10 Distribuciones bidimensionales

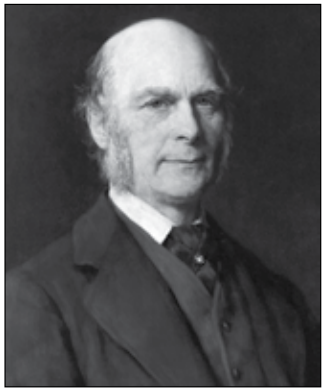
Qué es una distribución bidimensional

Un grupo de biólogos están estudiando una población de flamencos. Para ello, toman medidas de algunas de sus características anatómicas.

Si miden sus envergaduras (distancia entre los extremos de las dos alas extendidas), el conjunto de resultados es una distribución estadística de una variable (unidimensional). También es unidimensional la distribución de sus pesos.

Pero si atienden conjuntamente a ambas variables (envergadura y peso), se obtiene una **distribución bidimensional.** El grado de relación que existe entre ambas variables se llama **correlación.**

Galton (1822-1911)



Francis Galton (1822-1911).

Las nociones relativas a las distribuciones bidimensionales surgen de estudios realizados en Biología.

Aunque hay antecedentes interesantes, se puede considerar que el comienzo del estudio conjunto de dos variables se produce con **Francis Galton,** el cual, a instancias de su primo Charles Darwin, se interesó por la influencia que algunas características de los padres pudieran tener sobre las de los hijos.

No consideró que fuera factible experimentar con personas y carecía de datos suficientes para extraer conclusiones relativas a ellas. Por eso recurrió a la experimentación con guisantes (¡como **Mendel**!). En sus conclusiones acuñó el término **regresión.** El índice de correlación le sirvió para describir similitudes debidas al parentesco.

Fue consciente de que sus descubrimientos podían aplicarse a un amplio campo de problemas relativos a distintas ciencias.

Karl Pearson (1857-1936).

Pearson (1857-1936)

El continuador del trabajo de Galton fue **Karl Pearson,** quien por primera vez considera y describe el significado del coeficiente de correlación negativo. Diseñó y puso en práctica métodos matemáticos rigurosos con los que se pudo utilizar la correlación para inferir valores de una variable a partir de los de la otra. También extendió el estudio de la correlación a más de dos variables.

Nombre y apellidos: ..... Fecha: .....

1 Dos variables relacionadas. Correlación

ESTUDIANTES	NOTA EN M	NOTA EN F
a	7	6
b	6	4
c	8	7
d	3	4
e	6	5
f	9	6
g	4	2
h	10	9
i	2	1
j	5	6

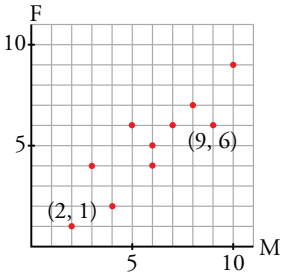
A la izquierda tienes las notas de diez estudiantes (*a*, *b*, *c*…) de una clase en dos asignaturas: matemáticas (M) y física (F).

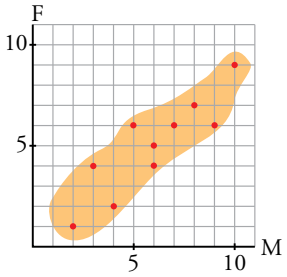
Hay dos variables. Cada individuo tiene, por tanto, dos valores asociados: su nota en M y su nota en F. Por eso se trata de una *distribución bidimensional*.

Nube de puntos

Si representamos a cada estudiante mediante un punto cuyas coordenadas son sus respectivas notas en M y en F, obtendremos la gráfica adjunta, llamada *nube de puntos* o *diagrama de dispersión*.

El punto (9, 6), por ejemplo, corresponde al estudiante *f*, y el (2, 1), al *i*.

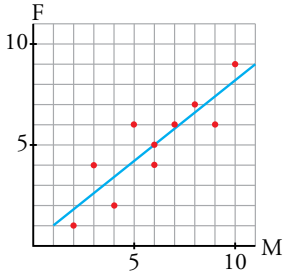




Correlación

Observando las notas de estos estudiantes, apreciamos una clara relación entre ellas: a notas bajas en una asignatura le corresponden, casi siempre, notas bajas en la otra; y otro tanto ocurre con las notas medias o altas.

Como consecuencia de esto, los puntos de la nube están en una franja estrecha. Diremos que hay *correlación* entre las dos variables.



Recta de regresión

Podemos trazar, a ojo, una recta que se amolde a la nube de puntos, como la que aparece en la gráfica de la izquierda.

Se llama *recta de regresión* y marca la tendencia de la nube.

En resumen:

- Si a cada uno de los individuos de un colectivo le asignamos dos valores, correspondientes a dos variables x e y, tenemos una **distribución bidimensional.** La representación gráfica de la distribución da lugar a un conjunto de puntos llamado **nube de puntos** o **diagrama de dispersión.**
- Cuando existe una cierta relación estadística entre los valores de la distribución, se dice que hay **correlación** entre las variables. Esta correlación se aprecia porque la nube de puntos es relativamente estrecha y, en tal caso, se puede trazar una recta que se amolda a ella. Se llama **recta de regresión.**

Piensa y practica

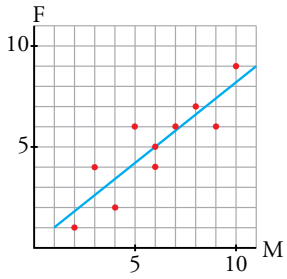
1. Identifica los restantes puntos del diagrama de dispersión del ejemplo de las notas en matemáticas y en física.

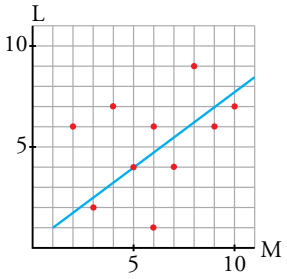
Nombre y apellidos: ..... Fecha: .....

La correlación puede ser más o menos fuerte

ESTUDIANTES	NOTA EN M	NOTA EN L
a	7	4
b	6	6
c	8	9
d	3	2
e	6	1
f	9	6
g	4	7
h	10	7
i	2	6
j	5	4

Veamos otra distribución bidimensional: las notas de los mismos diez estudiantes (*a*, *b*, *c*…) en matemáticas, M, y en la asignatura de lengua, L. Y vamos a comparar la nube de puntos de esta distribución bidimensional M-L con la que hemos visto en la página anterior, M-F.

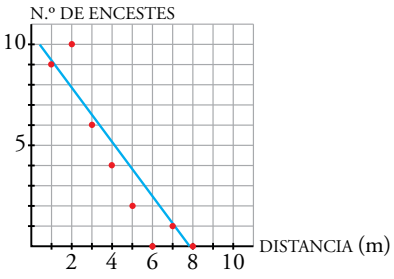




Es evidente que la correlación entre M y F es más fuerte que la correlación entre M y L, pues en la primera, los puntos están más apretados en torno a la recta de regresión que en la segunda.

La correlación entre dos variables puede ser más o menos fuerte según que los puntos de la nube estén más o menos próximos a la recta de regresión.

La correlación admite signo

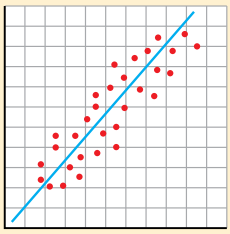


Una jugadora de baloncesto hace 10 lanzamientos a canasta desde una distancia de 1 m, otros 10 desde 2 m, y así sucesivamente hasta 8 m. En cada caso ha tomado nota del número de encestes. Al observar, en el margen, la nube de puntos, vemos que la correlación es fuerte pero negativa, pues a más distancia, menos encestes.

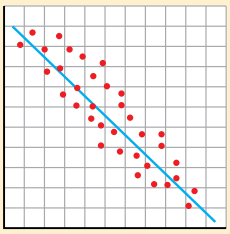
DISTANCIA (en m)	N.º DE ENCESTES
1	9
2	10
3	6
4	4
5	2
6	0
7	1
8	0

Una **correlación** es **positiva** cuando al aumentar una variable, *x*, tiende a aumentar la otra variable, *y*.

Una **correlación** es **negativa** cuando al aumentar *x*, tiende a disminuir *y*.



CORRELACIÓN POSITIVA



CORRELACIÓN NEGATIVA

El signo de la correlación coincide con el signo de la pendiente de la recta de regresión.

En la web

Diagramas de dispersión con diferentes grados de correlación.

Nombre y apellidos: ..... Fecha: .....

Ejercicio resuelto

Esta es la tabla de los 15 primeros clasificados en una liga de fútbol:

POS	J	G	E	P	F	C	PTOS
1. B. Senior	34	22	5	7	59	37	71
2. Dinamo	34	20	9	5	53	30	69
3. Kurgans	34	20	8	6	62	28	68
4. Colme	34	17	8	9	53	47	59
5. Roco	34	17	4	13	45	40	55
6. Malas	34	14	9	11	43	35	51
7. Leones	34	12	11	11	50	41	47
8. Rayos	34	13	8	13	42	43	47
9. Culebras	34	11	11	12	41	35	44
10. Mursi	34	12	8	14	45	47	44
11. Ramones	34	10	13	11	42	44	43
12. Mates	34	10	12	12	42	38	42
13. Potenkin	34	11	8	15	34	47	41
14. Azurro	34	11	7	16	44	60	40
15. Tigres	34	11	6	17	30	44	39

Las distintas columnas significan:

Pos. ***Posición final: 1.º, 2.º, 3.º…***

J. ***Partidos jugados***

G. ***Partidos ganados***

E. ***Partidos empatados***

P. ***Partidos perdidos***

F. ***Total de goles marcados (a favor)***

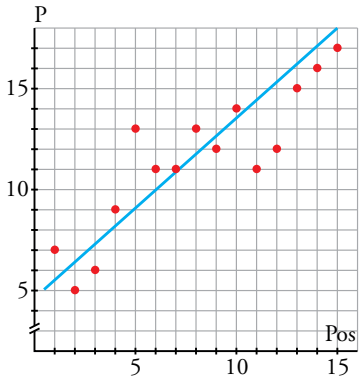
C. ***Total de goles recibidos (en contra)***

Ptos. ***Puntos***

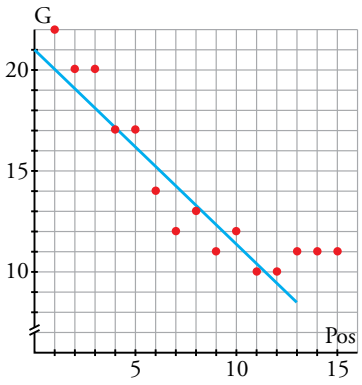
Analizar las siguientes distribuciones bidimensionales, representarlas y cuando la correlación sea fuerte, trazar la recta de regresión:

a) Pos-P b) Pos-G c) Pos-E

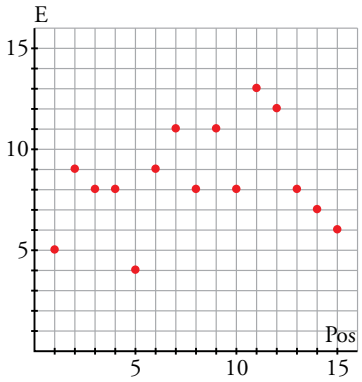
a) Si relacionamos la posición (Pos) con el número de partidos perdidos (P), es lógico que haya correlación positiva (cuanto más arriba en la tabla, 1.º, 2.º, 3.º…, menos partidos se han perdido). Al representarla, apreciamos en la nube de puntos una **correlación positiva fuerte.**



b) Al relacionar la posición (Pos) con los partidos ganados (G), se encuentra una **correlación negativa** (cuanto más abajo está en la tabla un equipo, …13.º, 14.º, 15.º, menos partidos ha ganado).



c) Entre la posición y los partidos empatados **no se aprecia correlación.** Lo mismo empatan pocos o muchos partidos los de arriba, los de abajo o los de en medio.



Piensa y practica

2. En cada una de las siguientes distribuciones bidimensionales, intenta, sin representarla, estimar si la correlación va a ser positiva o negativa, fuerte o débil. Luego, represéntala mediante la nube de puntos, trazando la recta de regresión, y corrobora o modifica tus estimaciones.

a) G - F b) Pos - F

c) F - C d) Pos - Ptos

3. Busca, en un periódico o en Internet, una tabla como la anterior, de actualidad, y estudia distribuciones como las que hemos visto en esta página.

Nombre y apellidos: ..... Fecha: .....

2 El valor de la correlación

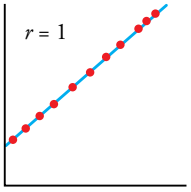
Al igual que para la media $(\bar{x})$ o para la desviación típica (σ), también hay una fórmula para hallar el valor de la correlación de una distribución bidimensional a partir de los datos de la tabla. En este curso no la vamos a hallar mediante la fórmula ni vamos a aprender a obtenerla con calculadora, pero sí vamos a familiarizarnos con la gama de valores que puede tomar.

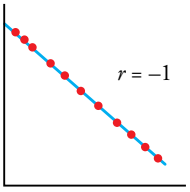
En la web

Ampliación teórica: explicación y cálculo del coeficiente de correlación.

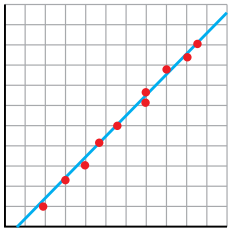
El valor de la correlación se denomina **coeficiente de correlación** y se designa con la letra r.

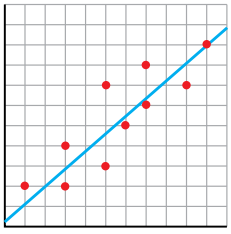
El mayor valor de r se da cuando los puntos están alineados (relación funcional). En tal caso, el valor de r es 1 o –1, según sea positiva o negativa. Por tanto, los valores que puede tomar r oscilan entre –1 y 1.

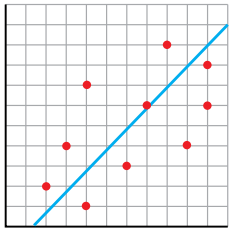


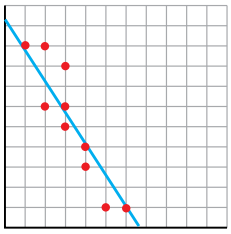


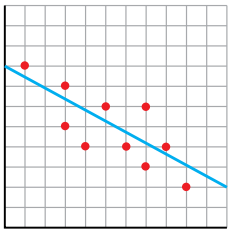
Observa las siguientes nubes de puntos y en cada una de ellas fíjate en la relación entre el valor de r y lo "apretados" o "separados" que se encuentran los puntos respecto a su recta de regresión.

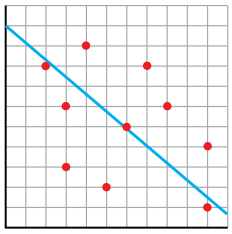

$r = 0{,}99$


$r = 0{,}86$


$r = 0{,}59$


$r = -0{,}92$


$r = -0{,}77$


$r = -0{,}41$

Correlación y pendiente

¡Atención! En una nube de puntos, el valor de la pendiente de la recta de regresión (1, 1/3, –1/2, –2) no tiene nada que ver con el valor de la correlación. Solo nos fijamos en el signo de la pendiente (positivo o negativo).

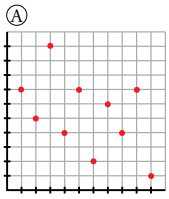
Piensa y practica

1. Los siguientes números son los valores absolutos de los coeficientes de correlación, r, de las distribuciones bidimensionales representadas a la derecha:

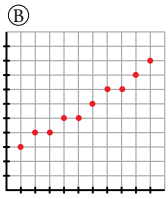
0,75 0,47 0,92 0,97

Asigna cada cual a la suya, cambiando el signo cuando convenga.

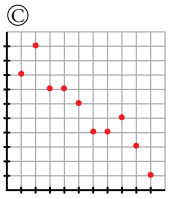
Ⓐ



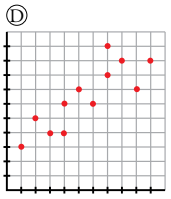
Ⓑ



Ⓒ



Ⓓ



104

Nombre y apellidos: ..... Fecha: .....

Ejercicio resuelto

Esta tabla muestra algunas características socioeconómicas de 10 países:

Pa	Po	Ex	RPC	IP	Me
A	316	9,8	56	15,1	5,6
B	202	8,5	12	21,4	1,9
C	127	0,38	38	16	2,3
D	121	2	16	13	2,1
E	92	0,33	44	11,3	4,9
F	64	0,64	43	7,9	4,2
G	61	0,3	35	6	3,8
H	38	0,3	14,4	10,6	2,2
I	14	1,3	2	48	0,04
J	4	1	2,2	40	0,13

Las columnas significan:

Pa. ***País***

Po. ***Población (en millones de personas)***

Ex. ***Extensión (en millones de km²)***

RPC. ***Renta per cápita (en miles de $)***

IP. ***Índice de pobreza (en %)***

Me. ***N.º de médicos por 1 000 habitantes***

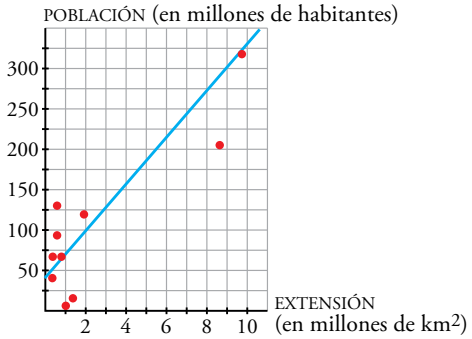
Analizar las distribuciones siguientes:

- **Ex - Po**
- **RCP - IP**
- **RCP - ME**

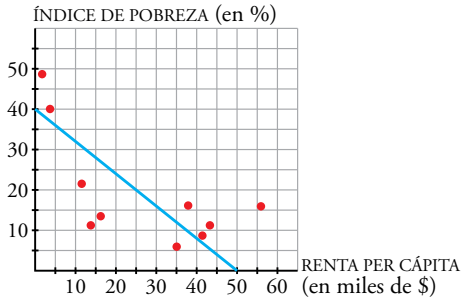
Sabiendo que las correlaciones son, no respectivamente, –0,68; 0,94 y 0,87, estimar cuál corresponde a cada una de ellas.

Es claro que a países más grandes suelen corresponderles poblaciones más numerosas, es decir, la correlación es positiva.

Si Australia estuviera en esta lista, se alejaría mucho de la recta de regresión, ya que es un país con una densidad de población anormalmente baja.



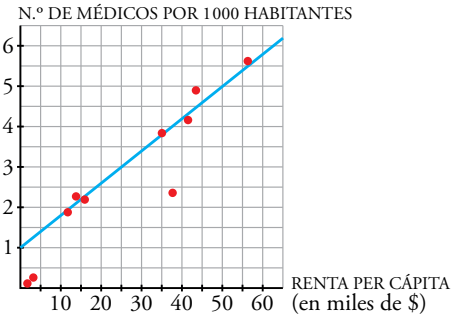
Es razonable que cuanta más renta per cápita tenga un país, menor sea el índice de pobreza, es decir, la correlación es negativa. Sin embargo, hay muchas excepciones, por lo que la correlación no es muy alta en valor absoluto.



La correlación entre la renta per cápita y el número de médicos por cada 1 000 habitantes es positiva y mucho más fuerte.

En este caso se ve más claramente que países mucho más ricos tienen más médicos por cada 1 000 habitantes.

Si Cuba estuviera en la lista se separaría mucho de la recta de regresión y bajaría la correlación, ya que, siendo un país económicamente débil, tiene una proporción muy alta de médicos.



A la vista de las tres gráficas, la correlación correspondiente a la primera (Ex - Po) es 0,87; a la segunda (RPC - IP), –0,68; y a la tercera (RPC - Me), 0,94.

Piensa y practica

2. Representa la nube de puntos y la recta de regresión de la distribución bidimensional IP - Me del ejercicio resuelto anterior.

3. Indica cuál de estos valores se ajusta mejor al valor de la correlación de la distribución del ejercicio 2.

0,5 –0,99 0,82 –0,77 0,99

Nombre y apellidos: ..... Fecha: .....

3 La recta de regresión para hacer estimaciones

En la web

Ampliación teórica: explicación y cálculo de la recta de regresión.

¿Sirve la recta de regresión para estimar el valor de y que le corresponde a un nuevo individuo del que se conoce el valor de x? Por supuesto que podemos hacer la estimación, pero, ¿qué grado de fiabilidad tendremos en ella?

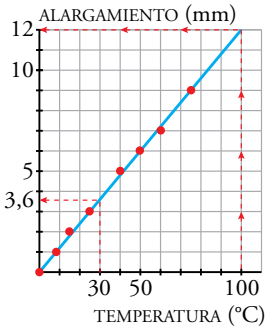
Parece razonable pensar que cuanto más fuerte sea la correlación, más fiable será la estimación, pero, ¿influyen otros factores? Antes de extraer conclusiones definitivas, veamos unos ejemplos.

Ejemplo 1

T	0	8	15	25	40	50	60	75
A	0	1	2	3	5	6	7	9

La longitud de un raíl de vía de tren a 0 °C es de 10 m. La tabla del margen nos muestra los alargamientos, A (en mm), a distintas temperaturas, T (en °C).

A partir de los datos de la tabla, nos preguntamos por el alargamiento que se obtendría para temperaturas de 30 °C y 100 °C.



Lo primero que hacemos es representar los datos en una nube de puntos. Observamos en el margen que se ajustan casi exactamente a una recta, recta de regresión trazada en azul. Por lo que damos por cierto que el coeficiente de correlación es muy próximo a 1.

Obtenemos la ecuación de la recta de regresión. Como pasa por (0, 0) y (50, 6), su ecuación es $y = \frac{6}{50}x \rightarrow y = 0{,}12x$.

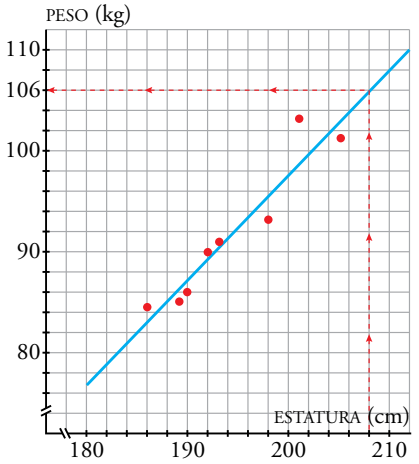
Para 30 °C obtenemos $\hat{y}(30) = 3{,}6$ mm, y para 100 °C, $\hat{y}(100) = 12$ mm. Ambas estimaciones pueden ser muy fiables, sobre todo la primera, ya que el valor de la temperatura está en el tramo de los valores controlados. En la segunda estimación la temperatura está fuera del intervalo de valores, pero poco alejada.

Ejemplo 2

E	P
186	85
189	85
190	86
192	90
193	91
198	93
201	102
205	101

Las estaturas, E (en cm), y los pesos, P (en kg), de 8 jugadores de baloncesto vienen dados en la tabla del margen.

Queremos estimar, mediante la recta de regresión, el peso de un nuevo fichaje cuya altura es de 208 cm. Para ello, representamos los datos y la recta de regresión y hallamos gráficamente el peso que corresponde a 208 cm:



$$\hat{y}(208) = 106$$

En este caso, la correlación no es tan alta como en el anterior, por ello, sería más prudente que dijéramos que el peso que corresponde a 208 cm es relativamente próximo a 106; por ejemplo, entre 102 kg y 110 kg.

Piensa y practica

1. Estima, con los datos del ejemplo 1, el alargamiento correspondiente a una temperatura de 45 ºC. ¿Consideras fiable la estimación?

2. Estima, con los datos del ejemplo 2, el peso de un nuevo jugador cuya estatura sea de 180 cm. ¿Consideras fiable la estimación?

Nombre y apellidos: ..... Fecha: .....

Ejercicios y problemas

Practica

1. Para cada uno de los siguientes casos:

— Di si se trata de una distribución bidimensional.

— Indica cuáles son las variables que se relacionan.

— Indica si se trata de una relación funcional o de una relación estadística.

a) Tamaño de la vivienda - Gasto en calefacción.

b) Número de personas que viven en una casa - Litros de agua consumidos en un mes.

c) Metros cúbicos de gas consumidos en una casa - Coste del recibo del gas.

d) Longitud de un palmo en un alumno - Número de calzado que usa.

e) Número de médicos por cada mil habitantes - Índice de mortalidad infantil.

f) Velocidad con que se lanza una pelota hacia arriba - Altura que alcanza.

2. En cada uno de los apartados del ejercicio anterior, estima si la correlación será positiva o negativa, fuerte o débil.

3. Estos son los resultados que hemos obtenido al tallar y pesar a varias personas:

ESTATURA (cm)	156	163	171	177	184
PESO (kg)	48	75	65	73	81

a) ¿Es una distribución bidimensional? ¿Cuáles son las variables que se relacionan? ¿Cuáles son los individuos?

b) Representa la nube de puntos.

c) ¿Es una relación estadística o funcional?

4. Las estaturas de 10 chicas (x_i) y las de sus respectivas madres (y_i) son:

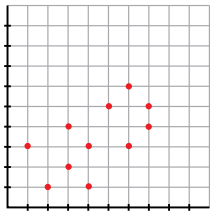
x_i	158	162	164	165	168	169	172	172	174	178
y_i	163	155	160	161	164	158	175	169	166	172

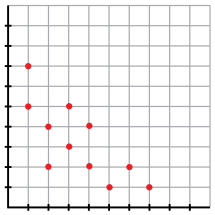
Representa los valores sobre papel cuadriculado mediante una nube de puntos, traza a ojo la recta de regresión y di si la correlación es positiva o negativa y más o menos fuerte de lo que esperabas.

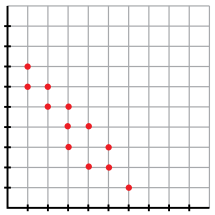
5. Representa la nube de puntos de la siguiente distribución y estima cuál de estos tres puede ser su coeficiente de correlación: $r = 0{,}98$; $r = -0{,}51$; $r = 0{,}57$.

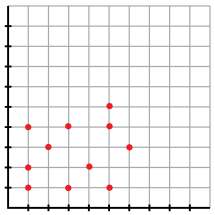
x	0	1	2	3	3	4	5	6	7	8	9
y	1	4	6	2	4	8	6	5	3	6	9

6. Los números 0,2; –0,9; –0,7 y 0,6 corresponden a los coeficientes de correlación de las siguientes distribuciones bidimensionales. Asigna a cada gráfica el suyo:

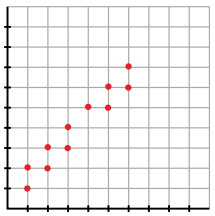
a) 

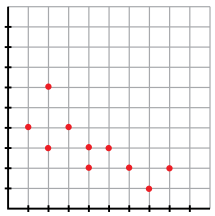
b) 

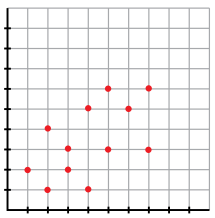
c) 

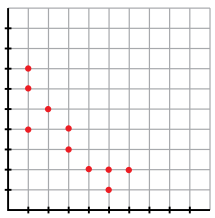
d) 

7. Los coeficientes de correlación de estas distribuciones bidimensionales son, en valor absoluto: 0,55; 0,75; 0,87 y 0,96. Asigna a cada una el suyo, cambiando el signo cuando proceda:

a) 

b) 

c) 

d) 

8. Traza la recta de regresión de las distribuciones a) y c) del ejercicio anterior y estima, en cada una de ellas, los valores que corresponden a $x = 0$ y a $x = 10$. ¿En cuál son más fiables las estimaciones?

Nombre y apellidos: ..... Fecha: .....

Ejercicios y problemas

Resuelve problemas

9. Se ha hecho un estudio con ratones para ver los aumentos de peso (en g) mensuales que producen ciertas sustancias A, B y C (en mg diarios):

SUSTANCIA	AUMENTO DE PESO SI LA SUSTANCIA ES A	AUMENTO DE PESO SI LA SUSTANCIA ES B	AUMENTO DE PESO SI LA SUSTANCIA ES C
1	3	2	3
2	1	2	3
3	3	1	2
4	5	3	0
5	6	0	1
6	4	3	–1
7	6	4	1
8	5	1	–2
9	7	3	–4
10	7	1	–2

Los resultados negativos quieren decir que en lugar de aumentar, el peso disminuye.

a) Representa la nube de puntos de cada distribución.

b) Indica si cada correlación es positiva o negativa.

c) Ordena las correlaciones de menos a más fuerte.

10. La correlación entre las temperaturas medias mensuales de una ciudad española y el tiempo que sus habitantes dedican a ver la televisión es de –0,89. ¿Te parece razonable este valor? Explica su significado.

11. Observa la siguiente tabla (2015):

	ESPERANZA DE VIDA AL NACER (en años)	MORTALIDAD INFANTIL POR 1000
ARGENTINA	76	13
BOLIVIA	67	43
BRASIL	72	23
COLOMBIA	74	19
CHILE	77	7
ECUADOR	73	20
PARAGUAY	73	31
PERÚ	72	19
URUGUAY	76	13
VENEZUELA	75	16

a) Representa la nube de puntos y di si la correlación que observas es positiva o negativa, fuerte o débil.

b) ¿Cuál de los siguientes valores será el coeficiente de correlación? –0,99; 0,5; 0,94; –0,92; 0,8

12. Se ha medido cada uno de los días de una semana la temperatura máxima, T, y el número de horas de sol, S, obteniéndose los siguientes resultados:

S	7	10	0	6	11	12	11
T	12	14	7	10	15	20	18

a) Traza a ojo la recta de regresión T-S.

b) Si el lunes siguiente a la medición hubo 9 horas de sol, ¿qué temperatura máxima cabe esperar que hiciera? ¿Qué fiabilidad tiene tu predicción?

Autoevaluación

1. Di en qué casos la correlación es positiva, en cuáles es negativa y en cuáles no ves correlación:

a) Altura de una persona - Tamaño de su perro.

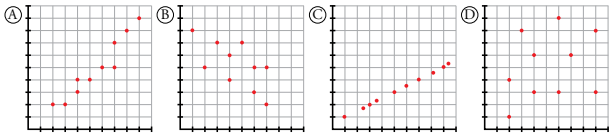
b) Distancia de un viaje de avión - Precio del billete.

c) Latitud de un lugar del hemisferio norte - Temperaturas medias anuales.

d) Altura - Presión atmosférica.

2. Asocia cada nube de puntos con una correlación:

$r = 1$ $r = -0{,}83$ $r = 0{,}97$ $r = 0{,}18$



3. Se han anotado a final de curso las notas de inglés y de francés de 10 estudiantes de ESO:

I	6	3	5	6	5	8	10	4	9	7
F	6	4	6	5	7	7	9	5	10	7

a) Representa los datos en una nube de puntos. Traza a ojo su correspondiente recta de regresión.

b) ¿Qué coeficiente de correlación le corresponde?

$r = 0{,}99$ $r = -0{,}86$ $r = 0{,}88$ $r = 0{,}63$

4. Sabiendo que la recta de regresión correspondiente a la actividad anterior tiene como ecuación $y = 1 + 0{,}85x$, estima qué nota obtendrán en francés 3 nuevos estudiantes cuyas notas en inglés fueron 1; 6,5 y 9,5.

Nombre y apellidos: ..... Fecha: .....