UNIDAD

11 Distribuciones estadísticas dobles

En muchos campos del conocimiento surge la necesidad de establecer relaciones entre dos conjuntos de datos, o dos variables estadísticas, aun sabiendo que tal relación no puede ser funcional, es decir, que no existe una fórmula que permita obtener los datos de uno de los conjuntos, o de una de las variables, a partir de los del otro, o de la otra variable.

Hay dos problemas fundamentales en el estudio de las relaciones entre dos variables estadísticas. El primero consiste en considerar una de las variables, la mejor conocida, como variable independiente y encontrar una función, en nuestro caso sólo hablaremos de la función lineal, que ilustre de modo aproximado la relación entre las dos variables y permita hacer predicciones para algunos datos desconocidos. A este problema se le conoce como Análisis de la Regresión o simplemente ajuste de los datos por la recta de regresión. El segundo de los problemas conduce al cálculo del coeficiente de correlación lineal que mide el grado de interdependencia lineal entre dos variables estadísticas, cuando los datos de ambas tienen la misma fiabilidad y no tiene mucho sentido tomar una de las variables como variable independiente.

El propósito de esta Unidad es, en primer lugar, encontrar la recta de regresión entre dos variables estadísticas y continuación, mediante el empleo de coeficiente de correlación, averiguar si el grado de relación entre las variables es lo suficientemente grande como para que la recta de regresión tenga alguna utilidad.

Los **objetivos** que nos proponemos alcanzar con el estudio de esta Unidad son los siguientes:

1. Identificar las variables estadísticas dobles como el estudio de dos características en cada individuo de una población.
2. Representar las variables estadísticas dobles por una nube de puntos.
3. Calcular la función lineal que mejor se aproxime a los puntos de una nube.
4. Analizar el grado de relación entre dos variables empleando el coeficiente de correlación.

ÍNDICE DE CONTENIDOS

1. Variables estadísticas dobles

En una población estudiaremos dos variables estadísticas: una variable que denominamos X y otra que denominamos Y, de modo que cada individuo de la población estará determinado por un par de datos (x_i, y_j), en el que x_i representa los valores o marcas de clase de la variable X e y_j representa los valores o marcas de clase de la variable Y.

Al estudio conjunto de dos características o variables estadísticas unidimensionales X e Y sobre una misma población se acostumbra a llamarlo **variable estadística bidimensional**.

Por ejemplo, en una evaluación de 30 alumnos se ha registrado el número de suspensos y el número de horas diarias que dedica cada uno al estudio, obteniéndose los siguientes resultados:

(0, 2) (2, 2) (5, 0) (2, 1) (1, 2) (1, 3) (0, 4) (4, 0) (2, 2) (2, 1) (1, 2) (0, 4) (1, 3) (4, 2) (1, 2) (2, 1) (1, 2) (0, 2) (0, 3) (2, 3) (2, 2) (2, 2) (1, 2) (6, 0) (3, 1) (2, 2) (1, 2) (3, 1) (4, 1) (1, 2)

Estamos ante dos variables. La variable X, la más fiable, cuenta el número de suspensos y sirve para explicar la variable Y, las horas diarias de estudio. El par (x_i, y_j) registra el número de suspensos, x_i , y el número de horas de estudio, y_i .

Los datos de una variable estadística bidimensional se distribuyen en tablas de frecuencias de doble entrada, así:

X \ Y	0	1	2	3	4	Totales
0	0	0	2	1	2	5
1	0	0	7	2	0	9
2	0	3	5	1	0	9
3	0	2	0	0	0	2
4	1	1	1	0	0	3
5	1	0	0	0	0	1
6	1	0	0	0	0	1
Totales	3	6	15	4	2	30

En la primera columna de la tabla hemos puesto los valores de la variable X y en la primera fila los valores de la variable Y, y en cada casilla figura la frecuencia absoluta f_{ij} del par (x_i, y_j). La última fila y la última columna presentan las llamadas **distribuciones marginales**. En la última fila figuran las frecuencias de la variable Y y en la última columna las frecuencias de la variable X. Las distribuciones de frecuencias bidimensionales se reflejan en tablas de doble entrada, que en el caso general sería así:

X \ Y	y_1	y_2	...	y_m	FRECUENCIAS VARIABLE X
x_1	f_{11}	f_{12}	...	f_{1m}	Σf_{1i}
x_2	f_{21}	f_{22}	...	f_{2m}	Σf_{2i}
...	...	...	...	...	...
x_n	f_{n1}	f_{n2}	..	f_{nm}	Σf_{ni}
FRECUENCIAS VARIABLE Y	Σf_{j1}	Σf_{j2}	...	Σf_{jm}	N

Sin embargo, cuando el número de datos u observaciones es pequeño, en vez de tablas de doble entrada, emplearemos tablas simples de dos filas, de modo que en cada columna figuren los valores, (x_i, y_j), correspondientes a las dos variables. En lo sucesivo sólo emplearemos tablas de dos filas (o de dos columnas, si las tablas las ponemos de pie).

Por ejemplo, las calificaciones de 12 alumnos en Matemáticas y Lengua son las siguientes:

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

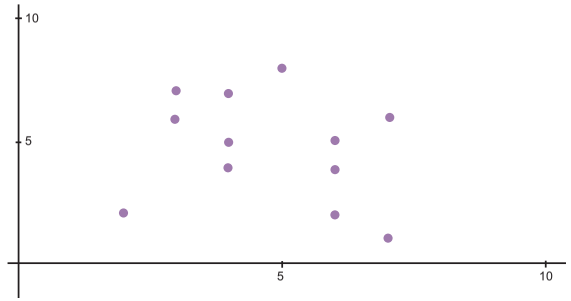
Estos datos se disponen en una tabla simple de dos filas así:

Matemáticas	2	4	4	6	4	6	3	6	5	7	3	7
Lengua	2	7	4	2	5	5	6	4	8	1	7	6

2. Diagrama de dispersión o nube de puntos

Cuando las variables X e Y de una distribución bidimensional son cuantitativas podemos representar los datos por puntos sobre unos ejes de coordenadas. En el eje de abscisas llevamos los valores de la variable X, que hemos considerado como variable independiente, y sobre el eje de ordenadas llevamos los valores de la variable Y, que hemos considerado como dependiente. Debe quedar claro que las dos variables no juegan el mismo papel, la que hemos denominado independiente es la que permite explicar el comportamiento de la otra, la denominada variable Y.

En el caso de las notas de Matemáticas y Lengua de 12 alumnos, del apartado anterior, si llevamos las calificaciones de Matemáticas sobre el eje de abscisas y las de Lengua sobre el eje de ordenadas obtenemos el siguiente gráfico:



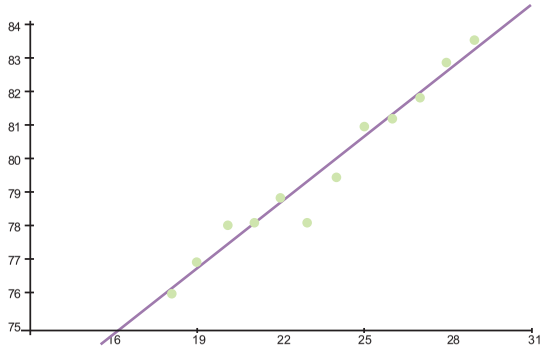
La representación gráfica de una distribución bidimensional se denomina **diagrama de dispersión o nube de puntos**. Cada punto tiene por coordenadas los valores que en cada individuo tienen las variables X e Y. La nube de puntos nos permite apreciar si existe una posible relación entre las variables.

En el diagrama anterior no parece que exista ninguna relación entre las dos variables, pero esto no siempre es así. Veamos otros ejemplos.

Un pediatra ha anotado las edades, en meses, y la altura en cm de 12 niños obteniendo los siguientes resultados:

meses	18	19	20	21	22	23	24	25	26	27	28	29
altura	76,1	77	78,1	78,2	78,8	78,2	79,5	81	81,2	81,8	82,8	83,5

La nube de puntos de esta distribución sería:

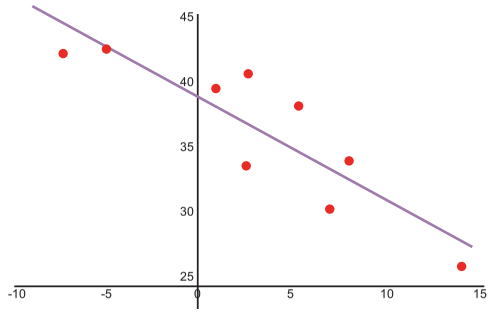


Hemos dibujado una recta entre los puntos porque todo sugiere que la relación entre las variables edades y alturas se aproxima a una relación lineal.

La tabla siguiente indica la media de las temperaturas mínimas en el mes de enero y las latitudes de algunas ciudades de Estados Unidos

	Temperatura	Latitud
Los Angeles	8.3	34.3
San Francisco	5.5	38.4
Washington	1	39.7
Miami	14.4	26.3
Atlanta	2.7	33.9
Chicago	-7.2	42.3
Nueva Orleans	7.2	30.8
Nueva York	2.7	40.8
Boston	-5	42.7

La nube de puntos correspondiente a esta distribución es la siguiente:



También hemos dibujado una recta que sugiere la existencia de una relación lineal, aunque no tan fuerte como en el caso anterior.

La nube de puntos permite apreciar si hay o no una relación entre las dos variables. El problema que se nos plantea ahora es el siguiente: **si la nube de puntos sugiere una relación lineal entre las variables, ¿cómo podemos encontrar la recta que mejor se ajusta a la nube de puntos?** Porque, evidentemente, podemos trazar varias rectas que pasen a través de los puntos del diagrama de dispersión. La respuesta a esta pregunta la veremos en apartado siguiente.

Actividades

1. Los pesos y las alturas de los jugadores de un equipo de fútbol están dados por la siguiente tabla:

***X*(peso kg)**	80	80	77	68	85	80	74	79	76	73	78
***Y*(altura cm)**	187	185	184	173	189	183	177	189	180	176	182

Dibuja el diagrama de dispersión.

2. El tiempo que tarda la sangre humana en coagular, según la temperatura, es la que figura en la tabla siguiente:

Temperatura en ºC	5	10	15	20	25	35	40	45
Tiempo segundos	45	38	32	28	24	19	22	21

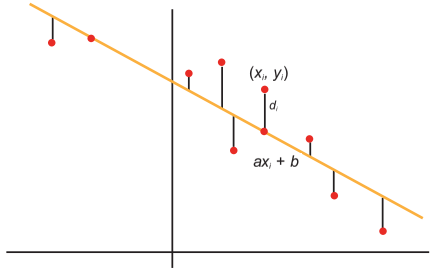
Dibuja el diagrama de dispersión.

3. Ajuste de la nube de puntos por una recta. Recta de regresión

Pretendemos encontrar una recta $y = ax + b$ que esté lo más próxima posible a los puntos de la nube. Podíamos hallar la pendiente, a, y la ordenada en el origen, b, de modo que la suma de las distancias de los puntos a la recta sea mínima, pero eso nos obligaría a emplear la función valor absoluto y es un poco incómodo.

Determinaremos a y b imponiendo como condición que la suma de los cuadrados de las distancias de los puntos a la recta sea mínima:

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2$$



Para hallar el mínimo de esta función hay que derivar e igualar la derivada a cero. Lamentablemente esta función tiene dos variables: *a* y *b*, y eso obliga a un método de derivación llamado derivación parcial, que no está entre los objetivos de este libro. En cualquier caso, se trata de derivar primero como si la incógnita fuese *a* e igualar a cero, y a continuación hacer lo mismo suponiendo que la incógnita fuese *b*, con lo que se obtiene un sistema de dos ecuaciones con dos incógnitas:

$$\sum y_i - a\sum x_i - nb = 0$$

$$\sum x_i y_i - a\sum x_i^2 - b\sum x_i = 0$$

o, dejando las incógnitas solas en el primer miembro,

$$a\sum x_i + nb = \sum y_i$$

$$a\sum x_i^2 + b\sum x_i = \sum x_i y_i$$

Las soluciones del sistema vienen dadas por:

$$a = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \quad \text{y} \quad b = \frac{\sum y_i - a\sum x_i}{n}$$

Es difícil memorizar estas fórmulas, pero si hacemos unas sencillas operaciones se convierten en otras más familiares. Dividimos el numerador y el denominador de la fórmula de *a* por n^2 y queda:

$$a = \frac{\frac{n\sum x_i y_i - \sum x_i \sum y_i}{n^2}}{\frac{n\sum x_i^2 - \left(\sum x_i\right)^2}{n^2}} = \frac{\frac{\sum x_i y_i}{n} - \frac{\sum x_i}{n}\cdot\frac{\sum y_i}{n}}{\frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2} = \frac{\frac{\sum x_i y_i}{n} - \bar{x}\cdot\bar{y}}{\frac{\sum x_i^2}{n} - \bar{x}^2} = \frac{\frac{\sum x_i y_i}{n} - \bar{x}\cdot\bar{y}}{s_x^2}$$

Aquí vemos que el denominador es igual a la varianza de la variable *X*. Por su parte la fórmula de *b* indica que:

$$b = \frac{\sum y_i - a\sum x_i}{n} = \frac{\sum y_i}{n} - a\frac{\sum x_i}{n} = \bar{y} - a\bar{x} \quad \text{o que} \quad a\bar{x} + b = \bar{y}$$ y, por tanto, la recta de regresión pasa por el punto $(\bar{x}, \bar{y})$, llamado **centro de gravedad de la nube de puntos**.

Sabiendo que la recta que buscamos pasa por el punto $(\bar{x}, \bar{y})$ y tiene como pendiente $a = \frac{\frac{\sum x_i y_i}{n} - \bar{x}\cdot\bar{y}}{s_x^2}$, entonces su ecuación es: $y - \bar{y} = \frac{\frac{\sum x_i y_i}{n} - \bar{x}\cdot\bar{y}}{s_x^2}\cdot(x - \bar{x})$

A la recta que mejor se ajusta a la nube de puntos la llamamos **recta de regresión**. Veremos ahora, en los ejemplos, que los ingredientes de la recta de regresión son muy fáciles de hallar con una calculadora científica sencilla.

Ejemplos

1. Hallar la ecuación de la recta de regresión correspondiente a la tabla de las edades y alturas de 12 niños registrados por un pediatra

meses	18	19	20	21	22	23	24	25	26	27	28	29
altura	76,1	77	78,1	78,2	78,8	78,2	79,5	81	81,2	81,8	82,8	83,5

Solución. Tenemos que encontrar los elementos de la ecuación:

$$y - \bar{y} = \frac{\frac{\sum x_i y_i}{n} - \bar{x}\cdot\bar{y}}{s_x^2}\cdot(x - \bar{x})$$

1º. Con las teclas MODE · ponemos la calculadora en modo estadístico, en la pantalla aparece SD, y ya están activas las teclas escritas en azul. Borramos los datos de la memoria con las teclas SHIFT SAC e introducimos los datos de la variable *X*:

18 DATA 19 DATA 20 DATA ... 29 DATA

Una vez introducidos los datos, con las teclas SHIFT $\bar{x}$ y las teclas SHIFT σ_n obtenemos $\bar{x} = 23{,}5$ y $s_x = 3{,}452$, que elevando al cuadrado resulta, $s_x^2 = 11{,}91$.

2º. Depués de borrar la memoria, introducimos los valores de *Y*

76.1 DATA 77 DATA 78.1 DATA ... 83.5 DATA

y con las teclas SHIFT $\bar{x}$ encontramos $\bar{y} = 79{,}683$

3º. Por último, después de borrar la memoria, introducimos $\sum x_i y_i$

18 × 76.1 DATA 19 × 77 DATA 20 × 78.1 DATA ... 29 × 83.5 DATA

y con las teclas SHIFT $\sum x$ obtenemos que $\sum x_i y_i = 22561{,}9$

4º. Escribimos la recta de regresión

$$y - 79{,}683 = \frac{\frac{22561{,}9}{12} - 23{,}5\cdot 79{,}683}{11{,}916}\cdot(x - 23{,}5)$$

Haciendo operaciones $y - 79{,}683 = 0{,}638 \cdot (x - 23{,}5)$

$$y = 0{,}638x + 64{,}679$$

Hay calculadoras científicas, más completas, que dan directamente la pendiente y la ordenada en el origen de la recta de regresión.

2. Hallar la ecuación de la recta de regresión correspondiente a la distribución de las temperaturas mínimas medias en el mes de enero y las latitudes de varias ciudades de Estados Unidos

	Temperatura	Latitud
Los Ángeles	8.3	34.3
San Francisco	5.5	38.4
Washington	1	39.7
Miami	14.4	26.3
Atlanta	2.7	33.9
Chicago	-7.2	42.3
Nueva Orleans	7.2	30.8
Nueva York	2.7	40.8
Boston	-5	42.7

Solución. Tenemos que encontrar los elementos de la ecuación:

$$y - \bar{y} = \frac{\frac{\sum x_i y_i}{n} - \bar{x} \cdot \bar{y}}{s_x^2} \cdot (x - \bar{x})$$

1º. Introducimos los datos de la variable *X*:

8.3 DATA 5.5 DATA 1 DATA ... - 5 DATA

Con las teclas SHIFT $\bar{x}$ y las teclas SHIFT σ_n obtenemos $\bar{x} = 3{,}288$ y $s_x = 6{,}266$, que elevando al cuadrado resulta, $s_x^2 = 39{,}267$.

2º. Depués de borrar la memoria, introducimos los valores de *Y*

34.3 DATA 38.4 DATA 39.7 DATA ... 42.7 DATA

y con las teclas SHIFT $\bar{x}$ encontramos $\bar{y} = 36{,}577$

3º. Por último, después de borrar la memoria, introducimos $\sum x_i y_i$

8.3 × 34.3 DATA 5.5 × 38.4 DATA 1 × 39.7 DATA ... 5 × 42.7 DATA

y con las teclas SHIFT $\sum x$ obtenemos que $\sum x_i y_i = 819{,}7$

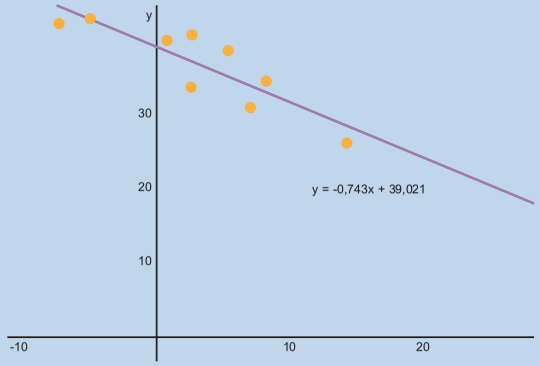
4º. Escribimos la recta de regresión

$$y - 36{,}577 = \frac{\frac{819{,}7}{9} - 3{,}288 \cdot 36{,}577}{39{,}267} \cdot (x - 3{,}288)$$

Haciendo operaciones, resulta la recta de regresión

$$y = -0{,}743x + 39{,}021$$

Gráficamente sería la recta de la figura:



La principal utilidad de la recta de regresión es hacer predicciones. Si quisiéramos saber cuál es la latitud de una ciudad de Estados Unidos cuya media de las temperaturas mínimas en el mes de enero es 4, 5º C, sustituimos *x* por 4,5 y obtenemos una estimación de la latitud:

$$y = -0{,}743 \cdot 4{,}5 + 39{,}021 = 35{,}677$$

La ciudad tendría 35,677 grados de latitud norte. Queda un problema por resolver: ¿qué fiabilidad proporciona la recta de regresión para hacer estimaciones? Eso lo sabremos conociendo el coeficiente de correlación lineal de las dos variables que estudiamos en el próximo apartado.

Actividades

3. El tiempo que tarda la sangre humana en coagular, según la temperatura, figura en la tabla siguiente:

Temperatura en ºC	5	10	15	20	25	35	40	45
Tiempo segundos	45	38	32	28	24	19	22	21

Halla la recta de regresión y estima el tiempo que tardará la sangre en coagular a 30º C.

4. Se ha medido las estaturas, en cm, de 12 madres y las de sus hijas, a partir de cierta edad y se han recogido los siguientes datos:

X estatura madres	166	168	165	156	170	167	154	169	167	158	172	175
Y estatura hijas	168	170	168	160	171	165	157	172	165	159	172	174

Hallar la recta de regresión.

5. Se ha anotado la potencia en caballos de vapor, la velocidad máxima que alcanzan y el peso en kilos de nueve modelos de automóvil:

	Cv	Km/h	Kg
Honda Civic	92	180	1020
Ford Scort	90	175	1133
Toyota Cel.	102	175	1360
Chevrolet B.	95	170	1360
Saab 9000	130	185	1587
Volvo 740	145	193	1587
Chrysler N.Y.	150	188	1814
Mercedes 500	322	265	2041
BMW 750IL	295	252	2041

a) Hallar la recta de regresión CV – Velocidad máxima, tomando como variable independiente CV. ¿Qué velocidad máxima alcanzaría un automóvil de 110 CV?

b) Hallar la recta de regresión CV – Peso, tomando como variable independiente CV. ¿Qué peso estimado tendría un automóvil de 200 CV?

6. Se han registrado las marcas olímpicas de tres especialidades de atletismo desde 1948 hasta 1992.

	salto de longitud	salto de altura	lanzamiento de disco
1948	7.82	1.98	52.78
1952	7.56	2.04	55.03
1956	7.82	2.11	56.34
1960	8.1	2.15	59.18
1964	8.05	2.17	61
1968	8.89	2.24	64.78
1972	8.22	2.22	64.78
1976	8.34	2.24	67.49
1980	8.54	2.35	66.64
1984	8.54	2.34	66.6
1988	8.71	2.37	68.81
1992	8.69	2.33	65.11

a) Halla la recta de regresión año olímpico – salto de altura. Estima las marcas olímpicas de salto de altura de las olimpiadas de Seúl (1996) y Sydney (2000).

b) Halla la recta de regresión año olímpico – salto de longitud. Estima las marcas olímpicas de salto de longitud de las olimpiadas de Seúl (1996) y Sydney (2000).

c) Halla la recta de regresión año olímpico – lanzamiento de disco. Estima las marcas olímpicas de lanzamiento de disco de las olimpiadas de Seúl (1996) y Sydney (2000).

7. Estima la latitud norte de una ciudad del continente americano que tuviese una temperatura mínima media en el mes de enero de 0º C. ¿Y la latitud de una ciudad que tiene de mínima media en el mismo mes –10º C? Comprueba en un atlas si esa ciudad pertenece a Estados Unidos. Recuerda la tabla:

	Temperatura	Latitud
Los Ángeles	8.3	34.3
San Francisco	5.5	38.4
Washington	1	39.7
Miami	14.4	26.3
Atlanta	2.7	33.9
Chicago	-7.2	42.3
Nueva Orleáns	7.2	30.8
Nueva York	2.7	40.8

8. Un fabricante de automóviles experimenta un tipo de frenos y registra a varias velocidades, en km/h, la distancia, en metros, que recorre el coche desde que se pisa el freno hasta que se detiene completamente. Los datos figuran en la tabla siguiente:

***X* km/h**	25	45	60	80	95	120	130	135
***Y* m**	6	14	28	45	65	85	94	108

Hallar la recta de regresión y estimar el recorrido antes de detenerse a una velocidad de 100 km/h.

4. Concepto de correlación

El grado de dependencia lineal entre dos variables se mide con el **coeficiente de correlación lineal**, y cuando la dependencia lineal es débil la recta de regresión carece de interés.

4.1. Covarianza

En primer lugar queremos averiguar si la relación entre dos variables es directa, es decir, cuando al aumentar la variable independiente aumenta también la variable dependiente, o si es inversa, qué ocurre cuando al aumentar la variable *X* disminuye la variable *Y*.

La **covarianza** es un parámetro que mide este tipo de relación y está definida como la media aritmética de los productos de la desviaciones de cada uno de los valores de las variables respecto a sus medias, se simboliza por s_{xy} y viene dada por:

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

La covarianza tiene una formulación más conocida si realizamos las operaciones indicadas

$$s_{xy} = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n} = \frac{\sum (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} \cdot \bar{y})}{n} = \frac{\sum x_i y_i}{n} - \bar{y} \frac{\sum x_i}{n} - \bar{x} \frac{\sum y_i}{n} + \bar{x} \cdot \bar{y} =$$

$$= \frac{\sum x_i y_i}{n} - \bar{y} \cdot \bar{x} - \bar{x} \cdot \bar{y} + \bar{x} \cdot \bar{y} = \frac{\sum x_i y_i}{n} - \bar{y} \cdot \bar{x}.$$

La covarianza resulta ser el numerador de la pendiente de la recta de regresión.

4.2. Coeficiente de correlación

La medida precisa de la relación de dos variables estadísticas lo proporciona el coeficiente de correlación lineal, representado por la letra *r*, y que está definido por la expresión siguiente:

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

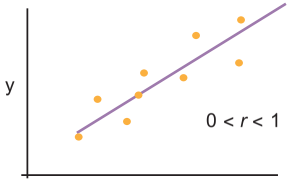
Es decir, es el cociente entre la covarianza y el producto de las desviaciones típicas de *X* e *Y*. Como la desviación típica de una variable estadística es siempre positiva, el signo del coeficiente de correlación depende del signo de la covarianza, y podemos afirmar:

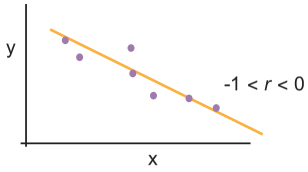
Covarianza positiva indica correlación directa.
Covarianza negativa indica correlación inversa.
Covarianza nula indica que no hay correlación entre la variables.

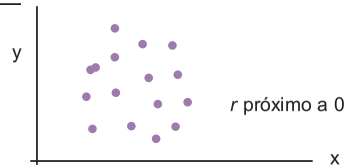
Se puede demostrar que el coeficiente de correlación es un número comprendido entre –1 y 1, y, en consecuencia, se pueden dar las siguientes situaciones:

- Que $r = 1$, entonces la relación entre las variables es funcional directa y la nube de puntos está sobre una recta de pendiente positiva.
- Que $0 < r < 1$, entonces hay una correlación directa entre las variables. Correlación fuerte cuando cuando *r* está próximo a 1 y débil cuando *r* se aproxima a 0.
- Que $r = 0$, entonces no existe ningún tipo de realción o dependencia entre las variables.
- Que $-1 < r < 0$, entonces hay correlación inversa entre las variables. Correlación fuerte cuando cuando *r* está próximo a -1 y débil cuando *r* está proximo a 0.
- Que $r = -1$, entonces la relación entre las variables es funcional inversa y la nube de puntos está sobre una recta de pendiente negativa.

En las figuras hemos ilustrado algunas de esta situaciones:







Resumiendo: la recta de regresión permite hacer previsiones o estimaciones, pero no debemos olvidar que estas estimaciones sólo son fiables cuando r toma valores próximos a –1 o a 1.

Ejemplos

1. Hallar el coeficiente de correlación lineal correspondiente a la tabla de las edades y alturas de 12 niños registrados por un pediatra

meses	18	19	20	21	22	23	24	25	26	27	28	29
altura	76,1	77	78,1	78,2	78,8	78,2	79,5	81	81,2	81,8	82,8	83,5

Solución. El coeficiente de correlación lineal viene dado por la fórmula:

$r = \dfrac{s_{xy}}{s_x \cdot s_y}$ donde $s_{xy} = \dfrac{\sum x_i y_i}{n} - \bar{x} \cdot \bar{y}$ y s_x y s_y son las desviaciones típicas de X e Y.

Ya sabemos, lo hemos calculado en el ejemplo 1, que

$$s_{xy} = \frac{\sum x_i y_i}{n} - \bar{x} \cdot \bar{y} = \frac{22561{,}19}{12} - 23{,}5 \cdot 79{,}68 = 7{,}6191$$

También conocemos que $\bar{x} = 23{,}5$ y $s_x = 3{,}45$.

Introducimos de nuevo los valores de Y

76.1 DATA 77 DATA 78.1 DATA ... 83.5 DATA

y con las teclas SHIFT $\bar{x}$ y las teclas SHIFT σ_n encontramos $\bar{y} = 79{,}68$ $s_y = 2{,}24$.

Luego

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{7{,}61}{3{,}45 \cdot 2{,}24} = 0{,}98.$$

Lo que indica un alto grando de correlación y las previsiones que se hagan con la recta de regresión son altamente fiables.

2. Hallar el coeficiente de correlación lineal de las calificaciones de 12 alumnos en Matemáticas y Lengua:

Matemáticas	2	4	4	6	4	6	3	6	5	7	3	7
Lengua	2	7	4	2	5	5	6	4	8	1	7	6

Solución.

1º Después de borrar la memoria, introducimos los datos de la variable Matemáticas, que llamaremos X,

2 DATA 4 DATA 4 DATA ... 7 DATA

Con las teclas SHIFT $\bar{x}$ y las teclas SHIFT σ_n obtenemos $\bar{x} = 4{,}75$ y $s_x = 1{,}68$.

2º Borramos la memoria e introducimos los datos de Lengua, variable Y

2 DATA 7 DATA 4 DATA ... 6 DATA

Con las teclas SHIFT $\bar{x}$ y las teclas SHIFT σ_n encontramos $\bar{y} = 4{,}75$ y $s_y = 2{,}12$.

3º Por último, introducimos

2 × 2 DATA 4 × 7 DATA 4 × 4 DATA ... 7 × 6 DATA

y con las teclas SHIFT $\sum x$ obtenemos que $\sum x_i y_i = 262$

Ahora, $s_{xy} = \dfrac{\sum x_i y_i}{n} - \bar{x} \cdot \bar{y} = \dfrac{262}{12} - 4{,}75 \cdot 4{,}75 = -0{,}7291$

entonces,

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{-0{,}7291}{1{,}68 \cdot 2{,}12} = -0{,}20$$

Lo que indica una correlación negativa pero muy débil.

Actividades

9. Halla el coeficiente de correlación lineal de la altura, en centímetros, de 12 madres y las de sus hijas, a partir de cierta edad, según los datos de la tabla.

X(estatura madres)	166	168	165	156	170	167	154	169	167	158	172	175
Y(estatura hijas)	168	170	168	160	171	165	157	172	165	159	172	174

10. Calcula el coeficiente de correlación lineal de las variables velocidad máxima - peso en kg, en los automóviles de la tabla:

	cv	**km/h**	**kg**
Honda Civic	92	180	1020
Ford Scort	90	175	1133
Toyota Cel.	102	175	1360
Chevrolet B.	95	170	1360
Saab 9000	130	185	1587

Volvo 740	145	193	1587
Chrysler N.Y.	150	188	1814
Mercedes 500	322	265	2041
BMW 750IL	295	252	2041

11. Calcula el coeficiente de correlación lineal de los pesos y las alturas de los jugadores de un equipo de fútbol que figuran en la tabla:

X(peso kg)	80	80	77	68	85	80	74	79	76	73	78
Y(altura cm)	187	185	184	173	189	183	177	189	180	176	182

12. En el cuadro siguiente aparecen las marcas en algunas especialidades de 10 atletas de decathlon.

	100 m	**salto longitud**	**salto altura**	**400 m**	**Disco**
A	10.43	8.08	2.07	48.51	48.56
B	10.44	8.01	2.03	46.97	46.56
C	10.7	7.76	2.07	48.05	49.36
D	11.06	7.79	2.03	48.43	46.58
E	10.89	7.49	2.03	47.38	46.9
F	10.5	7.26	2.11	47.63	49.7
G	10.96	7.57	1.97	48.72	48
H	10.96	7.43	2.04	48.19	49.88
I	10.87	7.42	2.1	49.75	51.2
K	10.69	7.88	2.1	47.96	43.96

Calcula el coeficiente de correlación lineal entre las variables 100 m – 400 m, salto de longitud – salto de altura y salto de altura – lanzamiento de disco.