Introducción

1. Las siguientes preguntas forman parte de una encuesta más amplia realizada a estudiantes de bachillerato.

1. Año de nacimiento
2. Sexo: H[ ] M[ ]
3. Estado civil: Soltero[ ] Casado[ ] Otros[ ]
4. Peso
5. Altura
6. Lugar de nacimiento: Comunidad autónoma[........] Fuera de España[ ]
7. Numero de hermanos (incluido el entrevistado)
8. Código postal del lugar de residencia
9. Gastos mensuales del entrevistado (en euros)

Clasifica las 11 preguntas de la encuesta en una de las casillas de la tabla siguiente, según sean de un tipo o de otro las variables que miden.

Cualitativas		
Cuantitativas	Discretas	
	Continuas	

2. Se desea estudiar la aceptación que tiene entre los vecinos de un barrio la programación de la sala cinematográfica situada en él. Para ello se está preparando un cuestionario al que tendrán que responder algunos de los vecinos. Da dos ejemplos de variables cualitativas, dos de variables cuantitativas continuas y dos de cuantitativas discretas que podrían incluirse en ese cuestionario.

3. 5.7.1 de Peña (1995) Vol. 1

4. 5.7.2 de Peña (1995) Vol. 1

5. 5.7.3 de Peña (1995) Vol. 1

Análisis exploratorio de datos. Descripción de una variable

1. Se preguntó a 40 personas por el medio de transporte utilizado para acudir al trabajo. Las posibles respuestas eran

Caminando (C), *Automóvil particular* (A),
Autobús (B), *Metro* (M), *Otros* (O).

Las respuestas obtenidas fueron las siguientes:

A	A	A	C	A	B	M	O	M	B
A	C	A	C	O	B	M	O	M	A
B	B	M	A	C	A	B	M	M	C
C	A	A	B	B	M	O	M	B	O

(a) Calcula la tabla de frecuencias absolutas y relativas de esta variable.
(b) Haz el diagrama de barras y el diagrama de sectores asociado.
(c) ¿Cuál es la clase modal?
(d) ¿Qué conclusiones pueden extraerse del estudio?

2. En una encuesta realizada entre las familias españolas se pregunta, entre otras cosas, por la clase socioeconómica a la que la familia pertenece. Las posibles respuestas son éstas:

1	Jornaleros
2	Empresarios agrarios sin asalariados
3	Obreros
4	Autónomos o sin clasificar
5	Clase media
6	Clase alta
7	Retirados

Las respuestas recogidas en 60 de los hogares entrevistados fueros las siguientes.

3	7	3	5	6	5	1	5	7	5	5	4
3	5	1	1	4	2	2	3	2	3	7	5
4	3	3	5	5	5	7	7	5	1	4	2
1	7	4	4	3	3	4	5	3	3	6	6
7	2	7	1	6	3	2	5	3	7	2	2

Visitas previas al parque	Número de visitantes
0	6.158
1	1.825
2	1.757
3	1.182
4	538
5	362

Tabla 1: Distribución de los clientes de un parque de atracciones según su asiduidad.

(a) Escribe la tabla de frecuencias asociada.
(b) Representa los datos mediante un diagrama de barras y un diagrama de sectores.
(c) Identifica la clase modal.

3. Los 6 pictogramas que aparecen en la figura 1 representan la cantidad de viviendas edificadas en tres capitales durante los últimos dos años.
(a) Señala si en cada uno de ellos los dibujos representan fielmente las cifras que los acompañan.
(b) Indica cuáles de ellos relacionarías con un diagrama de barras y cuáles con un diagrama de tarta.

4. Considera los datos sobre los estados europeos recogidos en la Tabla 1.2. de Moore (1997). Representa mediante un histograma las siguientes variables.

1. *Población.*
2. *PIB per cápita.*
3. *%PIB en educación pública.*

5. Se realizó una encuesta entre las personas que acudieron durante el mes de agosto a un parque de atracciones. Se les preguntaba cuántas veces habían acudido con anterioridad a ese mismo parque. Se obtuvieron los datos recogidos en la tabla 1.
(a) Hallar la distribución de frecuencias relativas.
(b) Escribe las frecuencias acumuladas.
(c) Representar mediante un histograma y un polígono de frecuencias.
(d) ¿Cuál es la moda de la distribución?

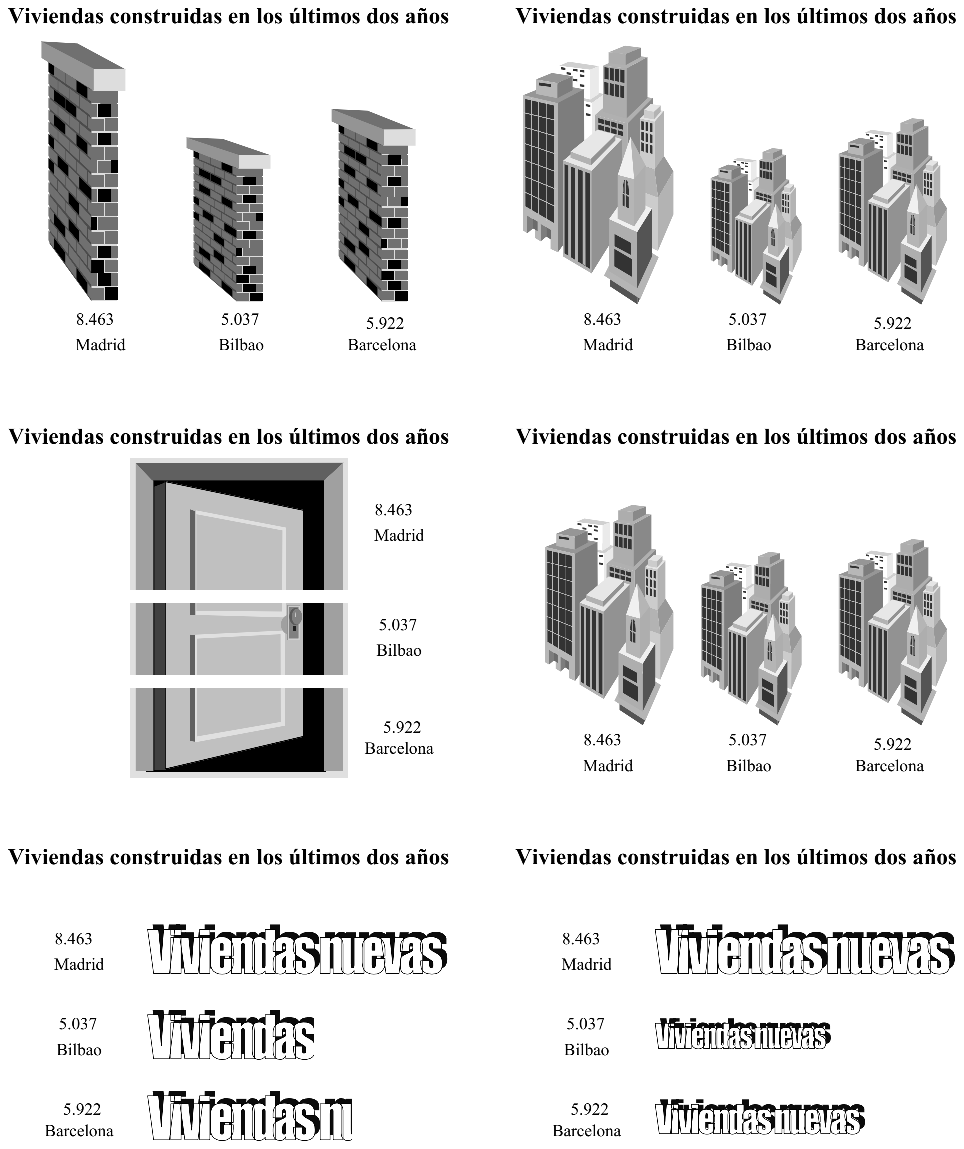


Figura 1: Diversos pictogramas que representan los mismos datos.

0	3	4	8	10	12	12	15	15	17	17
19	21	21	23	25	26	32	33	40	60	81

Tabla 2: Horas perdidas por bajas médicas.

Presupuesto dedicado a zonas verdes	Número de casos (n_i)
[0,20]	9
(20,40]	24
(40,60]	20
(60,80]	12
(80,100]	2
(100,200]	1

Tabla 3: Distribución de municipios según su presupuesto en zonas verdes.

(e) ¿Es esta variable simétrica?
(f) ¿Qué conclusiones pueden extraerse de los datos?

6. El conjunto de datos recogido en la Tabla 2 corresponde al número de horas que 22 trabajadores perdieron por bajas médicas el año pasado.
(a) Construye la tabla de frecuencias correspondiente a una partición del rango de valores en intervalos de anchura 10, indicando también las frecuencias absolutas y relativas acumuladas.
(b) Dibuja el histograma y el polígono de frecuencias asociados a las frecuencias relativas.
(c) Representa las frecuencias acumuladas mediante un gráfico del tipo histograma y mediante un polígono de frecuencias acumuladas.
(d) Señala alguna característica de los datos que se desprenda de esos gráficos.

7. La tabla 6 muestra las frecuencias correspondientes a la variable *"presupuesto anual dedicado a zonas verdes"* (medido en miles de euros) observada en los municipios de una determinada comarca turística.
(a) Completa la tabla incluyendo frecuencias relativas y frecuencias acumuladas absolutas y relativas.
(b) Dibuja un histograma a partir de la información de la tabla.
(c) ¿Qué proporción de municipios gasta más de 40 mil y menos de 100 euros en zonas verdes?

8. Los 301 municipios de una comunidad autónoma han sido agrupados según el número de habitantes. La tabla 4 recoge las frecuencias absolutas resultantes.
(a) Completa la tabla de frecuencias escribiendo las frecuencias relativas f_i, las frecuencias absolutas acumuladas N_i y las relativas acumuladas F_i.
(b) Representa gráficamente la distribución de la variable *número de habitantes* mediante un histograma y un polígono de frecuencias.
(c) ¿Qué forma tiene la distribución? ¿Qué conclusiones pueden extraerse de los datos?

Número de habitantes	n_i	N_i	f_i	F_i
$[0, 500)$	127			
$[500, 1000)$	74			
$[1000, 1500)$	30			
$[1500, 2000)$	22			
$[2000, 2500)$	15			
$[2500, 3000)$	11			
$[3000, 3500)$	3			
$[3500, 4000)$	4			
$[4000, 4500)$	2			
$[4500, 5000)$	2			
$[5000, 5500)$	1			
$[5500, 6000)$	2			
$[6000, 6500)$	5			
$[6500, 7000)$	1			
$[7000, 7500)$	1			
$[7500, 8000)$	0			
$[8000, 8500)$	0			
$[8500, 9000)$	1			

Tabla 4: Distribución de 301 municipios según número de habitantes.

9. Considera los datos sobre los estados europeos recogidos en la Tabla 1.2. de Moore (1997). Representa mediante un diagrama de tallo y hojas las siguientes variables.

1. *periódicos por 1000 habitantes*
2. *Televisores por 1000 habitantes*
3. *%PIB en educación pública.*

10. Considera los datos sobre los estados europeos recogidos en la Tabla 1.2. de Moore (1997). Selecciona las variables siguentes:

población, PIB per cápita,
periódicos por 1000 habitantes, %PIB en educación pública.

(a) Calcula la mediana cada variable.
(b) Calcula los cuartiles y los rangos intercuartílicos.
(c) Dibuja los diagramas de caja correspondientes.

11. Un conjunto de datos ha sido resumido en el siguiente diagrama de tallo y hojas.

```
Num.   Tallo|Hojas
casos
    4       0|0348
    9       1|022557779
    5       2|11356
    2       3|23
    1       4|0

    2 Extrm.|(603)(817)


           Anchura del tallo:  100
    N\'umero de casos por hoja:    1
```

(a) Para esos datos, construir la tabla de frecuencias que tenga intervalos de anchura 100.

(b) Identificar en el gráfico las *hojas* correspondientes a los datos 32 y 217.

(c) Indicar alguna característica de la distribución de los datos que se pueda deducir del gráfico.

12. Los siguientes datos corresponden a régimen de tenencia de viviendas en España:

Por compra (totalmente pagada)	6.158.320
Por compra (con pagos pendientes)	1.825.695
Por herencia o donación	1.182.109
Facilitadas (por otras personas, empresas u organismos)	538.572
En alquiler	1.757.469
Otras formas 362.504	

(a) Hallar la distribución de frecuencias relativas.

(b) Representar mediante un diagrama de Pareto y un diagrama de sectores la distribución de la variable.

(c) ¿Cuál es la clase modal?

(d) ¿Qué conclusiones pueden extraerse?

13. Los siguientes datos x_i corresponden al número de meses que llevan en paro 10 personas apuntadas a las listas del INEM:

4.4, 16.0, 3.0, 6.8, 0.8, 3.4, 6.3, 13, 3.0, 13.4

(a) Calcula la media $\bar{x}$.
(b) Escribe las desviaciones de los datos a la media (es decir, $x_i - \bar{x}$) y comprueba que suman 0.
(c) Calcula la desviación típica de los datos utilizando las dos expresiones equivalentes siguientes:

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad s_x^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2.$$

¿Con cuál resulta más sencillo el cálculo?

14. Se entrevistó a 5 matrimonios que recientemente habían adoptado un niño. Se les preguntó por el tiempo (en meses) que transcurrió desde que solicitaron la incorporación a la lista de espera hasta que se materializó la adopción. Las respuestas fueron

27, 15, 26, 26, 21.

(a) Calcula la media y la desviación típica.
(b) ¿Tiene sentido comparar la desviación típica de estos datos con la desviación típica de los datos del problema anterior?
(c) ¿Qué debe utilizarse para comparar la variabilidad (es decir, la dispersión) de ambos conjuntos de datos?
(d) ¿Cuál de los dos conjuntos de datos tiene mayor dispersión?

15. Se ha contado el número de oficinas municipales de información al consumidor abiertas al público en 25 ciudades. Los datos resultantes son

3, 6, 4, 2, 3, 4, 5, 4, 7, 3, 5, 4, 5, 4, 3, 3, 4, 3, 2, 4, 6, 1, 8, 2, 5.

(a) Construye la tabla de frecuencias de estos datos y dibuja un gráfico de barras.
(b) Calcula la media, la varianza, la desviación típica y el coeficiente de variación.

Sexo	Estatura	Sexo	Estatura	Sexo	Estatura	Sexo	Estatura	Sexo	Estatura
V	166	V	168	V	169	V	169	V	170
V	170	V	171	V	171	V	173	V	173
V	174	V	176	V	176	V	177	V	178
V	178	V	178	V	179	V	179	V	180
V	180	V	183	V	184	M	155	M	159
2	157	M	159	M	160	M	160	M	160
2	162	M	162	M	162	M	163	M	163
2	164	M	164	M	164	M	165	M	166
2	165	M	166	M	166	M	166	M	167
2	168	M	168	M	168	M	168	M	170
2	170	M	170	M	171	M	171	M	172
2	172	M	175	M	175				

Tabla 5: Sexo (V=varón, M=mujer) y estatura de 58 personas.

16. Calcula la media, la varianza, la desviación típica y el coeficiente de variación de los datos de la tabla 2.

17. Calcula la mediana de los conjuntos de datos de los ejercicios anteriores que se refieren a meses en paro y a oficinas de información al consumidor. Calcula también el rango intercuartílico de los datos.

18. Calcula la mediana, los cuartiles y el rango intercuartílico de los datos de la tabla 2.

19. Considera los datos de la tabla 1. ¿Cuál es el número medio de visitas previas al parque entre los visitantes encuestados?

20. Calcula la media, la varianza, la desviación típica y el coeficiente de variación de los datos de la tabla 6.

21. Calcula la media, la varianza, la desviación típica y el coeficiente de variación de los datos de la tabla 4.

22. Los datos de la tabla 5 corresponden a la estatura de 58 personas de ambos sexos con edades comprendidas entre los 20 y los 22 años.

(a) Escribe la tabla de frecuencias absolutas de la variable *estatura* asociada a los intervalos $[154, 156)$, $[156, 158), \ldots, [182, 184)$, $[148, 186)$.

(b) Usando como base la lista de intervalos anteriores haz tres histogramas de esa variable (todos con el mismo números de barras y abarcando el mismo rango de valores): uno para todos los individuos encuestados, otro para las mujeres (M) y otro para los varones (V).

(c) ¿Se aprecia bimodalidad en el primero de los histogramas?

(d) Si la respuesta es sí, ¿los dos histogramas restantes permiten asegurar que las dos modas corresponden a individuos de distinto sexo?

23. Se preguntó a 20 alumnos universitarios por el número de hermanos que tenían. Parte de la tabla de frecuencias se ha perdido y sólo disponemos de la información referente a frecuencias acumuladas:

x_i	N_i
0	4
1	12
2	18
3	18
4	20

(a) ¿Cuántos alumnos tienen como mucho 1 hermano?
(b) ¿Cuántos tienen al menos 2 hermanos y como mucho 3?
(c) ¿Qué proporción de alumnos tienen al menos dos hermanos?
(d) Reconstruye la tabla de frecuencias completa a partir de la información dada.

24. Se estudió la afluencia de clientes a las farmacias en horario de guardia nocturno. Se seleccionaron 60 farmacias de guardia y en cada una de ellas se anotó el número de clientes que acudieron entre las 2 y las tres de la madrugada. Las frecuencias relativas acumuladas son las siguientes.

Número de clientes	Frecuencia relativa acumulada
0	11.67
1	35.00
2	66.67
3	90.00
4	93.33
5	98.33
6	100.00

(a) Completa la tabla de frecuencias escribiendo frecuencias absolutas y relativas no acumuladas.
(b) Calcula la media y la mediana de estos datos.
(c) Cuál es la clase modal?

25. Asigna valores para la media, la desviación típica y los coeficientes de asimetría y apuntamiento de los histogramas representados en la figura 2. Usa valores de los que aparecen listados en la tabla.

Posibles valores

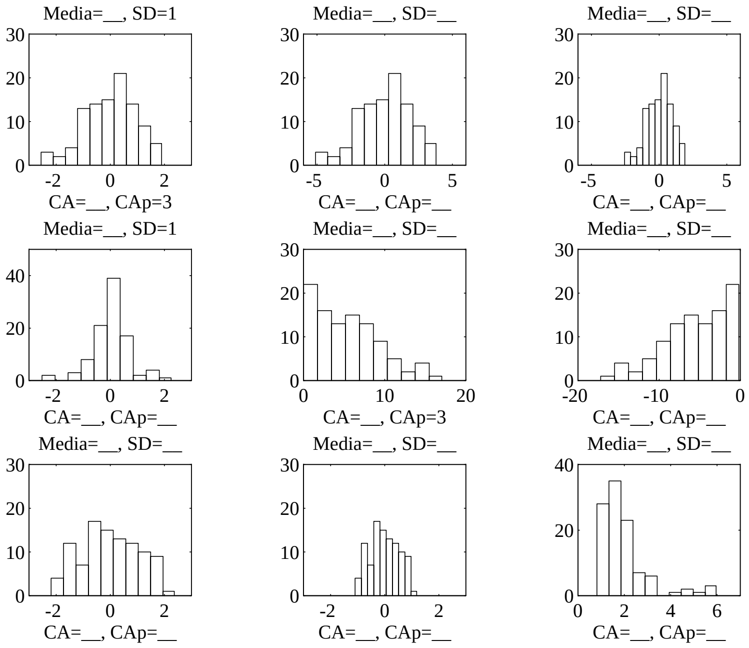


Figura 2: Histogramas de algunas variables.

Media	Desv. típica	Asimetría
0	1	simétrica
2	2	asimétrica
-5	0.5	
5	4	

26. Considera los datos de la tabla 5.

(a) Calcula la mediana y los cuartiles de todos los individuos.

(b) Calcula la mediana y los cuartiles de los datos referidos a mujeres.

(c) Calcula la mediana y los cuartiles de los datos que corresponden a los hombres.

(d) Dibuja en el mismos sitema de ejes coordenados tres diagrama de caja paralelos, uno para la población completa, otro para las mujeres y otro para los hombres.

(e) El gráfico que acabas de dibujar, ¿aporta una información acorde con la que se obtiene de representar las distribuciones de los datos mediante

histogramas?

27. Considera los datos de la tabla 1.

(a) Completa la tabla siguiente.

i	x_i	n_i	f_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	x_i^2
1	0	6.158				
2	1	1.825				
3	2	1.757				
4	3	1.182				
5	4	538				
6	5	362				

(b) Calcula la media de la variable *visitas previas* mediante la fórmula $\bar{x} = \sum_{i=1}^{6} x_i f_i$.

(c) Calcula la varianza de esa variable mediante la fórmula $s^2 = \sum_{i=1}^{6}(x_i - \bar{x})^2 f_i$.

(d) Calcula la media de los valores x_i^2:

$$\bar{x^2} = \sum_{i=1}^{6} x_i^2 f_i.$$

(e) Comprueba que la varianza de x también se puede calcular como

$$s^2 = \bar{x^2} - (\bar{x})^2 = \left(\sum_{i=1}^{6}(x_i - \bar{x})^2 f_i\right) - (\bar{x})^2.$$

28. Calcula la varianza de los datos listados en las tablas 6 y 4 usando la fórmula $s^2 = \bar{x^2} - (\bar{x})^2$.

29. Construye el histograma asociado a la distribución de frecuencias absolutas de los datos listados en la tabla 6. De esta forma el área total encerrada bajo el histograma será igual al número total de individuos.

(a) ¿Qué área encierra el histograma entre los valores 0 y 20 del eje de abscisas?

(b) ¿Qué área hay entre los valores 0 y 50?

(c) Calcula la mediana de ese conjunto de datos como el valor m tal que el histograma encierra entre 0 y m un área igual a la mitad del número de observaciones.

(d) ¿Cómo podrían definirse los cuartiles Q_1 y Q_3? Calcúlalos para estos datos.

30. Calcula la mediana y los cuartiles de los datos de la tabla 4 a partir del histograma que representa su distribución de frecuencias.

31. Demuestra por inducción en n las siguientes propiedades de los sumatorios.

$$\textbf{(a)}\ \sum_{i=1}^{n} a = an \quad \textbf{(b)}\ \sum_{i=1}^{n} ax_i = a\sum_{i=1}^{n} x_i$$

$$\textbf{(c)}\ \sum_{i=1}^{n}(x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

$$\textbf{(d)}\ \sum_{i=1}^{n}(x_i + a)^2 = \sum_{i=1}^{n} x_i^2 + 2a\sum_{i=1}^{n} x_i + na^2$$

$$\textbf{(e)}\ \sum_{i=1}^{n}(x_i + y_i)^2 = \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2 + 2\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

32. Utiliza las propiedades de los sumatorios para demostrar las siguientes identidades.

$$\textbf{(a)}\ \sum_{i=1}^{n}(x_i - \bar{x}) = 0 \quad \textbf{(b)}\ \frac{1}{n}\sum_{i=1}^{n}(x_i\bar{x}) = (\bar{x})^2$$

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

33. Un test de personalidad da resultados con distribución aproximadamente normal de media de 100 puntos y desviación típica de 20.

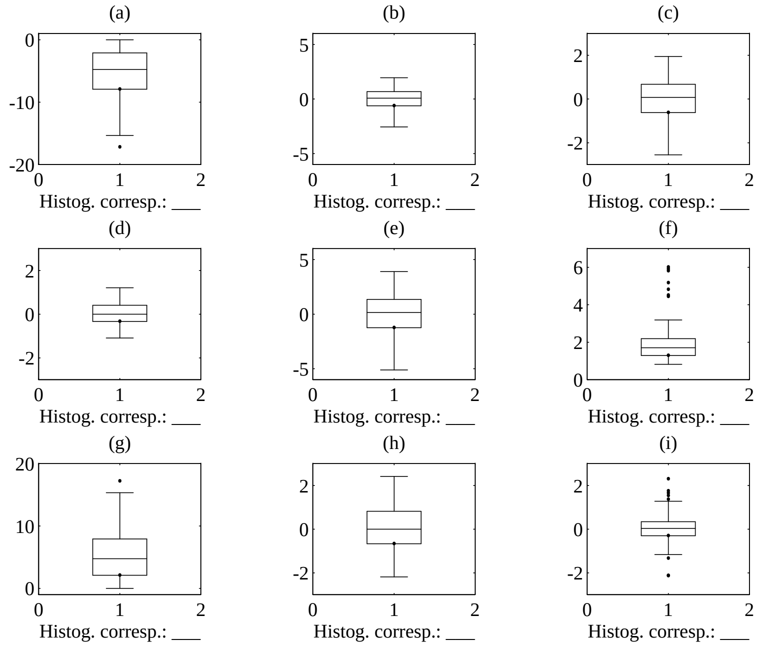
(a) Calcular la frecuencia de estudiante tales que su puntuación ...

1. es menor que 90,
2. está entre 80 y 110,
3. es mayor que 120.

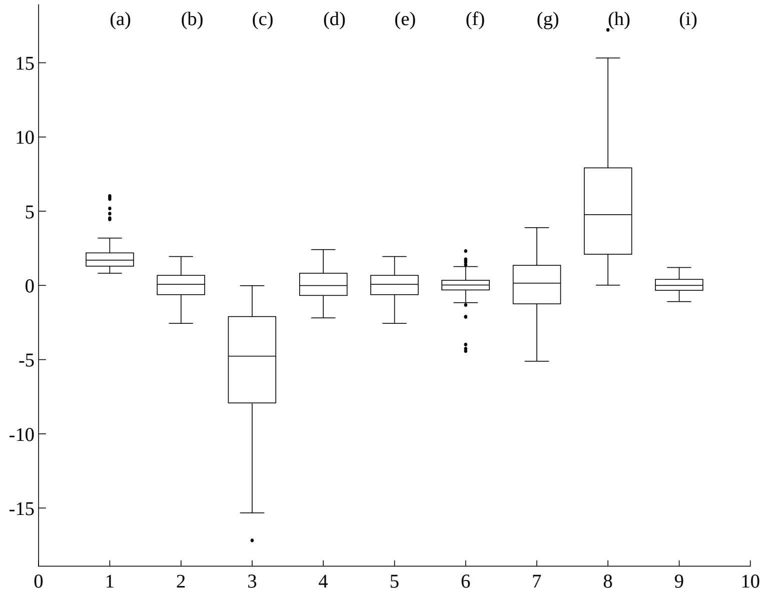
(b) Se pasa el test a una clase con 60 alumnos. Calcular la probabilidad de que la puntuación media de la clase:

1. es menor que 90,
2. está entre 80 y 110,
3. es mayor que 120.

34. Cada uno de los diagramas de caja de la figura corresponde a uno de los nueve histogramas que aparecían en uno de los problemas anteriores. Indica a qué histograma corresponde cada diagrama de caja.



35. Cada uno de los diagramas de caja de la figura corresponde a uno de los diagramas de caja del problema anterior. Indica a cuál.



Análisis exploratorio de datos. Descripción de dos variables

1. Escribe la tabla de distribución conjunta de las variables que aparecen en la tabla, según se indique.

(a) X frente Y.
(b) Z frente V.
(c) X frente Z.

X	Y	Z	V
53	55	12	12
53	56	11	13
54	55	11	12
55	53	11	13
55	53	13	12
55	55	12	11
56	51	13	12
56	52	13	12
56	54	10	13
56	56	10	11
57	51	12	12
57	52	12	11
57	53	13	12
57	55	12	11
58	51	11	12
58	52	11	11
57	54	10	11
59	51	12	11
59	52	12	12
59	51	10	12

2. La siguiente tabla de frecuencias absolutas conjuntas corresponde al cruce de las variables EGB(*Tipo de escuela donde estudiaste EGB*) y BUP (*Tipo de escuela donde estudiaste BUP*).

BUP / EGB	Escuela religiosa	Privada laica	Escuela pública
Escuela religiosa	18	2	7
Privada laica	2	15	6
Escuela pública	1	0	24

(a) Calcula las frecuencias absolutas marginales.

(b) Calcula la tabla de frecuencias relativas conjuntas y complétala con las frecuencias relativas marginales.

(c) Calcula las frecuencias relativas de la variable *BUP* condicionada a los distintos valores de la variable *EGB*.

(d) Calcula las frecuencias relativas de la variable *EGB* condicionada a los distintos valores de la variable *BUP*.

(e) ¿Qué porcentaje del total de estudiantes pasó de hacer la EGB en un colegio público a hacer el BUP en uno privado? ?Qué porcentaje representan esos estudiantes entre todos los que hicieron EGB en colegios públicos?

(f) ¿Crees que el tipo de colegio donde se estudia el BUP es independiente del tipo de colegio donde se estudió la EGB? Sustenta tu respuesta en los datos de las diferentes tablas.

(g) Podemos definir una medida de *fidelidad* a un tipo de colegio como el porcentaje de alumnos que habiendo estudiado en un tipo de colegio la EGB, siguen en el mismo tipo de institución sus estudios de BUP. Calcula los índices de fidelidad para cada uno de los tres tipos de colegios contemplados en la encuesta. ¿Cómo podrías explicar las diferencias en los grados de fidelidad observados?.

3. La siguiente tabla de frecuencias corresponde al presupuesto anual dedicado a zonas verdes (medido en millones de pesetas) por parte de municipios de dos comarcas, una de ellas turística y la otra eminentemente industrial.

	Presupuesto dedicado a zonas verdes					
	[0,2]	(2,4]	(4,6]	(6,8]	(8,10]	(10,20]
Comarca turística	9	24	20	12	2	1
Comarca industrial	13	25	11	1		

(a) ¿Qué dos variables se están estudiando conjuntamente con esta tabla? ¿Son las dos discretas?

(b) Calcula las distribuciones del gasto de los municipios en zonas verdes, condicionadas a la comarca a la que pertenecen.

(c) Calcula también las medias condicionadas.

(d) ¿Podemos afirmar que el presupuesto de un municipio dedicado a zonas verdes es independiente de si la comarca en que se encuentra es turística o industrial? ¿Por qué?

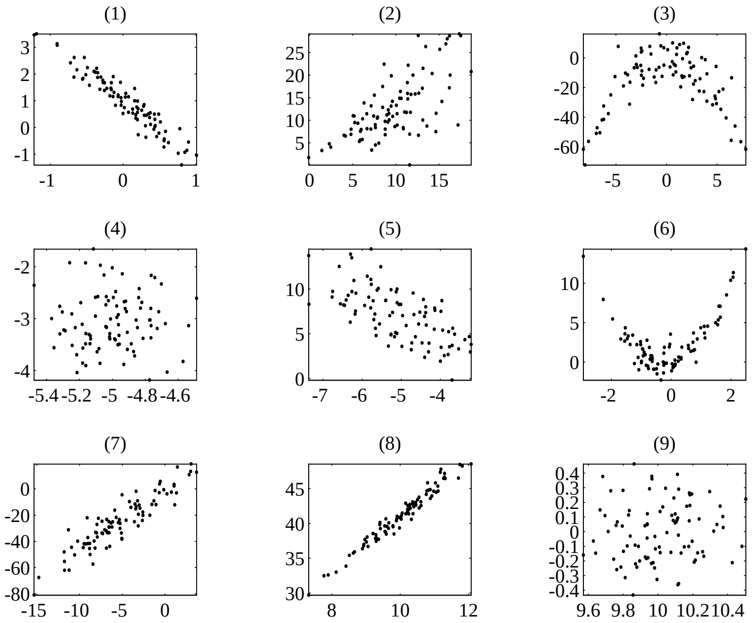
4. Se han observado dos variables estadísticas conjuntamente en 100 individuos. La información obtenida se ha resumido en la siguiente tabla incompleta de frecuencias relativas, en la que también aparecen algunas frecuencias relativas marginales.

$X\backslash Y$	1	3	5	
0	0,05	?	0,1	0,2
1	?	0,1	?	0,4
2	0,2	?	?	?
	?	0,3	0,2	

(a) Completa la tabla y calcula la media de X e Y.
(b) Calcula las distribuciones condicionadas de Y a valores de X y sus medias.

5. Considera las nubes de puntos siguientes. Indica si hay relación de dependencia entre la variable x (eje de abcisas) y la variable y (eje de ordenadas). En caso de haberla, ¿puede considerarse esta relación lineal? Asigna a las nubes de puntos en las que se observa relación lineal entre las variables una de las siguientes rectas (elige aquella que mejor reflejaría la información que aporta el gráfico de dispersión).

$(a) y = x$; $(b) y = 1 - 0.2x$; $(c) y = 2 + 4x$; $(d) y = 2 - 2x$; $(e) y = -3 - 2x$;



6. Calcular el coeficiente de correlación para los datos de la tabla.

x	2	3	4	5	6
y	5	7	8	13	14

Sin hacer cálculos, responder a las preguntas siguientes (puedes ayudarte de gráficos de dispersión).
¿Cómo cambiaría este ceficiente si ...
- sumamos 3 a la variable x?
- sumamos 3 a ambas variables?
- multiplicamos la variable x por 2?
- intercambiamos los valores de x por los de y?
- cambiamos el último valor de x por el de y?
- sumamos 10 a ambas variables pero sólo en el primer punto observado?

7. Calcular la recta de regresión para los datos de la tabla. Recordar que la recta de regresión es

$$\hat{y} = b_0 + b_1 x,$$

y que los coeficientes b_0 y b_1 se calculan como

$$b_1 = \frac{S_{xy}}{S_x^2} = r_{xy}\frac{S_y}{S_x},$$

$$b_0 = \bar{y} - b_1\bar{x}.$$

(**Nota:** aprovecha los cálculos hechos para el ejercicio anterior).

x	2	3	4	5	6
y	5	7	8	13	14

Sin hacer cálculos, responder a las preguntas siguientes (puedes ayudarte de gráficos de dispersión).
¿Cómo cambiaría la recta de regresión si ...
- sumamos 3 a la variable x?
- sumamos 3 a ambas variables?
- multiplicamos la variable x por 2?
- intercambiamos los valores de x por los de y?
- cambiamos el último valor de x por el de y?
- sumamos 10 a ambas variables pero sólo en el primer punto observado?

8. El crecimiento de los niños en sus priemros años muestra una relación lineal entre la edad y la estatura. Los siguientes datos corresponden a la estatura en centímetros de una niña entre los 3 y los 5 años.

Edad (meses)	36	48	51	54	57	60
Estatura (cm)	86	90	91	93	94	95

(a) Calcula las medias, varianzas y desviaciones típicas de la edad y la altura.

(b) Calcula la covarianza de estas dos variables y el coeficiente de correlación. Haz un diagrama de dispersión de los datos de la tabla.

(c) Calcula la recta de regresión que da la estatura (y) en función de la edad (x). Dibújala en el mismo gráfico en que has dibujado la nube de puntos.

(d) La tasa de crecimiento de esta niña es el número de centímetros que crece cada mes. De acuerdo con la recta de regresión calculada, ¿cuánto vale esta tasa de crecimiento?

(e) Calcula las desviaciones e_i de los puntos observados a la recta de regresión:

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i), i = 1, \ldots, 6.$$

$x(a,b,c)$	$y(a)$	$y(b)$	$y(c)$	$x(d)$	$y(d)$
10	8,04	9,14	7,46	8	6,58
8	6,95	8,14	6,77	8	5,76
13	7,58	8,74	12,74	8	7,71
9	8,81	8,77	7,11	8	8,84
11	8,33	9,26	7,81	8	8,47
14	9,96	8,10	8,84	8	7,04
6	7,24	6,13	6,08	8	5,25
4	4,26	3,10	5,39	19	12,50
12	10,84	9,13	8,15	8	5,56
7	4,82	7,26	6,42	8	7,91
5	5,68	4,74	5,73	8	6,89

Tabla 6: Datos publicados por T. Anscombe en 1973 en la revista *The american Statistician.*

(f) Calcula la varianza residual y la desviación típica residual:

$$S_r^2 = \frac{1}{6}\sum_{i=1}^{6} e_i^2, \quad S_r = \sqrt{S_r^2}.$$

(g) Calcula el coeficiente R^2, es decir, el porcentaje de la variabilidad de la estatura de la niña a lo largo del tiempo que queda explicado por la edad. Hazlo de las dos formas siguientes:

$$R^2 = 1 - \frac{S_r^2}{S_y^2}, \text{ o bien } R^2 = (r_{xy})^2,$$

y comprueba que se obtiene el mismo resultado.

(h) ¿Qué estatura crees que tenía esa niña a los 42 meses de edad?

9. Ejercicio para resolver con el ordenador. Los datos de la tabla 6 constituyen un ejemplo clásico en regresión.

(a) Comprueba que las cuatro regresiones

`y(a)` sobre `x(a,b,c)`,
`y(b)` sobre `x(a,b,c)`,
`y(c)` sobre `x(a,b,c)`,
`y(d)` sobre `x(d)`,

dan lugar a la misma recta, que el coeficiente de correlación es también el mismo y que el coeficiente de determinación también coincide.

(b) Haz ahora los gráficos de las cuatro nubes de puntos junto con la recta de regresión ajustada.

(c) Indica para cada conjunto de datos si crees el modelo de regresión es apropiado y, si no es ese el caso, indica por qué no lo crees así.

(d) Si hubieses visto primero los cuatro gráficos de dispersión, ¿habrías pensado que era sensato calcular la recta de regresión lineal para todos ellos?

10. Ejercicio para resolver con el ordenador. La tabla contiene los pesos y alturas de 57 personas entre 20 y 22 años de ambos sexos.

SEXO	PESO	ALTURA	SEXO	PESO	ALTURA	SEXO	PESO	ALTURA
H	62	170	H	57	166	H	82	190
H	73	180	H	75	176	H	70	178
H	68	173	H	70	178	H	65	170
H	68	173	H	84	183	H	75	168
H	60	169	H	68	179	H	82	177
H	72	179	H	56	171	H	80	180
H	61	174	H	62	171	H	63	178
H	78	184	H	58	169	M	56	166
M	57	164	M	56	171	M	57	168
M	60	175	M	65	170	M	51	157
M	52	162	M	56	164	M	50	164
M	57	162	M	62	168	M	65	171
M	48	165	M	51	160	M	63	175
M	49	163	M	55	166	M	58	165
M	59	167	M	47	165	M	52	163
M	50	160	M	60	168	M	57	172
M	48	155	M	53	175	M	56	170
M	50	159	M	60	168	M	53	170
M	53	160	M	62	172	M	56	162

(a) Realiza un gráfico de dispersión de la variable *PESO* frente a la variable *ALTURA*, señalando con símbolos diferentes para sexos distintos. ¿Qué conclusiones pueden extraerse de la diferencia de sexos?

(b) Calcula la covarianza entre peso y altura que corresponde a mujeres. Calcula también el coeficiente de correlación. ¿Son acordes estos valores con la información que se desprende de los gráficos de dispersión? Repite los cálculos para los datos relativos a hombres.

(c) Una persona que participó en el estudio del ejercicio anterior no contestó cuando se le preguntó por su peso. Si se sabe que mide 170 cm, ¿cuánto debe pesar? ¿Y si midiese 167 cm?

(d) Si además se sabe que la persona que mide 170 cm es mujer, ¿cuánto se diría que pesa?

(e) ¿Y qué se diría del peso de la persona que mide 167 cm si se sabe que es varón?

11. Los siguientes son ejemplos de comparación de medias de dos variables. Indicar en cada caso si se trata de comparación de medias en dos poblaciones independientes (los individuos en los que se mide una variable no están relacionados con ninguno de aquellos en los que se mide la otra; además, los tamaños de las dos poblaciones no tienen por qué coincidir) o si se trata de dos muestras con datos apareados (se han medido las dos variables en los mismos individuos, o bien se trabaja con pares de individuos relacionados y se mide una variable en cada miembro del par).

(a) Se queire saber si en los matrimonios, los sueldos medios de los maridos y los de las esposas son semejantes.

(b) Se desea comparar si la renta per cápita media es igual en dos comunidades autónomas distintas.

(c) Los vendedores de una empresa reciben un curso de nuevas técnicas de venta y se desea saber si las ventas medias antes y después del curso son iguales.

(d) Para los datos de la encuesta de prácticas, se quiere saber si la nota de selectividad fue, en promedio, igual para chicos que para chicas.

(e) Para los datos de la encuesta de prácticas, se compara la asistencia media al cine con la nivel medio de atracción por el cine.

La binomial y la normal

1. Calcula la esperanza, la varianza y la desviación típoca de las variables aleatorias con funciones de probabilidad dadas por las siguientes tablas.

(a)

x_i	1	2	3	4
p_i	0,25	0,25	0,25	0,25

(b)

x_i	$-\pi$	0	π
p_i	0,2	0,6	0,2

(c)

x_i	3
p_i	1

2. Para las variables aleatorias que se describen a continuación señala si siguen una distribución de Bernoulli, si son binomiales o si no responden a estos patrones.

(a) Se lanzan dos dados y se define X como la suma de puntos obtenidos.

(b) Se lanzan dos dados y se define X como la variable aleatoria que toma el valor 1 si la suma es par y 0 si la suma es impar.

(c) Se lanza una moneda hasta que sale la primera cara y se define X como el número de lanzamientos necesarios para que eso ocurra.

(d) Se define X como la variable aleatoria que vale 1 si al lanzar un dado aparece un número múltiplo de 3, y que vale 2 en caso contrario.

3. En un examen tipo test cada pregunta tiene tres posibles respuestas, sólo una de las cuáles es correcta. Cada pregunta contestada correctamente aporta un punto a la nota global del estudiante. Las preguntas no contestadas no suman ni restan puntos. Si una pregunta está mal contestada, se resta una cantidad m a la nota total. ¿Cuánto debe valer m para que un alumno sea indiferente entre contestar al azar una pregunta cuya respuesta no conoce o dejarla en blanco? ¿Cuál sería el valor de m si cada pregunta tuviese cuatro posibles respuestas?

4. Un vendedor inmobiliario sabe que cuando va con un cliente a ver un piso en venta, con probabilidad 0,08 esa visita acaba en una venta. El vendedor concierta 25 visitas esta semana.

(a) ¿Cuál es la probabilidad de que venda al menos un piso?

(b) ¿Con qué probabilidad venderá más de cuatro pisos?

(c) ¿Qué es más probable, que venda a lo sumo dos pisos o que venda entre 4 y 6 pisos?

5. Un obrero eventual consigue trabajar X horas a la semana, donde X es una variable aleatoria normal con media 25 y desviación típica de 5.

(a) Si cobra 20 euros por hora trabajada y ha de pagar un seguro médico de 100 euros semanales, ¿cuál es la distribución de sus ingresos netos durante una semana?

(b) Si expresa esos ingresos en pesetas ¿qué distribución tendrán?

6. Un estudiante ha obtenido la misma nota en los exámenes de matemáticas y de literatura: un 6. En el primer examen la media de las notas de su grupo fue de 7 y la desviación típica de 0,8. Por su parte, la nota media de las notas de literatura fue de 5,5, con una desviación típica de 0,2. Las calificaciones de ambas asignaturas se puede suponer que siguen distribuciones normales.

(a) Calcula la puntuación tipificada de ese estudiante en los dos exámenes.

(b) Calcula qué proporción de alumnos obtuvieron peores notas que ese estudiante en cada una de los exámenes.

(c) ¿Qué nota crees que tiene más valor para el estudiante, comparada con las notas del resto de sus compañeros, el 6 de matemáticas o el 6 de literatura?

7. El 30% de los votantes se inclinarán por el candidato A en las elección que se celebrarán dentro de 15 días. Hoy se hace una encuesta electoral en la que se entrevista a 150 votantes.

(a) ¿Con qué probabilidad habrá entre 40 y 50 encuestados a favor del candidato A?

(b) ¿Con qué probabilidad habrá más de un 40% de encuestados partidarios del candidato A?

8. La longitud del pie de los hombres adultos se distribuye según una variable aleatoria normal con media 24,5 cm y desviación típica de 1 centímetro. Un fabricante de zapatos saca al mercado un modelo del que presenta 4 tamaños, cada uno de los cuales es adecuado para pies con longitudes comprendidas en un intervalo de 1 cm. El zapato de menor talla es válido para longitudes de pie entre 23 cm y 24 cm, y los números siguientes cubren los intervalos de 24 a 25, de 25 a 26 y de 26 a 27.

(a) ¿Qué proporción de caballeros encontrarían un zapato de su número entre los fabricados de este modelo?

(b) Si el fabricante decide fabricar un número más, ¿qué debería hacer, el número inmediatamente anterior a los ya fabricados o el número inmediatamente posterior?

9. Un caminante llega a las puertas de una ciudad amurallada y pide permiso para entrar. El centinela le informa de que no es tan fácil acceder a esa ciudad. Para conseguirlo habrá de lanzar 20 veces un dado y sólo entrará si consigue exactamente cinco veces un cinco.

(a) ¿Con qué probabilidad logrará el caminante entrar en la ciudad?

(b) Un día llegan a la ciudad 100 caminantes, ¿cuántos esperas que logren entrar?

(c) ¿Con qué probabilidad entrarán más de 15 caminantes?

10. Encuentra números reales x que verifiquen las siguientes propiedades:

(a) $P\{X \leq x\} = 0,3669$, donde X es $N(0,\ 1)$.

(b) $P\{X \leq x\} = 0,5$, donde X es $N(12,\ 5)$.

(c) $P\{X \geq x\} = 0,7642$, donde X es $N(150,\ 15)$.

(d) $P\{|X| \leq x\} = 0,2360$, donde X es $N(0,\ 4)$.

11. Un agente de seguros tiene cita con un cliente importante a las diez de la mañana. El agente sabe que su cliente llega a las citas con un retraso medio de 5 minutos. Considerando que el retraso es una variable aleatoria normal con desviación típica $\sigma = 2$ minutos, el agente quiere llegar antes que su cliente con una probabilidad del 90%. ¿En qué momento debe llegar el agente?

Población y muestra. Distribuciones en el muestreo

1. Identifica la población, la muestra, el parámetro de interés y su estimador (cuando aparezcan) en los siguientes afirmaciones.

(a) Una encuesta que incluye al 5% del total de la población pone de manifiesto que la renta media anual de los españoles entre 25 y 30 años de edad es de 1.872.300 pesetas.

(b) Se entrevista un total de 1000 personas en edad de trabajar, de los cuales 185 se encuentran en paro.

(c) De los 18 delegados de clase que hay en la EUM, 7 son varones.

2. La empresa ELECTRICAFEIT, fabricante de maquinillas de afeitar eléctricas, anuncia desde hace un mes su nuevo modelo MAE en televisión durante la emisión de los partidos de fútbol. La rentabilidad de esta campaña depende de cuál sea la proporción de potenciales compradores de MAE que ven los partidos por televisión. Para estimar esa proporción, se seleccionan aleatoriamente 120 personas que han sido identificadas como potenciales compradores del modelo MAE y se les pregunta si vieron por televisión algún partido de fútbol durante el último mes.

¿Cuál es la población: A, B, C ó D?

A. Los espectadores de los partidos de fútbol retransmitidos por televisión durante el último mes.

B. Los espectadores de los partidos de fútbol retransmitidos por televisión que potencialmente pueden comprar una MAE.

C. Los potenciales compradores de una maquinilla del modelo MAE.

D. Los espectadores de los partidos de fútbol retransmitidos por televisión durante el último mes que potencialmente pueden comprar una MAE.

3. Se sabe que el tiempo de funcionamiento de un determinado tipo de bombillas sigue una distribución de probabilidad con desviación típica de 200 horas.

(a) Para una muestra de 36 bombillas, ¿con qué probabilidad excederá la media muestral a la media poblacional en más de 50 horas?

(b) Para una muestra de 36 bombillas, ¿con qué probabilidad distará más de 80 horas la media muestral de la media poblacional?

(c) ¿Qué tamaño debe tener la muestra para asegurar que, al menos con una probabilidad de 0.05, la media muestral distará de la media poblacional menos de 25 horas?

4. Se lanza una moneda no trucada n veces.

(a) ¿Con qué probabilidad se obtendrán entre un 40% y un 60% de caras, para $n = 10$, $n = 100$ y $n = 1000$?

(b) ¿Qué hecho ponen de manifiesto los resultados anteriores? ¿Cómo se llama esa ley?

5. Diez economistas elegidos aleatoriamente dieron las siguientes predicciones del porcentaje del crecimiento del Producto Interior Bruto (PIB) para el próximo año:

$$2.2\ \ 2.8\ \ 3.0\ \ 2.5\ \ 2.4\ \ 2.6\ \ 2.5\ \ 2.4\ \ 2.7\ \ 2.6$$

Proponer estimadores puntuales para las siguientes parámetros y calcular el valor que toman en la muestra anterior.

(a) La media poblacional.

(b) La varianza poblacional.

(c) La varianza de la media muestral.

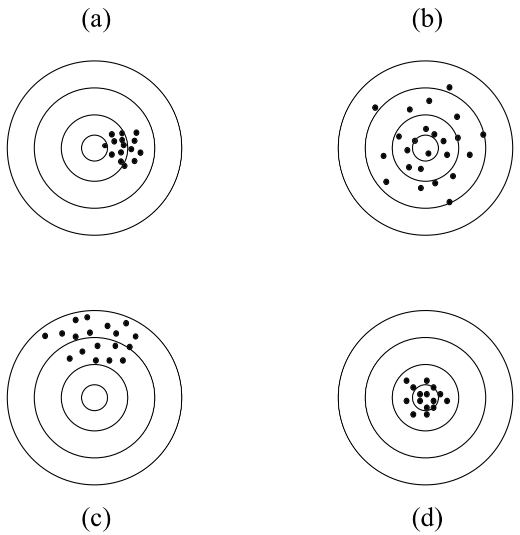
(d) La proporción poblacional de economistas que predicen un crecimiento del PIB de al menos un 2.5%.

(e) La varianza de la proporción muestral de economistas que predicen un crecimiento del PIB de al menos un 2.5%.

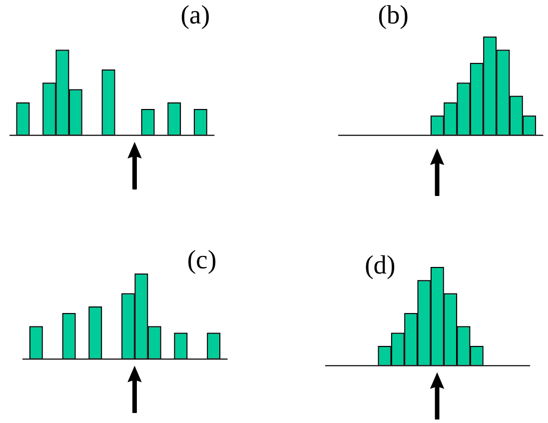
6. Un artículo del periódico daba cuenta de los resultados de una encuesta preelectoral: *"Sobre una muestra de 4800 entrevistas, [...], y con un nivel de confianza del 95.5%, el margen de error es del ± H% para datos globales y bajo un supuesto de máxima indeterminación (p=q=0.5)"*. ¿Cuál es el valor aproximado de H?

7. En un puesto de tiro al blanco hay cuatro escopetas de aire comprimido, etiquetadas con las letras a, b, c y d. He probado mi puntería con las cuatro escopetas y los resultados que he obtenido son los que se muestran en el gráfico siguiente.

1. ¿Cuáles de ellas tienen mal ajustado el punto de mira, es decir, con cuáles me desvío sistemáticamente del centro de la diana?
2. ¿Cuáles de ellas dispersan menos los impactos?
3. ¿Cuál es más precisa?
4. Si vuelvo mañana y sólo las escopetas (a) y (b) están libres, ¿con cuál me aconsejarías que disparara?



8. A continuación se muestran los histogramas de las distribuciones muestrales de cuatro estadísticos que intentan estimar el mismo parámetro. La posición del parámetro poblacional se señala con una flecha. Comparando las distribuciones entre sí indica cuál es la distribución con más y con menos sesgo, y también cuáles tienen mayor y menor varianza.



Técnicas de Decisión Avanzada I — **Enero-Marzo, 2000**
Pedro Delicado — **Lista de problemas 6**

Medias y proporciones. Estimación puntual y por intervalos

1. Estimar la proporción de estudiantes que están a favor de dedicar el 0.07% de ayuda al tercer mundo si una muestra aleatoria de 2000 estudiantes de las universidades españolas ha dado una proporción muestral de 0.72. Construir un intervalo de confianza del 95% y otro del 99%.

2. Se desea tomar una muestra aleatoria simple de votantes de un país. Calcular el tamaño muestral requerido si los intervalos de confianza de los votos al 95% quieren darse con una variación máxima de 5 puntos porcentuales.

3. Comenta la siguiente ficha técnica aparecida en un periódico junto a una encuesta electoral:

> *Encuesta realizada sobre una muestra de 2000 personas mayores de 18 años. Para un nivel de confianza del 95.5% y en la hipótesis más desfavorable de $p = q = 0.5$ el error de muestreo asciende a 2.2 puntos porcentuales para datos globales.*

4. Un test de personalidad da resultados con una distribución normal de media de 100 puntos y desviación típica de 20. Se pasa el test a una clase con 60 alumnos. Calcular la probabilidad de que la puntuación media de la clase:

(a) sea menor que 90,

(b) esté entre 80 y 110,

(c) sea mayor que 120.

5. En una muestra de 2000 personas se ha obtenido que en el año 1996 el gasto anual medio en alimentación es de 255.000 pts. por persona, con una desviación típica estimada de 187.000 pts.

(a) Calcula el error estándar de la media muestral.

(b) Calcula un intervalo de confianza para el gasto medio poblacional con un nivel del 95%.

6. Se lleva a cabo un sondeo de opinión en España y una de las preguntas hechas a las 1500 personas seleccionadas aleatoriamente es: *"¿Cree usted que el nivel de vida en España disminuirá apreciablemente debido a la entrada en funcionamiento del Euro?"*. Un total de 227 personas contestaron que sí y los restantes 1273 entrevistados dijeron que no lo creían así. Construye intervalos de confianza 95% para la proporción de españoles que contestarían afirmativamente a la anterior pregunta. Haz lo mismo para la proporción de respuestas negativas.

7. Para los nueve valores de la proporción poblacional $p = 0.1, 0.2, 0.3, ..., 0.9$, calcula el margen de error $(Z_{\alpha/2}\sqrt{p(1-p)/n})$ de un intervalo de confianza 95% para p, basado en una muestra de tamaño $n = 500$. Representa los resultados obtenidos como función de p.

8. Un analista financiero observa que la tasa de rendimiento (en %) de cierto activo es una variable aleatoria normal con desviación estándar 0.1. Los rendimientos observados en las dos últimas semanas fueron:

2.24, 2.01, 2.02, 1.83, 1.88, 1.99, 2.07, 2.12, 2.23, 2.02

(a) Hallar un intervalo de confianza para la media poblacional con un nivel de confianza de 95%.

(b) Manteniendo el mismo nivel de confianza, ¿cuál debería ser el tamaño de la muestra para disminuir el tamaño del intervalo calculado en (a) a la mitad?

9. Ejercicio para resolver con el ordenador.
Para estimar la velocidad media con la que los coches pasan por un punto de control de tráfico, se tomó una muestra aleatoria de 40 coches. Las velocidades registradas, en kilómetros por hora, fueron

49	83	58	65	68	60	76	86	74	53
71	74	65	72	64	42	62	62	58	82
78	64	55	87	56	50	71	58	57	75
58	86	64	56	45	73	54	86	70	73

Construye un intervalo de confianza 95% para la media poblacional de las velocidades de los coches cuando pasan por ese punto de control.

Contrastes de hipótesis sobre medias y proporciones. Diferencias de medias

1. La longitud media de los préstamos de la Biblioteca de la Universidad durante los cursos pasados fue de 16 días. Se toma una muestra de 100 libros este año y se obtiene una media de 14 días, con una desviación típica de 8 días. ¿Podemos concluir que se ha producido un cambio con relación al curso pasado? (NOTA: plantea la pregunta en términos de un contraste de hipótesis, especificando cuál es la hipótesis nula, cuál la alternativa, eligiendo un nivel de significación que te parezca adecuado, indicando cuál es la región de rechazo, si se rechaza o no H_0 con los datos del problema y cuál es el p-valor del contraste).

2. El restaurante francés *Chez Paquito* es uno de los más avanzados en cuanto a la tecnología que utiliza, no sólo en su cocina sino también en su gestión. Prueba de ello son los controles de calidad a que a menudo se somete. Por ejemplo, la última semana han realizado una prueba para determinar si las crepes que sirven tienen en media 30 cm. de diámetro, valor que el cheff considera ideal.

(a) En este caso, ¿cuál sería la hipótesis nula y cuál la alternativa?

(b) ¿La hipótesis alternativa es unilateral o bilateral?

(c) Explica cuáles serían los errores de tipo I y de tipo II en esta situación concreta.

(d) El gerente de *Chez Paquito* selecciona una muestra de 40 crepes de las preparadas en la cocina y obtiene que la media de los diámetros es de 30.5 cm. y que los datos presentan una desviación típica de 0.12 cm. Calcula el p-valor y comenta qué implicaciones tiene este valor en la calidad de la cocina de *Chez Paquito*.

3. Una compañía de seguros quiere lanzar un nuevo producto al mercado que consiste en un seguro de vida para personas con menos de 45 años que incluye una cláusula por la que el abonado dejará de pagar las primas cuando cumpla 75 años sin perder ningún derecho. Una de las razones por las que la empresa cree que este producto será rentable económicamente es porque se tiene la creencia en el sector de seguros de que la esperanza de vida es de 75 años. No obstante, como último paso antes de lanzar el nuevo producto la

empresa quiere asegurarse de ello y toma una muestra aleatoria de 50 historias de antiguos clientes de seguros de vida (ya fallecidos) para contrastar si su creencia está respaldada por los datos.

(a) ¿Cuáles son las hipótesis nula y alternativa?

(b) ¿Cómo se interpretan en este caso los errores de tipos I y II? ¿Cuál de los dos tipos de errores es más grave en este caso?

(c) En la muestra escogida, la media de las edades de fallecimiento fue de 76.8 años, con una desviación típica de 6.¿Cuál es el p-valor del contraste?

4. En las situaciones siguientes especifica en palabras cuál es la hipótesis nula y cuál la alternativa y cómo interpretas el error de tipo I y el error de tipo II.

(a) Una organización de consumidores decide probar una muestra de coches de una marca nueva en el mercado para contrastar si, tal como dice el fabricante, su consumo medio en autopista es de 8.3 litros por cada 100 km.

(b) El Departamento de Medio Ambiente de un gobierno analiza los efectos de un gas tóxico en el porcentaje de plantes de trigo muertas en una determinada zona. Los investigadores quieren saber si hay evidencia suficiente para rechazar que han muerto el 25%, como es habitual ante la presencia de este gas, o si esta vez ha sido superior el número de plantas muertas.

(c) Un sociólogo tiene la hipótesis de que, actualmente, hay un correlación negativa entre la edad y el porcentaje de votantes para cada grupo de edad.

5. Los siguientes datos son el resultado de medir el ritmo del pulso de cinco niños antes y después de ver un determinado programa de televisión.

Niño	Antes	Después
1	96	104
2	102	112
3	108	112
4	89	93
5	85	89

¿Se podría decir que este programa de televisión aumenta significativamente el ritmo del pulso de los niños?

6. Ejercicio para resolver con el ordenador.
Se pide a 12 estudiantes que prueben dos tipos de yogures, uno normal y otro tipo “mousse” y que los puntúen con una escala de 1 a 5, donde 1 significa *no me gusta nada* y 5 significa *me gusta muchísimo*. Los resultados son los siguientes.

Estudiante	yogur “mousse”	yogur normal
Montse	5	4
Nuria	4	4
Marta	3	1
Pau	5	5
Xavier	3	2
Pedro	2	1
Marcel	5	4
Antonio	3	1
Paco	3	5
Joan	3	2
Daniel	2	4
María	4	5

Suponiendo que las variables son normales queremos saber si hay una marcada preferencia a favor del yogur tipo mousse. ¿Qué se puede concluir para un nivel de significación del 5%?

7. Ejercicio para resolver con el ordenador.
Un educador cree que un nuevo método basado en una serie de actividades de lectura puede mejorar las habilidades lectoras de sus alumnos. Para comprobarlo, decide que la clase de primero A siga este método durante dos meses, mientras que la clase de primero B actuará como *grupo de control* (no recibirán ningún tratamiento especial). Al cabo de los dos meses todos los estudiantes pasan un test para medir la habilidad en la lectura. Los resultados son los siguientes:

Grupo de tratamiento (Primero A)
24 56 43 59 58 52 71 62 43 54 49 57 61 33 44 46 67 43 49 57 53
Grupo de control (Primero B)
42 46 43 10 55 17 26 60 62 53 37 42 33 37 41 42 19 55 54 28 20 48 85

(a) ¿Aportan los datos evidencia de una mejora?

(b) ¿Crees que el diseño de este experimento es el ideal? ¿Qué alternativa se te ocurre?

8. Ejercicio para resolver con el ordenador.
Los consumos (per-cápita) de vino en las provincias A y B son variables aleatorias independientes que pueden ser consideradas con distribución normal posiblemente con diferentes medias. Los consumos (en litros por día) de 20 individuos en la provincia A y de 25 individuos de la provincia B son los siguientes:

Provincia A

```
0.35  0.34  0.33  0.33  0.31  0.36  0.33  0.30  0.30  0.35
0.36  0.33  0.34  0.29  0.31  0.35  0.31  0.34  0.32  0.32
```

Provincia B

```
0.38  0.36  0.36  0.42  0.38  0.40  0.40  0.38  0.39
0.41  0.40  0.34  0.40  0.42  0.41  0.30  0.39  0.35
0.38  0.37  0.40  0.37  0.38  0.44  0.41
```

Analizar si los consumos medios pueden considerarse iguales. Contrastar estas hipótesis con un nivel de significación del 5% y suponiendo primero que las varianzas iguales en las dos poblaciones y, después, que las varianzas son distintas.

Técnicas de Decisión Avanzada I **Enero-Marzo, 2000**

Pedro Delicado

Lista de problemas 8

Contrastes de independencia en tablas de contingencia. Inferencia en el modelo de regresión lineal simple

1. Se lleva a cabo un muestreo entre los alumnos del segundo curso de una universidad de una licenciatura a la que se puede acceder sin haber cursado matemáticas en Bachillerato. Estos alumnos tuvieron una asignatura obligatoria de Matemáticas en el primer curso de Licenciatura. Se les pregunta si estudiaron Matemáticas en Bachillerato y cuál fue la calificación en la asignatura de Matemáticas en el primer curso. La siguiente tabla de contingencia corresponde al cruce de las variables MATEMÁTICAS BACHILLERATO y NOTA EN MATEMÁTICAS DE PRIMER CURSO.

	Susp.	Apr.	Not.	Exc.	M.H.
Mat.Bach.	6	23	20	6	2
No Mat.Bach.	8	6	4	0	0

(a) Calcula las frecuencias relativas conjuntas: f_{ij}

(b) Calcula las frecuencias marginales relativas: $f_{i\cdot}$, $f_{\cdot j}$

(c) Calcula las frecuencias absolutas esperadas: $E_{ij} = f_{i\cdot} f_{\cdot j} N$

(d) Calcula la contribución de cada celda de la tabla al estadístico χ^2:

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

(e) Calcula los residuos estandarizados de cada celda:

$$\frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$$

(f) Calcula el estadístico χ^2 para el contraste de independencia de las dos variables. Calcula los grados de libertad asociados a este estadístico.

(g) Realiza el contraste de independencia de las dos variables: especifica las hipótesis nula y alternativa, encuentra el p-valor del test e indica si se rechaza H_0 para $\alpha = 0.05$.

(h) ¿Cuáles son las celdas que más contribuyen al estadístico χ^2? ¿Podrías interpretar el resultado del contraste teniendo en cuenta las contribuciones de cada celda?

2. En una encuesta llevada a cabo entre estudiantes universitarios se preguntó a los encuestados sobre sus como lectores de periódicos y espectadores de televisión. Repite los apartados del ejercicio anterior para las variables LECTURA DE PERIÓDICOS (filas) y HORAS DE TV (columnas), cuya tabla de contingencia es la siguiente:

	$< \frac{1}{2}$ h.	1 h.	2 h.	> 3 h.
Todos los días	2	7	15	1
Muy de vez en cuando	6	16	19	2
Algunos días	2	2	2	1

3. Se ha realizado una encuesta entre militantes del Partido Radical (PR) para evaluar su opinión sobre la posible ruptura del pacto de gobierno del PR con el Partido Vanguardista (PV). La siguiente tabla de frecuencias absolutas conjuntas corresponde a las variables *AÑOS DE MILITANCIA* y *GRADO DE APOYO* a la continuidad pacto entre PR y PV.

```
    GRADO DE APOYO |
                   | MUY EN     EN        A       MUY A
                   | CONTRA   CONTRA    FAVOR     FAVOR     Row
AN~OS DE     Count |     1  |     2  |     3  |     4  | Total
MILITANCIA --------+--------+--------+--------+--------+------
                 1 |     3  |     3  |     5  |     6  |
  MENOS DE 1       |        |        |        |        |
                   |--------+--------+--------+--------+------
                 2 |     1  |     3  |     5  |     3  |
  ENTRE 1 Y 5      |        |        |        |        |
                   |--------+--------+--------+--------+------
                 3 |     4  |     2  |     3  |     5  |
  ENTRE 5 Y 10     |        |        |        |        |
                   |--------+--------+--------+--------+------
                 4 |     6  |     5  |     4  |     2  |
  ENTRE 10 Y 20    |        |        |        |        |
                   +--------+--------+--------+--------+------
     Column  Total |        |        |        |        | 60    100 %

      Chi-Square              Value        Significance
--------------------       -----------     ------------
Pearson                        9,5093          ,5801
```

1. Completa la tabla con las frecuencias marginales absolutas y relativas (expresadas estas últimas en porcentajes).

2. Calcula las frecuencias de la variable *GRADO DE APOYO* condicionadas a que el militante entrevistado lleva más de 10 años afiliado.

3. ¿Cuánto vale el estadístico χ^2? ¿Cuánto vale el p-valor asociado a ese estadístico?

4. Si las dos variables estudiadas fuesen independientes, ¿cuál sería el número esperado de encuestados con menos de un año de militancia que estarían muy en contra del pacto PR-PV?

5. Considera el contraste de hipótesis en el que la hipótesis nula es que las dos variables estudiadas en este problema son independientes. A la vista de la salida del ordenador, indica si tú rechazarías o no la hipótesis nula y di por qué.

4. Una empresa que se dedica a la administración de fincas urbanas ha hecho un estudio sobre los jóvenes de entre 25 y 30 años que alquilan pisos. Se seleccionaron aleatoriamente 60 personas asalariadas que vivían solas en una vivienda de alquiler. Se anotaron los datos de las variables *GASTO MENSUAL EN ALQUILER* y *SALARIO MENSUAL* (en miles de pesetas). Una vez procesados los datos con un programa informático se obtuvo la siguiente salida del programa. Se trata del análisis de la regresión para explicar la variable y =*SALARIO MENSUAL* en función de la variable x =*GASTO MENSUAL EN ALQUILER*.

1. Para cada una de las dos variables en estudio, escribe lo siguiente: media, desviación típica, varianza y error estándar de la media.

2. Da un intervalo de confianza al 95% para el valor de la media poblacional de la variable *ALQUILER MENSUAL*.

3. ¿Cuánto vale el coeficiente de correlación de las dos variables?

4. Escribe la recta de regresión estimada a partir de los datos. ¿Cuánto vale R^2 en esta regresión? ¿Y la desviación típica residual?

5. Se hace el contraste de hipótesis con un nivel de significación $\alpha = 0.05$ de la hipótesis nula $H_0 : \beta_1 = 1$ frente a la alternativa bilateral, donde β_1 es el coeficiente de la variable *ALQUILER* en la regresión. ¿Crees que se rechazará la hipótesis nula? Razona la respuesta.


```
                 Mean  Std Dev  Label

SALARIO   169,050    8,054  SALARIO MENSUAL
ALQUILER   59,967   10,596  GASTO MENSUAL EN ALQUILER
N of Cases =    60

Equation Number 1    Dependent Variable..  SALARIO

Multiple R           ,832
R Square             ,692
Standard Error      4,507

---------------------- Variables in the Equation ----------------------------

Variable             B         SE B     95% Confdnce Intrvl B         T  Sig T

ALQUILER         ,6323      ,057831      ,475680      ,788920    10,933 ,0000
(Constant)    131,1348    14,233626   102,668280   159,601320     9.213 ,0000
```

5. Ejercicio para resolver con el ordenador.
El partido político PX ha recogido los datos del número de diputados que obtuvo en doce provincias donde presentaba listas electorales. Son los siguientes.

Gasto electoral (mill. pts.)	19	9	15	11	12	6	6	17	12	9	9	17
Número de diputados	4	1	2	3	3	1	0	3	2	1	1	3

(a) Representa este conjunto de datos bivariante de la forma que consideres más conveniente.

(b) ¿Crees que hay relación entre la variable X =GASTO EN CAMPAÑA y la variable Y =NÚMERO DE DIPUTADOS? Si la respuesta es 'sí', indica de qué tipo es esta relación (lineal, no lineal, positiva, negativa, ...).

(c) La covarianza entre X e Y es $Cov(X,Y) = 4.17$ y la varianza de Y es $Var(Y) = 1.33$. Calcula el coeficiente de correlación entre X e Y.

(d) ¿Te parece que X e Y están muy correladas? ¿Crees que de ese valor de la correlación se puede concluir que hay una relación de *causa-efecto* entre el GASTO EN CAMPAÑA y el NÚMERO DE DIPUTADOS? ¿Crees que estamos ante un caso de correlación espuria y, en ese caso, qué variable (o variables) podrían estar afectando conjuntamente a X e Y?

```
* * * *   R E G R E S S I O N   * * * *

            Mean  Std Dev  Label

EDAD       36,582   12,778  EDAD
PROFESI     6,395    1,503  APOYO A LA PROFESIONALIZACION
N of Cases =    85

Equation Number 1    Dependent Variable..  PROFESI

Multiple R          -,864
R Square             ,747
Standard Error       ,761

---------------------- Variables in the Equation ----------------------------

Variable             B        SE B     95% Confdnce Intrvl B         T  Sig T
EDAD          -,101669     ,005736    -,112911     -,090426  -17,724  ,0000
(Constant)   10,114508     ,222126    9,679139    10,549875   45,535  ,0000
```

6. En un estudio de opinión sobre la conveniencia de la profesionalización de las fuerzas armadas, se ha registrado la variable $x = EDAD$ y, a partir de respuestas cualitativas, se ha construido un índice $y =$ *APOYO A LA PROFESIONALIZACIÓN* (con valores entre 0 y 10). Una vez procesados los datos con un programa informático se obtuvo la siguiente salida del programa. Se trata del análisis de la regresión para explicar la variable y en función de la variable x.

(a) Para cada una de las dos variables en estudio, escribe lo siguiente: media, desviación típica, varianza y error estándar de la media.

(b) Da un intervalo de confianza al 95% para el valor de la media poblacional de la variable *APOYO A LA PROFESIONALIZACION*.

(c) ¿Cuánto vale el coeficiente de correlación de las dos variables? Escribe la recta de regresión estimada a partir de los datos.

(d) ¿Cuánto vale R^2 en esta regresión? ¿Y la desviación típica residual? ¿Cómo interpretas ese valor de R^2?

(e) ¿Cuánto decrece el índice de apoyo a la profesionalización cuando la edad aumenta diez años?

7. Ejercicio para resolver con el ordenador.
Los datos siguientes relacionan la temperatura de ebullición del agua (en grados centígrados), con la presión barométrica (en mm de mercurio), y fueron tomados por el físico escocés Forbes el año 1857 en los Alpes y en Escocia.

Presión (mm)	768	769	770	773	774	775
Temperatura (^{o}C)	93.8	94.1	95.3	98.1	99.3	99.9

(a) Haz un gráfico de los datos, poniendo “Presión” en el eje vertical y “Temperatura” en el horizontal.

(b) Calcula el coeficiente de correlación entre las dos variables.

(c) ¿Es significativamente distinto de 0? (al nivel de $\alpha = 0.05$).

8. Ejercicio para resolver con el ordenador.
Los siguientes datos son las calificaciones medias obtenidas por un grupo estudiantes de un determinado instituto en los dos cursos de bachillerato (`Bach`) y en las pruebas de acceso a la universidad (`PAU`).

`Bach`	8.1	7.5	7.1	6.1	9.6	5.6	8.5	1.8	7.0	7.7	7.1	9.1	8.8	7.9	7.7
`PAU`	8.0	8.2	8.3	5.7	1.00	3.0	6.8	5.6	4.0	8.7	6.5	8.6	8.2	5.7	7.5

(a) Haz el diagrama de dispersión de los datos e identifica los casos atípicos (*outliers*).

(b) Calcula el coeficiente de correlación lineal entre unas notas y otras e interpreta su significado.

(c) Si tiene sentido, propón una recta que “a ojo” ajuste bien los datos.

(d) Determina la recta de regresión por el método de los mínimos cuadrados.

(e) Un estudiante del mismo instituto tiene una nota media de 8.0 en el Bachillerato. No se pudo presentar a las pruebas de acceso a la universidad. A la vista del comportamiento global de los estudiantes del instituto, ¿qué nota crees que habría sacado en esas pruebas de haberlas podido realizar? ¿Cómo interpretas esa hipotética calificación?