

U.D. 7: Estadística

IES Arroyo de la Miel

Índice

1. Conceptos estadísticos	2
2. Tablas de frecuencias	2
3. Gráficos estadísticos	3
4. Parámetros estadísticos	4
4.1. Parámetros de centralización	4
4.2. Parámetros de dispersión	5
4.3. Parámetros de posición	6
5. Variables estadísticas bidimensionales	7
5.1. Tablas simples y de doble entrada	7
5.2. Parámetros estadísticos	8
6. Correlación	9
7. Regresión	11

1. Conceptos estadísticos

Población: es el conjunto de todos los elementos cuyo conocimiento nos interesa y serán objetos de nuestro estudio.

Muestra: es un subconjunto, extraído de la población, cuyo estudio sirve para inferir características de toda la población.

Tamaño muestral: número de individuos que forman la muestra.

Individuo: es cada uno de los elementos que forman la población o la muestra.

Caracteres y variables: caracteres son los aspectos que deseamos estudiar en los individuos de una población. Cada carácter puede tomar distintos valores o modalidades. Una variable estadística recorre todos los valores de un cierto carácter.

Clasificación de las variables estadísticas:

- *Cualitativas:* no toman valores numéricos.
- *Cuantitativas discretas:* toman valores numéricos aislados.
- *Cuantitativas continuas:* pueden tomar todos los valores de un intervalo.

Clase: cada uno de los intervalos en los que se agrupan los datos.

2. Tablas de frecuencias

Las tablas de frecuencias sirven para ordenar y organizar los datos estadísticos. Con ellas, una masa amorfa de datos pasa a ser una colección ordenada y perfectamente inteligible.

Frecuencia absoluta (n_i): es el número de veces que aparece el resultado x_i .

Frecuencia relativa (f_i): es la proporción en que aparece cada valor. Se hallan dividiendo las frecuencias absolutas entre el número total de datos.

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

Frecuencia absoluta acumulada (N_i): nos dice el número de datos que hay igual o inferiores a uno x_i . Se obtiene con la suma de la frecuencia de ese valor con todas las anteriores.

Frecuencia relativa acumulada (F_i): es el resultado de dividir cada frecuencia acumulada por el número total de datos.

$$F_i = \frac{N_i}{N}$$

Cuando la variable toma muchos valores, la tabla que se obtiene es demasiado grande y, por tanto, poco clarificadora. La solución es agrupar los diferentes valores de la variable en intervalos (**intervalos de clase**). En este caso se define **marca de clase** (c_i) como el punto medio del intervalo.

Ejemplo:

Ejercicio 1 *La estatura, en centímetros, de 30 personas, viene dada por los siguientes datos:*

164 150 166 147 170 163 164 166 164 167 168 175 168 159 160 167 168
165 167 168 161 164 167 168 155 163 165 160 164 163

Elabora una tabla de frecuencias, agrupando los datos en intervalos de longitud 5.

3. Gráficos estadísticos

Diagrama de barras: se representa cada dato mediante una línea vertical o barra de altura proporcional a la frecuencia absoluta del dato. Las barras van separadas unas de otras y no se puede usar este tipo de gráficos para variables cuantitativas continuas.

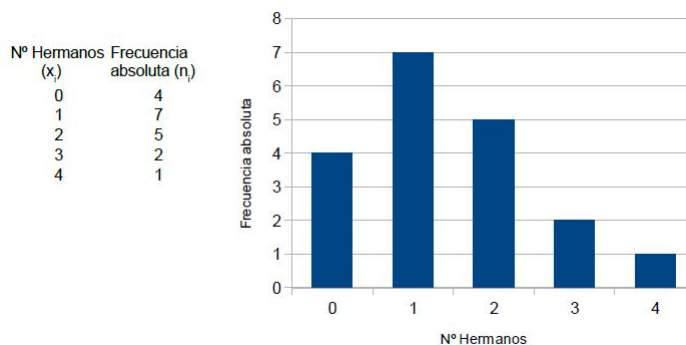
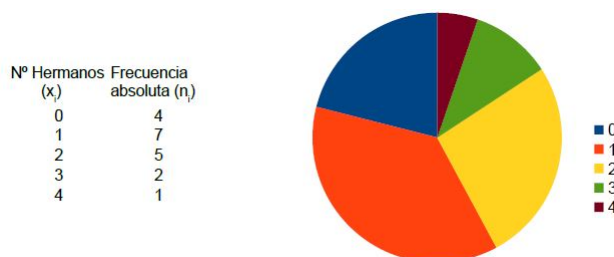


Diagrama de sectores: cada dato se representa mediante un sector circular cuya área es proporcional a la frecuencia del dato.



Histograma: se representa cada intervalo mediante un rectángulo cuya base coincide con la amplitud del intervalo y cuya área es proporcional a la frecuencia del intervalo.

Polígono de frecuencias: es la línea que une los extremos superiores de cada una de las barras del diagrama de barras o los puntos medios de los lados superiores de los rectángulos del histograma.

Ejercicio 2 *La estatura, en centímetros, de 30 personas, viene dada por los siguientes datos:*

164 150 166 147 170 163 164 166 164 167 168 175 168 159 160 167 168
165 167 168 161 164 167 168 155 163 165 160 164 163

Representa gráficamente el histograma y el polígono de frecuencias acumuladas.

4. Parámetros estadísticos

4.1. Parámetros de centralización

Los parámetros de centralización nos indican en torno a qué valor (centro) se distribuyen los datos. Son de este tipo la media, la mediana y la moda.

- a) **Media:** es el centro de gravedad de la distribución. Es decir, si las barras tuvieran peso, la media es el punto donde habrá que sostener la tablilla en que se sitúan. Análogamente ocurre con las distribuciones dadas mediante histogramas.

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

- b) **Mediana:** es aquel valor que, al ordenar las observaciones de menor a mayor, ocupa el lugar central. Para hallarla, buscamos en las frecuencias absolutas acumuladas aquella que supera $N/2$.
- c) **Moda:** es aquel valor que tiene mayor frecuencia.

Ejercicio 3 *La estatura, en centímetros, de 30 personas, viene dada por los siguientes datos:*

164 150 166 147 170 163 164 166 164 167 168 175 168 159 160 167 168
165 167 168 161 164 167 168 155 163 165 160 164 163

Halla la media, la moda y la mediana.

4.2. Parámetros de dispersión

Los parámetros de dispersión informan sobre cuánto se alejan del centro, los valores de la distribución. La mayoría de los parámetros de dispersión están asociados con un parámetro central al cual complementan:

- Mediana: $Me \rightarrow$ Desviación mediana, recorrido intercuartílico...
- Media: $\bar{x} \rightarrow$ Desviación media, varianza, desviación típica, coeficiente de variación...

- a) **Rango o recorrido:** es la diferencia entre el dato mayor y el dato menor.

- b) **Varianza**

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

- c) **Desviación típica**

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

- d) **Coeficiente de variación**

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

Al no tener unidades el coeficiente de variación, se usa para comparar la dispersión de las distintas distribuciones.

Ejercicio 4 La estatura, en centímetros, de 30 personas, viene dada por los siguientes datos:

164 150 166 147 170 163 164 166 164 167 168 175 168 159 160 167 168
165 167 168 161 164 167 168 155 163 165 160 164 163

Halla el recorrido, la varianza, la desviación típica y el coeficiente de variación.

4.3. Parámetros de posición

Los **cuantiles** son parámetros estadísticos que dividen los datos en partes iguales. Los cuantiles más frecuentes son los cuartiles, los deciles y los percentiles.

- a) **Cuartiles:** dividen a la muestra en cuatro partes iguales y se representan mediante Q_1, Q_2, Q_3 . El cuartil Q_i se corresponde con el primer dato de la muestra cuya posición supere el valor $\frac{i \cdot N}{4}$. Observar que $Q_2 = Me$
- b) **Deciles:** os deciles dividen a la muestra en diez partes iguales, y se representan por D_i . El decil D_i se corresponde con el primer dato de la muestra cuya posición supere el valor $\frac{i \cdot N}{10}$. Observar que $D_5 = Q_2 = Me$.
- c) **Percentiles:** os percentiles dividen la muestra en cien partes iguales y se representan por P_i . El percentil P_i se corresponde con el primer dato cuya posición supere el valor $\frac{i \cdot N}{100}$. Observar que $P_{50} = D_5 = Q_2 = Me$; $P_{25} = Q_1$; $P_{75} = Q_3$; $P_{10} = D_1$; $P_{20} = D_2, \dots$

Ejercicio 5 Las notas de matemáticas de una clase viene recogidas en la siguiente tabla:

Calificación	1	3	4	5	7	9
Nº de alumnos/as	2	5	5	10	6	2

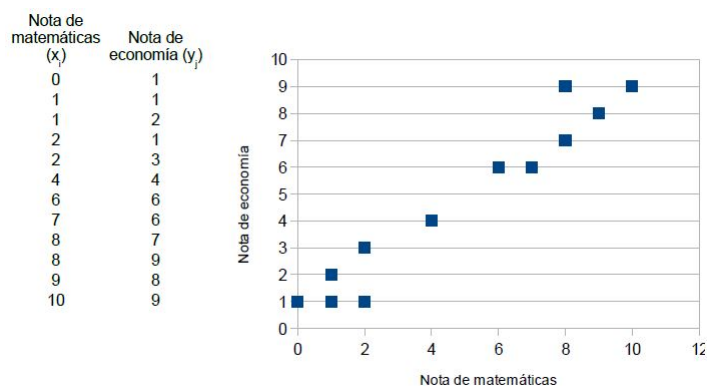
- a) Representa gráficamente la distribución.
- b) Halla la media, la moda y la mediana.
- c) Calcula la desviación típica.

- d) ¿Qué nota ocupa el percentil 30?
- e) Compara la dispersión con otra clase cuya media es 6,4 y tiene una desviación típica en la nota de 1,2.

5. Variables estadísticas bidimensionales

Se llama distribución bidimensional a un colectivo en el que cada uno de sus individuos lleva asociados los valores de dos variables x e y . Si interpretamos cada par de valores $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ como un punto del plano, el conjunto de todos ellos se llama nube de puntos o diagrama de dispersión.

Ejemplo:



Ejercicio 6 La siguiente tabla muestra los resultados que se han obtenido al preguntarle a un grupo de alumnos por su nota de matemáticas y de economía:

Nota de matemáticas	2	3	3	4	5	5	6	6	7	7	8	8	9
Nota de economía	2	2	3	5	5	6	6	7	7	7	7	8	9

Representa los datos gráficamente.

5.1. Tablas simples y de doble entrada

En la **tabla simple con frecuencias**, aparecen cada par de datos $(x_i; y_j)$ con su correspondiente frecuencia n_{ij} . Si no aparece la frecuencia absoluta de cada par se entiende que esta es 1.

x_i	y_j	n_{ij}
x_1	y_1	$n_{1,1}$
x_1	y_2	$n_{1,2}$
$\vdots$	$\vdots$	$\vdots$
x_2	y_1	$n_{2,1}$
$\vdots$	$\vdots$	$\vdots$

En la **tabla de doble entrada o tabla de contingencia** los datos se colocan por filas y columnas, y en el interior de la tabla se incluyen la frecuencias n_{ij} .

	x_1	x_2	x_3	$\cdots$	x_k	n_{*j}
y_1	$n_{1,1}$	$n_{2,1}$	$n_{3,1}$	$\cdots$	$n_{k,1}$	$\sum_{i=1}^k n_{i,1}$
y_2	$n_{1,2}$	$n_{2,2}$	$n_{3,2}$	$\cdots$	$n_{k,2}$	$\sum_{i=1}^k n_{i,2}$
y_3	$n_{1,3}$	$n_{2,3}$	$n_{3,3}$	$\cdots$	$n_{k,3}$	$\sum_{i=1}^k n_{i,3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_{i*}	$\sum_{j=1}^m n_{1,j}$	$\sum_{j=1}^m n_{2,j}$	$\sum_{j=1}^m n_{3,j}$	$\cdots$	$\sum_{j=1}^m n_{k,j}$	N

La última fila y la última columna de la tabla de doble entrada nos dan las distribuciones marginales de las variables X e Y.

Ejercicio 7 *A continuación se ofrece la tabla simple correspondiente a una variable bidimensional*

x_i	1	1	2	2	3	4	5	5	6
y_j	7	8	6	7	6	5	5	4	3
n_{ij}	4	5	6	8	7	6	5	3	3

- Elabora una tabla de doble entrada.
- Obtén las distribuciones marginales.
- Representa mediante una nube de puntos.

5.2. Parámetros estadísticos

- **Centro de gravedad:** es el punto $(\bar{x}, \bar{y})$.

- **Covarianza:**

$$s_{xy} = \frac{\sum_{i=1}^k \sum_{j=1}^m x_i y_j n_{ij}}{N} - \bar{x}\bar{y}$$

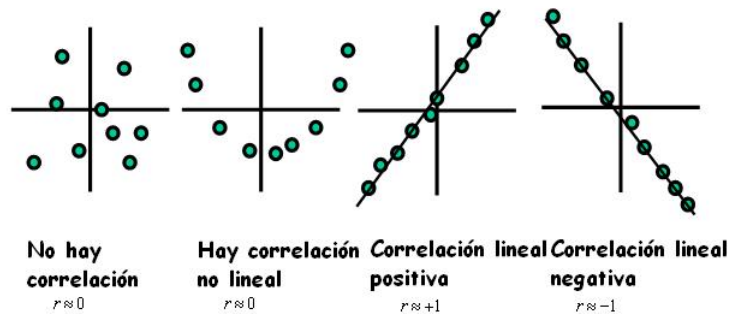
- **Coeficiente de correlación:**

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

La interpretación de este parámetro la veremos en el siguiente apartado.

6. Correlación

En las siguientes gráficas se aprecia que unos puntos están más alineados que otros, es decir, que en algunos de ellos hay mayor grados de asociación lineal.



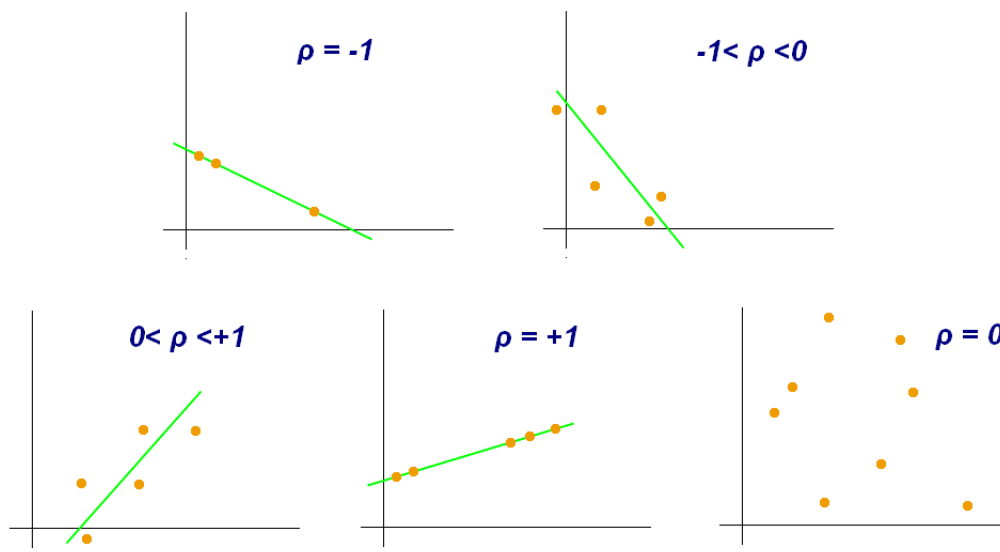
La relación que apreciamos entre dos variables se llama **correlación**. Esta puede ser:

- Lineal o curvilínea**, cuando los puntos están próximos a una línea recta o curva, respectivamente.
- Positiva o negativa**, cuando al aumentar los valores de una variable, se incrementan o disminuyen los valores de la otra, respectivamente.
- Funcional**, si los puntos representan exactamente la gráfica de una función.

En el caso de la **correlación lineal**, es decir, cuando los puntos se aproximan, en mayor o menor grado, a rectas (rectas de regresión). La correlación lineal puede ser más o menos fuerte según lo apretados que estén los puntos de la recta. También puede ser positiva o negativa, dependiendo de la pendiente de la recta.

La medida de asociación lineal más simple entre dos variables es la covarianza. El inconveniente de la covarianza como medida de la asociación lineal entre dos variables es que depende de las unidades de X e Y , por ello se define el coeficiente de correlación, cuyas propiedades son:

- Este número no tiene dimensiones y su valor está entre -1 y 1 .
- Los valores extremos -1 y 1 se alcanzan solamente si todos los datos se sitúan exactamente sobre una recta.
- Si la relación lineal es muy pequeña, el valor de r es próximo a cero.
- Si $r > 0$, la correlación es positiva.
- Si $r < 0$, la correlación es negativa.



Ejercicio 8 *Calcula el coeficiente de correlación de la siguiente distribución e interprétalo:*

x_i	1	1	2	2	3	4	5	5	6
y_j	7	8	6	7	6	5	5	4	3
n_{ij}	4	5	6	8	7	6	5	3	3

Ejercicio 9 La siguiente tabla muestra los resultados que se han obtenido al preguntarle a un grupo de alumnos por su nota de matemáticas y de economía:

Nota de matemáticas	2	3	3	4	5	5	6	6	7	7	8	8	9
Nota de economía	2	2	3	5	5	6	6	7	7	7	7	8	9

Halla el coeficiente de correlación.

7. Regresión

Partimos de la nube de puntos y queremos encontrar la recta que mejor de ajuste a la nube. A estas rectas se le llaman **rectas de regresión**. Como tenemos dos variables, según se tome x o y como variable independiente, tenemos dos rectas:

Recta de regresión de y sobre x :

$$y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x})$$

Recta de regresión de x sobre y :

$$x - \bar{x} = \frac{S_{xy}}{S_y^2}(y - \bar{y})$$

Propiedades de las rectas de regresión:

- Las dos rectas de regresión se cortan en el punto de las medias de las variables $(\bar{x}, \bar{y})$.
- El producto de las pendientes de las rectas es el cuadrado del coeficiente de correlación.

$$\frac{S_{xy}}{S_x^2} \frac{S_{xy}}{S_y^2} = r^2$$

- Las rectas de regresión se usan para predecir el valor de una variable cuando se conoce la otra, y se debe cumplir que el coeficiente de correlación sea próximo a -1 o a 1.

Ejercicio 10 *La siguiente tabla muestra los resultados que se han obtenido al preguntarle a un grupo de alumnos por su nota de matemáticas y de economía:*

<i>Nota de matemáticas</i>	2	3	3	4	5	5	5	7	7	7	8	8	9
<i>Nota de economía</i>	2	2	3	5	5	6	6	7	7	7	7	8	9

¿Qué nota se espera que tenga en economía un alumno que ha sacado un 6 en matemáticas?