

Responde siempre razonadamente:

1. Se ha realizado una encuesta a los trabajadores de una empresa sobre el sueldo anual. La siguiente tabla recoge los resultados obtenidos:

sueldo (miles de €)	[20, 25)	[25, 30)	[30, 35)	[35, 40)	[40, 45)
nº trabajadores	13	7	9	6	5

- a) Construye la tabla de frecuencias. **(1 punto)**
a) Representar los datos por medio de un histograma. **(0.75 puntos)**
b) Calcula las medidas de centralización (media, moda y mediana). **(1 punto)**
c) Calcula las medidas de dispersión (varianza, desviación típica, rango y coeficiente de variación). **(1 punto)**
2. Se ha realizado un estudio a 15 alumnos de bachillerato sobre el número de horas diarias que dedican al estudio (X) y el número de suspensos obtenidos en la primera evaluación(Y):

Nº horas de estudio	4	4	2	1	0
Nº suspensos	0	1	3	4	6
Frecuencia	5	4	3	2	1

- a) Calcula el coeficiente de correlación, indica los cálculos necesarios e interpreta su resultado. ¿Qué podemos afirmar sobre la dependencia lineal entre las variables?
b) Estima el número de horas diarias de estudio que se espera para un alumno que ha tenido 2 suspensos. ¿Es buena la estimación?
(3,75 puntos)
3. De una variable bidimensional (X, Y) sabemos que:
- La recta de regresión de X sobre Y es $x = 1,8 - 0,54 \cdot y$
 - $\bar{y} = 1,3$
 - $\sigma_y = 0,8$
 - $\sigma_x^2 = 0,7225$
- a) Calcula la media de la variable X. (0,75 puntos)
b) Calcula el coeficiente de correlación.
(2,5 puntos)

1. Se ha realizado una encuesta a los trabajadores de una empresa sobre el sueldo anual. La siguiente tabla recoge los resultados obtenidos:

suelo (miles de €)	[20, 25)	[25, 30)	[30, 35)	[35, 40)	[40, 45)
nº trabajadores	13	7	9	6	5

- a) Construye la tabla de frecuencias. (1 punto)

Al estar los datos agrupados en intervalos, calculamos la marca de clase x_i como el punto medio de cada uno de ellos.

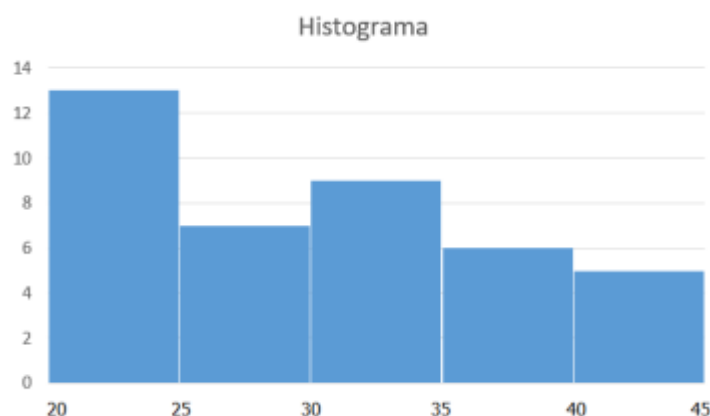
En la tabla de frecuencias indicaremos la marca de clase, las frecuencias absolutas y relativas y las frecuencias acumuladas.

intervalo	x_i	f_i	F_i	h_i	H_i	$x_i \cdot f_i$	$x_i^2 \cdot f_i$
[20, 25)	22,5	13	13	0,325	0,325	292,5	6581,25
[25, 30)	27,5	7	20	0,175	0,5	192,5	5293,75
[30, 35)	32,5	9	29	0,225	0,725	292,5	9506,25
[35, 40)	37,5	6	35	0,15	0,875	225	8437,5
[40, 45)	42,5	5	40	0,125	1	212,5	9031,25
		40				1215	38850

Podemos observar que hay un total de 40 datos.

Las dos últimas columnas nos sirven para calcular las medidas de centralización y de dispersión de los siguientes apartados.

- a) Representar los datos por medio de un histograma. (0.75 puntos)



- b) Calcula las medidas de centralización (media, moda y mediana). (1 punto)

- La media se calcula de la siguiente forma: $\bar{x} = \frac{\sum_{i=1}^{40} x_i \cdot f_i}{40} = \frac{1215}{40} = 30,375$ (30.375€)
- La moda es el dato que más se repite, es decir, el de mayor frecuencia absoluta. En este caso, observando la tabla de frecuencias podemos concluir que el intervalo modal sería [20, 25) y la moda la marca de clase correspondiente $M_o = 22,5$ (22.500€)
- La mediana es el dato central, es decir, el que deja por encima y por debajo el 50% de los datos. En este caso, al existir 40 datos no hay un dato central, sino que la mediana es la media de los datos situados en las posiciones 20 y 21. El dato 20 corresponde a la marca de clase 27,5 y el dato 31 corresponde a la marca de clase 32,5. $Me = \frac{27,5+32,5}{2} = 30$
El intervalo mediano es el que contiene la mediana: [30, 35)

- c) Calcula las medidas de dispersión (varianza, desviación típica, rango y coeficiente de variación). (1 punto)

- La varianza se calcula de la siguiente forma: $\sigma^2 = \frac{\sum x_i^2 \cdot f_i}{40} - \bar{x}^2 = \frac{38850}{40} - 30,375^2 \cong 48,609$
- La desviación típica es la raíz cuadrada de la varianza, por tanto: $\sigma = \sqrt{48,609} \cong 6,972$
- El rango es la diferencia entre el mayor valor y el menor valor que toma la variable, en este caso, en datos agrupados por intervalos, es la diferencia entre el extremo superior y el inferior. Rango $45 - 20 = 25$
- El coeficiente de variación se calcula $CV = \frac{\sigma}{\bar{x}} = \frac{6,972}{30,375} \cong 0,2295$ (22,95% dispersión)

2. Se ha realizado un estudio a 15 alumnos de bachillerato sobre el número de horas diarias que dedican al estudio (X) y el número de suspensos obtenidos en la primera evaluación(Y):

Nº horas de estudio	4	4	2	1	0
Nº suspensos	0	1	3	4	6
Frecuencia	5	4	3	2	1

- a) Calcula el coeficiente de correlación, indica los cálculos necesarios e interpreta su resultado. ¿Qué podemos afirmar sobre la dependencia lineal entre las variables? (3,75 puntos)

El coeficiente de correlación es una medida que indica el grado de dependencia lineal que existe entre ambas variables y toma siempre un valor comprendido entre -1 y 1.

La forma de calcularlo es a partir de la covarianza y de las desviaciones típicas de las variables:

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

Se trata de datos de una variable bidimensional, que relaciona el número de habitaciones (X) y el número de personas que habitan en el hogar (Y) de 6 inquilinos.

El número medio de horas de estudio:

$$\bar{x} = \frac{20 + 16 + 6 + 2}{15} = \frac{44}{15} \cong 2,933 \text{ (aproximadamente 3 suspensos)}$$

El número medio de suspensos es:

$$\bar{y} = \frac{4 + 9 + 8 + 6}{15} = \frac{27}{15} \cong 1,8$$

Calculamos las desviaciones típicas. Para ello, calculamos la varianza de X y la varianza de Y:

$$\sigma_x^2 = \frac{\sum x_i^2 \cdot f_i}{15} - \bar{x}^2 = \frac{158}{15} - 2,933^2 \cong 1,931 ; \text{ Luego, } \sigma_x = \sqrt{1,931} \cong 1,39$$

$$\sigma_y^2 = \frac{\sum y_i^2 \cdot f_i}{15} - \bar{y}^2 = \frac{99}{15} - 1,8^2 \cong 3,36 ; \text{ Luego, } \sigma_y = \sqrt{3,36} \cong 1,833$$

Calculamos la covarianza de X y de Y:

$$\sigma_{xy} = \frac{\sum x_i \cdot y_i}{6} - \bar{x} \cdot \bar{y} = \frac{16+18+8}{15} - 2,933 \cdot 1,8 = \frac{42}{15} - 2,933 \cdot 1,8 \cong -2,4794$$

Por tanto, el coeficiente de correlación es: $r = \frac{-2,4794}{1,39 \cdot 1,833} \cong -0,9731$ Hay una dependencia lineal fuerte y negativa entre ambas variables, la nube de puntos se ajusta muy bien a una recta. Es decir, a medida que aumenta el número de horas de estudio (X) disminuye el número de suspensos (Y).

- b) Estima el número de horas diarias de estudio que se espera para un alumno que ha tenido 2 suspensos. ¿Es buena la estimación?

La ecuación de la recta de regresión de X sobre Y es la siguiente:

$$x = \frac{\sigma_{xy}}{\sigma_y^2} \cdot y + \bar{x} - \frac{\sigma_{xy}}{\sigma_y^2} \cdot \bar{y}$$

Sustituyendo:

$$x = \frac{-2,4794}{3,36} \cdot y + 2,933 - \frac{-2,4794}{3,36} \cdot 1,8 = -0,738 \cdot y + 4,2613$$

Así, para un alumno con $y = 2$ suspensos, se esperan: $x = -0,738 \cdot 2 + 4,2613 \simeq 2,79$
(aproximadamente 2 horas, 47 minutos y 24 segundos)

Como el coeficiente de correlación es un valor muy cercano a 1, podemos concluir que la estimación es buena y fiable.

b) De una variable bidimensional (X, Y) sabemos que:

- La recta de regresión de X sobre Y es $x = 1,8 - 0,54 \cdot y$
- $\bar{y} = 1,3$
- $\sigma_y = 0,8$
- $\sigma_x^2 = 0,7225$

a) Calcula la media de la variable X. (0,75 puntos)

b) Calcula el coeficiente de correlación

a) Sabemos que el punto $(\bar{x}, \bar{y})$ pertenece a la recta de regresión, es decir:

$$\bar{x} = 1,8 - 0,54 \cdot \bar{y} = 1,8 - 0,54 \cdot 1,3 = 1,098$$

Otra forma es:

La ecuación de la recta de regresión de X sobre Y es la siguiente:

$$x = \frac{\sigma_{xy}}{\sigma_y^2} \cdot y + \bar{x} - \frac{\sigma_{xy}}{\sigma_y^2} \cdot \bar{y}$$

Y sabemos que, la ecuación es: $x = 1,8 - 0,54 \cdot y$

Por tanto, identificando coeficiente a coeficiente:

$$\frac{\sigma_{xy}}{\sigma_y^2} = -0,54 ; \bar{x} - \frac{\sigma_{xy}}{\sigma_y^2} \cdot \bar{y} = 1,8 \rightarrow$$

$$\bar{x} - (-0,54) \cdot 1,3 = 1,8 \rightarrow \bar{x} + 0,702 = 1,8 \rightarrow \bar{x} = 1,098$$

b) Sabemos que:

$$\frac{\sigma_{xy}}{\sigma_y^2} = -0,54 ; \frac{\sigma_{xy}}{0,8^2} = -0,54 ; \sigma_{xy} = -0,3456$$

Por tanto, el coeficiente de correlación es

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = -\frac{0,3456}{0,85 \cdot 0,8} = -0,5082$$