

Proyecto MaTeX



Estadística Descriptiva

Fco Javier Glez Ortiz

Directorio

- **Tabla de Contenido**
- **Inicio Capítulo**

Copyright © 2002 gonzaleof@unican.es
Actualizado el: 26 de septiembre de 2003

Versión 2.00

Tabla de Contenido

- 1. Introducción
- 2. Tipos de variables
- 3. Métodos de Estadística Descriptiva básica
- 4. Tablas
- 5. Gráficos
 - 5.1. Pictogramas
 - 5.2. Diagramas de barras
 - 5.3. Histograma y polígono acumulativo
- 6. Estadísticos
 - 6.1. Estadísticos de tendencia central
 - 6.2. Estadísticos de dispersión
 - 6.3. Estadísticos de posición
- Soluciones a los Ejercicios

1. Introducción

La Estadística Descriptiva nace de la necesidad de extraer y resumir la 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 grandes de datos por la mera visión de listados de dichos datos.

2. Tipos de variables

Atendiendo a su naturaleza, las variables bajo estudio pueden clasificarse según los siguientes tipos:

- Variables **cualitativas**: Son aquellas que *no* toman valores numéricos; por ejemplo, el color de una tela.
- Variables **cuantitativas**: Son aquellas que toman valores numéricos. A su vez, entre éstas pueden clasificarse en
 - Variables **cuantitativas discretas**: Si solamente pueden tomar valores enteros. Por ejemplo, el número de coches pro-

ducidos en una factoría.

- Variables **cuantitativas continuas**: Si pueden tomar cualquier valor dentro de un intervalo de la recta real. Por ejemplo, la resistencia a tracción de una barra de acero.

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

Desde otro punto de vista, las variables pueden clasificarse atendiendo al número de observaciones que se realizan en el mismo individuo. Así podemos encontrar:

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

El estudio de las variables uni y bidimensionales es propio de la Estadística Descriptiva básica.

Cuestión 2.1 Responde a las siguientes cuestiones sobre el tipo de variable cuando clasificamos a los alumnos de una clase:

1. la variable deporte que practican es:

(a) *Cualitativa* (b) *Discreta* (c) *continua*

2. la variable número de hermanos es:

(a) *Cualitativa* (b) *Discreta* (c) *Continua*

3. la variable tiempo que ven la televisión en una semana es:

(a) *Cualitativa* (b) *Discreta* (c) *Continua*

4. la variable peso es:

(a) *Cualitativa* (b) *Discreta* (c) *Continua*

5. la variable color de su pelo es:

(a) *Cualitativa* (b) *Discreta* (c) *Continua*

6. la variable altura es:

(a) *Cualitativa* (b) *Discreta* (c) *Continua*

3. Métodos de Estadística Descriptiva básica

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

- **Tablas:** Se construyen a partir de un listado exhaustivo de los datos muestrales. Las tablas pueden usarse con cualquier tipo de variables y permiten realizar un resumen inicial de la información contenida en la muestra.
- **Gráficos:** Pueden construirse a partir de un listado exhaustivo de datos, o bien a partir de tablas. Dependiendo del tipo de la variable bajo estudio, puede variar el tipo de gráfico que convenga usar.
- **Estadísticos:** 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 es cualitativa.

4. Tablas

La ventaja fundamental que proporcionan las tablas, frente al listado exhaustivo de los datos, es que facilitan enormemente la comprensión de la información contenida en la muestra.

En la tabla 1 se registran la edad de 132 pasajeros de un avión, que corresponden a 132 **datos sin agrupar**.

1	3	6	7	12	13	13	14	14	14	16	17
18	19	20	23	23	23	24	27	27	28	28	29
29	29	29	30	31	32	33	33	34	34	34	34
35	36	36	36	36	36	37	37	37	39	39	40
42	43	44	44	44	44	45	45	45	45	45	45
46	46	46	47	47	47	47	47	48	48	49	49
49	50	50	50	50	51	51	52	52	52	53	53
54	55	55	55	55	56	56	57	57	58	58	58
59	60	60	60	61	61	62	63	63	64	64	65
65	65	66	68	70	71	71	71	71	71	72	73
76	77	77	78	78	79	79	79	86	88	93	97

Tabla 1: Las edades ordenadas de 132 pasajeros de un avión

La variable edad es cuantitativa discreta, pero deseamos clasificar

a los pasajeros en las categorías, ‘jóvenes’ si la edad es inferior a 20, ‘adultos’ entre 20 y 65 y ‘jubilados’ el resto. Después de un recuento diseñamos una tabla de frecuencias.

Los datos de la muestra pueden resumirse como se indica:

Edad	Frecuencia Absoluta	Frecuencia Relativa
jóvenes	14	0,106
adultos	93	0,704
jubilados	25	0,190
Totales	132	1

Tabla 2: Clasificación de los pasajeros por tramos de edad.

Debe observarse que la **frecuencia absoluta** f_i de cada tramo de edad es, por definición, *el número* de individuos que cumplen ese tramo. La **frecuencia relativa** h_i es, también por definición, *la proporción* individuos que cumplen ese tramo y se obtiene dividiendo la correspondiente frecuencia absoluta por el número total de datos en la muestra (por ejemplo, la frecuencia relativa 0,704 se obtiene dividiendo 93

entre 132); no obstante y a pesar de que no recomendamos esta práctica, no es extraño encontrar publicaciones en las que las frecuencias relativas se expresen en *tantos por ciento* (en este caso, se escribiría 70.4 % en lugar de 0,704). Al número total de datos en la muestra se le suele designar por la letra n (en este ejemplo, $n = 132$).

Ejemplo. 4.1. Con los datos de la tabla 1 confeccionar una tabla de frecuencias absolutas y relativas en 5 clases de longitud 20 años.

Solución: Después de un recuento diseñamos la tabla de frecuencias.

Edades	f_i	h_i
$0 \leq x < 20$	14	0,106
$20 \leq x < 40$	33	0,250
$40 \leq x < 60$	50	0,379
$60 \leq x < 80$	31	0,235
$80 \leq x < 100$	4	0,030
	$n = 132$	1

que consiste en una distribución de frecuencias para **datos agrupados**. Así por ejemplo la frecuencia absoluta 33 corresponde al número de pasajeros con edad entre 20 y 40, y la frecuencia relativa 0,250

corresponde al cociente entre los 33 y el número total de 132. Puede observarse que la suma de las frecuencias absolutas $\sum f_i = 132$ es el número de datos, y la suma de las frecuencias relativas $\sum h_i = 1$ es siempre la unidad. \square

Conviene formalizar las siguientes definiciones importantes que han aparecido a lo largo de esta sección:

Definición 1 (Tamaño muestral) *El tamaño de la muestra es el número de ítems o individuos de los que se han obtenido los datos de la muestra. El tamaño de la muestra suele designarse por la letra n .*

Definición 2 (Clase) *Clase es cada intervalo usado para agrupar los datos de la muestra cuando el número de datos diferentes entre sí es muy grande.*

Es necesario usar clases cuando la variable observada es cuantitativa continua, pero también puede serlo cuando es discreta si el número de datos diferentes es muy grande. Siempre que sea posible deben usarse clases de igual anchura.

Definición 3 (Marca de clase) *Marca de clase es el punto medio de la clase.*

Definición 4 (Frecuencia absoluta) *Frecuencia absoluta de un dato es el número de veces que ocurre dicho dato en la muestra. Frecuencia absoluta de una clase es el número de datos de la muestra que pertenecen a dicha clase. Se indica con f_i .*

Definición 5 (Frecuencia relativa) *Frecuencia relativa de un dato o una clase es el cociente entre su frecuencia absoluta y el tamaño de la muestra. Se indica con h_i .*

De acuerdo con esta definición, las frecuencias relativas son proporciones y sus valores deben estar en el intervalo cerrado $[0, 1]$. Aunque las frecuencias relativas se expresan a veces en tanto por ciento, este uso debe considerarse informal.

Cuestión 4.2 Observa la siguiente tabla de frecuencias y responde:

Edades	f_i	h_i
$0 \leq x < 20$	14	0,106
$20 \leq x < 40$	33	0,250
$40 \leq x < 60$	50	0,379
$60 \leq x < 80$	31	0,235
$80 \leq x < 100$	4	0,030
	$n = 132$	1

- El tamaño de la muestra es:
(a) 100 (b) 132 (c) Otro valor
- La clase con 31 pasajeros es:
(a) $[20; 40)$ (b) $[40; 60)$ (c) $[60; 80)$
- La marca de clase de $[20; 40)$:
(a) 20 (b) 30 (c) 40
- La proporción de pasajeros de 20 a 40 años es:
(a) 33 (b) 0,25 (c) Otro valor

Cuestión 4.3 La tabla siguiente (incompleta) resume las notas obtenidas por 80 alumnos de un instituto en selectividad. Responder

Calificación	f_i	h_i
Suspensos	...	0,375
Aprobados	20	...
Notables	16	...
Sobresalientes

1. El número de Suspensos es:

(a) 20

(b) 30

(c) 40

2. El número de Sobresalientes es:

(a) 10

(b) 12

(c) 14

3. La proporción de Notables es:

(a) 0,20

(b) 0,25

(c) 16

4. La suma de las frecuencias absolutas f_i es:

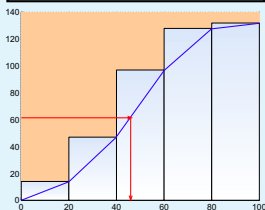
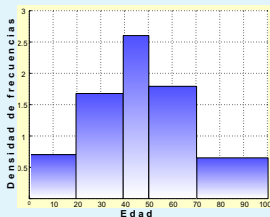
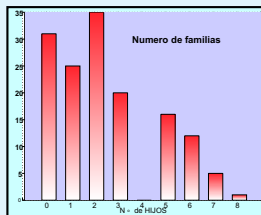
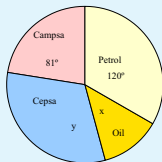
(a) 100

(b) 80

(c) Otro valor

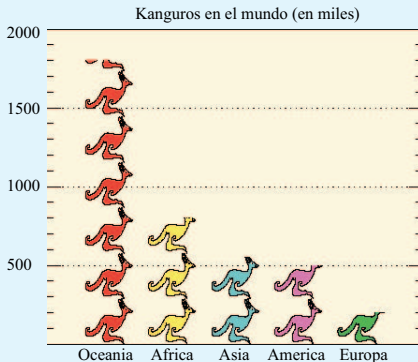
5. Gráficos

Los gráficos suelen construirse a partir de tablas, pero pueden también construirse a partir de un listado exhaustivo de datos. En general, el tipo de gráfico usado depende del tipo de la variable observada.



5.1. Pictogramas

Con el nombre de pictogramas suelen denominarse un conjunto de gráficos de aspecto más o menos atractivo que sirven para transmitir de forma sencilla la información contenida en una muestra.



Entre los pictogramas pueden incluirse los **gráficos de sectores**.

Ejemplo. 5.1. En un hipermercado se han producido las siguientes ventas en euros: Juguetes 125, Plantas 175, Discos 250, Alimentación 450. Realizar un diagrama de sectores

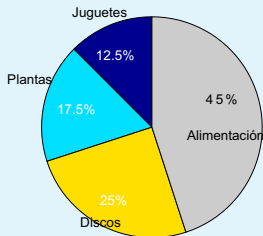
Solución: Almacenamos la información en una tabla

Variable	f_i	h_i	Ángulo
Juguetes	125	0,125	$0,125 \times 360^\circ = 45^\circ$
Plantas	175	0,175	$0,175 \times 360^\circ = 63^\circ$
Discos	250	0,250	$0,250 \times 360^\circ = 90^\circ$
Alimentación	450	0,450	$0,450 \times 360^\circ = 162^\circ$
	1000	1	360°

Por ejemplo, para el caso de las Plantas con 175 euros, para hallar el ángulo del sector

$$\frac{175}{1000} = 0,175 \quad 0,175 \times 360^\circ = 63^\circ$$

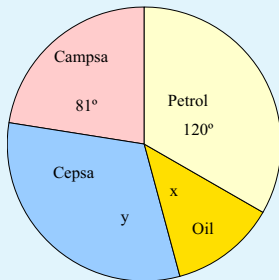
Análogamente los demás.



EJERCICIO 5.1. Un fabricante de ordenadores vende sus productos en cuatro países de acuerdo a la tabla. Realiza un diagrama de sectores.

país ventas	España	Francia	Alemania	Reino Unido
	5 %	15 %	15 %	65 %

EJERCICIO 5.2. El diagrama de sectores inferior muestra las ventas de empresas de petróleo. ¿Cuál es el porcentaje de ventas que tiene la empresa Campsa?. Si la empresa Oil tiene un 12,5% del total de ventas, calcula los valores de los ángulos x e y .

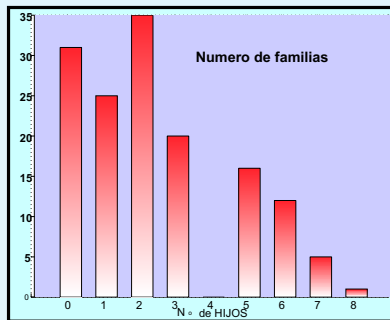


5.2. Diagramas de barras

La tabla registra el número de hijos de 145 familias, que corresponde a una variable cuantitativa discreta :

hijos	0	1	2	3	4	5	6	7	8
frecuencia	31	25	35	20	0	16	12	5	1

En el diagramas de barras aparecen en la escala de ordenadas las frecuencias absolutas o relativas con las que ha ocurrido cada dato; estas frecuencias vienen indicadas por barras de longitudes proporcionales a dichas frecuencias; en la escala de abscisas del gráfico aparecen los datos. Por ejemplo, con 2 hijos hay 35 familias.



5.3. Histograma y polígono acumulativo

El gráfico usado más frecuentemente para describir variables cuantitativas continuas es el **histograma**. La parte izquierda de la figura 1 muestra el histograma correspondiente a la tabla del ejemplo 4.1. Las frecuencias de los intervalos vienen representadas por rectángulos cuya altura¹ es proporcional a la frecuencia del intervalo. Sin embargo, puesto que la variable es ahora continua, los rectángulos tienen que aparecer compartiendo sus lados verticales con los rectángulos contiguos. Además, el orden en que aparecen las clases es el natural de los datos, independientemente de cuáles sean las respectivas frecuencias de las clases. En la escala del eje de ordenadas pueden aparecer las frecuencias absolutas o relativas². En la escala del eje de abscisas

¹En realidad es más exacto decir que, en los histogramas, el *área* de los rectángulos debe ser proporcional a la frecuencia del intervalo. Sin embargo, esta precisión *solamente es estrictamente necesaria* cuando se usan clases de diferentes anchuras.

²Si las clases tuvieran diferentes anchuras, la escala de ordenadas debería mostrar las frecuencias relativas por unidad de amplitud, que se obtienen dividiendo las frecuencias relativas por las anchuras de las clases.

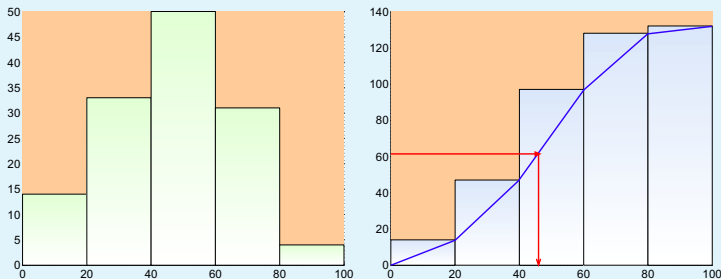


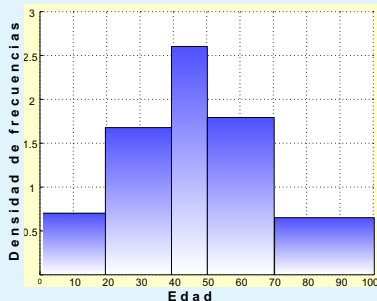
Figura 1: Histograma (izquierda) y polígono acumulativo (derecha) para las edades de los pasajeros. El tamaño de la muestra es $n = 132$.

pueden aparecer las marcas de clases, los extremos de las clases, o cualquier otra marca que cubra el rango de los datos.

Ejemplo. 5.2. Es este ejemplo se pide un histograma cuando los datos de las edades se registran en clases de longitud diferente.

Solución: Si hiciésemos un histograma con las frecuencias absolutas o relativas el gráfico sería engañoso. Lo correcto es, que la escala de ordenadas debe mostrar la densidad de frecuencias, que se obtienen dividiendo las frecuencias por las anchuras de las clases.

Edades	f_i	densidad
$0 \leq x < 20$	14	$28 : 20 = 0,70$
$20 \leq x < 40$	33	$33 : 20 = 1,65$
$40 \leq x < 50$	26	$26 : 10 = 2,60$
$50 \leq x < 70$	39	$39 : 20 = 1,80$
$70 \leq x < 100$	20	$20 : 30 = 0,66$



EJERCICIO 5.3. En un hospital se han tomado los pesos (en kg) de 50 recién nacidos:

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

Agrupar en 5 intervalos de amplitud 0.5 kg. y dibujar el histograma.

6. Estadísticos

Definición 6 (Estadístico) *Estadístico es cualquier función de los datos de la muestra.*

Si los datos de la muestra se designan por x_1, x_2, \dots, x_n , cualquier función de los datos es un estadístico. Por tanto, solamente existen estadísticos cuando la variable observada X es cuantitativa. El poder de síntesis de los estadístico es muy grande, ya que cada estadístico resume el conjunto de todos los datos en un único valor. En contrapartida, la información suministrada por cada estadístico tiene que ser forzosamente menor que la información suministrada por toda la muestra. Dependiendo del tipo de información que proporcione un estadístico puede clasificarse en alguno de los siguientes tipos:

1. *Estadísticos de tendencia central o de localización.* Los más importantes son la media, la moda y la mediana muestrales.
2. *Estadísticos de dispersión.* Los más importantes son la varianza y la desviación típica muestrales, aunque también puede citarse el coeficiente de variación.

3. *Estadísticos de posición.* Los más importantes son los cuantiles muestrales y sus variantes: deciles, cuartiles y percentiles.

6.1. Estadísticos de tendencia central

La dimensión de los estadísticos de tendencia central siempre coincide con la de la variable observada. Por ejemplo, si estuviéramos observando el peso expresado en kilos de los individuos de una muestra, los estadísticos de localización también se expresarían en kilos.

Definición 7 *Media aritmética* La media aritmética de los datos está dada por:

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

En el caso de que haya datos repetidos la expresión es:

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

Ejemplo. 6.1. Obtener la nota media de un alumno con las calificaciones: 5, 7, 4 y 2. .

Solución: La nota media es la media aritmética de las calificaciones, es decir

$$\bar{x} = \frac{5 + 7 + 4 + 2}{4} = \frac{18}{4} = 4,5$$



Ejemplo. 6.2. En una clase de 10 alumnos se han registrado las siguientes las calificaciones: 6 alumnos un 5, 3 alumnos un 7 y un alumno un 9. Obtener la nota media.

Solución: Como ahora las puntuaciones se repiten, preparamos una tabla de frecuencias para realizar los cálculos,

x_i	f_i	$x_i f_i$
5	6	30
7	3	21
9	1	9
	$n = 10$	60

La nota media de la clase es

$$\bar{x} = \frac{\sum_{i=1}^n x_i f_i}{n} = \frac{60}{10} = 6$$



Definición 8 *Mediana muestral* Se llama mediana muestral Med a cualquier valor que deje tantos valores por debajo como encima

$$fr(x_i \leq Med) \geq 0,5, \quad fr(x_i \geq Med) \geq 0,5,$$

donde $fr(x_i \leq Med)$ es la frecuencia relativa de los datos menores o iguales que Med y $fr(x_i \geq Med)$ es la frecuencia relativa de los datos mayores o iguales que Med .

La mediana siempre existe, pero puede no ser única.

Ejemplo. 6.3. Un alumno ha obtenido las calificaciones: 5, 7, 4, 6 y 2. Obtener la nota mediana.

Solución: Ordenamos las calificaciones, es decir

$$2, 4, 5, 6, 7$$

Como hay un número impar de datos, la nota que deja la mitad de los valores por encima y por debajo es el 5. $Med = 5$ □

Ejemplo. 6.4. Un alumno ha obtenido las calificaciones: 8, 5, 7, 4, 9 y 2. Obtener la nota mediana.

Solución: Ordenamos las calificaciones, es decir

$$2, 4, 5, 7, 8, 9$$

En este caso, como hay un número par de datos, cualquier número entre 5 y 7 se podría tomar como mediana, pero por convenio se toma el punto medio entre las centrales 5 y 7. Es decir

$$Med = \frac{5 + 7}{2} = 6$$

□

Ejemplo. 6.5. Obtener la mediana para los datos de las edades de los pasajeros que están en la tabla 1 .

Solución: Como la mitad de los datos son $N/2 = 132/2 = 66$ y el número de datos es par, tomamos el valor central entre las edades que ocupan el lugar 66 y 67, que corresponde a 47 años. Es decir $Med = 47$ años. Significa que la mitad de los pasajeros tienen una edad inferior o igual a 47 años y la otra mitad una edad superior o igual a 47 años.

□

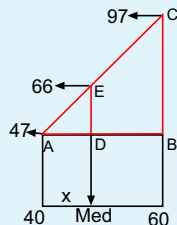
Ejemplo. 6.6. Obtener la mediana para los datos de las edades agrupados por clases como en la tabla del ejemplo 4.1.

Solución: Preparamos la tabla añadiendo las frecuencias acumuladas F_i .

Edades	f_i	F_i
$0 \leq x < 20$	14	14
$20 \leq x < 40$	33	47
$40 \leq x < 60$	50	97
$60 \leq x < 80$	31	128
$80 \leq x < 100$	4	132
	$n = 132$	

Calculamos la mitad de los datos $N/2 = 132/2 = 66$. Como hay 47 pasajeros hasta 40 años, y 97 pasajeros hasta 60 años, la mediana estará entre 40 y 60.

Para calcularla se interpola con la proporción de triángulos en rojo.



$$\frac{AB}{AD} = \frac{BC}{ED}$$

$$\left. \begin{array}{l} 20 \longrightarrow 50 \\ x \longrightarrow 19 \end{array} \right\} x = \frac{20 \cdot 19}{50} = 7,6$$

$$Med = 40 + x = 40 + 7,6 = 47,6 \text{ años}$$

□

Ejemplo. 6.7. En la siguiente distribución de notas, hallar Me .

x_i	200 – 240	240 – 280	280 – 320	320 – 360	360 – 400
f_i	57	82	73	31	15
F_i	57	139	212	243	258

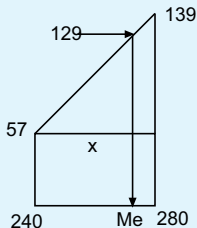
Solución:

La Me deja por debajo el 50 % de los datos, es decir $0,5 \times 258 = 129$. Mirando en las frecuencias acumuladas F_i corresponde a la clase 240-280. Interpolamos de acuerdo al gráfico.

En $280-240=40$ puntos se reparten $139 - 57 = 80$ observaciones y en x puntos se reparten $129 - 57 = 72$ observaciones, es decir

$$\left. \begin{array}{l} 40 \longrightarrow 82 \\ x \longrightarrow 72 \end{array} \right\} x = \frac{40 \cdot 72}{82} = 35,12$$

$$Med = 240 + x = 240 + 35,12 = \mathbf{275.12} \text{ puntos}$$



Definición 9 *Moda y clase modal* La moda muestral es el dato de mayor frecuencia relativa. La clase modal es la clase que tiene mayor frecuencia relativa por unidad de amplitud.

Ejemplo. 6.8. Hallar la moda Mo de los puntos obtenidos en un test.

x_i	200 – 240	240 – 280	280 – 320	320 – 360	360 – 400
f_i	57	82	73	31	15

Solución: La clase modal corresponde a la de mayor frecuencia (cuando las clases tienen la misma amplitud). En este caso, 240 – 280. Se toma como moda la marca de clase. $Mo = 260$ puntos. \square

Ejemplo. 6.9. Hallar la moda en la tabla del ejemplo 4.1.

Solución: La clase modal corresponde a la de mayor frecuencia, [40; 60) y se toma como moda la marca de clase, $Mo = 50$ años. \square

EJERCICIO 6.4. Los sueldos en euros en una empresa son: 1 director con 1800 euros ; 3 jefes con 1500; 6 encargados con 900 y 9 operarios con 480. Calcula el sueldo medio, la moda y la mediana.

6.2. Estadísticos de dispersión

Ejemplo. 6.10. Dos grupos uno de letras y otro de ciencias de 8 alumnos cada uno han realizado un test, obteniendo las siguientes puntuaciones. Compararlos estadísticamente

Grupo A	46	48	49	50	50	51	52	54
Grupo B	10	18	30	50	50	70	82	90

Solución: Resulta que si resumimos estos grupos con las medidas de centralización la media, moda y mediana coinciden.

$$\bar{x}_A = \bar{x}_B = Mo_A = Mo_B = Me_A = Me_B = 50$$

Con este ejemplo se aprecia la necesidad de las medidas estadísticas de dispersión o variabilidad, pues el grupo B tiene puntuaciones más dispersas o alejadas respecto de la media. \square

Como muestra el ejemplo anterior se necesitan estadísticos que registren la dispersión o variabilidad de los datos.

Definición 10 *Varianza muestral* Se llama varianza muestral al estadístico definido por:

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

Operando se obtiene la expresión más cómoda

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

Definición 11 *Desviación típica* Se llama desviación típica muestral a la raíz cuadrada, s , de la varianza muestral.

Ejemplo. 6.11. Hallar la media, varianza y desviación típica de las notas 3, 5, 7.

Solución:

$$\bar{x} = \frac{3 + 5 + 7}{3} = 5$$

$$s^2 = \frac{3^2 + 5^2 + 7^2}{3} - 5^2 = 2,67 \implies s = \sqrt{2,67} = 1,63$$

□

Ejemplo. 6.12. Hallar la media, varianza y desviación típica de las edades de la tabla 4.1.

Solución: Preparamos la tabla añadiendo las columnas x_i , $x_i f_i$ y $x_i^2 f_i$.

Clases	x_i	f_i	$x_i f_i$	$x_i^2 f_i$
$0 \leq x < 20$	10	14	140	1400
$20 \leq x < 40$	30	33	990	29700
$40 \leq x < 60$	50	50	2500	125000
$60 \leq x < 80$	70	31	2170	151900
$80 \leq x < 100$	90	4	360	32400
Σ		132	6160	340400

$$\bar{x} = \frac{\sum x_i f_i}{n} = \frac{6160}{132} = 46,67$$

$$s_x^2 = \frac{\sum x_i^2 f_i}{n} - \bar{x}^2 = \frac{340400}{132} - 46,67^2 = 400,7$$

$$s_x = \sqrt{400,7} = 20,02$$

□

Para practicar realiza los siguientes ejercicios.

EJERCICIO 6.5. Hallar la media y la desviación típica de las siguientes distribuciones de datos:

(a) 1, 4, 8, 9, 10.

(b) 3.2, 4.7, 5.1, 5.2, 6.3

EJERCICIO 6.6. Hallar la media y la desviación típica de la siguiente distribución de datos:

x_i	3	4	5	7	9
f_i	1	5	4	2	3

EJERCICIO 6.7. Hallar la media y la desviación típica de la siguiente distribución de datos:

Clases	10 – 20	20 – 30	30 – 40	40 – 50
f_i	1	5	3	1

Definición 12 *Coeficiente de variación* El coeficiente de variación se define como el cociente

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

supuesto que $\bar{x} \neq 0$.

Ejemplo. 6.13. El peso medio de los alumnos de una clase es 58,2 kg y su desviación típica 4.0 kg. Por otra parte la altura media es de 175 cm y su desviación típica es 5,0 cm. Calcula el coeficiente de variación y compara la dispersión de ambos grupos.

Solución: Como las variables tienen unidades distintas. Una mide kg y la otra cm, para comparar la dispersión acudimos al coeficiente de variación que no tiene unidades.

$$CV_{pesos} = \frac{4}{58,2} = 0,0687$$

$$CV_{alturas} = \frac{5}{175} = 0,0286$$

luego hay más dispersión en los pesos que en las alturas



6.3. Estadísticos de posición

Empecemos por el siguiente ejemplo. Las notas ordenadas de matemáticas y de física de una clase de 10 alumnos son:

matemáticas	4	5	6	7	7.8	7.9	8.1	9	9	10
física	2	2.5	3	3.5	3.8	4	4.5	5.5	6	7

En rojo se han puesto las notas obtenidas por el alumno Javier. ¿En que asignatura ha obtenido mejor resultado Javier?

En matemáticas, como hay 4 notas iguales o inferiores a un 7, tenemos que $4/10 = 0,40(40\%)$ de la clase han obtenido una nota igual o inferior a 7, este es el cuantil 0,4, $c_{0,4}$ o percentil 40 p_{40} . En física hay 7 notas iguales o inferiores a un 4.5, tenemos que $7/10 = 0,70(70\%)$ de la clase han obtenido una nota igual o inferior a 4,5, este es el cuantil 0,7, $c_{0,7}$ o percentil 70 p_{70} .

De todo ello vemos que dentro de su grupo la nota de física es mejor que la nota de matemáticas, pues ocupa una posición superior dentro del grupo de datos.

Definición 13 *Cuantil de orden α* Se llama cuantil muestral de orden α a cualquier valor C_α que verifique las dos desigualdades siguientes

$$fr(x_i \leq C_\alpha) \geq \alpha, \quad fr(x_i \geq C_\alpha) \geq 1 - \alpha,$$

para cualquier α entre 0 y 1.

Por tanto, la mediana es el cuantil muestral $C_{0,5}$. Otros cuantiles importantes son los siguientes:

Definición 14 *Cuartiles* Se llama primer cuartil muestral Q_1 al cuantil $C_{0,25}$. El segundo cuartil muestral es la mediana. Se llama tercer cuartil muestral Q_3 al cuantil $C_{0,75}$.

Definición 15 *Deciles* Se llaman deciles muestrales a los cuantiles $C_{0,1}, C_{0,2}, \dots, C_{0,9}$.

Definición 16 *Percentiles* Se llaman percentiles muestrales a los cuantiles indicados en porcentaje. Así el cuantiles $C_{0,15}$ es el percentil 15 p_{15} , etc.

Ejemplo. 6.14. Hallar los cuartiles Q_1, Q_2 de las calificaciones ya ordenadas de un grupo:

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

Solución: Con este ejemplo se explica la técnica para calcular cuantiles de datos sin agrupar.

Como Q_1 deja por debajo el 25 % de los datos, se calcula

$$0,25 \times (n + 1) = 0,25 \times (31) = 7,75 \text{ datos}$$

y se toma entre el séptimo y el octavo el valor

$$Q_1 = x_7 + 0,75(x_8 - x_7) = 1 + 0,75(1 - 1) = 1$$

Como $Q_2 = Me$ deja por debajo el 50 % de los datos, se calcula

$$0,5 \times (n + 1) = 0,5 \times (31) = 15,5 \text{ datos}$$

y se toma entre el decimoquinto y el decimosexto el valor

$$Q_2 = x_{15} + 0,5(x_{16} - x_{15}) = 3 + 0,5(3 - 3) = 3$$

□

Ejemplo. 6.15. Hallar Q_1, Q_3 de la siguiente distribución de notas,

x_i	1	2	3	4	5	6	7	8	9	10
f_i	7	15	41	52	104	69	26	13	19	14
F_i	7	22	63	115	219	288	314	327	346	360

Solución:

- Como Q_1 deja por debajo el 25 % de los datos, se calcula

$$0,25 \times (n + 1) = 0,25 \times (361) = 90,25 \text{ datos}$$

y se toma entre el 90-dato y el 91-dato el valor

$$Q_1 = x_{90} + 0,25(x_{91} - x_{90}) = 4 + 0,25(4 - 4) = 4$$

- Como Q_3 deja por debajo el 75 % de los datos, se calcula

$$0,75 \times (n + 1) = 0,75 \times (361) = 270,75 \text{ datos}$$

y se toma entre el 270-dato y el 271-dato el valor

$$Q_3 = x_{270} + 0,75(x_{271} - x_{270}) = 6 + 0,75(6 - 6) = 6$$

□

Ejemplo. 6.16. Hallar los percentiles p_{15}, p_{40}, p_{80} de la tabla 1.

Solución:

- Como p_{15} deja por debajo el 15 % de los datos, se calcula

$$0,15 \times (n + 1) = 0,15 \times (133) = 19,95 \text{ datos}$$

y se toma entre el 19-dato y el 20-dato el valor

$$p_{15} = x_{19} + 0,95(x_{20} - x_{19}) = 24 + 0,95(27 - 24) = \mathbf{26.85}$$

- Para p_{40} , $0,40 \times (n + 1) = 0,40 \times (133) = 53,20$ datos y se toma entre el 53-dato y el 54-dato el valor

$$p_{40} = x_{53} + 0,2(x_{54} - x_{53}) = 44 + 0,2(44 - 44) = \mathbf{44}$$

- Para p_{80} , $0,8 \times (n + 1) = 0,8 \times (133) = 106,4$ datos y se toma entre el 106-dato y el 107-dato el valor

$$p_{80} = x_{106} + 0,4(x_{107} - x_{106}) = 64 + 0,4(64 - 64) = \mathbf{64}$$

□

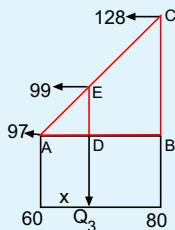
Ejemplo. 6.17. Obtener el tercer cuartil Q_3 para los datos de las edades agrupados por clases como en la tabla 4.1.

Solución: Preparamos la tabla añadiendo las frecuencias acumuladas F_i .

Edades	f_i	F_i
$0 \leq x < 20$	14	14
$20 \leq x < 40$	33	47
$40 \leq x < 60$	50	97
$60 \leq x < 80$	31	128
$80 \leq x < 100$	4	132
	$n = 132$	

Calculamos el 0,75 de los datos $0,75 \times n = 99$. Como hay 97 pasajeros hasta 60 años, y 128 pasajeros hasta 80 años, el Q_3 estará entre 60 y 80.

Para su cálculo se interpola con la proporción de triángulos en rojo.



$$\frac{AB}{AD} = \frac{BC}{ED}$$

$$\left. \begin{array}{l} 20 \longrightarrow 31 \\ x \longrightarrow 2 \end{array} \right\} x = \frac{20 \cdot 2}{31} = 1,29$$

$$Q_3 = 60 + x = 60 + 1,29 = \mathbf{61.29} \text{ años}$$



Definición 17 *Extremos: Mínimo y Máximo* Los extremos muestrales son el menor dato o mínimo y el mayor dato o máximo de la muestra.

A partir de los cuartiles pueden definirse el siguiente estadístico de dispersión:

Definición 18 *Rango intercuartílico* El rango intercuartílico es la diferencia entre el tercer y el primer cuartil, es decir, $RIC = C_{0,75} - C_{0,25}$.

A partir de los extremos se define el rango de la muestra:

Definición 19 *Rango* El rango muestral es la diferencia entre el máximo y el mínimo.

EJERCICIO 6.8. Un dentista anota el número de caries de cada uno de los 100 niños de un colegio. La información se almacena en una tabla:

Nº de caries	0	1	2	3	4
f_i	25	20	x	15	y
h_i	0.25	0.2	z	0.15	0.05

1. Completa la tabla , hallando x, y, z .
2. Haz un diagrama de sectores.
3. Calcula el número medio de caries.

EJERCICIO 6.9. Para comprobar la resistencia de unas varillas de nailon se someten 250 varillas a un test de resistencia. Se comprueba si se rompen cuando se aplica una fuerza sobre 5 puntos diferentes de la varilla. El número de roturas sufridas por cada varilla aparece en la tabla

Nº de roturas	0	1	2	3	4	5
Nº de varillas	141	62	31	14	1	1

1. Calcula el número medio de roturas por varilla y el porcentaje de varillas que sufren más de 2 roturas.
2. Calcula la moda, mediana y varianza de la distribución.

EJERCICIO 6.10. La tabla registra el tiempo en horas que dedicaron 62 personas a ver la televisión un fin de semana.

Nº horas	$[0; 0,5)$	$[0,5; 1,5)$	$[1,5; 2,5)$	$[2,5; 4)$	$[4; 8)$
Nº de personas	10	10	18	12	12

1. Dibujar el histograma correspondiente.
2. Calcula la media y la desviación típica.

Índice alfabético

coeficiente de variación, 35

cuantil, 37

 cuartil, 38

 decil, 38

 percentil, 38

desviación típica, 33

frecuencia, 9

 absoluta, 9, 12

 relativa, 9, 12

gráfico, 15

 barras, 19

 histograma, 20

 pictograma, 16

media aritmética, 25

mediana, 27

moda, 31

rango, 43

variable, 3

 continua, 4

 cualitativa, 3

 cuantitativa, 3

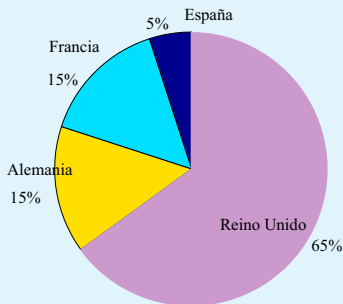
 discreta, 3

varianza, 33

Soluciones a los Ejercicios

Ejercicio 5.1.

país	España	Francia	Alemania	Reino Unido
ventas	5 %	15 %	15 %	65 %
ángulos	$0,05 \times 360^\circ = 18^\circ$	54°	54°	234°



Ejercicio 5.1

Ejercicio 5.2. Para la empresa Campsa se tiene que

$$\frac{81^{\circ}}{360^{\circ}} = 0,225 \implies 22,5 \%$$

Como la empresa Oil tiene un 12,5 % del total,

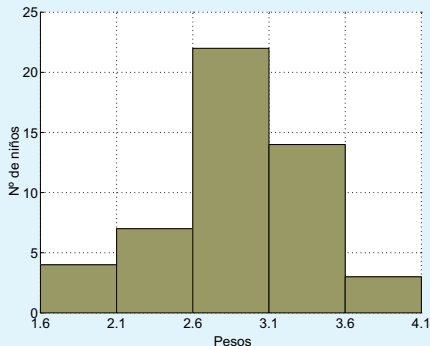
$$x = 0,125 \times 360^{\circ} = 45^{\circ}$$

La suma total debe ser 360° , luego $y = 114^{\circ}$.

Ejercicio 5.2

Ejercicio 5.3. Entre el valor máximo $x_{max} = 3,9$ y el valor mínimo $x_{min} = 1,8$ hay 2.1 kg . Con 5 intervalos de 0.5 se abarca un rango de 2.5. La diferencia de 0.4 se reparte entre ambos extremos, luego empezamos con un valor de 1.6

clase	f_i
[1,6; 2,1)	4
[2,1; 2,6)	7
[2,6; 3,1)	22
[3,1; 3,6)	14
[3,6; 4,1)	3
	50



Ejercicio 5.3

Ejercicio 6.4. Preparamos una tabla de frecuencias

Puesto	x_i	f_i	$x_i f_i$
operarios	480	9	4320
encargados	900	6	5400
jefes	1500	3	4500
director	1800	1	1800
		$n = 19$	16020

El sueldo medio es

$$\bar{x} = \frac{16020}{19} = 843,16$$

La moda es el sueldo de 480 euros pues el de mayor frecuencia. Para hallar la mediana, como $0,5 \times (n + 1) = 10$ el décimo sueldo de menor a mayor corresponde a $Me = 900$ euros.

Ejercicio 6.4

Ejercicio 6.5(a)

$$\bar{x} = \frac{1 + 4 + 8 + 9 + 10}{5} = 6,4$$

$$s^2 = \frac{1^2 + \dots + 10^2}{5} - 6,4^2 = 11,44 \implies s = \sqrt{11,44} = 3,38$$



Ejercicio 6.5(b)

$$\bar{x} = \frac{3,2 + \cdots + 6,3}{5} = \frac{24,5}{5} = 4,9$$

$$s^2 = \frac{3,2^2 + \cdots + 6,3^2}{5} - 4,9^2 = \frac{125,07}{5} = 1,004 \implies s = \sqrt{1,004} = 1$$



Ejercicio 6.6.

x_i	f_i	$x_i f_i$	$x_i^2 f_i$
3	1	3	9
4	5	20	80
5	4	20	100
7	2	14	98
9	3	27	243
Σ	15	84	530

$$\bar{x} = \frac{\sum x_i f_i}{n} = \frac{84}{15} = 5,6$$

$$s_x^2 = \frac{\sum x_i^2 f_i}{n} - \bar{x}^2 = \frac{530}{15} - 5,6^2 = 3,97$$

$$s_x = \sqrt{3,97} = 1,99$$

Ejercicio 6.6

Ejercicio 6.7.

x_i	f_i	$x_i f_i$	$x_i^2 f_i$
15	1	15	225
25	5	125	3125
35	3	105	3675
45	1	45	2025
Σ	10	290	9050

$$\bar{x} = \frac{\sum x_i f_i}{n} = \frac{290}{10} = 29$$

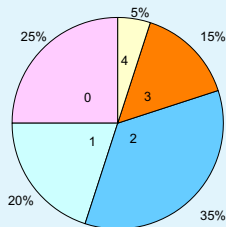
$$s_x^2 = \frac{\sum x_i^2 f_i}{n} - \bar{x}^2 = \frac{9050}{10} - 29^2 = 64$$

$$s_x = \sqrt{64} = 8$$

Ejercicio 6.7

Ejercicio 6.8.

1. Como $\sum_i h_i = 1$, entonces $z = 0,35$. Como $x = z \cdot 100 = 35$ y como $\sum_i f_i = 100$ entonces $y = 5$.
2. Tenemos los porcentajes, luego hallamos los ángulos de los sectores.



$$0,25 \times 360^\circ = 90^\circ$$

$$0,20 \times 360^\circ = 72^\circ$$

$$0,35 \times 360^\circ = 126^\circ$$

$$0,15 \times 360^\circ = 54^\circ$$

$$0,05 \times 360^\circ = 18^\circ$$

3.

$$\bar{x} = \frac{0 \cdot 25 + 1 \cdot 20 + \dots + 4 \cdot 5}{100} = 1,55$$

Ejercicio 6.8

Ejercicio 6.9.

1.

x_i	f_i	$x_i f_i$	$x_i^2 f_i$
0	141	0	0
1	62	62	62
2	31	62	124
3	14	42	126
4	1	4	16
5	1	5	25
	250	175	353

$$\bar{x} = \frac{175}{250} = 0,7$$

Con más de 2 roturas hay
 $14 + 1 + 1 = 16$ varillas, luego
 $\frac{16}{250} \times 100 = 16,4\%$

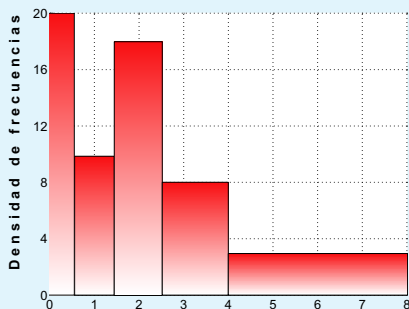
2. La moda $Mo = 0$. Como $N/2 = 125$, la mediana $Med = 0$. Para la varianza se tiene

$$S_x^2 = \frac{\sum x_i^2 f_i}{N} - \bar{x}^2 = \frac{353}{250} - 0,7^2 = 0,922$$

Ejercicio 6.9

Ejercicio 6.10.

clase	x_i	f_i	d.f.	$x_i f_i$	$x_i^2 f_i$
[0; 0,5)	0,25	10	20	2,5	0,625
[0,5; 1,5)	1	10	10	10,0	10,0
[1,5; 2,5)	2	18	18	36,0	72,0
[2,5; 4)	3,25	12	8	39,0	126,75
[4; 8)	6	12	3	72,0	432,0
		62		159,5	641,375



$$\bar{x} = \frac{159,5}{62} = 2,57 \text{ horas}$$

$$S_x^2 = \frac{\sum x_i^2 f_i}{N} - \bar{x}^2$$

$$= \frac{641,375}{62} - 2,57^2 = 3,74$$

$$S_x = 1,93$$