexo, Presión arterial (mmHg), Estatura (cm), Peso, Estado de dolor tras la toma de un fármaco (Peor, Igual, Mejor), Provincia, Edad, Número de preguntas acertadas en un test, Grupo sanguíneo.

Distribución de frecuencias

El instrumento más utilizado para organizar datos es **la distribución de frecuencias**.

La distribución de frecuencias nos va a permitir:

Organizar los datos de una forma racional con el fin de poder extraer información de los mismos de forma rápida.

Nos va a facilitar los cálculos de los estadísticos que veremos en temas posteriores

Frecuencias I

Definiciones

❖ **La frecuencia absoluta** de un dato, n_i , es el número de veces que dicho dato se repite en el conjunto de la muestra.

❖ **La frecuencia relativa** de un dato, f_i , es el número de veces que dicho dato se repite en el conjunto de la muestra, comparado con el número total de datos, N ,

$$f_i = n_i / N$$

Frecuencias II

Definiciones

❖ **La frecuencia absoluta acumulada N_i** , Suma de las frecuencias absolutas de los valores *inferiores o iguales a x_i* .

$$N_i = N_{i-1} + n_i$$

❖ **La frecuencia relativa acumulada F_i** , es el cociente entre la frecuencia absoluta acumulada y el número de observaciones

$$F_i = N_i / N$$

El subíndice servirá para indicar la posición que ocupa un valor dentro de un conjunto de valores:

i adoptará valores de 1 (primer valor de la serie), **hasta k** (enésimo valor o último valor de la serie).

El gobierno desea averiguar si el número medio de hijos por familia ha descendido respecto de la década anterior. Para ello ha encuestado a 50 familias respecto al número de hijos, y ha obtenido los siguientes datos:

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

Población, variable, tipo de variable. Determinar la tabla de frecuencias.

La población objeto de estudio es el **conjunto de familias** de un determinado país.

Variable: **número de hijos** por familia

tipo de variable: discreta, ya que el número de hijos solo puede tomar determinados valores enteros (es imposible tener medio o un cuarto de hijo).

Para construir la tabla de frecuencias tenemos que ver cuantas familias tienen un determinado número de hijos. Podemos ver que el número de hijos, toma los valores existentes entre 0 hijos, los que menos y 6 hijos, los que más y tendremos:

xi	ni	Ni	fi	Fi
0	2	2	0.04	0.04
1	4	6	0.08	0.12
2	21	27	0.42	0.54
3	15	42	0.30	0.84
4	6	48	0.12	0.96
5	1	49	0.02	0.98
6	1	50	0.02	1
	N = 50		1	

Frecuencias

Frecuencia absoluta n_i : Número de veces que aparece repetido el valor x_j de la variable estadística X

$$0 \leq n_i \leq N \quad ; \quad \sum_{i=1}^k n_i = N$$

Frecuencia relativa n_i : Frecuencia absoluta dividida por el número de datos muestrales

$$f_i = \frac{n_i}{N} \quad 0 \leq f_i \leq 1 \quad \sum_{i=1}^k f_i = \sum_{i=1}^k \frac{n_i}{N} = 1$$

Frecuencia absoluta acumulada N_i : Frecuencia absoluta dividida por el número de datos muestrales

$$N_i = \sum_{j=1}^i n_j$$

$$N_i = N_{i-1} + n_i \quad ; N_1 = n_1 \quad ; N_k = N$$

Frecuencia relativa acumulada:

$$F_i = \frac{N_i}{N} = \frac{\sum_{j=1}^i n_j}{N} = \sum_{j=1}^i \frac{n_j}{N} = \sum_{j=1}^i f_j$$
$$F_K = 1$$

Distribuciones de Frecuencias

Tabla de frecuencias de una variable **cualitativa**

Categorías de la variable	Frecuencias absolutas n_i	Frecuencias relativas f_i
c_1	n_1	f_1
c_2	n_2	f_2
c_3	n_3	f_3
..	.	.
.	.	.
c_k	n_k	f_k

Categorías de la variable	Frecuencias absolutas n_i	Frecuencias relativas f_i
PARTIDO A	35	0,35
PARTICO B	20	0,20
PARTIDO C	12	0,12
INDECISO	7	0,07
NO VOTARA	26	0,26
TOTAL	100	1

Distribuciones de Frecuencias

Tabla de frecuencias de una variable **discreta**

Valores de la variable X_i	Frecuencias absolutas n_i	Frecuencias relativas f_i	Frecuencias absolutas acumuladas N_i	Frecuencias relativas acumuladas F_i
X_1	n_1	f_1	N_1	F_1
X_2	n_2	f_2	N_2	F_2
X_3	n_3	f_3	N_3	F_3
..
.
.
X_k	n_k	f_k	N_k	F_k

si $f_i=0.25$ esto quiere decir que la variable X_i se repite en un 25% de la muestra

Agrupamiento en intervalos de clase

Cuando la variable es continua o el rango demasiado grande

¿Cómo obtener, a partir de los datos, una tabla de frecuencias agrupada?

1. **Determinar el recorrido:** Valor mayor, menos valor menor de los datos. $Re = X_n - X_1$
2. **Decidir el nº de intervalos:** A partir de la raíz cuadrada del número de datos, redondeando el número de intervalos. Más de 5 y menos de 20
3. **División** entre el Recorrido y el número de intervalos que hayamos decidido. Se puede redondear también. $L = R/N$
4. **Determinar los extremos.** Evitar coincidencia con datos, añadimos una cifra decimal.
5. **Calcular las marcas de clase:** valor medio entre los límites inferior y superior de cada intervalo.

algunas observaciones

La tabla de frecuencias **no puede** presentar intervalos de clase con **frecuencia nula**,

Siempre que sea posible las clases deberán tener **la misma longitud**, con el fin de no enmascarar la realidad del fenómeno.

El **extremo superior** del último intervalo debe ser **mayor** que el mayor valor observado.

Cuando el dato más pequeño (grande) se encuentra muy alejado del resto, se dirá que se trata de una observación **anómala** o extrema, (**outlier**). $Q_1 \pm 1,5 * RI$

Los extremos de clase de los intervalos se deben definir con precisión, de forma que los **intervalos sean contiguos**, pero **no solapados**. $I_1 = [x_m, x_m + L)$

La agrupación de los datos en intervalos de clase supone una **pérdida de información**

Intervalos de clase $[a_{i-1}, a_i)$	Marcas de Clase c_i	Frecuencias absolutas n_i	Frecuencias relativas f_i	Frecuencias absolutas acumuladas N_i	Frecuencias relativas acumuladas F_i
$a_0 - a_1$	c_1	n_1	f_1	N_1	F_1
$a_1 - a_2$	c_2	n_2	f_2	N_2	F_2
$a_2 - a_3$	c_3	n_3	f_3	N_3	F_3
.
.
.
$A_{k-1} - a_k$	c_k	n_k	f_k	N_k	F_k

Ejemplo 2:

Un nuevo hotel va a abrir sus puertas en cierta ciudad. Antes de decidir el precio de sus habitaciones, el gerente investiga los precios por habitación de 40 hoteles de la misma categoría de esa ciudad. Los datos obtenidos en decenas de euros fueron

3,9	4,7	3,7	5,6	4,3	4,9	5,0	6,1	5,1	4,5
5,3	3,9	4,3	5,0	6,0	4,7	5,1	4,2	4,4	5,8
3,3	4,3	4,1	5,8	4,4	4,8	6,1	4,3	5,3	4,5
4,0	5,4	3,9	4,7	3,3	4,5	4,7	4,2	4,5	4,8

Población: hoteles de una ciudad.

Variable: precio.

Tipo de variable: continua.

Existen muchos valores diferentes por tanto es bueno agrupar la serie en intervalos.

La manera de hacerlo sería la siguiente: primero, calculamos el recorrido $Re = x_n - x_1 = 6.1 - 3.3 = 2.8$

Cuando no se nos dice nada el nº de intervalos, se obtiene calculando la raíz cuadrada del nº de datos observado. Veremos que la raíz cuadrada de 40 es igual a 6.32 por lo tanto tomaremos 6 intervalos.

Como el recorrido es 2.8 si lo dividimos por el nº de intervalos tendremos la amplitud de cada uno de ellos y así: $2,8/6 = 0,46$.

Importante: La amplitud es de 0,46 por lo que además de no ser muy fácil de operar, puede que no cubra el rango de la variable.
Lo podemos evitar, tomando un valor superior, en este caso 0,5:

$[a_{i-1}, a_i)$	ci	ni	Ni	fi	Fi
[3,25,3,75)	3,50	3	3	0.075	0.075
[3,75,4,25)	4,00	8	11	0.2	0.275
[4,25,4,75)	4,50	14	25	0.35	0.625
[4,75,5,25)	5,00	6	31	0.15	0.775
[5,25,5,75)	5,50	4	35	0.1	0.875
[5,75,6,25)	6,00	5	40	0.125	1
		N= 40		1	

Fallecidos por sexo en la CM

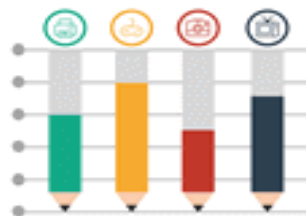
Edad	Mujeres	Hombres	Total
0-9	2	8	10
10-19	3	1	4
20-29	5	9	14
30-39	22	24	46
40-49	55	126	181
50-59	197	480	677
60-69	493	1.198	1.691
70-79	1.428	2.778	4.206
80-89	3.946	4.652	8.598
90>	3.942	2.400	6.342
Edad no confirmada*	134	152	286
Total general	10.227	11.828	22.055

*Sin fecha de nacimiento en el certificado de defunción

Algunas observaciones

- ❑ Cuando el volumen de datos es importante, las tablas pueden resultar confusas.
- ❑ Un gráfico es generalmente más intuitivo, aunque contenga la misma información que la tabla.
- ❑ Las tablas y los gráficos contienen menos información que el conjunto de datos.

LOREM IPSUM DOLOR SIT AMET



Lorem ipsum dolor sit amet, consectetur adipiscing elit.

60
Lorem ipsum dolor sit amet.

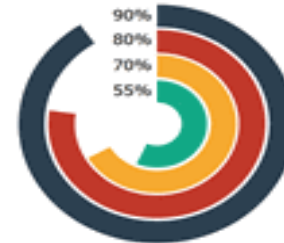
80
Lorem ipsum dolor sit amet.

50
Lorem ipsum dolor sit amet.

70
Lorem ipsum dolor sit amet.

LOREM IPSUM DOLOR SIT AMET

90%
Lorem ipsum dolor sit amet, consectetur adipiscing elit.



80%
Lorem ipsum dolor sit amet, consectetur adipiscing elit.

70%
Lorem ipsum dolor sit amet, consectetur adipiscing elit.

55%
Lorem ipsum dolor sit amet, consectetur adipiscing elit.

LOREM IPSUM DOLOR SIT AMET

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

10%
Donec ornare fringilla augue sed placerat. Vivamus convallis.



10%
Donec ornare fringilla augue sed placerat. Vivamus convallis.



15%
Donec ornare fringilla augue sed placerat. Vivamus convallis.



15%
Donec ornare fringilla augue sed placerat. Vivamus convallis.

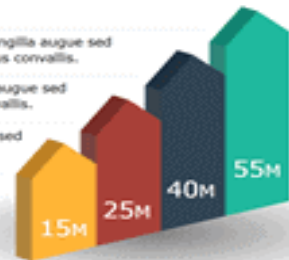


LOREM IPSUM DOLOR SIT AMET

15M
Donec ornare fringilla augue sed placerat. Vivamus convallis.

25M
Donec ornare fringilla augue sed placerat. Vivamus convallis.

40M
Donec ornare fringilla augue sed placerat. Vivamus convallis.



LOREM IPSUM DOLOR SIT AMET

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Lorem ipsum dolor sit amet, consectetur adipiscing elit.

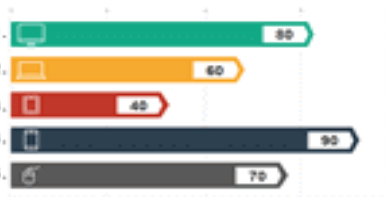
DATO 01. **80**

DATO 02. **60**

DATO 03. **40**

DATO 04. **90**

DATO 05. **70**



LOREM IPSUM DOLOR SIT AMET

INDICADOR 01.

INDICADOR 02.

75
50
65

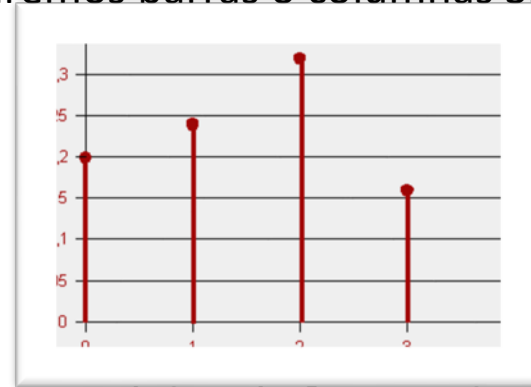


75
50
65

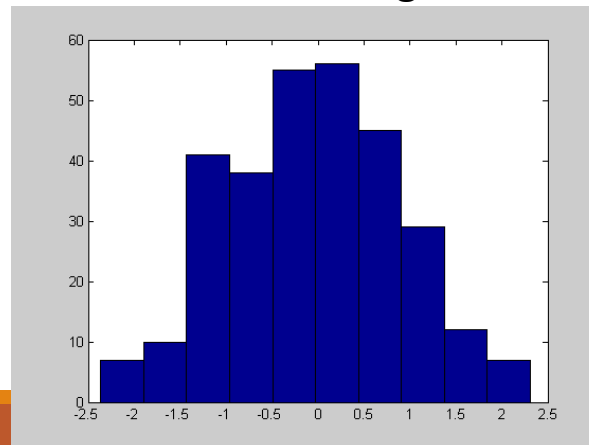
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Diagrama de Barras e Histograma

Diagrama de barras: se utiliza para frecuencias absolutas o relativas, acumuladas o no, de una **VARIABLE DISCRETA**. En el eje de abscisas, situaremos los diferentes valores de la variable. En el eje de ordenadas la frecuencia. Levantaremos barras o columnas **SEPARADAS** de altura correspondiente a la frecuencia adecuada.

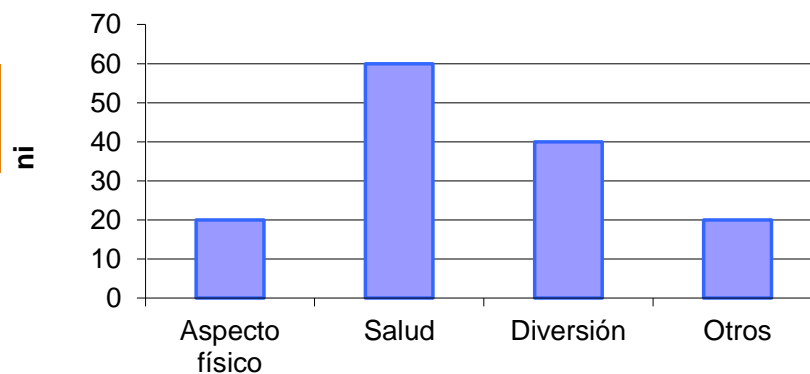


Histograma de frecuencias: Igual que el anterior en cuanto al tipo de frecuencias que se pueden utilizar. La diferencia : es para variables **CONTINUAS**. Si la amplitud del intervalo es la misma, elevaremos columnas **UNIDAS**, a altura la frecuencia correspondiente. Si la amplitud del intervalo es diferente, el área del rectángulo columna será proporcional a la frecuencia representada.



Representación gráfica de una variable. Histograma

Diagrama de rectángulos



Frecuencias absolutas
o relativas eje ordenadas

Variables:

nominales

ordinales

cuantitativas discretas.

Motivos para practicar deporte

Valores de la variable en el eje de abscisas

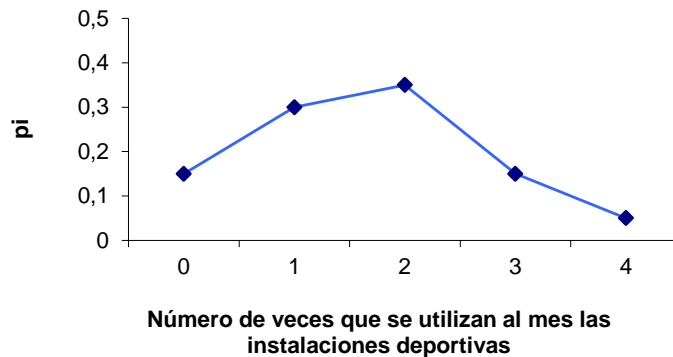
Representación gráfica de una variable

Variables:

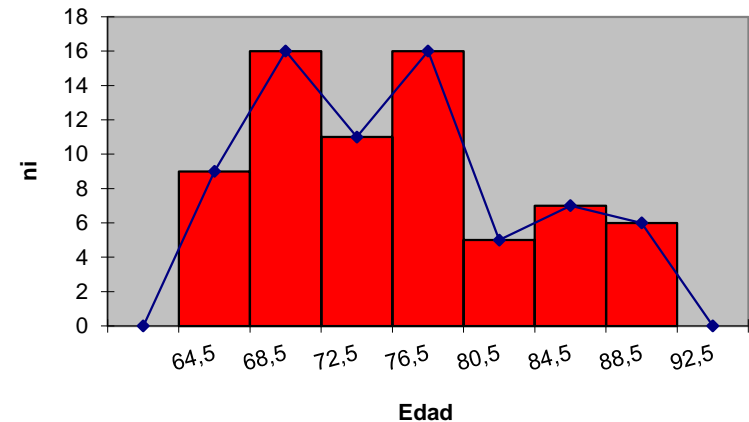
Cuantitativas discretas

Cuantitativas continuas

Polígono de frecuencias



Polígono para datos agrupados en intervalos

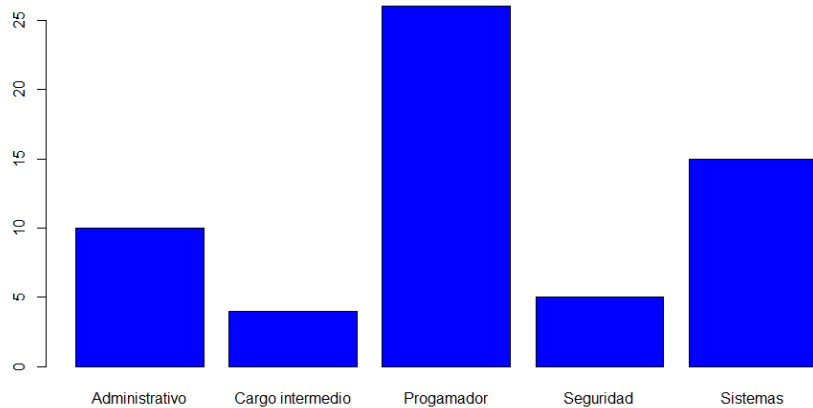


Discretas: unen puntos que resultan de levantar una línea vertical sobre cada valor de la variable, cuya altura corresponde a su frecuencia absoluta o relativa

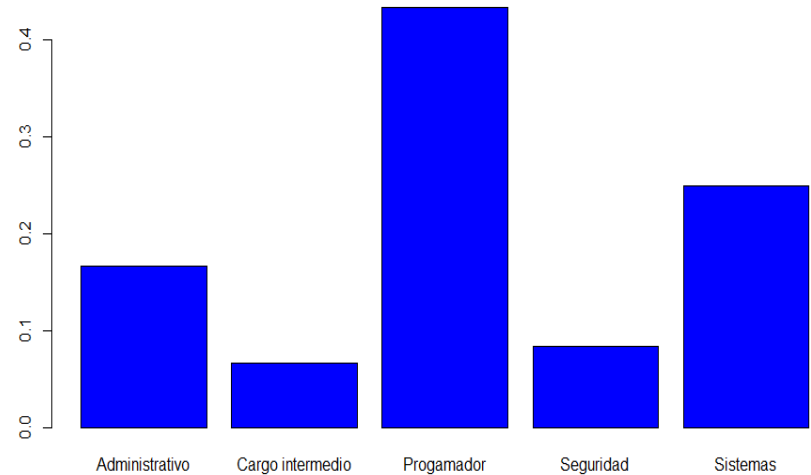
Continuas: unen los puntos medios de la base superior de los rectángulos del histograma

El diagrama de barras es un gráfico que representa sobre el horizontal las diferentes categorías o valores de la variable y para cada una de ellas levanta una barra de longitud igual a la frecuencia absoluta o relativa. De este modo la altura de las barra se puede apreciar sobre el eje y.

Con frecuencias absolutas



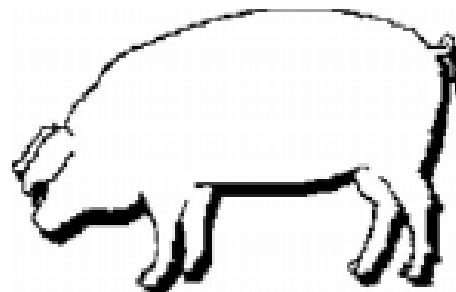
Con frecuencias relativas



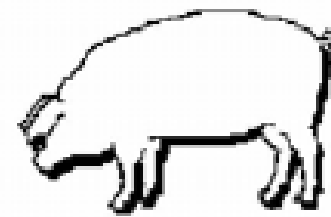
Gráficos para variables cualitativas

- **Pictogramas**

- Fáciles de entender.
- Cada modalidad debe ser proporcional a la frecuencia.



100 Kg

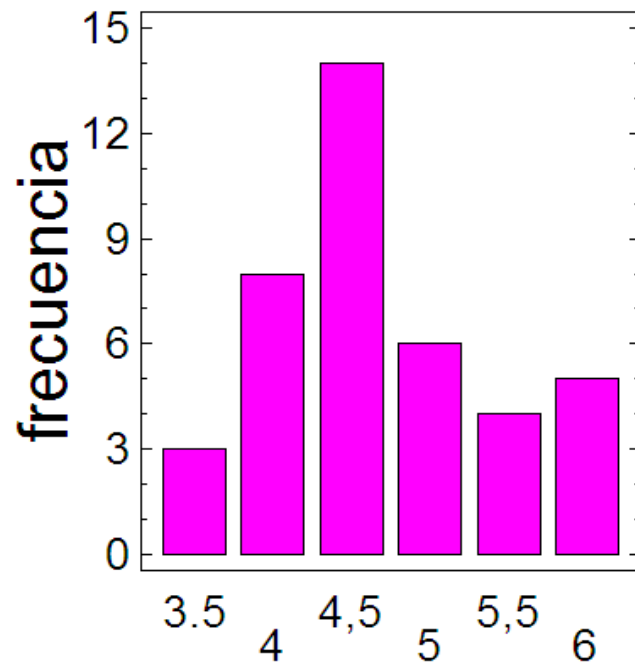


50Kg

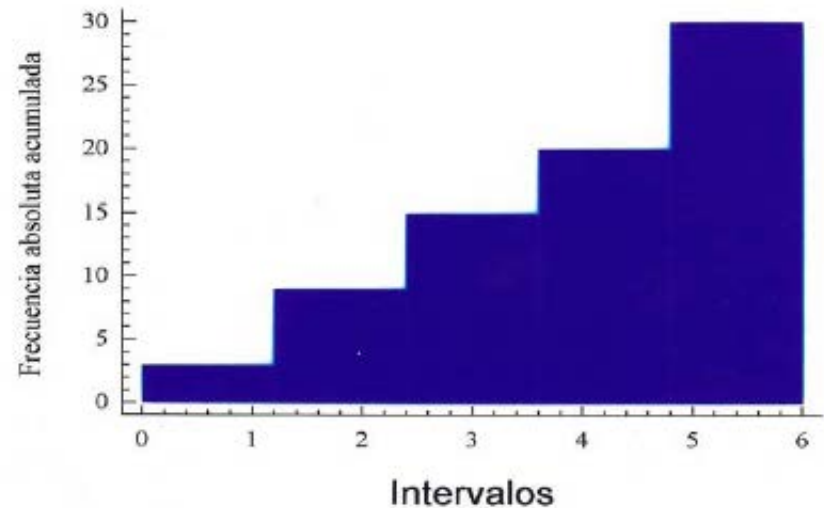
Histograma

La apariencia del histograma puede cambiar si se modifica el número de clases.

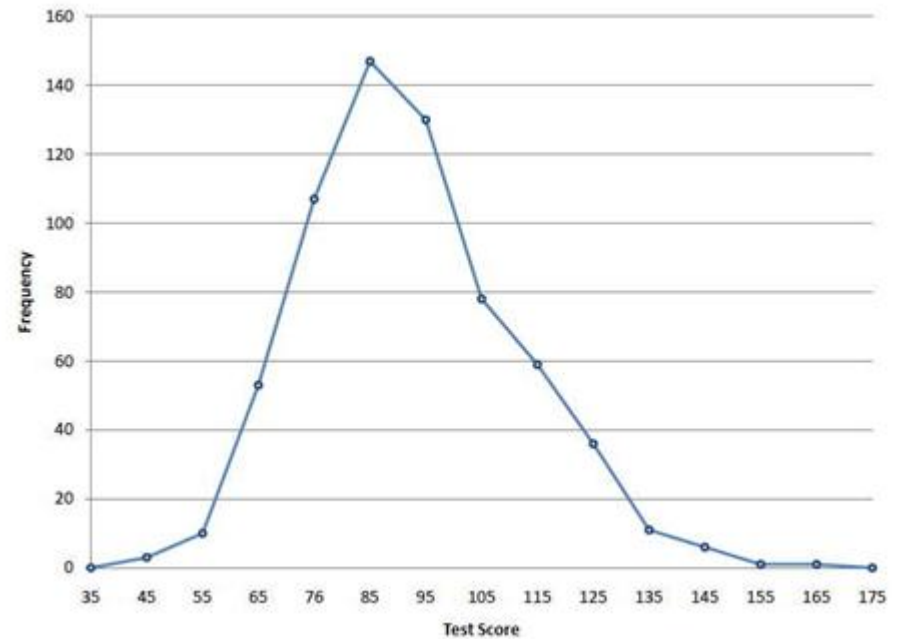
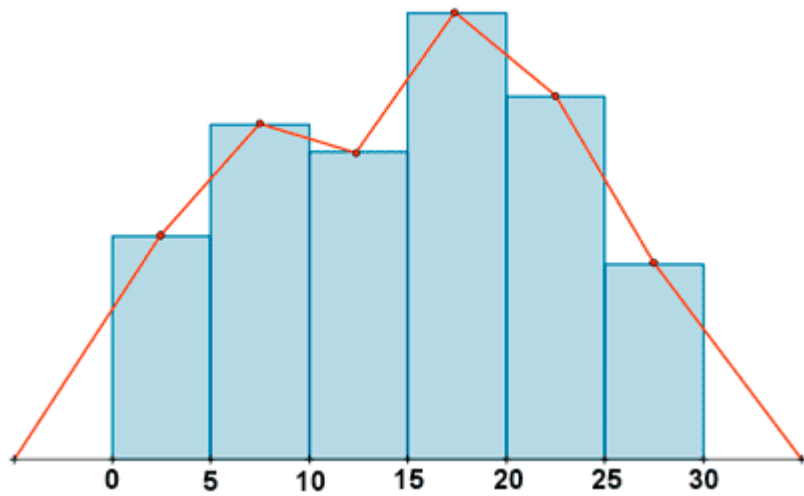
Precio hoteles



Histograma de frecuencias acumuladas



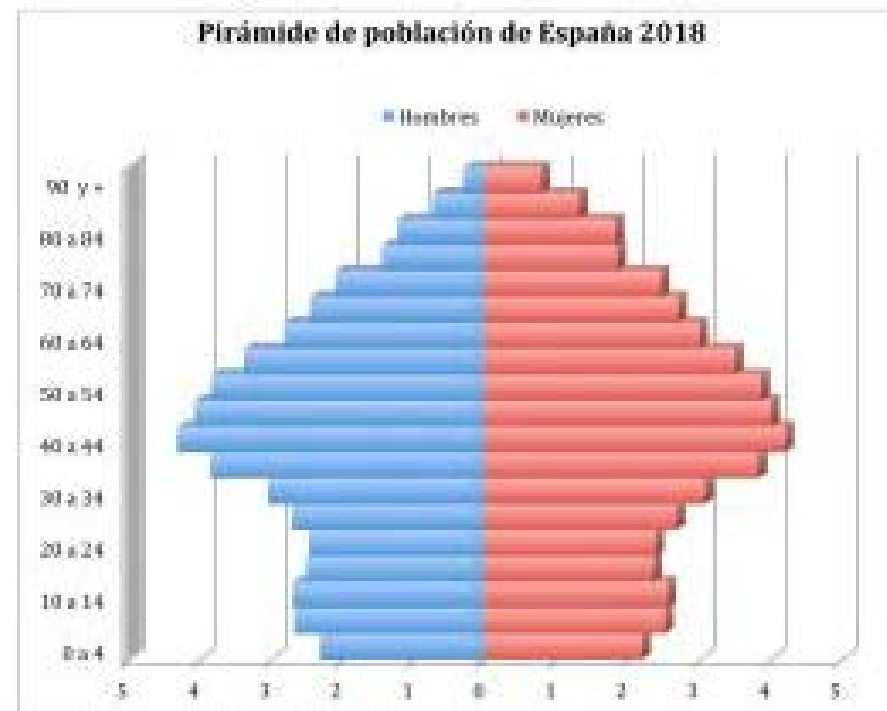
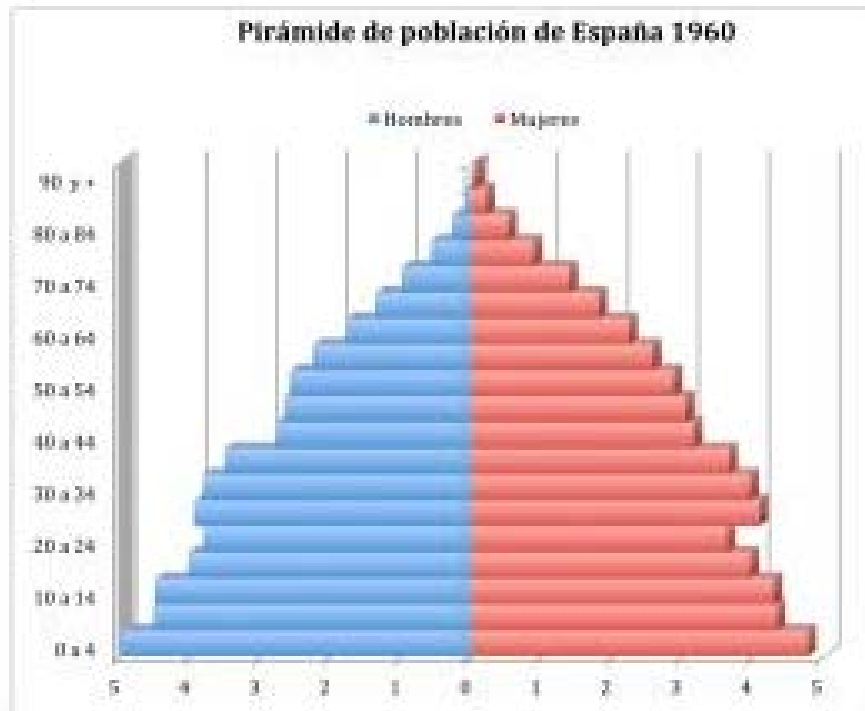
Polígono de frecuencia



Construcción de gráficos

- a) En el eje de **abscisas** se sitúan los **valores de la variable** y en el de **ordenadas** las **frecuencias**.
- b) En el eje de **abscisas** los **valores de la variable** se sitúan en **orden ascendente** de izquierda a derecha y en el de ordenadas de abajo a arriba.
- c) Si se representan dos variables en el mismo gráfico y los tamaños de las muestras son diferentes se **utilizarán** frecuencias o porcentajes relativos.
- d) Para la interpretación de una representación gráfica es imprescindible incluir en la misma cierta información, en concreto: variable representada en el eje de abscisas, tipo de frecuencia en el eje de ordenadas y grupo al que se corresponde cada gráfica en el caso de representar dos variables en una leyenda.
- e) Se puede cortar el eje de abscisas cuando los valores de la variable que aparecen en la muestra son muy grandes. **Conviene ser mucho más cautos a la hora de hacer lo mismo con el eje de ordenadas**, pues la representación gráfica podría dar una idea engañosa de los resultados obtenidos

Pirámide de población



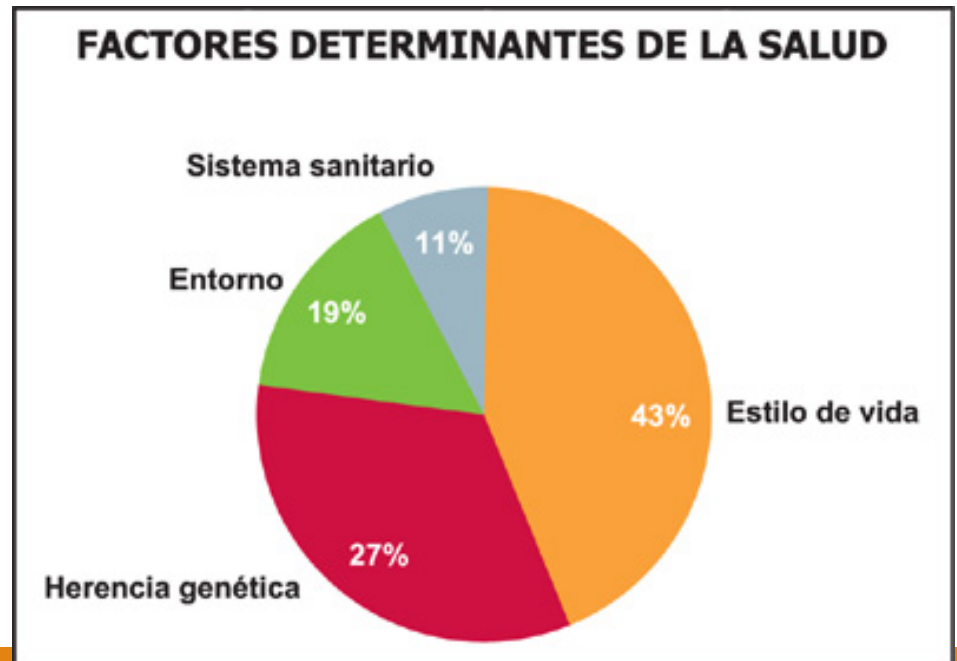
<https://www.populationpyramid.net/es/esp%C3%B1a/2019/>

Diagrama de sectores

El área de cada sector es proporcional a la frecuencia que se quiera representar, sea absoluta o relativa.

Para calcularlo podemos decir que el área depende del ángulo central, mediante la siguiente proporción:

$$n_i/N = \alpha/360 \rightarrow \alpha = f_i * 360$$

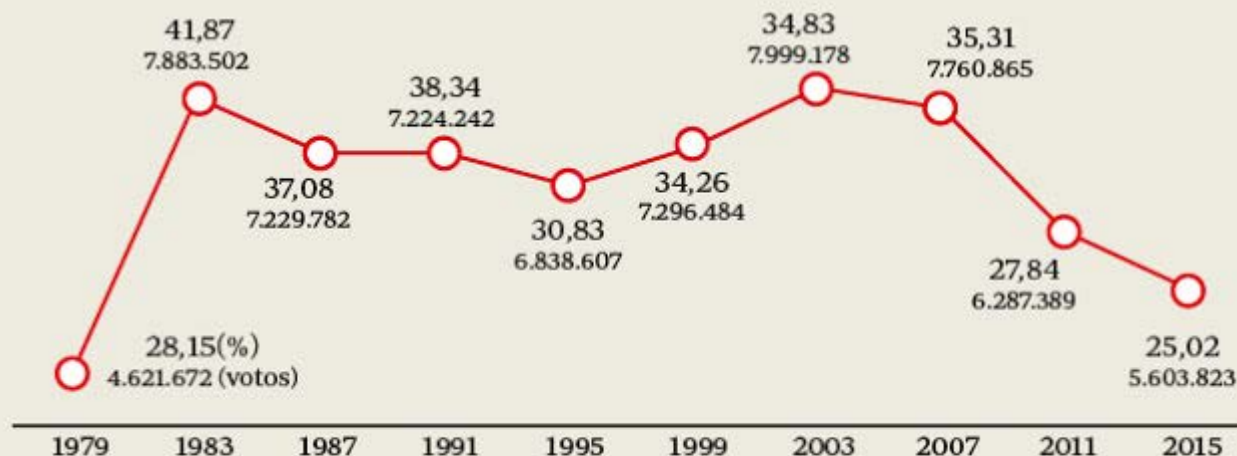


Ganas de empezar el cuatrimestre

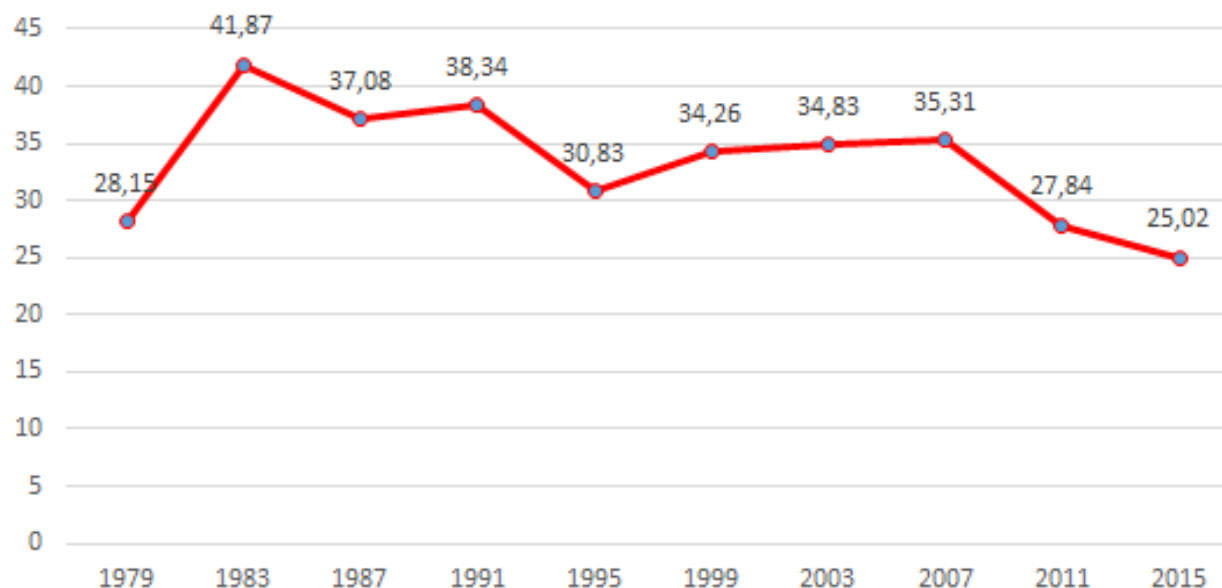


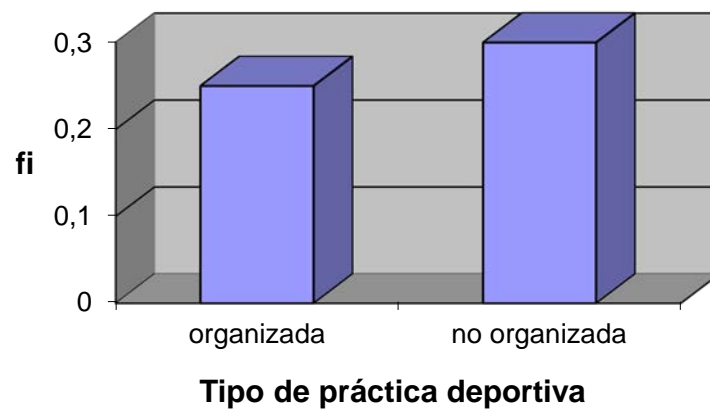
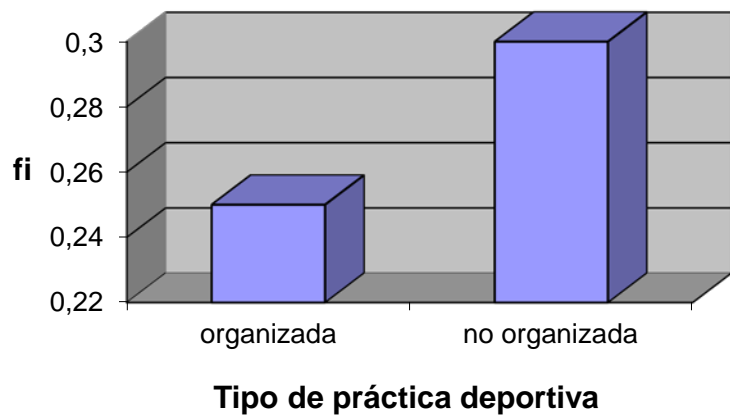
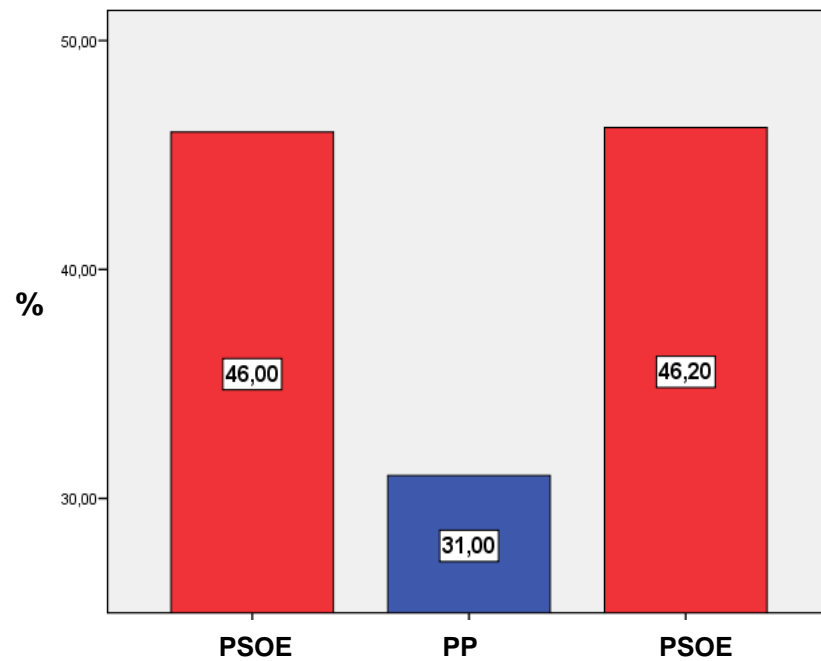
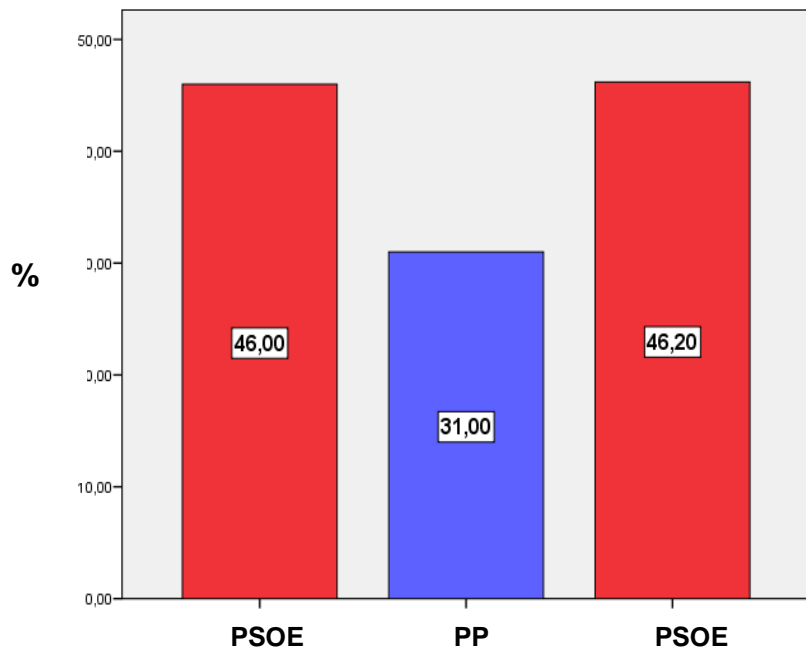


Pérdida de votos municipales



Voto PSOE municipales (%)





Un diagrama de **Pareto** es un diagrama de barras, en el que los datos aparecen ordenados por el valor de sus **frecuencias**.

Está basado en la relación 80/20.

Ventajas

- Permite centrarse en los aspectos cuya mejora tendrán más impacto, optimizando por tanto los esfuerzos.
- Proporciona una visión sencilla y rápida de la importancia relativa de los problemas.
- Ayuda a evitar que empeoren algunas causas al tratar de solucionar otras menos significativas.
- Su visión gráfica del análisis es fácil de comprender

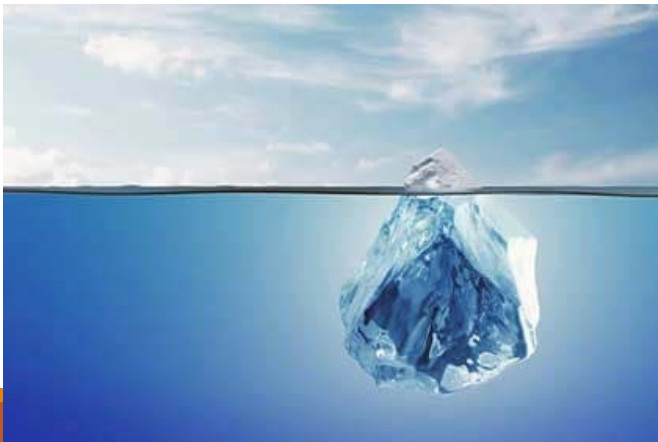


Diagrama de Pareto

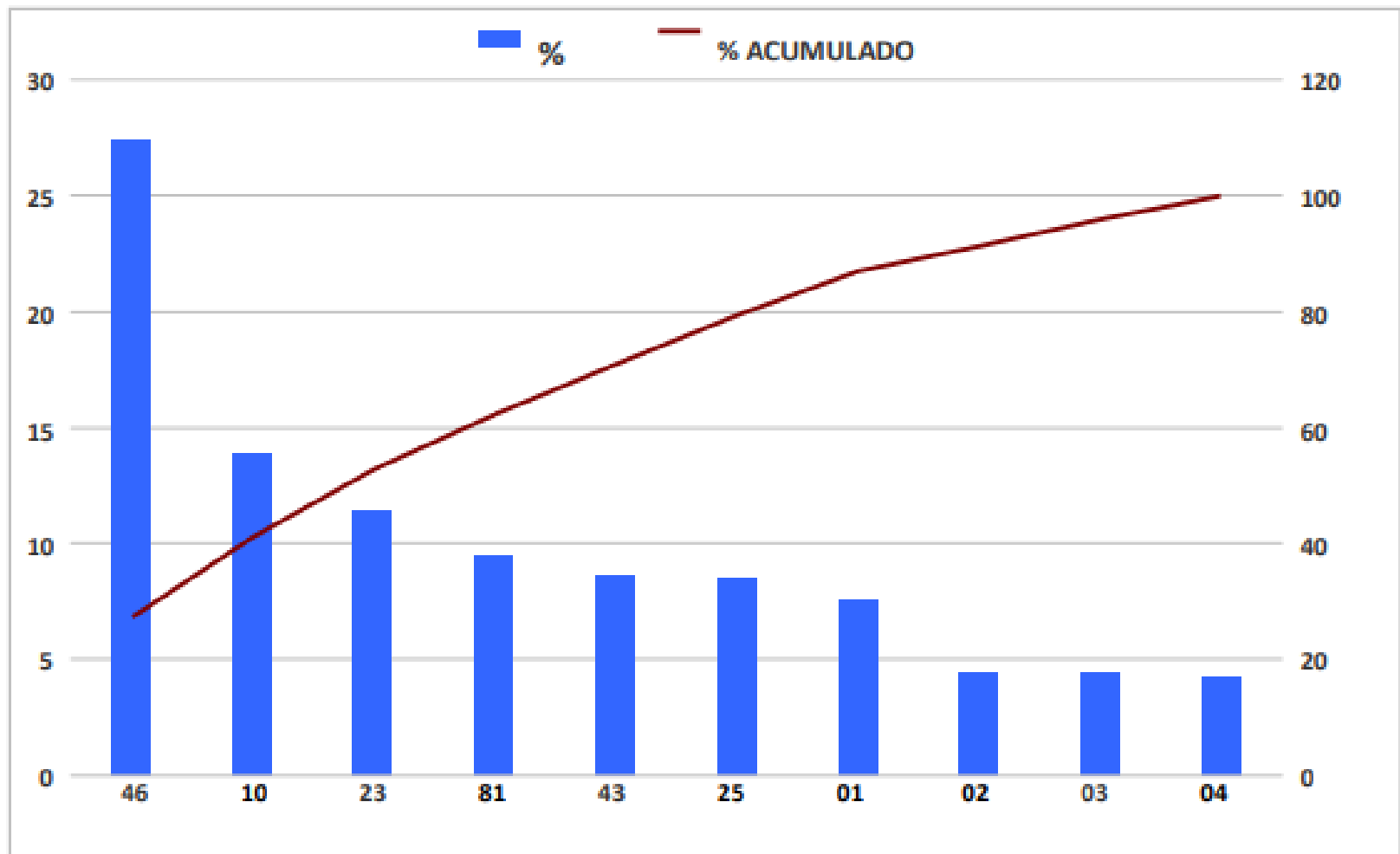
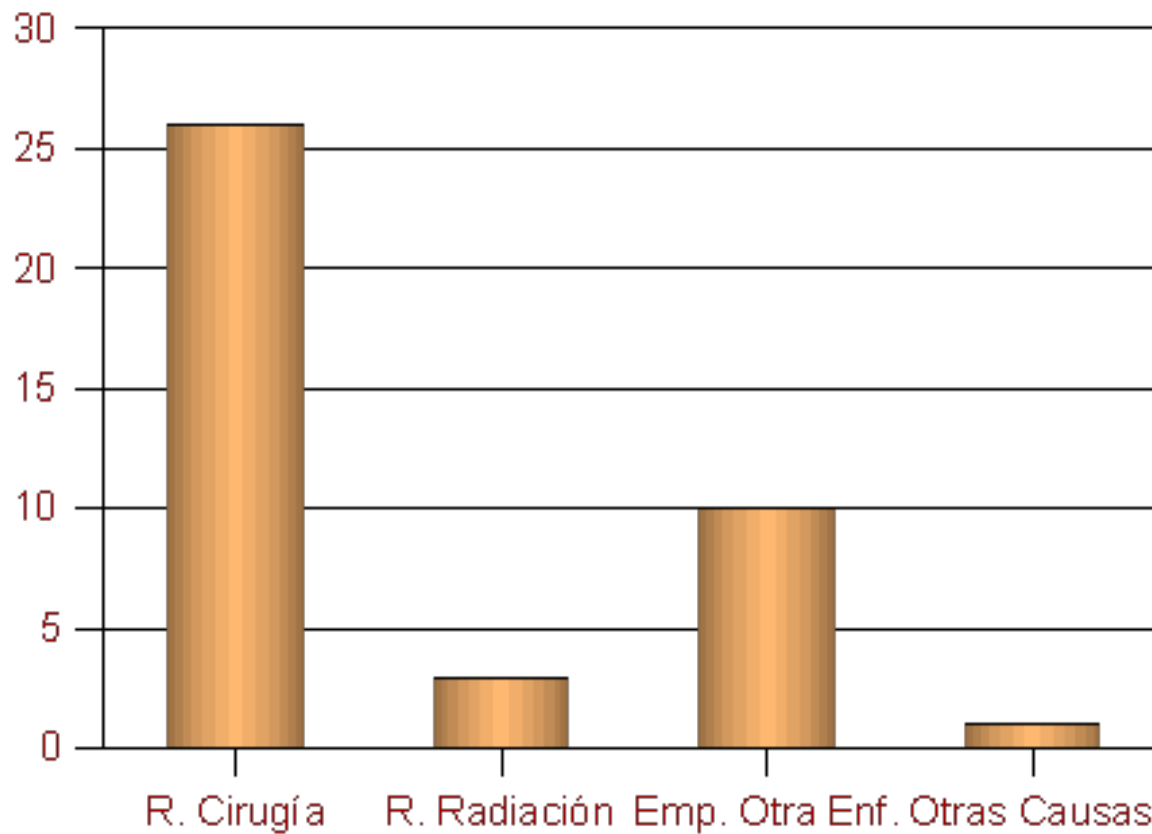


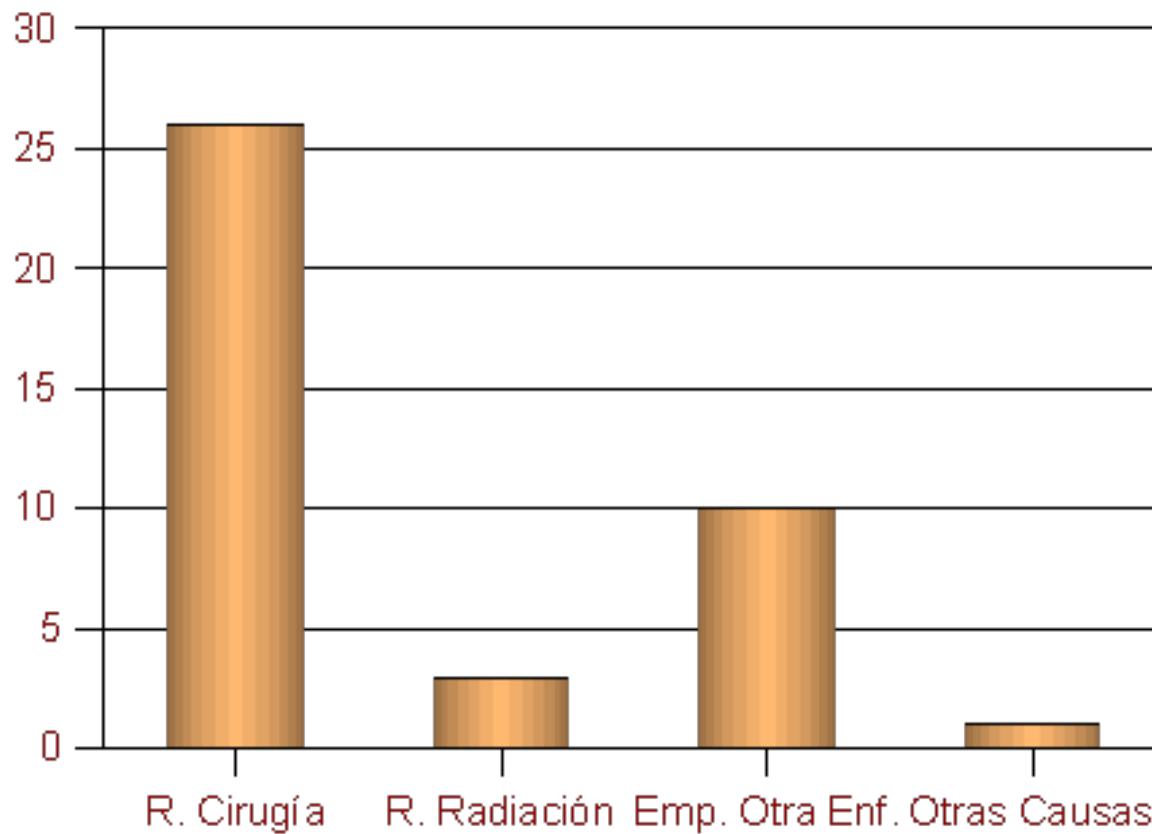
Diagrama de Rectángulos (variables cualitativas)



En un histograma conviene observar, al menos:

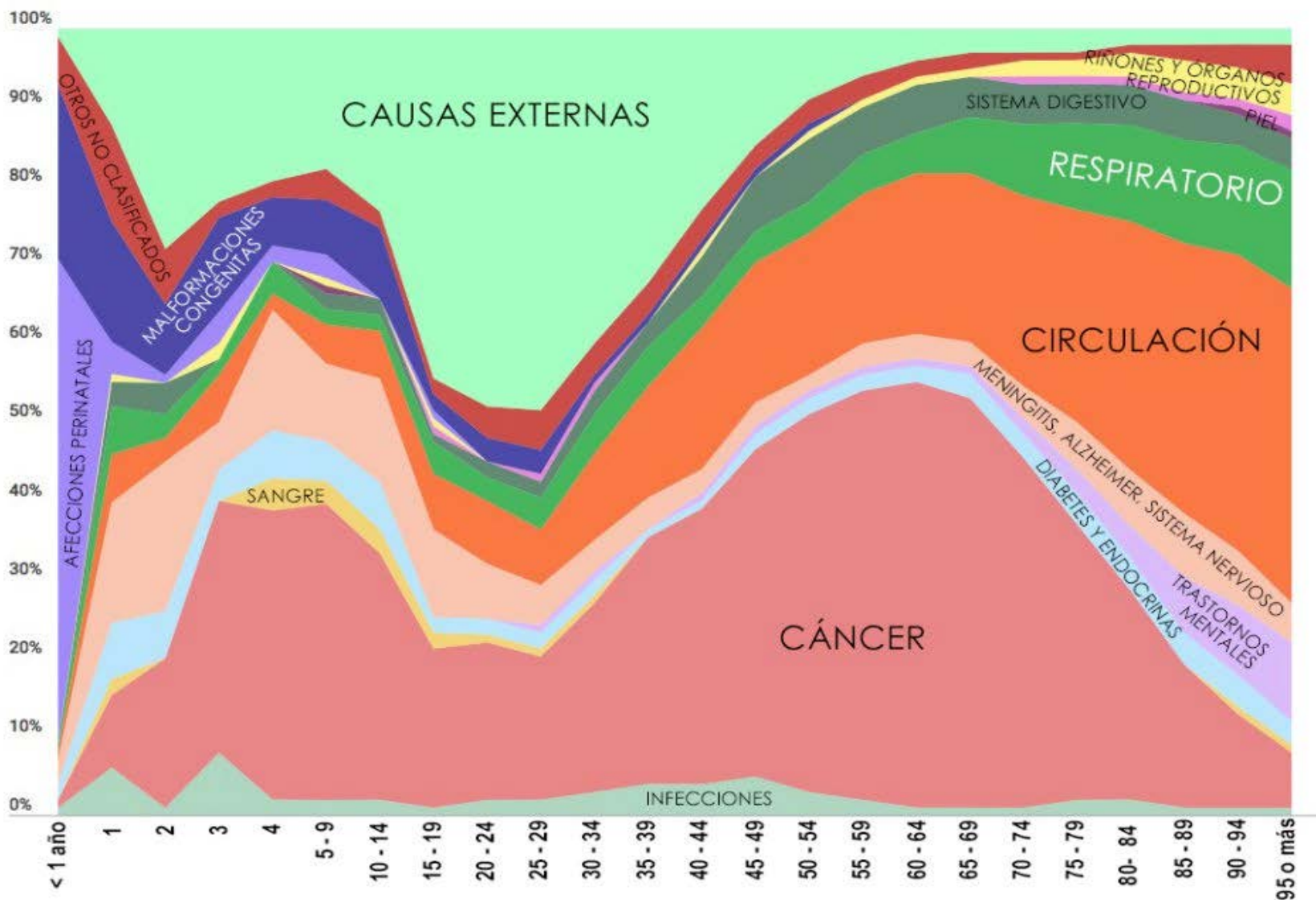
- 1.- Las zonas de concentración de los datos, una o varias.
- 2.- La variabilidad de los datos.
- 3.- La simetría.
- 4.- La existencia de cortes.
- 5.- Los posibles puntos atípicos.

Diagrama de Rectángulos (variables cualitativas)



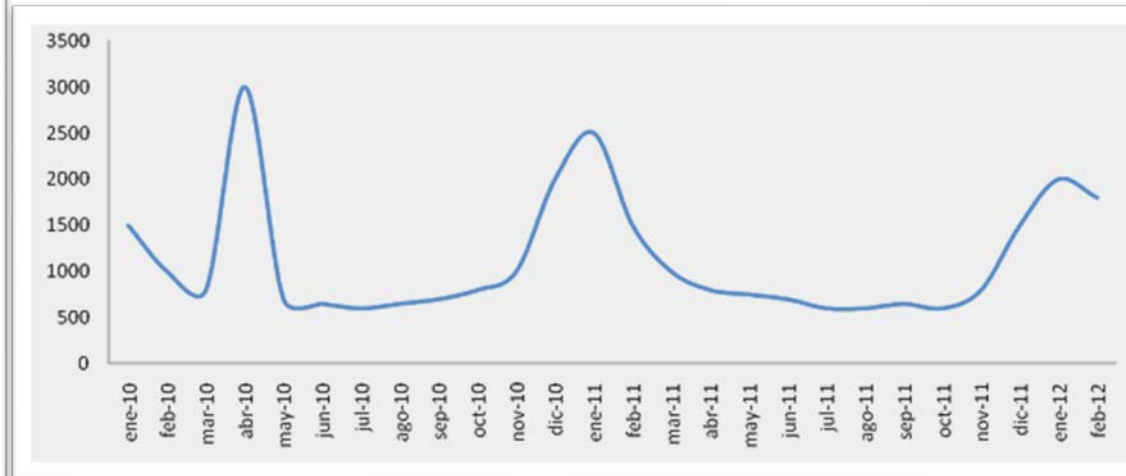
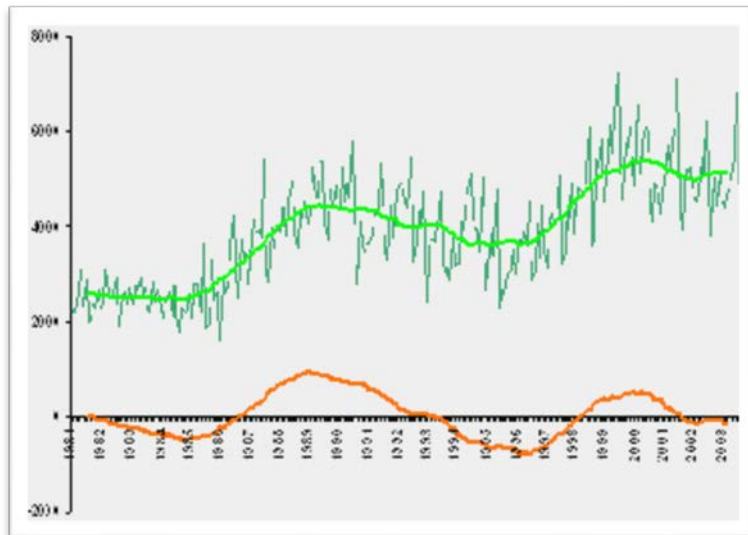


De qué mueren los españoles según su edad

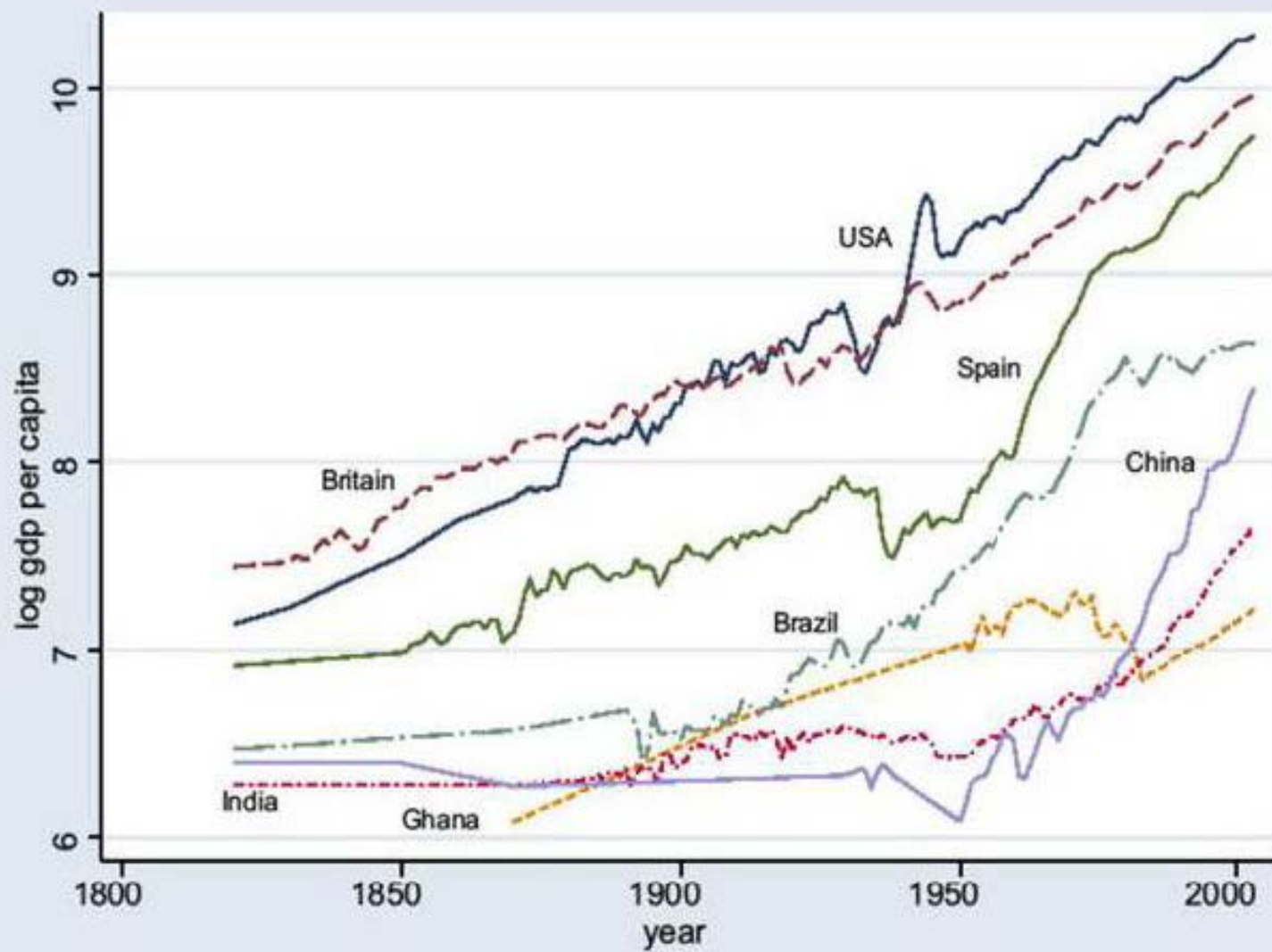


SERIES TEMPORALES

Colección de observaciones de una variable recogidas secuencialmente en el tiempo. Estas observaciones se suelen recoger en instantes de tiempo equiespaciados.

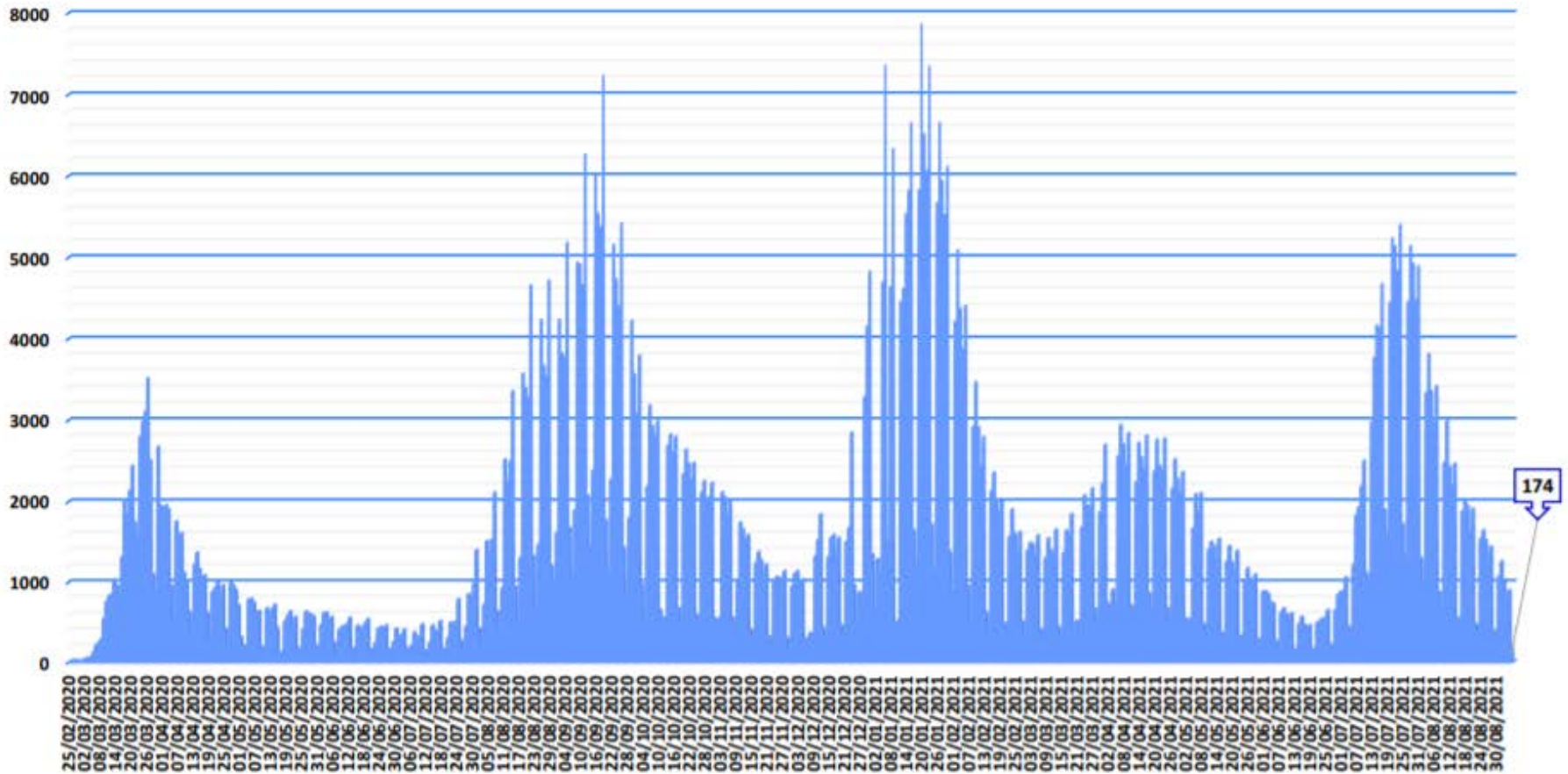


SERIES TEMPORALES



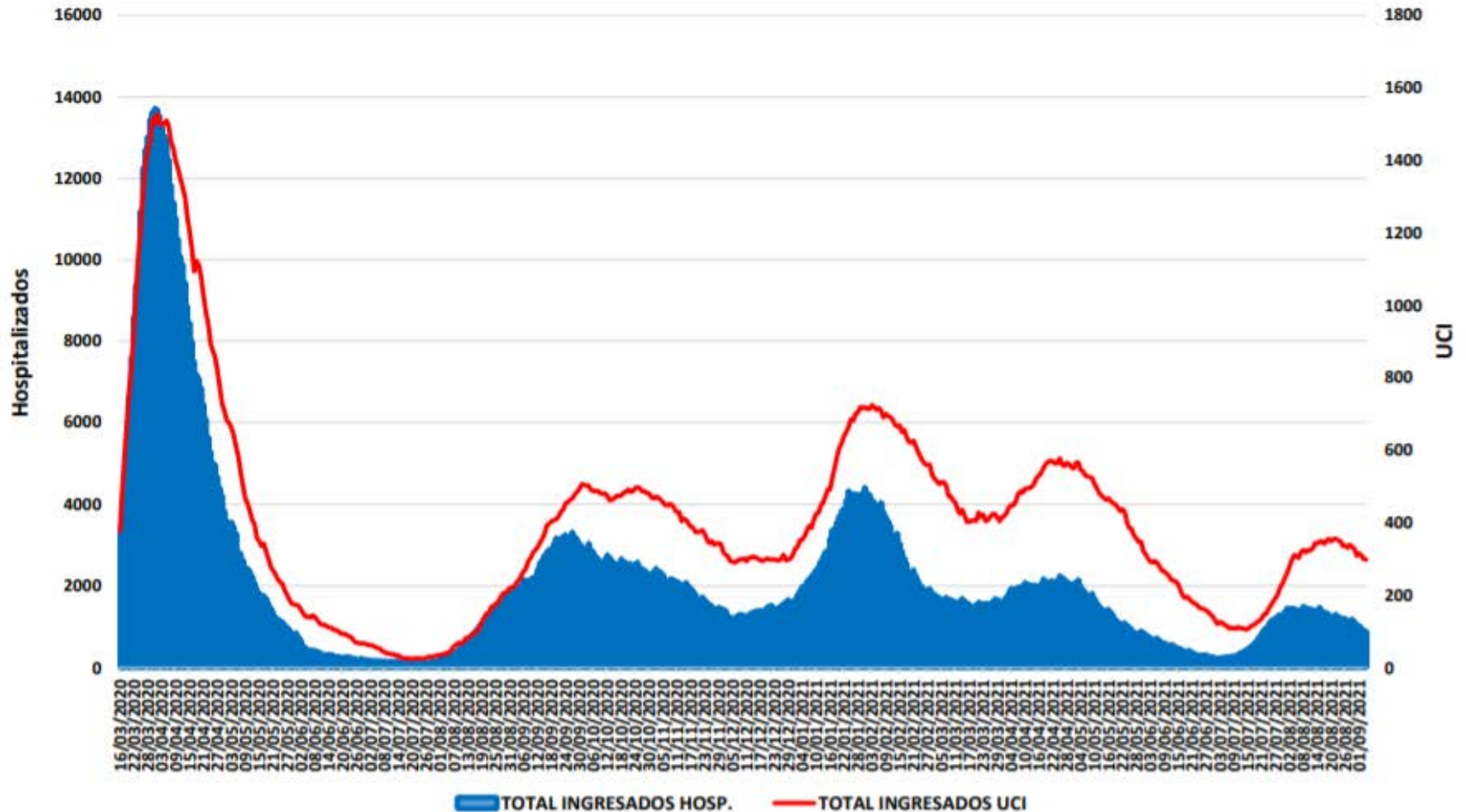
SERIES TEMPORALES

Evolución casos positivos de Covid-19



SERIES TEMPORALES

Evolución casos hospitalizados y UCI

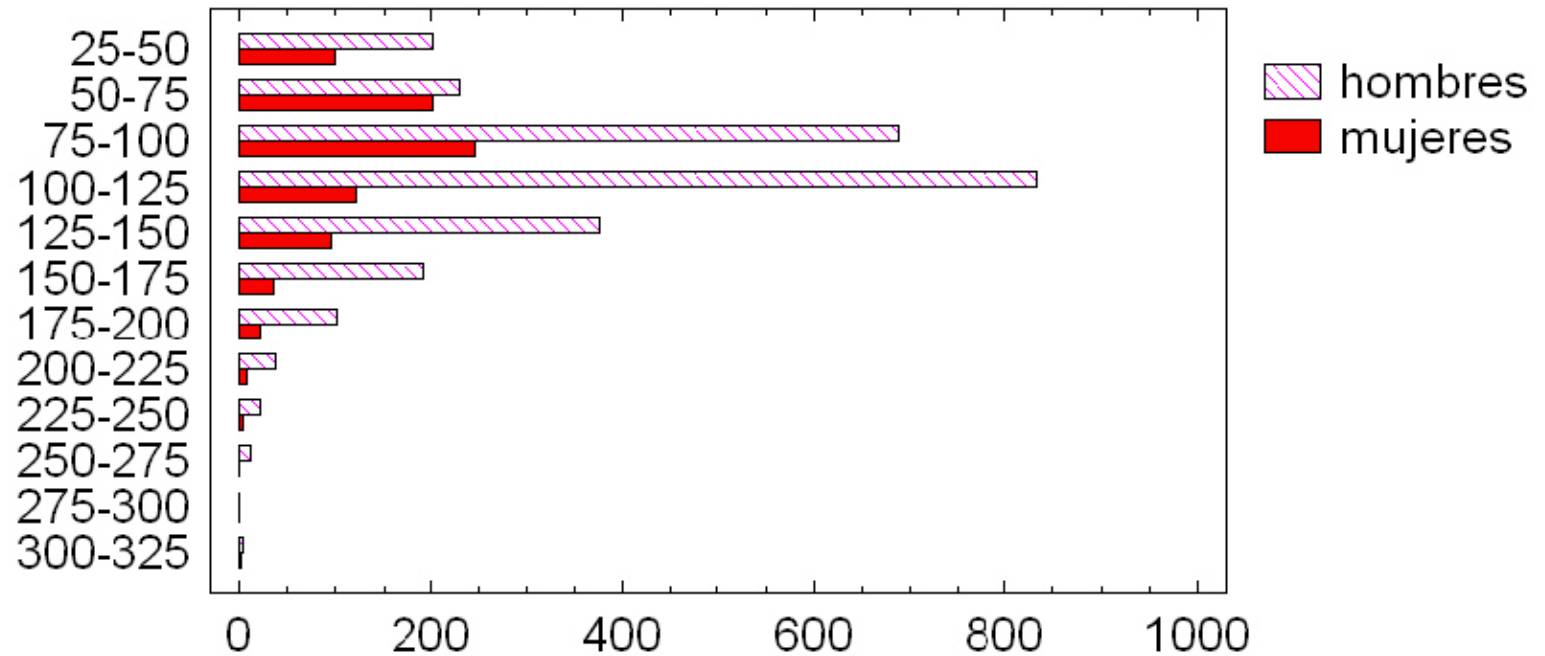


Fuente: SERMAS y hospitales privados.

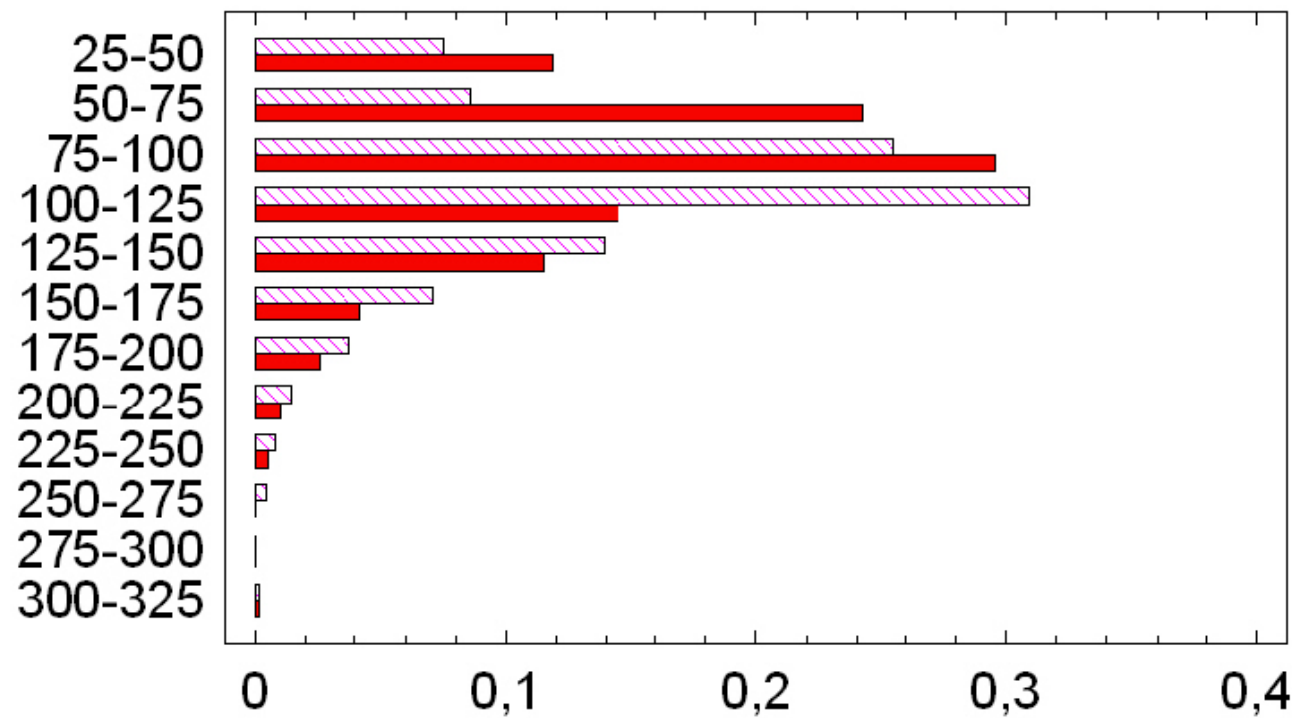
observación

El aspecto de los gráficos y, consecuentemente, el resultado de un análisis puede depender de algunas elecciones que haga el investigador. Considérese el siguiente ejemplo:

- *Se afirma, en general, que los salarios de las mujeres son inferiores al de los hombres. Para comprobar esta conjetura se han recogido los salarios de 833 mujeres y 2694 hombres, que trabajan en una gran empresa.*



¿Se considera admisible la hipótesis?



¿y ahora?

MEDIDAS NUMÉRICAS

¿Cual es la localización del centro de los datos?

(“**medidas de centralizacion o de tendencia central**”)

- Indican el valor medio de los datos o cómo se distribuyen estos

Media

Mediana

Moda

cuartiles, deciles y percentiles

- ¿Cómo varían los datos? (“**medidas de dispersión**”)

- Indican la variabilidad de los datos.

Recorrido

Varianza

Desviación típica

Introducción

- La distribución de frecuencias refleja **toda** la información disponible, lo que en general es demasiada información. Es, por tanto, necesario **resumir** la información disponible.
- Existen diferentes medidas que proporcionan una descripción global de
 - la variable.
- Son características deseables para las medidas de posición:
 - Que utilicen **todas y cada una** de las observaciones disponibles.
 - Que sean sencillas de **calcular**.
 - Que sean fáciles de **interpretar**.
 - Que sean **únicas** para cada distribución de frecuencias.

Medidas de Centralización

Media aritmética: Es el valor que tendría cada observación si todas las observaciones tuviesen el mismo valor

$$\bar{X} = \frac{1}{n} \sum_{i=1}^k n_i X_i = \sum_{i=1}^k f_i X_i$$

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

Suma de todos los valores dividido por el número de datos.

Datos (# de clases a las que no vas): **2 8 3 4 1**

Media = $(2+8+3+4+1)/5 = 3.6$

¡No puedes redondear!

Es el centro de gravedad de los datos

$$\overline{X} = \frac{\sum_{i=1}^j x_i n_i}{N} = \sum_{i=1}^j x_i f_i$$

Datos Agrupados en marcas de clase

$$\overline{X} = \frac{\sum_{i=1}^j c_i n_i}{N} = \sum_{i=1}^j c_i f_i$$

Para **datos agrupados**, se usan las **marcas de clase** para calcular la media en vez de los valores individuales, que son desconocidos.

Observaciones sobre la media

VENTAJAS

Definición **objetiva**

Utiliza **todas** las observaciones

Interpretación **sencilla**

Es **única**

Es el centro de **gravedad**

Fácil de calcular

Manipulación algebraica

INCONVENIENTE

Muy sensible a observaciones anómalas (**outliers**)

PROPIEDADES

1.- La suma de dos variables tiene por media la suma de las medias.

$$Z=X+Y \quad \rightarrow \quad \bar{z} = \bar{x} + \bar{y}$$

2.-Al multiplicar una variable por una constante, la media queda multiplicada por esta constante. $Y=C \cdot X$

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

Ejemplo: Trabajamos con una tabla en euros $X= \{3.45, 12.34, 6.25, 7.89, 12.10, 10.1, 9.10, 9.80\}$ Aplicamos una transformación lineal $y=100 \cdot x$
 $Y=\{345, 1234, 625, 789, 1210, 1010, 910, 980\}$

PROPIEDADES

3.- La suma de todas las desviaciones respecto a la media aritmética es cero
(importante porque le quita un grado de libertad al sistema)

$$\sum_{i=1}^N (x_i - \bar{x}) = 0 \quad \Rightarrow \quad \sum_{i=1}^N x_i - N\bar{x} = 0$$

4.- Cualquier variable x puede transformarse en una variable y mediante una transformación lineal

$$Y = a + bx$$

$$\bar{y} = \frac{\sum_{i=1}^N y_i}{N} = \frac{\sum_{i=1}^N (bx_i + a)}{N} = \frac{bN\bar{x} + Na}{N} = a + b\bar{x} \Rightarrow \bar{x} = \frac{\bar{y} - a}{b}$$

Ejemplo: Medidas de la gravedad (m/s^2) con péndulo simple

x_i	y_i
9.77	-3
9.78	-2
9.80	0
9.81	+1
9.83	+3
10.25	+45

Cambio de variable
 $y = 100x - 980$

$$\bar{y} = 7.33$$

$$\bar{x} = \frac{\bar{y} + 980}{100} = 9.873 m/s^2$$

Ejemplo

Calcula la media aritmética de la siguiente distribución.

x_i	n_i
300	20
600	40
900	60
1200	50
1500	30

xi	ni	Ni	fi	Fi	Xi x ni	Xi x fi
0	2	2	0,04	0,04	0	0
1	4	6	0,08	0,12	4	0,08
2	21	27	0,42	0,54	42	0,84
3	15	42	0,3	0,84	45	0,9
4	6	48	0,12	0,96	24	0,48
5	1	49	0,02	0,98	5	0,1
6	1	50	0,02	1	6	0,12
	N = 50		1		126	2,52

$[a_{i-1}, a_i)$	ci	ni	Ni	fi	Fi	Ci x ni	Ci X fi
[3,25,3,75)	3,50	3,00	3,00	0,075	0.075	10,50	0,2625
[3,75,4,25)	4,00	8,00	11,00	0,200	0.275	32,00	0,8
[4,25,4,75)	4,50	14,00	25,00	0,350	0.625	63,00	1,575
[4,75,5,25)	5,00	6,00	31,00	0,150	0.775	30,00	0,75
[5,25,5,75)	5,50	4,00	35,00	0,100	0.875	22,00	0,55
[5,75,6,25)	6,00	5,00	40,00	0,125	1,00	30,00	0,75
		N= 40		1		187,50	4,6875

vs 4,685

OTRAS MEDIAS

- **Media geométrica** (no permite valores nulos)
-

$$x_G = \sqrt[N]{x_1 \cdot x_2 \cdot x_3 \cdot x_4 \cdot \dots x_N} = \sqrt[N]{x_1^{n_1} \cdot x_2^{n_2} \cdot x_3^{n_3} \cdot x_4^{n_4} \cdot \dots x_K^{n_k}}$$

- **Media armónica** (no permite valores nulos) la inversa de la media de las inversas.

$$x_A = \frac{N}{\sum_{i=1}^N \frac{1}{x_i}} = \frac{N}{\sum_{i=1}^k \frac{n_i}{x_i}}$$

- **Media cuadrática**

$$\overline{x_Q} = \sqrt{\frac{\sum_{i=1}^N x_i^2}{N}} = \sqrt{\frac{\sum_{i=1}^k n_i \cdot x_i^2}{N}}$$

CARACTERÍSTICAS DE LAS DIFERENTES MEDIAS

- La media armónica se ve poco afectada por valores más altos que el resto y muy afectada por valores pequeños.
- La media cuadrática evita los efectos del signo.
- La media geométrica no es válida cuando hay datos nulos o número impar de datos negativos.

$$x_A \leq x_G \leq \bar{x} \leq x_Q$$

Problema armónico



En la subida, la velocidad media es de 15 km/h, los organizadores calculan que la media del trayecto será de 30 km por hora. ¿ A que velocidad rodarán en la segunda mitad de la etapa?. Los dos tramos son de igual longitud.

Mediana

Me Otro nombre: percentil 50

- ✓ Es el valor de la variable que deja el mismo número de datos antes y después que él una vez ordenados estos.
- ✓ Mide el valor central de una serie de valores. Con los datos ordenados de menor a mayor, el 50% de los datos son inferiores a Me y el 50% son superiores a Me

Cálculo de la mediana muestral

Datos no repetidos

Ordenar los datos de menor a mayor.

Para un **número impar** de datos, la mediana es el valor del medio. El que ocupa la posición $(n+1)/2$

Datos (# de clases a las que no vas): **2 8 3 4 1**

Datos ordenados: **1 2 3 4 8**



Mediana

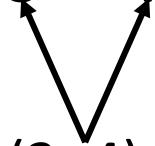
Mediana

- Es el valor de la distribución, ordenado de forma ascendente, que deja a **la mitad** de las observaciones a su izquierda y a la otra mitad a su derecha.
- La forma de obtenerla es ordenar los datos, y la mediana será el valor que ocupe la posición **central** si el número de observaciones es impar. Si el número de observaciones es par, entonces no existe un único valor en la posición central; entonces se suele definir la mediana como la media de los dos valores de las posiciones centrales.
- También puede definirse como el valor de la distribución cuya frecuencia absoluta acumulada es **$N/2$** (o su frecuencia relativa acumulada es **50%**).
- La mediana es la medida de posición más representativa en el caso de datos en **escala ordinal**.
- Cuando la media aritmética resulta poco representativa a causa de la presencia de **ouliers** en la distribución, se suele tener en cuenta la mediana.

Para un **número par**, la mediana es la media de los dos valores centrales.

Datos (# de clases a las que no vas): **2 8 3 4 1 8**

Datos ordenados: **1 2 3 4 8 8**


$$\text{Mediana} = (3+4)/2 = 3.5$$

Variable Discreta con valores repetidos

Si $N/2 = N_j \rightarrow \text{Me} = (x_j + x_{j+1})/2$

Si $N/2 \neq N_j \rightarrow \text{Me} = \text{primer valor de } x_j \text{ con } N_j > N/2$

x_i	N_i
1	6
2	13
3	17
4	19
5	20

Me=2

x_i	N_i
1	6
2	10
3	15
4	17
5	20

Me=2.5

x_i	N_i
1	6
2	10
3	15
4	17
5	21

Me=3

Ejemplo

Calcule la mediana de la siguiente distribución.

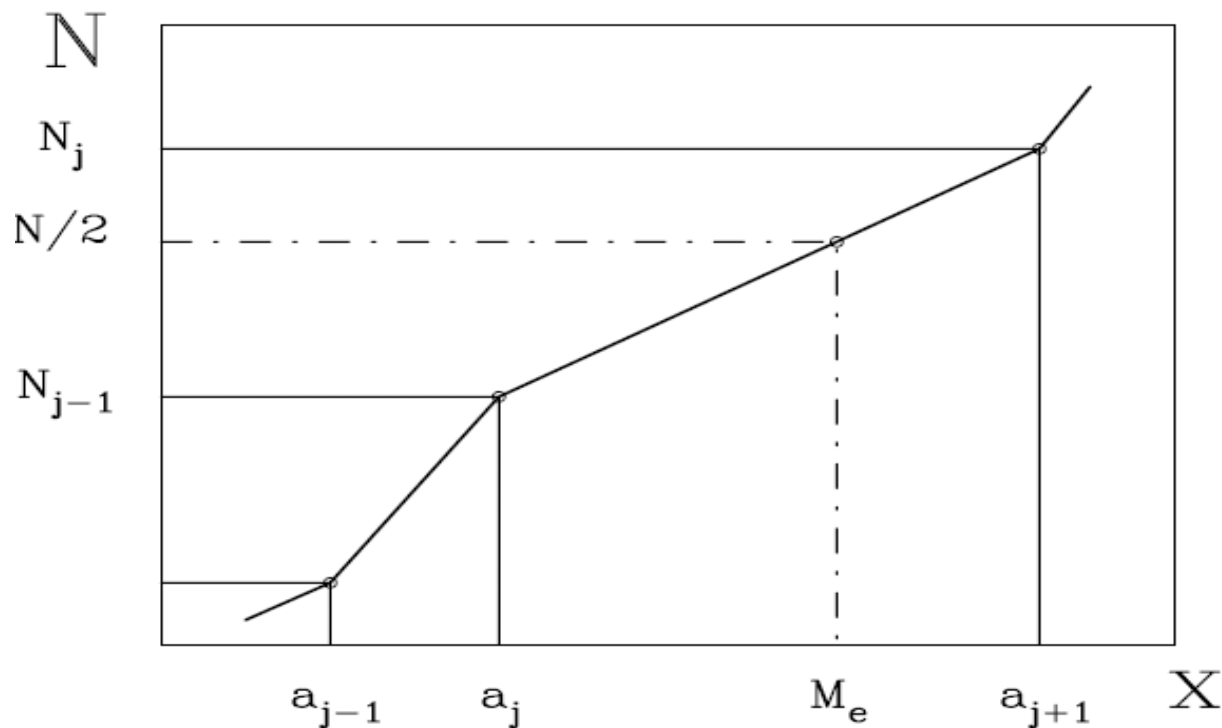
X_i	n_i	N_i
300	20	20
600	40	60
900	60	120
1200	50	170
1500	30	200

$$\frac{N}{2} = \frac{200}{2} = 100 \implies \text{posición } 100 - 101 \implies Me = 900$$

$[a_{i-1}, a_i)$	ci	ni	Ni	fi	Fi	Ci x ni	Ci X fi
[3,25,3,75)	3,50	3,00	3,00	0,075	0.075	10,50	0,2625
[3,75,4,25)	4,00	8,00	11,00	0,200	0.275	32,00	0,8
[4,25,4,75)	4,50	14,00	25,00	0,350	0.625	63,00	1,575
[4,75,5,25)	5,00	6,00	31,00	0,150	0.775	30,00	0,75
[5,25,5,75)	5,50	4,00	35,00	0,100	0.875	22,00	0,55
[5,75,6,25)	6,00	5,00	40,00	0,125	1,00	30,00	0,75
		N= 40		1		187,50	4,6875

$$Me = 4,50$$

Datos agrupados en intervalos de clase (variable continua)



Por igualdad entre pendientes (interpolación lineal), se llega a la siguiente ecuación

$$Me = a_j + \frac{N/2 - N_{j-1}}{n_j} (a_{j+1} - a_j)$$

$[a_{i-1}, a_i)$	ci	ni	Ni	fi	Fi	Ci x ni	Ci X fi
[3,25,3,75)	3,50	3,00	3,00	0,075	0.075	10,50	0,2625
[3,75,4,25)	4,00	8,00	11,00	0,200	0.275	32,00	0,8
[4,25,4,75)	4,50	14,00	25,00	0,350	0.625	63,00	1,575
[4,75,5,25)	5,00	6,00	31,00	0,150	0.775	30,00	0,75
[5,25,5,75)	5,50	4,00	35,00	0,100	0.875	22,00	0,55
[5,75,6,25)	6,00	5,00	40,00	0,125	1,00	30,00	0,75
		N= 40		1		187,50	4,6875

$$Me = a_2 + \frac{N/2 - N_2}{N_3 - N_2} (a_3 - a_2) = 4,57$$

Mediana

Ventajas:

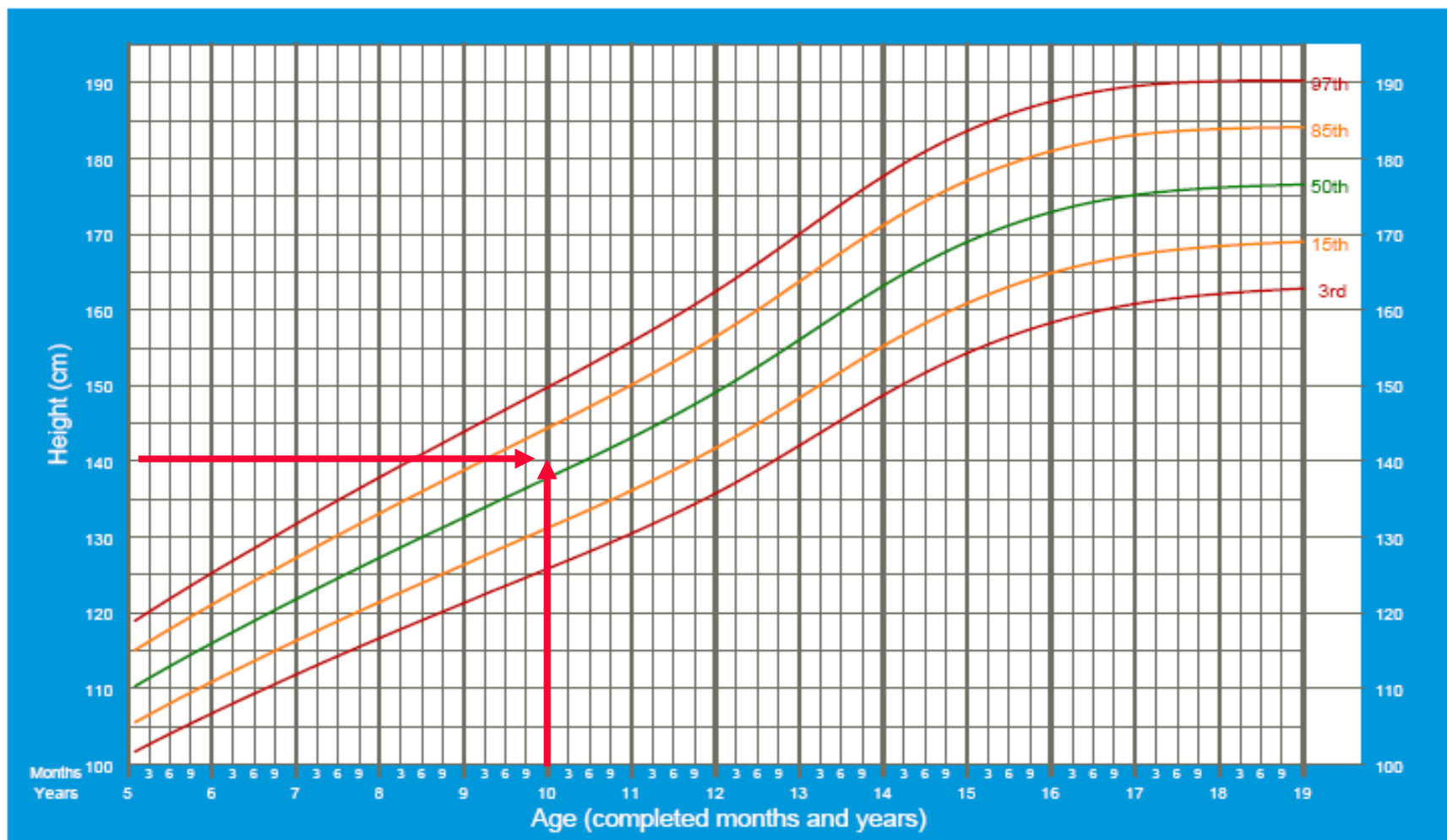
- Los valores extremos **no** afectan a la mediana tanto como a la media;
- Es fácil de **calcular** e **interpretar**;
- Se puede calcular incluso con **datos cualitativos**.

Desventajas:

- Es necesario ordenar la muestra, por lo que solo se puede usar con datos cualitativos en **escala ordinal**.
- **No usa** toda la información disponible en la muestra.

Height-for-age BOYS

5 to 19 years (percentiles)



2007 WHO Reference

cuantiles

Los cuantiles son medidas que dividen en partes iguales la distribución. Los más utilizados son:

Los cuartiles: Son tres valores que dividen la distribución en cuatro partes iguales, es decir, en cuatro intervalos dentro de cada cual están incluidos el 25% de los valores de la distribución.

Los deciles: Son los nueve valores que dividen la distribución en diez partes que incluyen al 10% de los valores cada una.

Los percentiles: Son los noventa y nueve puntos que dividen la distribución en cien partes iguales.

CUARTILES, DECILES Y PERCENTILES

Definición:

Se denomina **percentil** α de un conjunto ordenado de datos, al menor dato que es mayor o igual que el α % de todos ellos, se representa por P_{α} .

1. Si el $100p\%$ de n , donde n es el número de datos, es un entero, k , entonces $Q_p = (x_{(k)} + x_{(k+1)})/2$
2. Si el $100p\%$ de n no es un entero, lo redondeamos al entero siguiente, k , y entonces $Q_p = x_{(k)}$

x_i	N_i
1	6
2	13
3	17
4	19
5	20

Cuartiles: dividen la muestra en cuatro partes iguales, cada una con, un 25% . Hay 3 cuartiles ($Q_{1/4}, Q_{2/4}, Q_{3/4}$) son los percentiles 25, 50 y 75

Percentiles: P_k dividen la muestra en 100 partes iguales. Hay 99 percentiles.

$$Q_{1/4} = 1$$

$$Q_{1/2} = M_e = 2$$

$$Q_{3/4} = 3$$

$$P_{50} = D_5 = Q_{1/2} = M_e$$

Moda

- La **moda** es el valor con mayor frecuencia de la distribución.
- Para **datos sin agrupar**: Para obtener la moda en datos sin agrupar simplemente se busca el valor con **mayor frecuencia**.
- La moda puede no tener un único resultado por lo que existen distribuciones con varias modas (distribuciones **multimodales**)

Moda

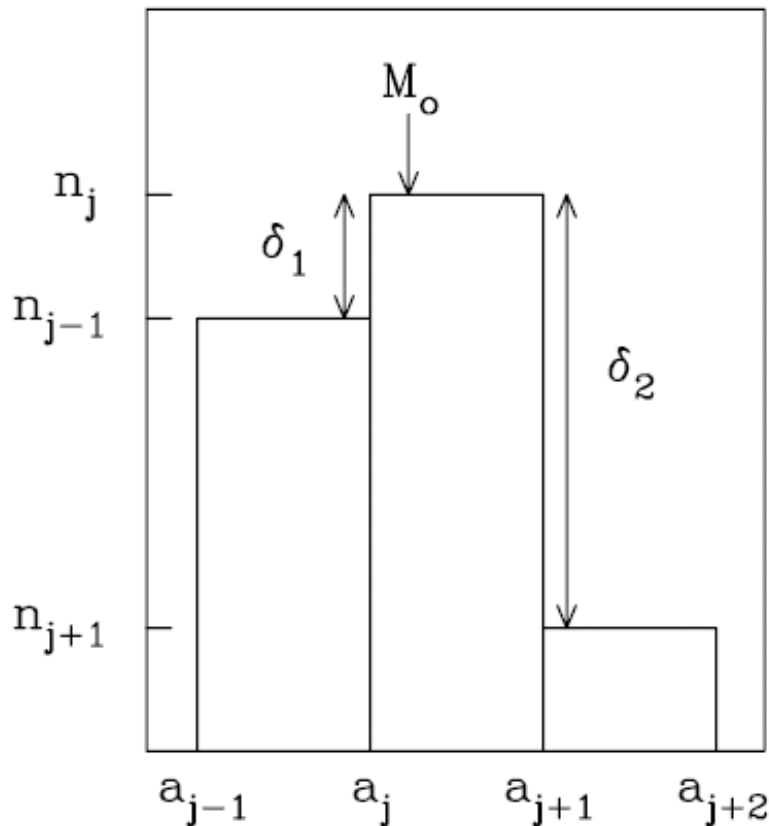
Valor que ocurre de forma más frecuente. Valor de la variable que tiene una frecuencia máxima

Un conjunto de datos puede tener muchas modas → distribución bimodal, trimodal...

Apropiado para todos los tipos de datos, pero más útil para datos categóricos o discretos con sólo unos pocos valores posibles.

x_i	n_i	f_i	N_i	F_i
1	6	0.30	6	0.30
2	<u>7</u>	0.35	13	0.65
3	4	0.20	17	0.85
4	2	0.10	19	0.95
5	1	0.05	20	1.00

DATOS AGRUPADOS EN INTERVALOS DE CLASE



$$M_o = a_j + \frac{\delta_1}{\delta_1 + \delta_2} (a_{j+1} - a_j)$$

$$Mo = a_j + \frac{(n_j - n_{j-1})(a_{j+1} - a_j)}{(n_j - n_{j-1}) + (n_j - n_{j+1})}$$

- ✗ La moda estará más próxima a a_j cuanto menor sea la diferencia de frecuencias con el intervalo anterior

MODA

Ventajas:

- Es la medida de posición **más representativa** para datos en **escala nominal**.
- Excepto en el caso de muestras muy pequeñas, la moda no se ve afectada por elementos outliers.
- Se puede calcular incluso con intervalos abiertos.

Desventajas:

- No se usa toda la información disponible en la muestra.
- A veces el hecho de que un elemento se repita más que el resto es casualidad. Por eso no se suele utilizar en el caso de variables numéricas.
- Si la distribución es multimodal se hace difícil de interpretar la moda.

Ejercicio de repaso

En la tabla siguiente se listan los datos medidos por James Short en 1763 sobre la paralaje del Sol en segundos de arco. El paralaje es el ángulo subtendido por la Tierra vista desde el Sol. Se midió observando tránsitos de Venus desde diferentes posiciones y permitió la primera medida de la distancia Tierra-Sol, que es la unidad básica de la escala de distancias en el Sistema Solar (la unidad astronómica).

8.63	10.16	8.50	8.31	10.80	7.50	8.12
8.42	9.20	8.16	8.36	9.77	7.52	7.96
7.83	8.62	7.54	8.28	9.32	7.96	7.47

MEDIDAS DE CENTRALIZACIÓN

X_i : 4 5 5 6 6 7 7 7 8 8

Media aritmética

$$\frac{4 + 5 + 5 + 6 + 6 + 7 + 7 + 7 + 8 + 8}{10} = 6,3$$

Mediana

6,5

Moda

7

MEDIDAS DE DISPERSIÓN

- Dada una distribución de frecuencias, ¿hasta qué punto las medidas de tendencia central son **representativas** o representan adecuadamente la información de la muestra?
- Cuanto más cerca estén las observaciones de la medida de tendencia central, más representativa será.
- Será menos representativa si se observa mucha dispersión a su alrededor.

MEDIDAS DE DISPERSIÓN

- La media de ambos conjuntos de puntuaciones es la misma, 26
- Estos dos conjuntos de datos, a pesar de tener la misma tendencia central, son bastante diferentes.

Xi: 6 16 22 26 30 36 46

Yi: 23 24 25 26 27 28 29



- Más parecidas entre sí
- Menor variabilidad
- Menor dispersión

ESTADÍSTICOS PARA CUANTIFICAR ESTA PROPIEDAD



MEDIDAS DE DISPERSIÓN

Muestra 1: 6 16 22 26 30 36 46

Muestra 2: 23 24 25 26 27 28 29

Puntuaciones diferenciales:

Muestra 1: -20 -10 -4 0 4 10 20

Muestra 2: -3 -2 -1 0 1 2 3

➤ Si la dispersión o variabilidad es muy grande, la medida de tendencia central no será representativa.

MEDIDAS DE DISPERSIÓN

Indican la **variabilidad** de los datos en torno al valor promedio

Recorridos

También se llama (1) **rango** = $X_{\text{máx}} - X_{\text{min}}$. Diferencia entre mayor y menor valor

(2) **Recorrido** intercuartílico

$$R_I = Q_{3/4} - Q_{1/4}$$

Rango que ocupa en 50% central de los datos

(3) **Recorrido semiintercuartílico**

$$R_{SI} = \frac{Q_{3/4} - Q_{1/4}}{2}$$

Estás medidas nos dan una idea de la dispersión de la muestra pero no utilizan ninguna medida de posición central, por lo que no pueden utilizarse para analizar la representatividad de ninguna medida en concreto

DESVIACIÓN MEDIA

Comparación de cada valor con **una medida de centralización**

Es la desviación media con respecto a la media aritmética

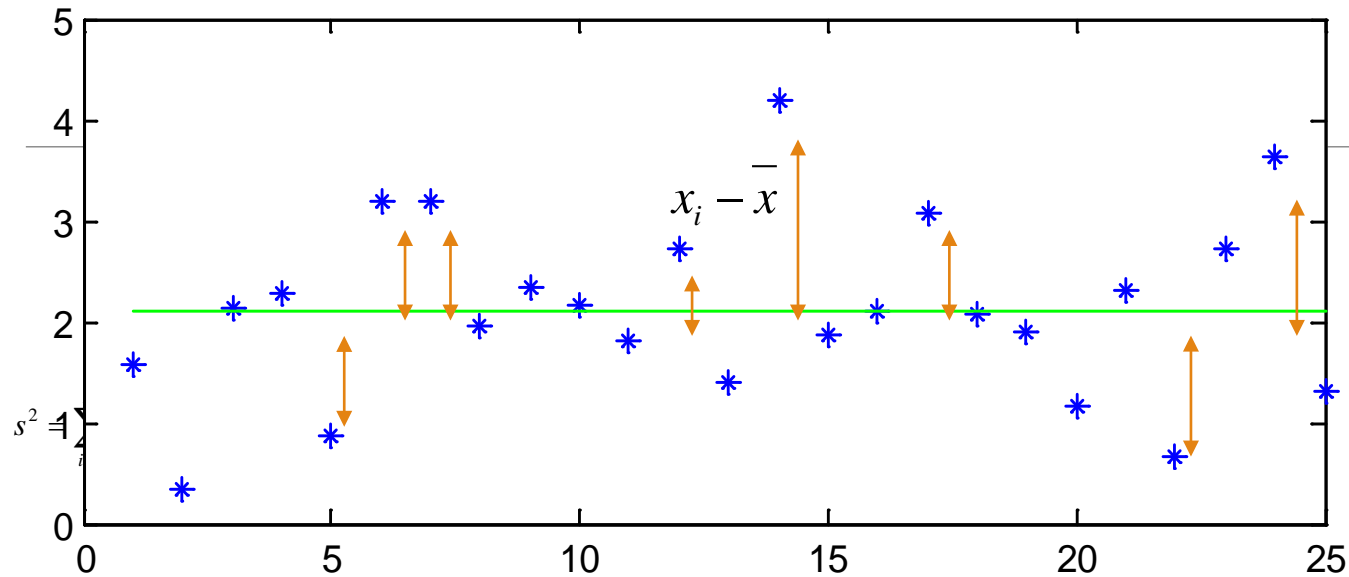
$$D_{\bar{x}} = \sum_{i=1}^k \frac{|x_i - \bar{x}| n_i}{N} \quad \text{¡ojo en valor absoluto!}$$

Existe también la desviación media respecto a la mediana

$$D_{Me} = \sum_{i=1}^k \frac{|x_i - Me| n_i}{N}$$

la desviación media se hace mínima al calcularla con la mediana
utiliza el valor absoluto, función que no es derivable y no resulta muy adecuada para determinados cálculos.

Varianza y Desviación típica



varianza

$$s^2 = \sum_{i=1}^N \frac{(x_i - \bar{x})^2}{N} = \sum_{i=1}^k \frac{(x_i - \bar{x})^2 n_i}{N} = \frac{1}{N} \sum_{i=1}^K x_i^2 n_i - \bar{X}^2$$

$$s = \sqrt{s^2} = \sqrt{\sum_{i=1}^N \frac{(x_i - \bar{x})^2}{N}} = \sqrt{\sum_{i=1}^k \frac{(x_i - \bar{x})^2 n_i}{N}} \quad \text{Desviación típica (conserva unidades)}$$

Muestra 1:

6

16

22

26

30

36

46

Muestra 2:

23 24 25 26 27 28 29

Puntuaciones diferenciales:

Muestra 1:

-20 -10 -4 0 4 10 20

Muestra 2:

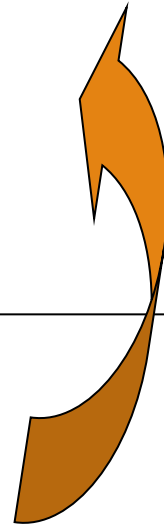
-3 -2 -1 0 1 2 3

$$S_x = \sqrt{S_x^2} = \sqrt{147,429} = 12,142$$

$$S_x^2 = 147,429$$

$$S_y = \sqrt{S_y^2} = \sqrt{4} = 2$$

$$S_y^2 = 4$$



Varianza

- La varianza se expresa en las **unidades de la variable** elevadas al cuadrado y esto dificulta su interpretación.
- La desviación típica se mide en las **mismas unidades** que las observaciones, lo que la hace más sencilla de interpretar.
- Los **valores extremos** tienen una fuerte influencia tanto en la desviación típica como en la varianza, ya que su desviación respecto a la media se eleva al cuadrado.
- La varianza nunca es **negativa**
- La varianza es la desviación **cuadrática óptima**
- La varianza **permanece invariante ante cambios de origen**
$$S^2 (x + a) = S^2 (x)$$
- En cambios de escala, la varianza quedará multiplicada por el cuadrado de la constante que define el cambio de escala
$$S^2 (kx) = k^2 S^2 (x)$$

Diagrama de cajas y bigotes

(boxplots o box and whiskers)

Un diagrama de caja y bigotes representa los valores de los cuartiles, del máximo y el mínimo de los datos no atípicos (Min y Max), así como los valores atípicos.

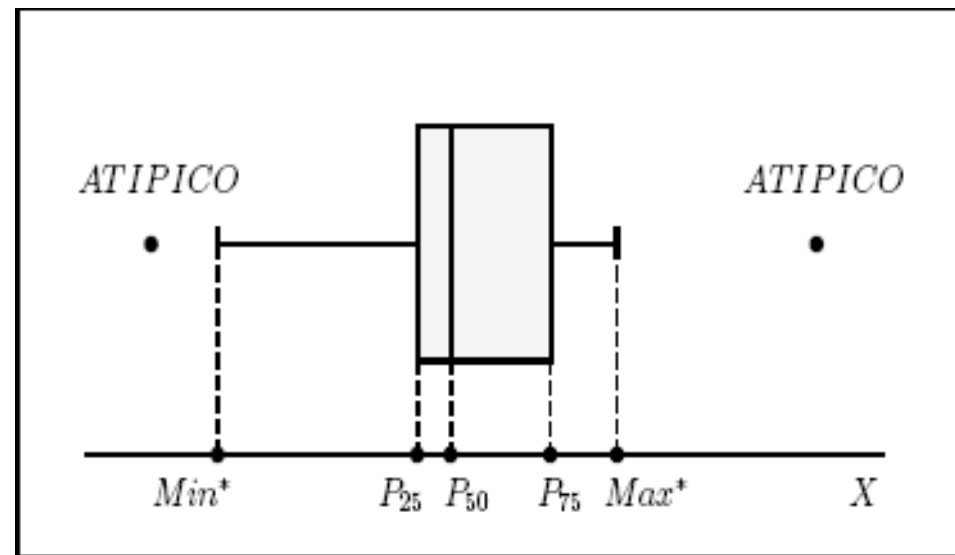


Diagrama de cajas y bigotes (boxplots o box and whiskers)

Esta figura muestra cómo el criterio del rango intercuartílico considera atípicos aquellos valores que se alejan del P_{25} , o del P_{75} , más de $1,5 \times R_I$, por la izquierda o por la derecha, respectivamente.

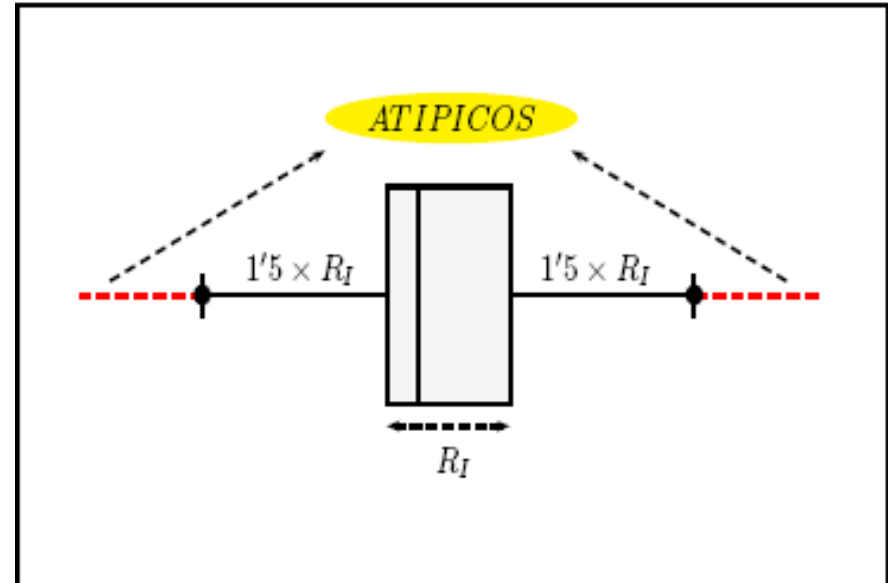
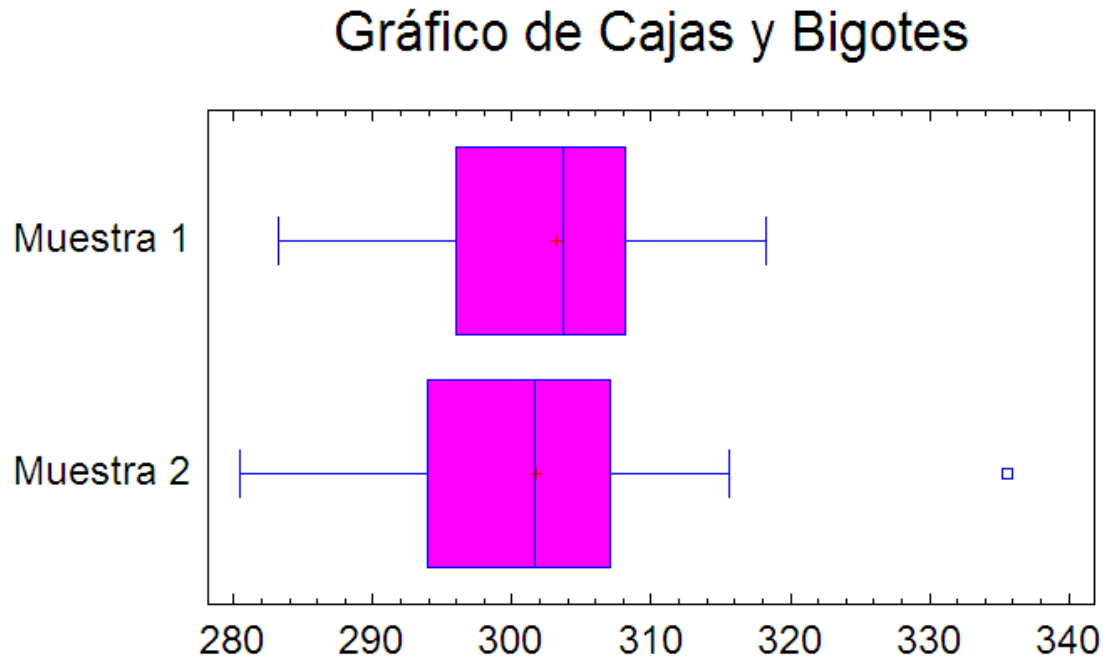


Diagrama de cajas y bigotes

(boxplots o box and whiskers)

Una aplicación del diagrama de cajas es la comparación de muestras.



Desigualdad de Chebychef

La desigualdad de Chebychef establece que si un conjunto de datos tiene media y desviación típica, para todo $k \neq 1$, en el intervalo: $(\bar{x} - ks, \bar{x} + ks)$ se encuentra, al menos, el $(1 - 1/k^2) \times 100\%$ de los datos.

Como consecuencia, para cualquier conjunto de datos en los intervalos $(\bar{x} - 2s, \bar{x} + 2s)$ y $(\bar{x} - 3s, \bar{x} + 3s)$ se encuentran, como mínimo, el 75% ó el 88,88% de los datos, respectivamente.

Desigualdad de Chebychef (ejemplo)

En la empresa A el salario medio anual de los empleados es 35.000 euros y la desviación típica 5.000 euros.

En la empresa B el salario medio anual de los empleados es 35.000 euros y la desviación típica 1000 euros.

¿En cuál de las dos empresas preferiría trabajar?

DISCUSIÓN

- La desviación típica **no es una medida robusta**. Es muy sensible a observaciones extremas.
- utiliza todas las observaciones
- Fácil de obtener computacionalmente
- La desviación típica es más sensible que la desviación media a datos extremos.
- El rango **intercuartílico** es que el que nos da una medida más aproximada de la desviación de los datos. Pero la relación con la normal hace que se use más la desviación típica.

Comparación de dispersiones

- Tanto la varianza como la desviación típica carecen de escala. Son medidas absolutas.
- Cuando se desea comparar variabilidades entre dos conjuntos de datos conviene tener en cuenta la magnitud de los mismos, no siendo razonable comparar variabilidades de conjuntos de datos muy heterogéneos.
- **Problema:** No está definido si la media es cero

COEFICIENTES DE VARIACIÓN

Coeficiente de variación de Pearson: medida relativa de variabilidad

$$CV = \frac{s}{\overline{x}} \quad (\text{en } 100\%)$$

Es siempre, en valor absoluto, menor que 1.

observación: El coeficiente de variación mide cuántas veces contiene la desviación típica de un conjunto de datos a su media.

El inverso es el llamado coeficiente “señal-ruido”

$$SR = \frac{\overline{x}}{s}$$

Ejemplo

Con un micrómetro, se realizan mediciones del diámetro de una tuerca, que tienen una media de 4.03 mm y una desviación estándar de 0.012 mm; con otro micrómetro se toman mediciones de la longitud de un tornillo que tiene una media de 1.76 pulgadas y una desviación estándar de 0.0075 pulgadas. ¿Cuál de los dos micrómetros presenta una variabilidad relativamente menor?.

$$\frac{0.012}{4.03} \times 100 = 0.3\% \qquad \frac{0.0075}{1.76} \times 100 = 0.4\%$$

En consecuencia, las mediciones hechas por el primer micrómetro exhiben una variabilidad relativamente menor con respecto a su media que las efectuadas por el otro.

xi	ni	Ni	fi	Fi	Xi x ni	Xi x fi	xi ² x ni
0	2	2	0,04	0,04	0	0	0
1	4	6	0,08	0,12	4	0,08	4
2	21	27	0,42	0,54	42	0,84	84
3	15	42	0,3	0,84	45	0,9	135
4	6	48	0,12	0,96	24	0,48	96
5	1	49	0,02	0,98	5	0,1	25
6	1	50	0,02	1	6	0,12	36
	N = 50		1		126	2,52	380

$$S^2 = (380 - (126)^2/50)/50 = 1,25 = (380/50) - (2.52)^2$$

MEDIDAS DE ASIMETRÍA

Coeficientes de asimetría: permiten caracterizar hacia que lado de la curva se encuentra la cola de la distribución

Asimétrica por la derecha o positiva: $M_o \leq Me \leq \bar{X}$

Asimétrica por la izquierda o negativa: $M_o \geq Me \geq \bar{X}$

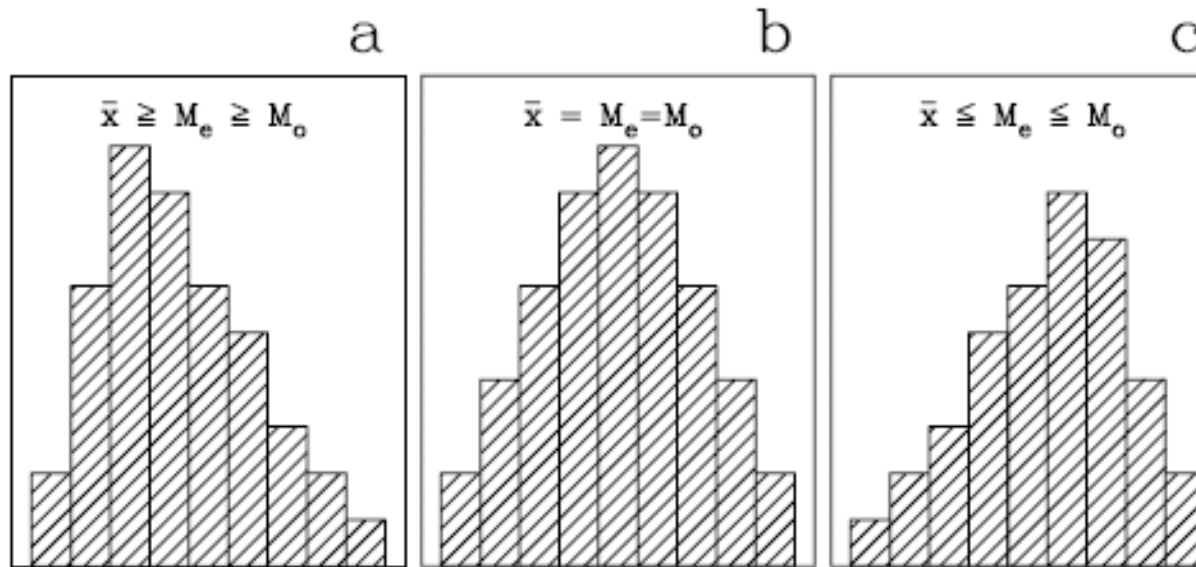
Simétrica: $M_o = Me = \bar{X}$

MEDIDAS DE ASIMETRÍA Y CURTOSIS

- Es importante analizar la forma de la distribución, para entender mejor el comportamiento de la variable.
- La **asimetría** mide si la distribución es simétrica y si no lo es cuanto dista de serlo.
- La **curtosis** mide la concentración de valores alrededor de la media aritmética.

ASIMETRÍA

- Es un indicador que permite evaluar el **grado de simetría** (o asimetría) de la distribución sin representarlas gráficamente.
- Si una distribución es simétrica, la distancia media de los valores a la media para los valores que inferiores a esta, es igual a la distancia media de los valores superiores a la media.
- **Asimetría negativa**: la cola izquierda es más larga y la masa de la distribución se concentra a la derecha.
- **Asimetría positiva**: la cola derecha es más larga y la masa de la distribución se concentra a la izquierda.



$A_p > 0 \rightarrow$ Media mayor que moda: sesgada hacia la derecha

$A_p < 0 \rightarrow$ Media menor que moda: sesgada hacia la izquierda

Coeficiente de **Asimetría de Pearson**

$$A_p = \frac{\bar{x} - M_o}{s}$$

MEDIDAS DE ASIMETRÍA

Coeficientes de asimetría

Simétrica: cuando los valores de la variable equidistantes, a uno y otro lado , del valor central tienen la misma frecuencia.

Momento de orden 3 respecto de la media

$$M_3 = \frac{\sum_{i=1}^k (x_i - \bar{x})^3 n_i}{N}$$

-Coeficiente de **Asimetría de Fisher**

$$g_1 = \frac{m_3}{s^3}$$

$g_1 > 0$ asimetría positiva (hacia la derecha)
 $g_1 = 0$ simétrica
 $g_1 < 0$ asimetría negativa (hacia la izda.)

Una distribución simétrica tiene $g_1 = 0$, pero eso no significa que una distribución con $g_1 = 0$ sea necesariamente simétrica.

MEDIDAS DE APUNTAMIENTO (CURTOSIS)

Evalúan el **agrupamiento** en torno al valor central

- Las medidas de curtosis se aplican a distribuciones con forma de campana, es decir, a distribuciones simétricas o ligeramente asimétricas y unimodales.
- Las medidas de curtosis se centran analizar la concentración de valores en la "zona central" de la distribución.
- Una mayor curtosis significa que una mayor parte de la varianza es el resultado de desviaciones extremas, en lugar de desviaciones frecuentes de tamaño modesto.

MEDIDAS DE APUNTAMIENTO (CURTOSIS)

Evalúan el **agrupamiento** en torno al valor central

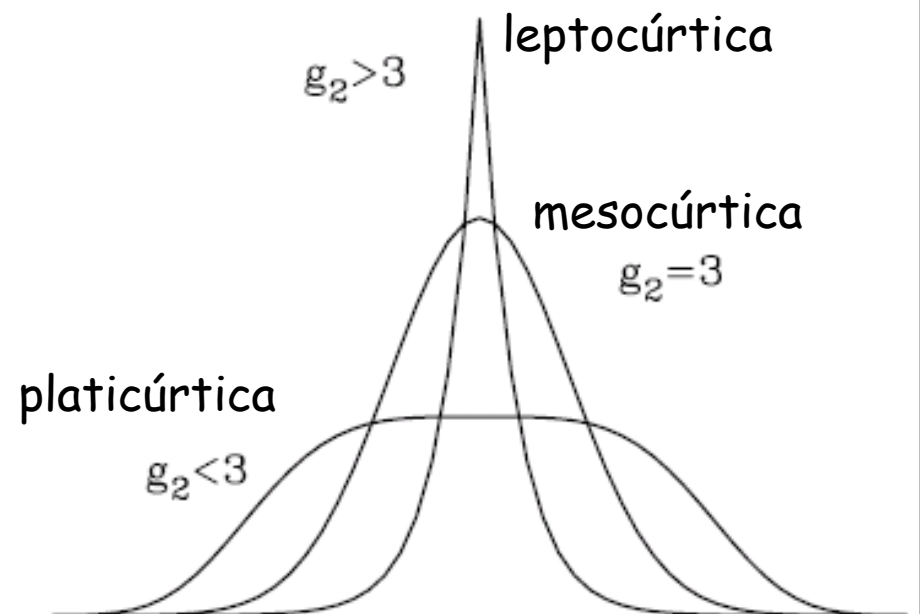
Gran apuntamiento: **leptocúrtica** $g_2 > 3$

Distribución normal : **mesocúrtica** $g_2 = 3$

Aplanado : **platicúrtica** $g_2 < 3$

$$g_2 = \frac{m_4}{s^4}$$

$$m_4 = \frac{\sum_{i=1}^k (x_i - \bar{x})^4 n_i}{N}$$



ALGUNOS EJEMPLOS

$[A_{i-1}, a_i)$	n_i
7,5-9	3
9-10,5	8
10,5-12	10
12-13,5	10
13,5-15	1
15-16,5	2

Hallar el percentil 70

x_i	n_i
0	5
1	6
2	8
3	4
4	2

Calcular la varianza