TEMA 18. DISTRIBUCIONES BIDIMENSIONALES

1. REPASO DE ESTADÍSTICA BÁSICA

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 (f_i) o acumulada (F_i).
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

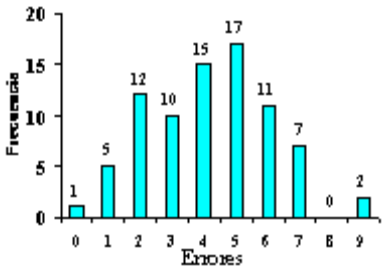
La tabla de la izquierda indica el número de errores cometidos por 80 personas al realizar un determinado test.
La tabla de abajo es la apropiada para variables continuas. En ella se indica el tiempo de espera de 80 personas que han tomado un autobús

Intervalo	M.c.	f_i	%	F_i	%a
[0, 2)	1	4	5	4	5
[2, 4)	3	15	18,75	19	23,75
[4, 6)	5	26	32,5	45	56,25
[6, 8)	7	21	26,25	66	82,5
[8, 10)	9	14	17,5	80	100
Totales		80	100	80	100

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 adjunta, que se corresponde con la primera tabla del ejemplo anterior.



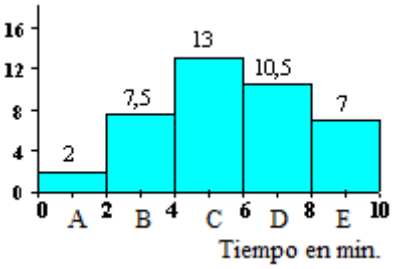
Los diagramas de barras son apropiados para datos medidos en escala nominal u ordinal.

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, en este ejemplo, como la base (la amplitud del intervalo) es 2, la altura de cada rectángulo se halla dividendo la frecuencia que representa entre 2.

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



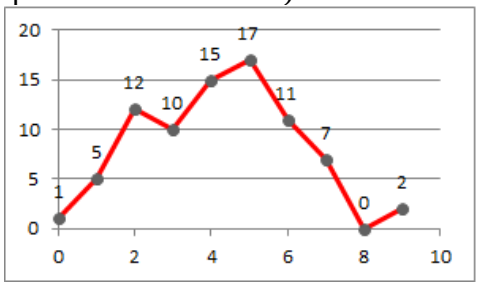
→ En la figura adjunta, el segundo rectángulo [B], que representa una frecuencia de 15, tiene altura 7,5 = 15/2; el 3º, [C], 13 = 26/2; …

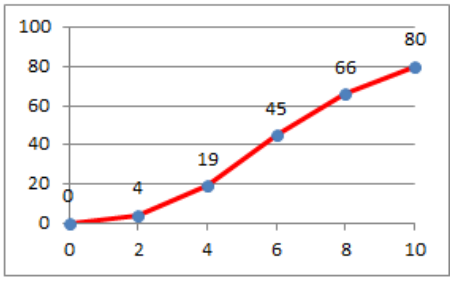
Este histograma puede resumir el tiempo de espera de 80 personas que han tomado un autobús. El intervalo [C] = [4, 6], indica que 26 personas han esperado entre 2 y 4 minutos.

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.



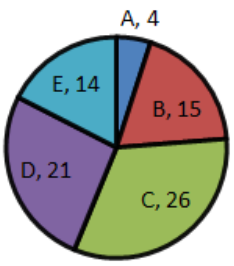


La poligonal (simple) de la izquierda representa los datos del diagrama de barras de arriba; la de la derecha, acumulada, se corresponde con el histograma anterior.

→ Muchos de estos gráficos pueden hacerse con Excel: véase el Problema n. 2.

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.

→ La amplitud, por ejemplo, de B, se halla así: $\frac{360}{80}\cdot 15 = 4,5\cdot 15 = 67,5$.

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:

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

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

Ejemplos:

a) Las calificaciones de una alumna a lo largo del curso en los distritos exámenes de Matemáticas han sido: 7, 9, 6, 8, 10, 9, 5, 7 y 7.

→ Si todos los exámenes tienen el mismo peso, su nota media será:

$\bar{x} = \dfrac{7+9+6+8+10+9+5+7+7}{9} \approx 7,56$.

b) Las calificaciones de 85 alumnos en un examen fueron las que se dan en la siguiente tabla:

Nota: x_i	1	2	3	4	5	6	7	8	9	10
N.º de alumnos: f_i	2	3	6	10	22	15	12	7	5	3

La nota media será:

$\bar{x} = \dfrac{\Sigma x_i f_i}{\Sigma f_i} = \dfrac{1\cdot2+2\cdot3+3\cdot6+4\cdot10+5\cdot22+6\cdot15+7\cdot12+8\cdot7+9\cdot5+10\cdot3}{2+3+6+10+22+15+12+7+5+3} = \dfrac{481}{85} \approx 5,66$.

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

Se calcula igual que la media para datos agrupados: el peso (p_i) sustituye a la frecuencia (f_i).

Ejemplos:

En un concurso oposición, para cada uno de los participantes, se valoran los siguientes aspectos, con los pesos que se indica:

méritos, 25 %; examen teórico: 35 %; examen práctico: 40%.

Cada uno de esos apartados se puntúa entre 0 y 10.

Si los opositores A, B y C obtuvieron los puntos que se indican (respectivamente en los aspectos considerados), ¿en qué orden se clasificaron?

Opositor A: 4, 7, 10. Opositor B: 7, 7, 8. Opositor C: 10, 7, 6.

(Observa que los puntos directos suman, respectivamente, 21, 22 y 23).

→ Las calificaciones ponderadas son:

$\bar{x}_p(A) = \dfrac{4\cdot25+7\cdot35+10\cdot40}{25+35+40} = \dfrac{745}{100} = 7,45$. $\bar{x}_p(B) = \dfrac{7\cdot25+7\cdot35+8\cdot40}{25+35+40} = \dfrac{740}{100} = 7,40$.

$\bar{x}_p(C) = \dfrac{10\cdot25+7\cdot35+6\cdot40}{25+35+40} = \dfrac{735}{100} = 7,35$.

El orden es el dado: 1º, A; 2º, B; 3º, C.

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. (La moda, en muchos casos, no pasa de ser un simple dato anecdótico).

Medidas de posición

Indican la situación, en términos porcentuales, de algunos elementos de la distribución. Los datos deben estar 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: Son los valores de las posiciones correspondientes al 25 %, al 50 % y al 75 % de los datos.
Deciles: Son los valores correspondientes al 10 %, 20 %, ... y 90 % de los datos.
Percentiles (o centiles): dan el valor de la posición correspondiente a cualquier porcentaje.
En muchos casos, su cálculo requiere aplicar la interpolación. (Véase el Problema n. 7).

Medidas de dispersión

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

La varianza. Se obtiene aplicando la siguiente fórmula:

$$\sigma^2 = \frac{\sum (x_i - \overline{x})^2}{n} = \frac{\sum x_i^{\,2}}{n} - \overline{x}^2 .$$ → Para datos agrupados: $\sigma^2 = \frac{\sum (x_i - \overline{x})^2 \, f_i}{\sum f_i}$; $\sigma^2 = \frac{\sum x_i^{\,2} f_i}{\sum f_i} - \overline{x}^2$.

→ Es la media de las desviaciones al cuadrado de cada dato con relación a la media: $(x_i - \overline{x})^2$.

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

$$\sigma = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n}} \; ; \; \sigma = \sqrt{\frac{\sum x_i^{\,2}}{n} - \overline{x}^2} \; ; \; \sigma = \sqrt{\frac{\sum f_i (x_i - \overline{x})^2}{\sum f_i}} \; ; \; \sigma = \sqrt{\frac{\sum x_i^{\,2} f_i}{\sum f_i} - \overline{x}^2} \; .$$

→ La desviación típica es una medida de la desigualdad de los datos estudiados: a mayor desigualdad corresponde mayor desviación típica. (Se mide en la misma unidad que los datos).

Ejemplo:

a) Para hallar desviación típica de los valores 1, 4, 5, 6, 14:

1) Se halla su media, $\overline{x} = \frac{1+4+5+6+14}{5} = 6$.

2) Aplicando la fórmula $\sigma = \sqrt{\frac{\sum x_i^{\,2}}{n} - \overline{x}^2}$ →

$$\sigma = \sqrt{\frac{1^2 + 4^2 + 5^2 + 6^2 + 14^2}{5} - 6^2} = \sqrt{\frac{274}{5} - 36} = \sqrt{18{,}8} \approx 4{,}34 .$$

x_i	$x_i^{\,2}$
1	1
4	16
5	25
6	36
14	196
$\sum x_i = 36$	$\sum x_i^{\,2} = 274$

b) Datos agrupados.

Las calificaciones de 85 alumnos en un examen fueron las que se dan en la siguiente tabla:

Nota: x_i	1	2	3	4	5	6	7	8	9	10	Sumas
N.º de alumnos: f_i	2	3	6	10	22	15	12	7	5	3	85
Producto $(x_i \cdot f_i)$	2	6	18	40	110	90	84	56	45	30	481
Valores $x_i^{\,2}$	1	4	9	16	25	36	49	64	81	100	––
Producto $(x_i^{\,2} \cdot f_i)$	2	12	54	160	550	540	588	448	405	300	3059

Por tanto, aplicando las fórmulas anteriores:

$$\overline{x} = \frac{481}{85} \approx 5{,}66 \; ; \; \sigma = \sqrt{\frac{3059}{85} - (5{,}66)^2} = \sqrt{3{,}95} \approx 1{,}99 .$$

Uso de la calculadora

Cada calculadora puede presentar pautas diferentes: consulta el manual de la tuya.

Las pautas de uso para una de las calculadoras más frecuentes son:

(1) Hay que poner la calculadora en el "modo estadístico": SD.

–Pulsar [MODE] [2] (en la pantalla aparecerá SD).

–Borrar los datos de memoria: Pulsar [SHIFT] [AC]. (Posiblemente debas volver a pulsar ON).

(2) Introducir los datos, sumando con la tecla [M+]

Por ejemplo, para datos simples: x_1 [M+] x_2 [M+] … x_n [M+].

(3) Los parámetros se obtienen pulsando:

[SHIFT] [2] [1] [=] → $\bar{x}$ (media); [SHIFT] [2] [2] [=] → σ_n (desv. típica)

- Pulsando [SHIFT] [1] se obtienen otros resultados: ($\sum x_i^2$, con 1; $\sum x_i$, con 2; n, con 3).

Ejemplo:

Utilizando la calculadora, comprueba que para los datos 1, 4, 5, 6, 14, se obtienen los parámetros que se indican:

$\sum x_i^2 = 274$; $\sum x_i = 30$; $n = 5$; $\bar{x} = 6$; $\sigma = 4{,}335896678$.

–Para datos agrupados:

(1) Poner la calculadora en el modo SD y borrar los datos de memoria: [MODE] [2] [SHIFT] [AC].

(2) Introducir los datos como sigue:

x_1 [SHIFT] [,] f_1 [M+] x_2 [SHIFT] [,] f_2 [M+] … x_n [SHIFT] [,] f_n [M+].

(3) Los parámetros se obtienen pulsando:

[SHIFT] [2] [1] [=] → $\bar{x}$ (media); [SHIFT] [2] [2] [=] → σ_x (desv. típica)

- Pulsando [SHIFT] [1] se obtienen otros resultados: ($\sum x_i^2$, con 1; $\sum x_i$, con 2; n, con 3).

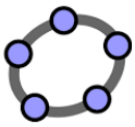
Ejemplo:

Utilizando la calculadora, comprueba para el siguiente conjunto de datos, los resultados que se indican.

Nota: x_i	1	2	3	4	5	6	7	8	9	10
N.º de alumnos: f_i	2	3	6	10	22	15	12	7	5	3

$\sum x_i^2 = 3059$; $\sum x_i = 481$; $n = 85$; $\bar{x} = 5{,}658823529$; $\sigma = 1{,}991469698$).

Nota: Utilizando **GeoGebra** también se pueden obtener esos parámetros:



Para datos simples (los mismos del Ejemplo de arriba), teclear:

media(1,4,5,6,14) → sale 6; de(1,4,5,6,14) → sale 4,34. (de ≡ desv. estándar)

Para datos agrupados (los mismos de Ejemplo anterior), teclear:

media({1,2,3,4,5,6,7,8,9,10}, {2,3,6,10,22,15,12,7,5,3}) → sale 5,66.

de({1,2,3,4,5,6,7,8,9,10}, {2,3,6,10,22,15,12,7,5,3}) → sale 1,99.

El coeficiente de variación

Es una medida de la *dispersión relativa* del conjunto de datos. Se define como: $CV = \frac{\sigma}{\bar{x}}$. (Sirve comparar dos conjuntos: es más disperso el que tiene un CV mayor).

El coeficiente de variación suele darse en porcentajes: $CV = \frac{\sigma}{\bar{x}} \cdot 100$. → Véase el Problema n. 10.

Ejemplo:

El coeficiente de variación de los conjuntos de datos dados en los ejemplos anteriores es:

$$CV\,(\text{Ej.1º}) = \frac{4{,}34}{6} \approx 0{,}723 \rightarrow 72{,}3\ \%; \qquad CV\,(\text{Ej.2º}) = \frac{1{,}99}{5{,}66} \approx 0{,}352 \rightarrow 35{,}2\ \%.$$

2. DISTRIBUCIONES BIDIMENSIONALES

Las distribuciones bidimensionales estudian a la vez dos características (dos variables aleatorias) de una población. Ambas variables deben ser cuantitativas; si se denotan genéricamente por X e Y, sus valores se dan en forma de pares, (x_i, y_i), asimilables a puntos del plano.

- Para que el estudio tenga consistencia, los datos deben obtenerse aleatoriamente; aumentando la fiabilidad del resultado cuando lo hace el número de pares considerado, el tamaño muestral.

Ejemplos:
a) En una población se pueden estudiar a la vez las variables: X = "estatura en cm"; Y = "número de zapato". Algunos de esos pares podrían ser: (180, 43); (165, 40); (195, 45); … Normalmente los individuos altos tienen mayor talla de pie.

b) En el conjunto de los alumnos y alumnas de un instituto se pueden estudiar las variables: X = "tiempo diario dedicado a las *redes sociales*"; Y = "notas obtenidas en Matemáticas". Algunos de pares podrían ser: (1, 7); (3, 2); (2, 6), (2, 8), … Aunque no sea determinante, si un estudiante dedica mucho tiempo a las redes sociales dedicará menos al estudio y, consecuentemente, sus notas en Matemáticas y en otras asignaturas serán peores.

c) Para los distintos países se pueden estudiar las variables: X = "porcentaje del PIB invertido en investigación"; Y = "porcentaje de población aficionada al ajedrez". Algunos pares de valores pueden ser: (2, 6); (3, 8); (2,3, 5); (1,3, 7), …

Correlación entre las variables
Al estudiar distribuciones bidimensionales, el objetivo es determinar si existe relación estadística entre las dos variables consideradas; es decir, ver si los cambios en una de las variables influyen en los cambios de la otra. Cuando sucede esto, se dice que ambas variables están correlacionadas o que hay correlación entre ellas. En este caso, la variable Y podría estimarse (*deducirse*) a partir de la X.

Si las variables aumentan o disminuyen conjuntamente, la correlación es directa. Si, por el contrario, al aumentar una de ellas disminuye la otra, la correlación será inversa.
Si la correlación es *fuerte*, a partir de una variable puede estimarse la otra con una fiabilidad (probabilidad) alta. Si la correlación es débil, la estimación de una variable a partir de la otra es poco fiable.
→ Aquí se estudiará solo la correlación lineal, que utiliza la ecuación de una recta para hacer la estimación.

Ejemplos:
a) La correlación entre la "estatura de las personas" y el "número de zapato" puede suponerse que será directa y fuerte: a más altura de una persona suele corresponder una mayor talla de zapato.

b) Las variables "tiempo dedicado a las redes sociales" y "nota obtenida en Matemáticas" están inversamente correlacionadas: a mayor tiempo dedicado a redes sociales suele corresponder una menor nota en Matemáticas. Para determinar si la correlación es fuerte hay que hacer un estudio detallado, pues es posible que se presenten excepciones significativas.

c) Las variables "porcentaje del PIB invertido en investigación" y "porcentaje de población aficionada al ajedrez" es posible que estén poco correlacionadas; hay países en los que la tradición ajedrecista está muy arraigada por motivos culturales.

Relación funcional y *relación* estadística

- Dos variables X e Y están relacionadas funcionalmente cuando conocida X se puede saber con exactitud el valor de Y. La relación funcional se cumple siempre: globalmente y para cada valor particular.

Por ejemplo, el tiempo que tarda en impactar contra el suelo un objeto que se deja caer puede saberle exactamente, pues depende de la altura inicial: la fórmula $9{,}8t^2 = 2h$, permite hallar el tiempo *t* que tarda en impactar en el suelo un objeto que se deja caer desde una altura *h*.
Así, un cuerpo que se deja caer desde una altura de 20 m tarda en impactar 2,02 s.

- Dos variables X e Y está relacionadas estadísticamente (correlacionadas) cuando conocida X se puede estimar aproximadamente el valor de Y. La relación estadística se cumple en general; para cada valor particular la respuesta puede ser múltiple.

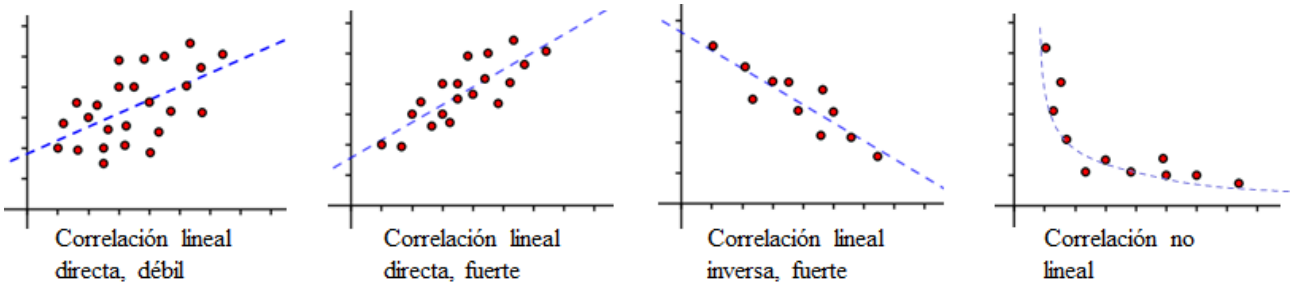
Por ejemplo, la altura de un niño "depende" de su edad: normalmente los niños de 4 años son más altos que los de 3 años; pero no en todos los casos. Globalmente puede admitirse que la relación "más años → más alto" es cierta; pero no siempre se cumple: se tiene una *certeza probable.*

Diagramas de dispersión

El primer paso para determinar el sentido y el grado de la correlación entre dos variables consiste en representar gráficamente, en el plano cartesiano, los pares de valores conocidos. Estos gráficos, que reciben el nombre de diagramas de dispersión, permiten visualizar la posición de los datos en el plano. La forma de la nube de puntos asociada a cada diagrama permitirá establecer conjeturas sobre la correlación existente entre las variables estudiadas.

En general, dependiendo de la forma de la nube de puntos, puede asegurarse:

- Una nube de puntos alargada indica correlación lineal: los puntos se distribuyen en torno a una línea recta. La estrechez de la nube expresa que la correlación es fuerte.
- Si la recta que se ajusta a la nube tiene pendiente positiva, la correlación será directa: al crecer la variable X, lo hace también la variable Y.
- Una recta con pendiente negativa, indica que la correlación es inversa, al crecer X, disminuye Y.

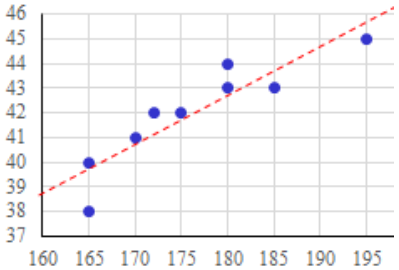


Aquí se estudiará solamente la correlación de tipo lineal, cuando los puntos de la nube se distribuyen de alguna manera en torno a una línea recta.

Ejemplo:

En la tabla siguiente se da la estatura (en cm) y el número de zapato de nueve personas:

Estatura	180	165	195	170	172	165	185	180	175
N.º zapato	43	40	45	41	42	38	43	44	42



El diagrama de dispersión es el adjunto; se ha dibujado con Excel. (Observa que los ejes no se cortan en el punto (0, 0); se hace esta modificación buscando claridad).
La correlación entre ambas variables es directa y parece *fuerte*.
La línea recta se ha ajustado de manera aproximada.

3. COEFICIENTE DE CORRELACIÓN LINEAL

Hasta ahora se ha calificado la correlación como fuerte o débil a partir de la nube de puntos; a partir de su apariencia visual, de un modo intuitivo.

La confirmación cuantitativa (objetiva) de estas conjeturas se obtiene a partir del cálculo de un coeficiente, llamado de correlación, que mide la dependencia estadística entre las variables consideradas.
Este coeficiente se halla a partir de los parámetros (media y desviación típica) de cada una de las variables consideradas, por separado y conjuntamente.

Parámetros de una distribución bidimensional
Cuando estos parámetros se consideran para cada una de las variables se llaman marginales.

- Medias marginales para cada una de las variables X e Y. Valen:

$\bar{x} = \frac{\sum x_i}{n}$; $\bar{y} = \frac{\sum y_i}{n}$ → n es el número de observaciones

El punto $(\bar{x}, \bar{y})$ se llama centro medio de la distribución.

- Varianzas y desviaciones típicas

$$\sigma_x^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{\sum x_i^2}{n} - \bar{x}^2; \quad \sigma_y^2 = \frac{\sum (y_i - \bar{y})^2}{n} = \frac{\sum y_i^2}{n} - \bar{y}^2 .$$

Las desviaciones típicas marginales, σ_x y σ_y, son la raíz cuadrada de las varianzas.

- La covarianza: La covarianza es un parámetro estadístico conjunto, pues, en su cálculo intervienen las dos variables a la vez. Se define como la media aritmética de los productos de las diferencias de los valores de cada variable respecto de su media marginal. Por tanto, vale:

$$\sigma_{xy} = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n} \rightarrow \sigma_{xy} = \frac{\sum x_i \cdot y_i}{n} - \bar{x} \cdot \bar{y} .$$

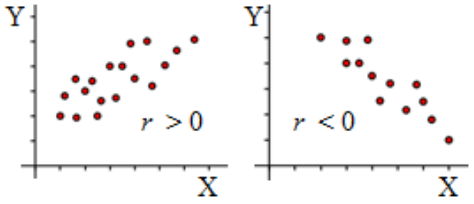
Si $\sigma_{xy} > 0$, la correlación es directa; si $\sigma_{xy} < 0$, la correlación es inversa.

El coeficiente de correlación lineal
Da una medida de la fuerza de la correlación entre las dos variables estudiadas.

Su valor es $r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$: es la razón entre la covarianza de las variables X e Y y el producto de sus desviaciones típicas marginales.

- El coeficiente de correlación cumple:

1) El valor de r no cambia al hacerlo la escala de medición.
2) El signo de r es el mismo que el de la covarianza: si $r > 0$, la correlación es directa; si $r < 0$, la correlación es inversa.
3) El valor de r está entre –1 y +1: $-1 \le r \le 1$
4) Si $|r|$ toma valores cercanos a 1, la correlación es fuerte.



5) El cuadrado de r, r^2, indica la proporción de la variación en la variable Y que puede ser explicada por los cambios de la variable X. A r^2 se le llama coeficiente de determinación.

Ejemplo:
Si $r = 0{,}8$, el coeficiente de determinación vale $r^2 = 0{,}8^2 = 0{,}64$. Esto significa que el 64 % de la variación de Y puede ser explicada a partir de la variación de X.

Cálculo del coeficiente de correlación lineal

→ Método manual:

Los cálculos que siguen solo tienen sentido didáctico: son largos y falibles.

Para el caso de las variables estatura y número de zapato pueden organizarse los datos como sigue:

	Estatura (X)	N. zapato (Y)	x_i^2	y_i^2	$x_i \cdot y_i$
	180	43	32400	1849	7740
	165	40	27225	1600	6600
	195	45	38025	2025	8775
	170	41	28900	1681	6970
	172	42	29584	1764	7224
	165	38	27225	1444	6270
	185	43	34225	1849	7955
	180	44	32400	1936	7920
	175	42	30625	1764	7350
Sumas	1587	378	280609	15912	66804

Aplicando las fórmulas (redondeado a centésimas):

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1587}{9} = 176,33...;$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{378}{9} = 42;$$

$$\sigma_x = \sqrt{\frac{\sum x_i^2}{n} - \bar{x}^2} \Rightarrow$$

$$\sigma_x = \sqrt{\frac{280609}{9} - (176,33...)^2} \approx 9,24$$

$$\sigma_y = \sqrt{\frac{\sum y_i^2}{n} - \bar{y}^2} = \sqrt{\frac{15912}{9} - 42^2} = 2;\ \sigma_{xy} = \frac{\sum x_i y_i}{n} - \bar{x}\cdot\bar{y} = \frac{66804}{9} - 176,33\cdot 42 \approx 16,81.$$

Por tanto, el coeficiente de correlación vale: $r = \dfrac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \dfrac{16,81}{9,24\cdot 2} \approx 0,91.$

La correlación entre las variables "estatura" y "número de zapato" es directa y bastante fuerte. Como $r^2 = 0,91^2 = 0,8281$, puede deducirse que el 82,81 % de la variación en el número de zapato de las personas depende de la estatura.

→ Calculadora:

Las medias y desviaciones típicas marginales se pueden calcular una a una, con la calculadora en el MODO SD; o a la vez, poniéndola en el MODO REG Lin (LR *Linear regression*). (Tu calculadora puede ser diferente: lee sus instrucciones). Con la que tengo a mano (Casio *fx–82MS*) hay que proceder como sigue:

(1) Poner la calculadora en el MODO REG Lin.

–Pulsar [MODE] [3] [Lin] [1] (en la pantalla aparecerá REG).

–Borrar los datos de memoria: Pulsar [SHIFT] [AC].

(2) Introducir los pares de datos (x_i, y_i), tecleando: x_1 ['] y_1 [M+] x_2 ['] y_2 [M+] … x_n ['] y_n [M+]

(3) Los parámetros se obtienen pulsando:

[SHIFT] [2] [1] [=] → $\bar{x}$; [SHIFT] [2] [2] [=] → σ_x (desv. típica de X)

[SHIFT] [2] [▸] [1] → $\bar{y}$; [SHIFT] [2] [▸] [2] → σ_y.

[SHIFT] [2] [▸] [▸] [3] → r (coeficiente de correlación lineal).

- Pulsando [SHIFT] [1] se obtienen otros resultados: ($\sum x_i^2$, con 1; $\sum x_i$, con 2; n, con 3).
- Y con [SHIFT] [1] [▸] [1] → ($\sum y_i^2$, con 1; $\sum y_i$, con 2; $\sum x_i y_i$, con 3).

Para el ejemplo anterior hay que teclear:

180 ['] 43 [M+] 165 ['] 40 [M+] … 175 ['] 42 [M+]

Los resultados que se obtienen deben ser los indicados arriba (allí redondeados).

→ En Excel, introduciendo los datos de manera ordenada, se puede dibujar y obtener la ecuación de la "línea de tendencia". Además, proporciona el valor de r^2 (da 0,8138, como veremos más adelante). En la solución del Problema 19 se detalla el proceso.

4. RECTA DE REGRESIÓN LINEAL

Cuando la nube de puntos correspondiente a una distribución bidimensional tiene forma alargada, se le puede asociar una "línea de tendencia" recta, llamada recta de regresión (de Y sobre X). La ecuación de esa recta permite hacer estimaciones de la variable Y a partir de la X.

- La recta de regresión es la que mejor se ajusta a la nube de puntos. Se conoce con el nombre de recta de regresión mínimo cuadrática, pues es la que minimiza la suma de los cuadrados de los errores. (El error es la diferencia entre el valor real, y el valor teórico obtenido con la recta).
Es una recta ideal que asignaría a cada valor x_i de la variable X el promedio de los y_i correspondientes a x_i: (ver Problema n. 23). En consecuencia, debe pasar por el punto $(\bar{x}, \bar{y})$, centro de gravedad de la distribución bidimensional: (ver Problemas n. 25 y 26).
Las estimaciones son más fiables cuando el valor de X está cercano a la media $\bar{x}$ y cuando $|r|$ está próximo a 1. (Esa fiabilidad puede medirse en términos de probabilidad, aunque aquí no se hará).
Su ecuación es:

$$y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x}).$$

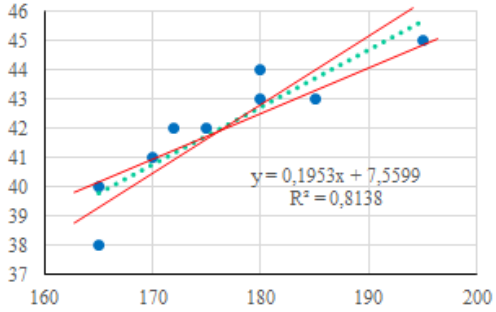
Recuerda que la ecuación general de una recta es $y = mx + n$, En las calculadoras (ordenadores) está recta suele expresarse en la forma y = A + Bx, siendo $B = \frac{\sigma_{xy}}{\sigma_x^2}$ y $A = \bar{y} - \frac{\sigma_{xy}}{\sigma_x^2}\cdot\bar{x}$; valores que también se obtienen de manera automática: teclas [SHIFT] [2] [►] [►] [1 → A]; [2 → B].

- La recta de regresión de X sobre Y (que no es la misma que la de Y sobre X) permite estimar los valores de Y a partir de los de la variable X. Su ecuación es: $x - \bar{x} = \frac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y})$. (Problema n. 22).

Ejemplo:
La ecuación de la recta de regresión correspondiente a correlación "estatura/n.º de zapato" estudiada antes es:

$$y - 42 = \frac{16,81}{9,24^2}(x - 176,33) \Rightarrow y = 0,1969x + 7,263.$$

→ Con la calculadora se obtiene:
A = 7,559895833; B = 0,1953125; r = 0,902109775.



En el gráfico adjunto, cuyos resultados se han obtenido automáticamente con Excel, se observan pequeñas diferencias en los coeficientes de la recta (son debidas al redondeo).
La recta de regresión es la de trazo discontinuo verde; las trazadas en rojo son aproximaciones.

→ La ecuación de la recta permite inferir (deducir estadísticamente) la talla de zapato que usará un individuo de una determinada altura.
Así, por ejemplo, si el individuo mide 190 cm puede esperarse que use el número
$y = 0,1969 \cdot 190 + 7,263 = 44,674$ → talla 45, aprox.
Si mide 162 cm, usará el número $y = 0,1969 \cdot 162 + 7,263 = 39,1608$ → talla 39, aprox.

La fiabilidad de estas inferencias (deducciones) depende del coeficiente de determinación, que en Excel se denota por R^2 y que, en este caso, vale 0,8138; lo que significa que el 81,38 % de la variación en el número de zapato de las personas depende de su estatura. El resto, hasta el 100 % dependerá de otros factores (estatura de la madre, alimentación, …).

5. INTERPRETACIÓN CONJUNTA: COEFICIENTE DE CORRELACIÓN, RECTA DE REGRESIÓN

Como se ha indicado anteriormente, el análisis de regresión puede hacerse:

- Visualmente.

Representando los puntos de la distribución se obtiene el diagrama de dispersión, la nube de puntos: su forma permite deducir si hay correlación lineal y aproximar su intensidad.

- Algebraicamente.

Calculando el coeficiente de correlación lineal r (y el de determinación r^2); si el valor absoluto de r se acerca a 1, la correlación es fuerte. En este caso, la recta de regresión puede utilizarse para realizar estimaciones del fenómeno estudiado.

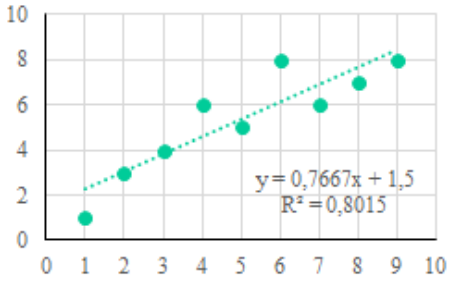
→ A continuación se comentan tres supuestos. (Los cálculos pueden hacerse *a mano* o utilizando herramientas informáticas. Lo importante no es hacer esos cálculos, sino saber interpretar los resultados. Aquí se han hecho con Excel: quizás puedas repetirlos, para practicar.

Ejemplos:

a) Datos

X	Y
1	1
2	3
3	4
4	6
5	5
6	8
7	6
8	7
9	8

Nube de puntos



Interpretación

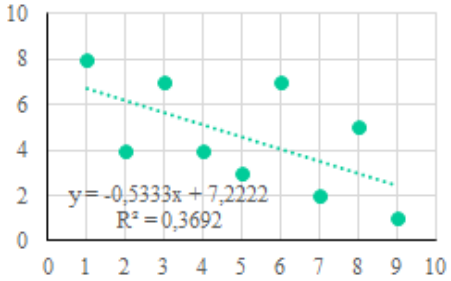
La nube de puntos es estrecha y "creciente", lo que indica una correlación positiva fuerte.
El coeficiente de correlación vale $r = \sqrt{0,8015} \approx 0,895$. La correlación entre ambas variables es muy fuerte.
La recta de regresión asigna valores fiables a Y a partir de la X.
Por ejemplo, a x = 7,5 le asigna, $y = 0,7667 \cdot 7,5 + 1,5 \approx 7,25$

---

b) Datos

X	Y
1	8
2	4
3	7
4	4
5	3
6	7
7	2
8	5
9	1

Nube de puntos



Interpretación

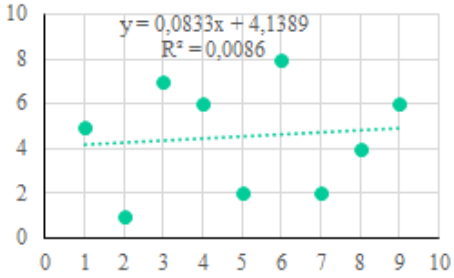
La nube de puntos no es estrecha, aunque sigue una tendencia decreciente. Esto sugiere una correlación negativa y no muy fuerte. Las variaciones de X *solo* explican el 36,92 % de las de Y, pues $r^2 = 0,3692 \rightarrow r \approx -0,608$. La recta de regresión asigna valores *poco* fiables a Y a partir de X. (A veces, un *poco* puede ser importante: dependerá de la naturaleza del problema).

---

c) Datos

X	Y
1	5
2	1
3	7
4	6
5	2
6	8
7	2
8	4
9	6

Nube de puntos



Interpretación

La nube de puntos no sigue ninguna tendencia claramente definida. Se debe admitir que no hay correlación lineal entre las variables X e Y.
El análisis puede darse por terminado: no se puede inferir nada de una variable a partir de la otra.

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]
N.º 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 2007/2018, se da en la siguiente tabla:

Año	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Miles de turismos	1634	1185	971	1000	818	711	742	890	1094	1230	1342	1425

a) Tomando como base 100 el número de turismos matriculados en el año 2007 expresa en números índices la variación de la serie.
b) Representa los datos mediante una poligonal simple usando Excel.

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/as, 50 operarios/as y 8 vendedores/as. Los sueldos mensuales, en euros, de cada categoría son los siguientes: directivos/as, 4000 €; operarios/as, 1400 €; vendedores/as, 2000 €.
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/as, 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/as de primero.

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

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.

8. 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.

9. Se ha realizado una encuesta a los 40 empleados de una empresa para saber cuánto 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.

10. 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

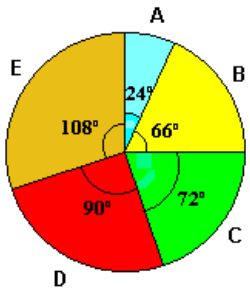
Halla los rendimientos medios para el quinquenio de cada cereal. ¿Qué cereal es más fiable?

11. A un congreso asisten seis mujeres cuyas edades son:

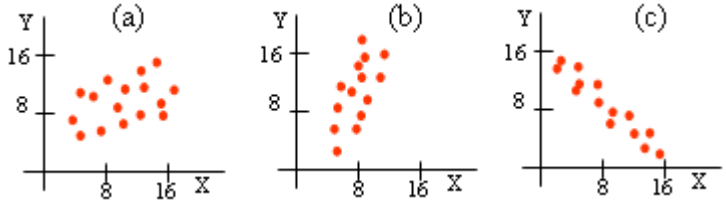
27 34 38 42 33 36 (años)

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.

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



13. a) Asocia las rectas de regresión: $y = -x + 16$, $y = 2x - 12$ e $y = 0{,}5x + 5$ a las nubes de puntos siguientes:



b) Asigna los coeficientes de correlación lineal $r = 0{,}4$, $r = -0{,}85$ y $r = 0{,}7$, a las mismas nubes de puntos.

14. Se han tomado ocho medidas de la temperatura de una batería y de su voltaje, y se obtuvieron los siguientes datos:

X: temperatura	10,0	10,0	23,1	23,5	34,0	34,5	45,0	45,6
Y: voltaje	430	425	450	460	470	480	495	510

a) Sin efectuar cálculos, razona cuál de las siguientes ecuaciones es la de la recta de regresión de *Y* sobre *X* para los datos anteriores:

$y = 350 - 2{,}1x$; $y = 460 - 2{,}1x$; $y = 406 + 2{,}1x$.

b) Para 25 grados, ¿qué voltaje sería razonable suponer?

15. El número de horas de estudio de una asignatura y la calificación obtenida en el examen correspondiente fue, para 7 personas, la siguiente:

Estudio (h)	5	8	10	12	15	17	18
Calificación	3	6	5	6	9	7	9

a) Dibuja la nube de puntos y traza, aproximadamente, la recta de regresión asociada.
b) Indica el carácter y estima la fuerza de la correlación.

16. Calcula, paso a paso (sin utilizar la calculadora en modo estadístico), el coeficiente de correlación y la recta de regresión asociada a los datos del problema anterior.

17. a) Calcula la recta de regresión de Y sobre X en la distribución siguiente realizando todos los cálculos intermedios.

X	10	7	5	3	0
Y	2	4	6	8	10

b) ¿Cuál es el valor que correspondería según dicha recta a X = 7?

18. El departamento de control de calidad de una empresa de instalación de componentes electrónicos desea determinar la relación entre las semanas de experiencia de sus trabajadores y el número de componentes rechazados a esos trabajadores la semana anterior.

Trabajador examinado	A	B	C	D	E	F	G	H	I	J
Semanas de experiencia (X)	7	8	10	1	4	5	15	18	4	8
Número de rechazos (Y)	22	35	15	42	26	30	16	20	31	23

a) Representa el diagrama de dispersión correspondiente a esos datos. ¿Sugiere la gráfica alguna asociación lineal?
b) ¿Cómo calificarías la correlación?

19. Para los datos del problema anterior, halla con ayuda de la calculadora:
a) Las medias y desviaciones típicas marginales.
b) La covarianza.
c) El coeficiente de correlación lineal.
d) La recta de regresión de Y sobre X.
e) El número de rechazos que hay que esperar para una persona con 20 semanas de experiencia.
f) Detalla la solución con Excel.

20. Se midieron los valores de concentración de una sustancia A en suero fetal y los valores de su concentración en suero materno. Se obtuvieron los siguientes datos en una muestra de seis embarazadas a término:

Madre (X)	8	4	12	2	7	9
Feto (Y)	6	4	8	1	4	5

a) Calcula el coeficiente de correlación lineal.
b) Halla la expresión de la recta que permita estimar los valores fetales a partir de los maternos.

21. En seis alumnas de bachillerato se observaron dos variables: X = puntuación obtenida en un determinado test e Y = nota en un examen de Matemáticas. Los resultados se indican en la siguiente tabla:

Test: X	110	90	140	120	120	100
Examen: Y	6	5	9	7	8	6

a) Halla la recta de regresión.
b) Sabiendo que una alumna obtuvo 130 puntos en el test, pero no realizó el examen de Matemáticas, predice, si es posible, la nota que hubiese obtenido.

22. La altura, en cm, de 8 padres y del mayor de sus hijos varones, son:

Padre (X)	170	173	178	167	171	169	184	175
Hijo (Y)	172	177	175	170	178	169	180	187

a) Calcula la recta de regresión que permita estimar la altura de los hijos dependiendo de la del padre; y la del padre conociendo la del hijo.
b) ¿Qué altura cabría esperar para un hijo si su padre mide 174 cm? ¿Y para un padre, si su hijo mide 190 cm?

23. Los años de siete árboles y el diámetro de su tronco, en cm, se dan en la siguiente tabla:

Años	2	4	5	8	10	14	20
Diámetro	10	15	17	20	23	25	27

a) Calcula, utilizando la recta de regresión, el diámetro que se puede predecir para árboles de 10 y 20 años.
b) Compara el resultado anterior con los valores observados en la tabla. Razona el porqué de las diferencias.

24. El número de bacterias por unidad de volumen, presentes en un cultivo después de un cierto número de horas, viene expresado en la siguiente tabla:

X: N.º de horas	0	1	2	3	4	5
Y: N.º de bacterias	12	19	23	34	56	62

Calcula:
a) Las medias y desviaciones típicas de las variables, número de horas y número de bacterias.
b) La covarianza de la variable bidimensional.
c) El coeficiente de correlación, dando una interpretación del resultado.
d) La recta de regresión de Y sobre X.

25. Un conjunto de datos bidimensionales (x, y) tiene un coeficiente de correlación $r = 0{,}8$. Las medias marginales valen: $\bar{x} = 2$; $\bar{y} = 4$. Indica si alguna de las siguientes ecuaciones puede corresponder a la recta de regresión de Y sobre X:

$y = -2x + 8$; $y = 0{,}8x + 2$, $y = 1{,}5x + 1$.

27. Una compañía de seguros de automóvil sospecha que el número de accidentes está en función de la edad del conductor. Para ello elige 100 personas de cada grupo de edad y contabiliza los accidentes totales del último año. Los datos fueron:

Edad	20	25	30	35	40	45
N.º accidentes	10	11	9	7	4	5

a) Representa gráficamente la nube de puntos asociada a estos datos. ¿Qué correlación se observa?
b) Halla, sin utilizar la calculadora en el modo REG, el coeficiente de correlación lineal entre las variables medidas. Comenta su valor.

28. Se está experimentado la resistencia a la rotura de una determinad fibra textil. Para ello se ha medido el diámetro de la fibra y el peso que soporta hasta la rotura, obteniéndose los siguientes datos:

Diámetro en mm (X)	1	1,2	1,4	1,6	1,8	2
Peso a la rotura en kg (Y)	12,5	18	25	32	41	52

a) Representa el diagrama de dispersión asociado a esos datos. ¿Sugiere la gráfica alguna asociación lineal?
b) ¿Cómo calificarías la correlación?

29. Con los datos del problema anterior, halla:
a) La recta de regresión de Y sobre X, y determina la resistencia a la rotura de una fibra de 2,5 mm de diámetro.
b) La recta de regresión de X sobre Y, y determina el diámetro mínimo de una fibra para que soporte más de 60 kg.

30. La siguiente tabla indica las horas de asistencia a un curso de informática y las notas obtenidas por seis alumnos:

Horas de asistencia (X)	2	1	7	5	4	3
Notas (Y)	2	1	10	6	5	3

a) Representa la nube de puntos.
b) Halla el coeficiente de correlación entre X e Y, e interprétalo.
c) Halla la recta de regresión; y represéntala.
d) Si una persona asistiera seis horas al curso, ¿qué nota cabe esperar para ella?

31. El número de faltas de asistencia a clase en los dos últimos meses, en la asignatura de Matemáticas, de ocho estudiantes de 1º de bachillerato, y su calificación final en dicha asignatura han sido:

Faltas (X)	1	2	4	4	6	7	8	12
Calificaciones (Y)	8	9	6	4	3	5	3	3

a) Dibuja la nube de puntos asociada. ¿Qué tipo de correlación se observa entre las variables estudiadas?
b) Calcula, indicando todos los pasos intermedios, el coeficiente de correlación y la recta de regresión de *Y* sobre *X*.