ESTADÍSTICA BÁSICA RESUMEN

Caracteres y escalas de medición

Al hacer un trabajo estadístico hay que decidir los caracteres (las propiedades) que desean estudiarse. Un carácter puede ser cuantitativo o cualitativo.
Los valores que toma un carácter pueden medirse en distintas escalas: nominal, ordinal, de intervalo, o de proporción.

- La escala nominal consiste en situar a cada individuo o elemento en una u otra clase dada (por ejemplo, hombre/mujer; lugar de nacimiento). Pertenecer a una u otra clase no significa ser mejor o peor, indica que son distintos.
- La escala ordinal sitúa los posibles valores en orden (primero, segundo, ...), sin que la *distancia* entre dos posiciones consecutivas sea necesariamente constante, fija. En esta escala puede distinguirse, además, entre mayor y menor. Por ejemplo, la posición de los equipos de fútbol en el Campeonato de Liga; o las categorías profesionales en una empresa.

Las escalas nominal y ordinal son apropiadas para caracteres cualitativos.

- La escala de intervalo permite asignar a cada individuo un número para así indicar su posición exacta a lo largo de una escala continua. Por ejemplo, la temperatura medida en grados Celsius, donde 10 ºC significa más calor que 5 ºC, pero no el doble de calor.
- La escala de proporción (o proporcional) es la más perfecta. En ella existe un cero absoluto y, además, tiene sentido hablar de doble o mitad (un ejemplo de esta medida sería la longitud). Las escalas de intervalo y proporcional se usan para medir caracteres cuantitativos.

Tablas de frecuencias

Se utilizan para facilitar la lectura e interpretación de grandes conjuntos de datos. Los datos suelen agruparse, indicando su frecuencia absoluta o relativa; simple o acumulada.
La agrupación puede hacerse también en intervalos de clase. El punto medio de cada uno de esos intervalos sería el valor que representa a todos; se llama marca de clase (M.c.).

Ejemplos:

x_i	f_i	F_i
0	1	1
1	5	6
2	12	18
3	10	28
4	15	43
5	17	60
6	11	71
7	7	78
8	0	78
9	2	80
Totales	80	80

Intervalo	M.c.	f_i	%	F_i	%a
[0, 3)	1,5	0	0	0	0
[3, 6)	4,5	8	6,2	8	6,2
[6, 9)	7,5	31	23,8	39	30
[9, 12)	10,5	51	39,2	90	69,2
[12, 15)	13,5	33	25,4	123	94,6
[15, 18)	16,5	7	5,4	130	100
Totales		130	100	130	100

Diagrama de tallo y hojas

Es otro método de organización y visualización de un conjunto numérico de datos. Cada número del conjunto se representa por una *hoja* y un *tallo*. La hoja es la cifra de las unidades; el tallo es la cifra de las decenas. Una raya vertical separa el tallo de las hojas.

Ejemplo: El diagrama adjunto es el asociado a los datos
45 47 24 42 43 51 58 60 45 47 47 67 82 51 55

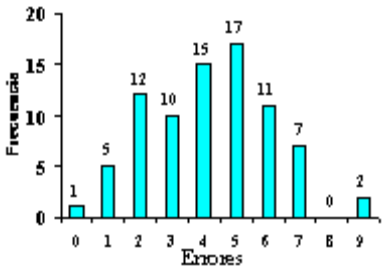
2 | 4
4 | 2 3 5 5 7 7 7
5 | 1 1 5 8
6 | 0 7
8 | 2

2 | 4 representa 24 corbatas

Gráficos estadísticos

Diagramas de barras

Son gráficos que representan cada valor de la variable mediante una barra proporcional a la frecuencia con que se presenta. Las barras deben estar separadas, como en la figura del ejemplo, que se corresponde con la primera tabla del ejemplo anterior.



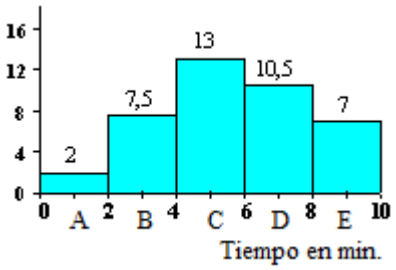
Los diagramas de barras son apropiados para datos medidos en escala nominal u ordinal.

Este gráfico puede ser el "resumen visual" del número de errores cometidos por 80 personas al realizar un determinado test.

Histogramas

Se usan para variables agrupadas en intervalos, asignando a cada intervalo un rectángulo de superficie proporcional a su frecuencia. Por tanto, como la base, la amplitud del intervalo es 2, la altura de cada rectángulo se halla dividendo la frecuencia que representa entre 2. (En la figura adjunta, el segundo rectángulo [B] tiene altura 7,5 = 15/2; el 3º, [C], 13 = 26/2;… Este histograma puede resumir el tiempo de espera de 80 personas que han tomado un autobús. El intervalo [C] = [4, 6], indica que 26 personas han esperado entre 2 y 4 minutos).

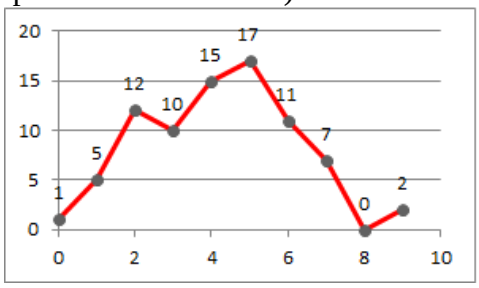
Intervalo	Mc	f_i
[0, 2)	1	4
[2, 4)	3	15
[4, 6)	5	26
[6, 8)	7	21
[8, 10)	9	14
Totales		80

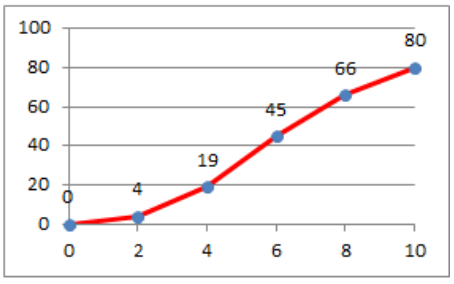


Los histogramas son apropiados para variables continuas (medidas en escala de intervalo o de proporción); por eso las barras van unidas y tienen la anchura indicada por el intervalo.

Poligonal de frecuencias

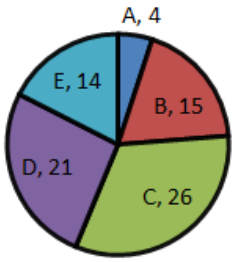
Los histogramas, y algunos diagramas de barras, también se pueden representar por una poligonal de frecuencias, que es la línea que une los puntos correspondientes a las frecuencias de cada valor (extremos superiores de las barras). Pueden ser simples o de frecuencias acumuladas.





La poligonal (simple) de la izquierda representa los datos del diagrama de barras; la de la derecha, acumulada, se corresponde con el histograma.

Diagrama de sectores. En estos gráficos, cada suceso viene representado por un sector circular de amplitud proporcional a su frecuencia. La amplitud de cada sector se halla mediante una regla de tres.



Este ejemplo se corresponde con los datos del histograma de arriba. La amplitud de cada sector, en grados, es:

A, 4: 18º; B, 15: 67,5º; C, 26: 117º; D, 21: 94,5º; E, 14: 63º ,

Medidas de centralización

Están relacionadas con el promedio de los datos estudiados, y dan una idea de los valores más representativos para todo el conjunto.

La media aritmética

Se calcula sumando el valor de todos los datos y dividiendo por el número de ellos. Esto es:

$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$; o más breve: $\bar{x} = \frac{\Sigma x_i}{n}$

- Para datos agrupados: $\bar{x} = \frac{\Sigma x_i f_i}{\Sigma f_i}$, donde f_i es el número de veces que se repite el valor x_i.

Media ponderada: $\bar{x}_p = \frac{\Sigma x_i p_i}{\Sigma p_i}$, siendo p_i el peso del dato x_i.

La mediana: es el valor del dato que ocupa el lugar intermedio. Los datos deben estar ordenados.
La moda. Es el valor que se presenta con mayor frecuencia.

Medidas de posición

Indican la situación, en términos porcentuales, de algunos elementos de la distribución. Los datos estén ordenados de menor a mayor.

Amplitud, rango o recorrido. Es la diferencia entre los valores de los datos máximo y mínimo.
La información que proporciona es imprecisa, pues sólo tiene en cuenta los valores extremos.
Cuartiles, deciles y percentiles

- Cuartiles: Valores de las posiciones correspondientes al 25 %, al 50 % y al 75 % de los datos.
- Deciles: Valores correspondientes al 10 %, 20 %, ... y 90 % de los datos
- Percentiles (o centiles) dan el valor de la posición correspondiente a cualquier porcentaje.

En todos los casos, su cálculo requiere aplicar la interpolación.

Medidas de dispersión

Dan una idea del *alejamiento* de los datos respecto de la media.

Varianza y desviación típica

- La varianza:

$s^2 = \frac{\sum (x_i - \bar{x})^2}{n}$ $s^2 = \frac{\sum x_i^2}{n} - \bar{x}^2$ Alternativas: $s^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i}$ $s^2 = \frac{\sum x_i^2 f_i}{\sum f_i} - \bar{x}^2$

- La desviación típica es la raíz cuadrada de la varianza. En consecuencia:

$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$; $s = \sqrt{\frac{\sum x_i^2}{n} - \bar{x}^2}$ $s = \sqrt{\frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i}}$

Ejemplo: La desviación típica de los conjuntos de datos: (a) 1, 4, 5, 6, 14 y (b) 1, 2, 3, 10, 14, es, respectivamente $s = \sqrt{18{,}8} = 4{,}34$ y $s = \sqrt{26} = 5{,}1$.

El coeficiente de variación. Es una medida de la *dispersión relativa* de dos conjuntos de datos. Se define como: $CV = \frac{s}{\bar{x}}$

- El coeficiente de variación suele darse en porcentajes: $CV = \frac{s}{\bar{x}} \cdot 100$. Un CV mayor del 30 % indica que la media es poco representativa como medida del promedio.

PROBLEMAS PROPUESTOS

1. En la siguiente tabla se dan los datos correspondientes a las notas de Matemáticas de 60 alumnos de 1º Bachillerato.

Notas	IN: [1, 5)	SF: [5, 6)	BI: [6, 7)	NT: [7, 9)	SB: [9, 10]
Nº de alumnos	20	13	12	10	5

a) Haz una tabla de frecuencias y porcentajes, simple y acumulada.
b) Dibuja el correspondiente histograma.
c) Representa los datos mediante un diagrama de sectores y mediante una poligonal acumulativa.

2. El número de turismos matriculados en España, para el período 1996–2005, se da en la siguiente tabla:

Año	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
Miles de turismos	911	1.016	1.193	1.406	1.381	1.426	1.332	1.382	1.517	1.529

a) Tomando como base 100 el número de turismos matriculados en el año 1996, expresa en números índices la variación de la serie.
b) Representa los datos mediante una poligonal simple

3. La precipitación (P) y la temperatura media mensual (T) registradas en Soria a lo largo del año son:

Mes	E	F	M	A	M	J	J	A	S	O	N	D
P (mm)	44	45	48	47	62	55	32	31	47	46	49	55
T (ºC)	1,3	3,1	5,6	7,5	10,6	15,6	18,1	18,1	15	9,4	5,6	3,1

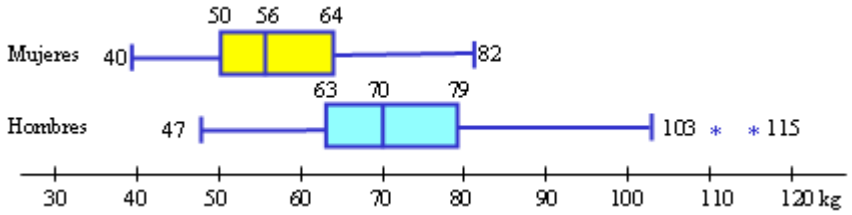
Representa gráficamente estos datos mediante un climograma.

4. Siete estudiantes han leído este curso el siguiente número de libros: 3 4 5 6 5 7 5
Para estos datos, determina:
a) La media b) La mediana c) La moda d) El rango

5. En una empresa hay 3 directivos, 50 operarios y 8 vendedores. Los sueldos mensuales, en euros, de cada categoría son los siguientes: directivos, 4.000; operarios, 1.400; vendedores, 2.000.
a) Halla la moda, la mediana y la media de los sueldos.
b) ¿Qué medida es más representativa del promedio?

6. En primero de bachillerato de un centro escolar hay tres grupos, A, B y C, con 30, 35 y 25 alumnos, respectivamente. La nota media en Matemáticas fue, también respectivamente, de 5,3, 6,5 y 5,6. Halla la nota media de Matemáticas de todos los alumnos de primero.

7. El gráfico siguiente representa los pesos (en kg) de un grupo similar de hombres y mujeres.



a) Indica los valores de las medianas respectivas.
b) ¿Cuánto vale en cada caso el rango intercuartílico?
c) ¿Hay algún elemento extraño? ¿Cuál es su peso?
d) ¿Qué porcentaje de mujeres pesa entre 40 y 50 kg?
e) ¿Dónde se da más homogeneidad de pesos, entre los más flacos o entre los más pesados?

8. El cociente intelectual de los 210 alumnos de un centro de bachillerato se da en la tabla adjunta:

Intervalo	[82, 90)	[90, 98)	[98, 106)	[106, 114)	[114, 122)	[122, 130)	[130, 138)	[138, 146
Frecuencia	12	32	49	54	30	17	11	5

a) Calcula los cuartiles y el rango intercuartílico.
b) Halla la diferencia entre los deciles 3 y 6.
c) Calcula la puntuación necesaria para pertenecer al 15 % de alumnos con mayor cociente intelectual.

9. Se ha preguntado a 50 mujeres sobre su número de hijos, obteniéndose los resultados:

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

Construye la tabla de frecuencias y calcula la media, varianza y desviación típica.

10. Se ha realizado una encuesta a los 40 empleados de una empresa para saber cuanto tiempo tardan en llegar desde su casa hasta su puesto de trabajo. Las respuestas, en minutos, son las siguientes:

30 42 37 50 15 35 90 65 38 45 30 12 78 20 35 41 25 32 85 25
41 28 50 30 20 60 14 36 48 32 27 30 76 30 51 28 25 22 17 10

a) Construye la tabla de frecuencias agrupando los datos en intervalos.
b) Calcula la mediana, la moda, la media y la desviación típica.

11. Halla la media y la desviación típica de los datos correspondientes al diagrama de tallo y hojas adjunto.

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

3 | 7 representa 37 kilos

12. Los rendimientos medios (en kilogramos por hectárea) en España, para los cereales que se indican, fueron:

Año	2010	2011	2012	2013	2014
Trigo	2150	3100	2300	2830	2840
Maíz	9450	9220	9720	9510	9110

Halla los rendimientos medios para el quinquenio de cada cereal. ¿Qué cereal es más fiable?

13. A un congreso asisten seis mujeres cuyas edades son: 27 34 38 42 33 36, años
a) Calcula la media y varianza de sus edades.
b) Cinco años después coinciden las mismas mujeres. A partir de los cálculos anteriores, halla la nueva media y varianza de sus edades.

14. El siguiente gráfico representa un total de 600 elementos. ¿Cuál es la frecuencia de cada categoría?