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 permite 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 nombre de recta de regresión se debe al médico inglés Francis Galton (1822-1917), especialista en varios campos como la herencia genética, la psicología y estadística.

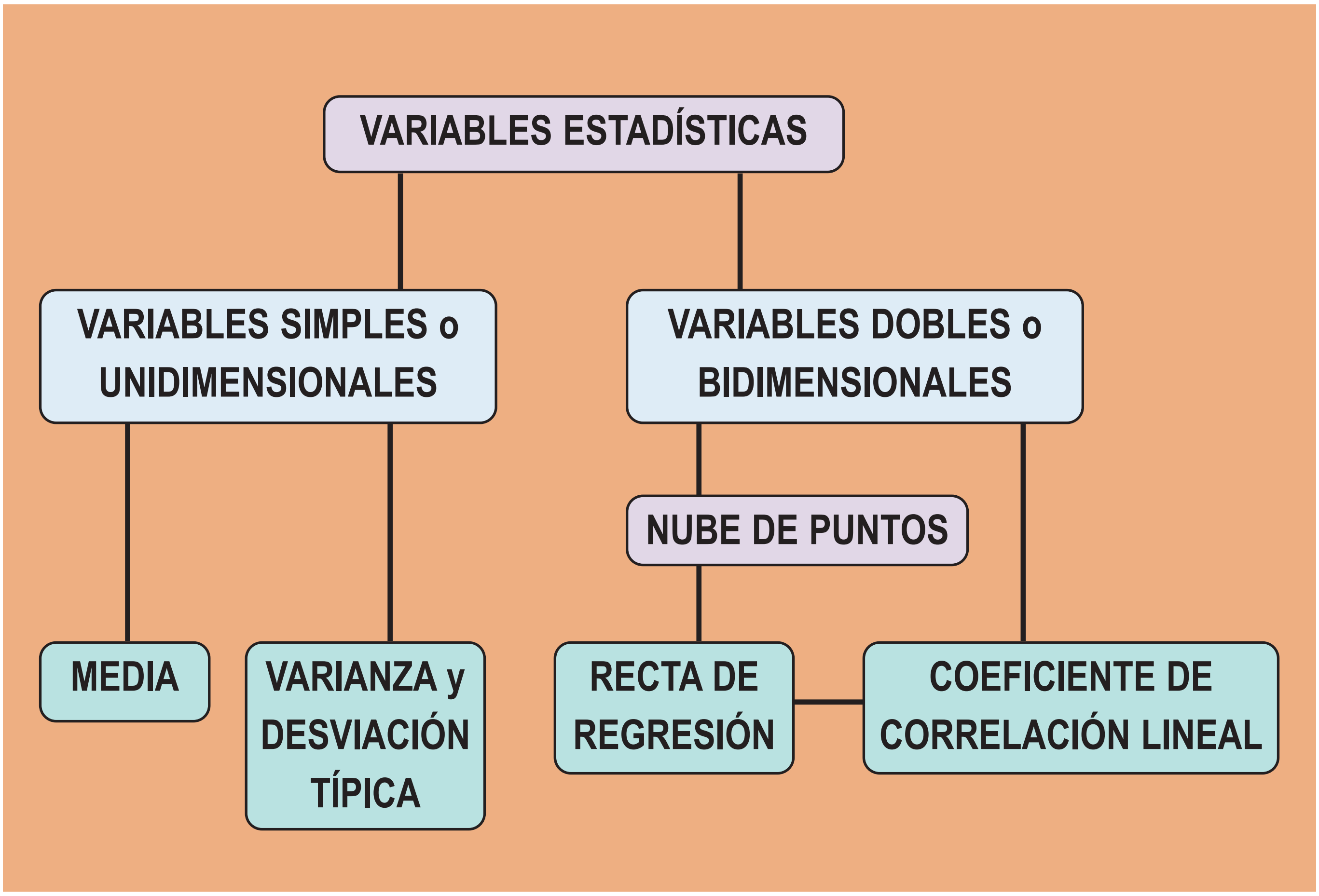
Francis Galton (Wikimedia Commons)

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 esencial de esta Unidad es, en primer lugar, encontrar la recta de regresión entre dos variables estadísticas y, a continuación, mediante el empleo del 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.

En esta Unidad didáctica nos proponemos alcanzar los **objetivos** siguientes:

1. Conocer y entender el concepto de variable estadística unidimensional.
2. Conocer y entender el concepto de variable estadística bidimensional.
3. Utilizar el diagrama de dispersión para reconocer si existe relación entre dos variables estadísticas.
4. Determinar la recta de regresión y entender cuál es la función lineal que mejor aproxima la relación entre las variables estadísticas.
5. Entender que la covarianza es el instrumento para conocer si la relación entre las variables es directa o inversa.
6. Calcular el coeficiente de correlación lineal y entender que su utilidad radica en dar sentido a la recta de regresión.



ÍNDICE DE CONTENIDOS

1. Variables estadísticas

La estadística estudia una característica o carácter de un conjunto de individuos llamado población. El carácter en estudio puede adoptar diversas modalidades; todas esas modalidades que puede adoptar un determinado carácter se llaman **variable estadística**.

La variable estadística puede ser cuantitativa o cualitativa. Por ejemplo, en la población de los empleados de una empresa edad, estatura, peso, sueldo, etc., son caracteres cuantitativos, es decir, son caracteres contables o medibles, mientras que la marca del automóvil que conducen, el nombre de su equipo de fútbol favorito, etc., son caracteres cualitativos.

Los números que sirven para contar o medir un carácter cuantitativo, y que pueden variar con cada individuo de la población, constituyen lo que se llama una **variable estadística cuantitativa**.

A veces es necesario disponer de un valor numérico que represente la diversidad de valores de una variable estadística. A los valores numéricos que cumplen esta función se les llama parámetros centrales o medidas de centralización de una distribución, y, como sabemos, el más importante es **la media.**

Si n datos de una característica se distribuyen en k valores de la variable y éstos son x_1, x_2,...,x_k y f_1, f_2,...,f_k, las frecuencias respectivas, en donde $f_1 + f_2 + ... + f_k = n$, entonces la media aritmética se calcula con la fórmula:

$$\bar{x} = \frac{x_1 \cdot f_1 + x_2 \cdot f_2 + ... + x_k \cdot f_k}{n} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{n}.$$

La media aritmética es la medida o parámetro de centralización más empleado.

Puede ocurrir que dos distribuciones de datos (dos variables estadísticas) de la misma característica, en poblaciones distintas, tengan los parámetros centrales muy parecidos o iguales y, sin embargo, que las distribuciones sean completamente diferentes. Por esta razón se necesitan otras medidas que nos indiquen el grado de variación de los datos respecto a la media. Estas nuevas medidas, que llamamos parámetros o medidas de dispersión, informan del grado de dispersión o de concentración de los datos en relación a la media. Los más utilizados son la varianza y la desviación típica.

La **varianza** es la media aritmética de los cuadrados de las desviaciones de todos los valores de la variable estadística respecto de la media. La varianza se simboliza por s^2.

Si los valores de la variable son x_1, x_2,...,x_n, y no se repite ninguno, entonces

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n} = \frac{\sum_{i=1}^{n} (x_n - \bar{x})^2}{n}$$

que también se puede expresar como:

$$s^2 = \frac{\sum_{i=1}^{n} (x_n - \bar{x})^2}{n} = \frac{\sum_{i=1}^{n} x_n^2 - \sum_{i=1}^{n} 2x_n \cdot \bar{x} + \sum_{i=1}^{n} \bar{x}^2}{n} = \frac{\sum_{i=1}^{n} x_n^2}{n} - \frac{2\bar{x}\sum_{i=1}^{n} x_n}{n} + \frac{\sum_{i=1}^{n} \bar{x}^2}{n} = \frac{\sum_{i=1}^{n} x_n^2}{n} - 2\bar{x}^2 + \bar{x}^2 = \frac{\sum_{i=1}^{n} x_n^2}{n} - \bar{x}^2.$$

Si los valores vienen con frecuencias, la varianza se calcula con las siguientes fórmulas que, como vimos anteriormente sin las frecuencias, son equivalentes:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot f_i}{n} \quad \text{ó} \quad s^2 = \frac{\sum_{i=1}^{n} x_i^2 \cdot f_i}{n} - \bar{x}^2 .$$

La varianza es siempre positiva, pero tiene un pequeño inconveniente: no se expresa en las mismas unidades que los datos ya que está al cuadrado (si los datos son cm la varianza serán cm^2). Para subsanar este problema se define la **desviación típica** de una variable estadística como la raíz cuadrada positiva de la varianza, y ésta ya sí tiene las mismas unidades que los datos.

La representamos por la letra *s* y viene expresada por:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot f_i}{n}} = \sqrt{\frac{\sum_{i=1}^{k} x_i^2 \cdot f_i}{n} - \bar{x}^2}$$

La desviación típica es el parámetro de dispersión más utilizado.

Ejemplos

1. Se han registrado las distancias en km, desde su casa al trabajo, de 40 empleados de una empresa situada en un polígono industrial, y han resultado los siguientes datos:

   3, 5, 10, 15, 20, 25, 3, 6, 12, 18, 23, 28, 4, 8, 14, 15, 22, 25, 8, 19, 10, 20, 9, 12, 16, 11, 15, 13, 18, 23, 10, 17, 22, 12, 15, 25, 14, 18, 20, 22.

   Halla la media, la varianza y la desviación típica.

   Solución:

Clases	Marcas de clase x_i	f_i	$x_i \cdot f_i$
[0, 5)	2,5	3	7,5
[5, 10)	7,5	5	37,5
[10,15)	12,5	10	125
[15,20)	7,5	10	175
[20,25)	22,5	8	180
[25,30)	27,5	4	110
			$\sum x_i \cdot f_i = 635$

   Se construye una tabla de distribución de frecuencias. Se agrupan en primer lugar los datos en clases o intervalos de amplitud 5 km que incluyan el extremo izquierdo del intervalo, pero no el derecho. Calculamos el centro de la clase, que como recordarás se denomina **marca de clase**. En la tabla x_i y f_i son las marcas de clase y las frecuencias.

   La media de las distancias buscada será: $\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{635}{40} = 15{,}87$ kilómetros.

   La varianza es la media de los cuadrados de las desviaciones o separaciones de cada uno de los valores de la variable respecto a la media aritmética. La desviación típica de una distribución estadística es la raíz cuadrada de la varianza:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot f_i}{n}} = \sqrt{\frac{\sum_{i=1}^{k} x_i^2 \cdot f_i}{n} - \bar{x}^2} .$$

Multiplicando la columna de las marcas de clase por la última columna de la tabla anterior se obtiene una nueva columna encabezada por $x_i^2 \cdot f_i$.

x_i	f_i	$x_i \cdot f_i$	$x_i^2 f_i$
2,5	3	7,5	18,75
7,5	5	37,5	281,25
12,5	10	125	1562,5
17,5	10	175	3062,5
22,5	8	180	4050
27,5	4	110	3025
	$\sum f_i = 40$		$\sum x_i^2 \cdot f_i = 12000$

La varianza es $s^2 = \frac{\sum f_i (x_i - \bar{x})^2}{n} = \frac{\sum x_i^2 \cdot f_i}{n} - \bar{x}^2 = \frac{12000}{40} - 15{,}87^2 = 48{,}1431$ y la desviación típica

$$s = \sqrt{s^2} = \sqrt{48{,}1431} = 6{,}9385...$$

Actividades

1. Se toma el pulso a un grupo de 30 personas, obteniéndose los datos siguientes:

72, 66, 81, 74, 57, 58, 74, 62, 73, 65, 78, 75, 84, 72, 69, 76, 65, 79, 76, 68, 82, 71, 77, 72, 56, 62, 83, 63, 70, 73.

Calcula la media, la varianza y la desviación típica.

2 Las temperaturas máximas en una ciudad durante el mes de mayo han sido:

26, 26, 30, 30, 31, 27, 28, 27, 27, 26, 26, 37, 27, 28, 28, 29, 27, 26, 25, 22, 28, 28, 27, 28,29, 30, 26, 25, 27, 29, 37.

Calcula la temperatura máxima media, la varianza y la desviación típica.

2. Variables estadísticas dobles

En una población estudiaremos dos variables estadísticas: una que llamamos X y otra 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 de la variable X e y_j representa los valores de la variable Y.

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

Por ejemplo, en una evaluación de 30 alumnos se registra el número de asignaturas suspensas y el número de horas diarias que dedica cada uno al estudio obteniéndose:

(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:

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 *X* y en la primera fila los valores de *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 *Y* y en la última columna las frecuencias de *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 verticales).

Por ejemplo, las calificaciones de 12 estudiantes 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).

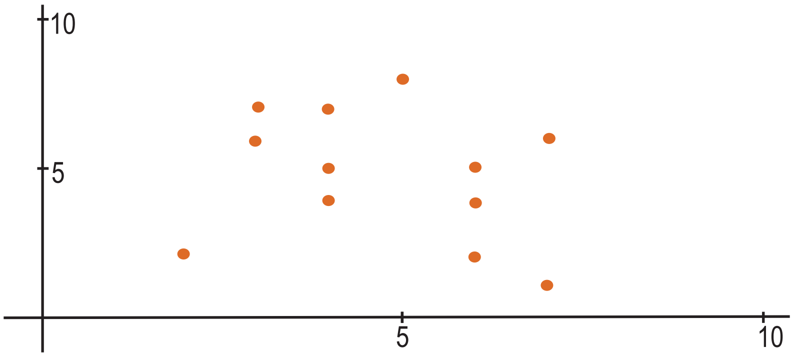
Estos datos se disponen en una tabla simple de dos filas así: .

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

3. 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, pues la independiente *X* es la que permite explicar el comportamiento de la dependiente *Y*.

En el caso de las notas de Matemáticas y Lengua de 12 estudiantes, 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:



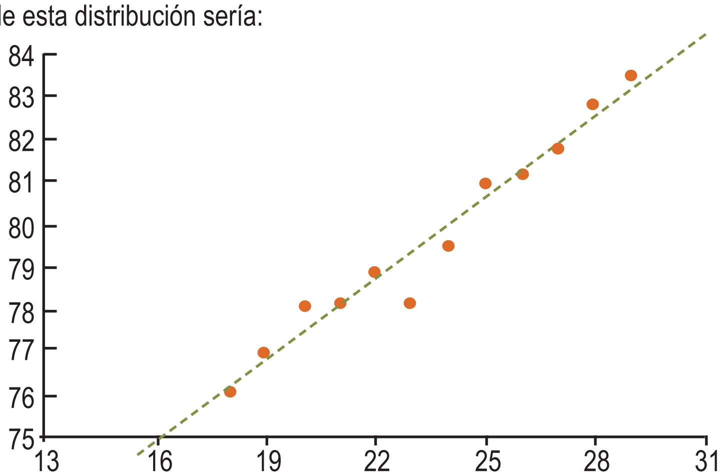
La representación gráfica de una distribución bidimensional se denomina **diagrama de dispersión o nube de puntos**. Cada punto tiene por coordenadas los valores que en cada individuo tienen las variables *X* e *Y*. La nube de puntos nos permite apreciar si existe una posible relación entre las variables.

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:

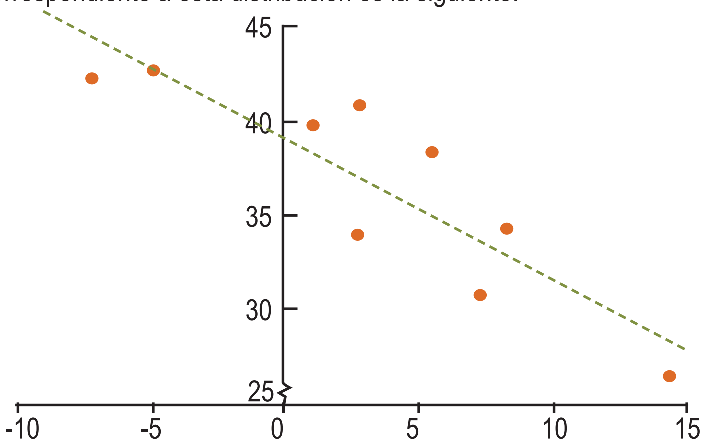


Hemos dibujado una recta entre los puntos porque todo sugiere que la relación entre las variables edades y alturas se aproxima a una relación lineal.

La tabla siguiente indica el valor medio de las temperaturas mínimas en el mes de enero y las latitudes de algunas 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

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 el apartado siguiente.

Actividades

3. Los pesos y las alturas de los jugadores de un equipo de fútbol vienen 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.

4. 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

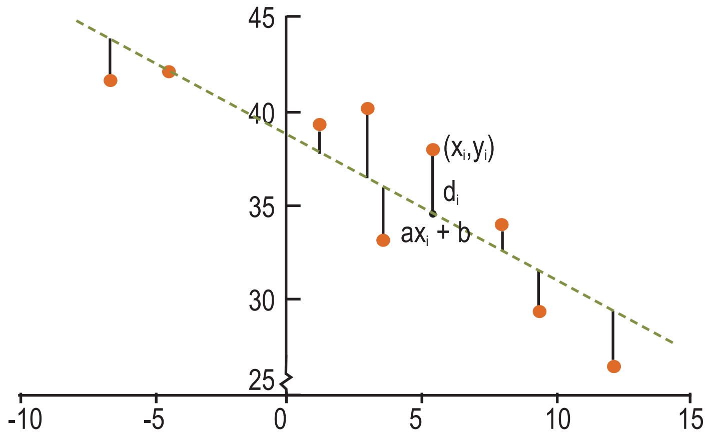
Dibuja el diagrama de dispersión.

4. 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 sumamente 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 o, mejor dicho, la suma de los cuadrados de las diferencias entre las ordenadas del punto y de 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, lo que obliga a usar 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 única variable fuese a e igualar a cero, y a continuación hacer lo mismo suponiendo que la única variable fuese b, con lo que se obtiene un sistema de dos ecuaciones con dos incógnitas:

$$\begin{cases} \sum y_i - a\sum x_i - nb = 0 \\ \sum x_i y_i - a\sum x_i^2 - b\sum x_i = 0 \end{cases}$$

Si dejamos solas las incógnitas, a y b, en el primer miembro, el sistema queda así:

$$\begin{cases} a\sum x_i + nb = \sum y_i \\ a\sum x_i^2 + b\sum x_i = \sum x_i y_i \end{cases}$$

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 al hacer unas sencillas operaciones se convierten en otras más familiares. Dividimos numerador y denominador de la fórmula de *a* por n^2 y queda:

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

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

$$b=\frac{\sum y_i-a\sum x_i}{n}=\frac{\sum y_i}{n}-a\frac{\sum x_i}{n}=\overline{y}-a\overline{x}$$

o que $a\overline{x}+b=\overline{y}$ y , por tanto, la recta de regresión pasa por el punto $(\overline{x},\overline{y})$, llamado **centro de gravedad de la nube de puntos**.

Sabiendo que la recta que buscamos pasa por el punto $(\overline{x},\overline{y})$ y tiene como pendiente $a=\dfrac{\dfrac{\sum x_iy_i}{n}-\overline{x}\cdot\overline{y}}{s_x^2}$, entonces podemos aplicar la fórmula punto-pendiente y la ecuación de esta recta será:

$$y-\overline{y}=\frac{\dfrac{\sum x_iy_i}{n}-\overline{x}\cdot\overline{y}}{s_x^2}\cdot(x-\overline{x})$$

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

2. Halla la ecuación de la recta de regresión correspondiente a la tabla de las edades (en meses) y alturas (en cm) 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: $y-\overline{y}=\dfrac{\dfrac{\sum x_iy_i}{n}-\overline{x}\cdot\overline{y}}{s_x^2}\cdot(x-\overline{x})$

Mostraremos un procedimiento para hallarlos con una de las calculadoras más empleadas (las SVPAM); cualquier otra calculadora tiene un procedimiento semejante, pero conviene consultar el manual de esa calculadora.

1º. Ponemos la calculadora en el modo SD (Standard deviation) con la tecla MODE y elegimos el número 2 (SD). Borramos los datos que pueda haber en la memoria estadística con las teclas SHIFT CLR, elegimos el número 1 (Scl), pulsamos = y AC para borrar la pantalla. Introducimos los datos de la variable X:

18 DT 19 DT 20 DT ... 29 DT

Una vez introducidos los datos, con las teclas SHIFT S-VAR y 1 ($\overline{x}$) = y aparece en pantalla 23,5, luego $\overline{x}$ = 23,5. Borramos la pantalla con AC y pulsando SHIFT S-VAR y 2 ($x\sigma n$) =, vemos que s_x = 3,45, que elevado al cuadrado resulta, s_x^2 = 11,91.

2º. Después de borrar la memoria estadística, introducimos los valores de *Y*:

76.1 DT 77 DT 78.1 DT ... 83.5 DT

Con SHIFT S-VAR y 1 ($\overline{x}$) = encontramos que $\overline{y}$ = 79,68

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

18 SHIFT ; 76.1 DT 19 SHIFT ; 77 DT 20 SHIFT ; 78.1 DT ... 29 SHIFT ; 83.5 DT

y con las teclas SHIFT S-SUM y 2 ($\sum x$) = obtenemos que $\sum x_i \cdot y_i$ = 22561,19.

4º. Escribimos la recta de regresión: $y - 79{,}68 = \dfrac{\dfrac{22561{,}19}{12} - 23{,}5 \cdot 79{,}68}{11{,}91} \cdot (x - 23{,}5)$

Haciendo operaciones tenemos: $y - 79{,}68 = 0{,}64 \cdot (x - 23{,}5)$ ó $y = 0{,}64x + 64{,}64$

La misma calculadora nos da directamente la pendiente y la ordenada en el origen de la recta de regresión como indicamos en la nota del ejemplo 3.

3. Halla la ecuación de la recta de regresión correspondiente a la distribución de las temperaturas medias mínimas 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: $y - \overline{y} = \dfrac{\dfrac{\sum x_i y_i}{n} - \overline{x} \cdot \overline{y}}{s_x^2} \cdot (x - \overline{x})$

1º. Ponemos la calculadora en el modo SD (Standard deviation) con la tecla MODE y elegimos el número 2 (SD) y borramos los datos que pueda haber en la memoria estadística con las teclas SHIFT CLR, elegimos el número 1 y pulsamos =, y AC para borrar la pantalla. Introducimos los datos de la variable *X*:

8.3 DT 5.5 DT 1 DT ... -5 DT

Con las teclas SHIFT S-VAR y 1 = y las teclas SHIFT S-VAR y 2 =, obtenemos $\overline{x}$ = 3,28 y s_x = 6,26; ésta última elevada al cuadrado es: s_x^2 = 39,26.

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

34.3 DT 38.4 DT 39.7 DT ... 42.7 DT

y con las teclas SHIFT S-VAR y 1 =, encontramos que $\overline{y}$ = 36,57

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

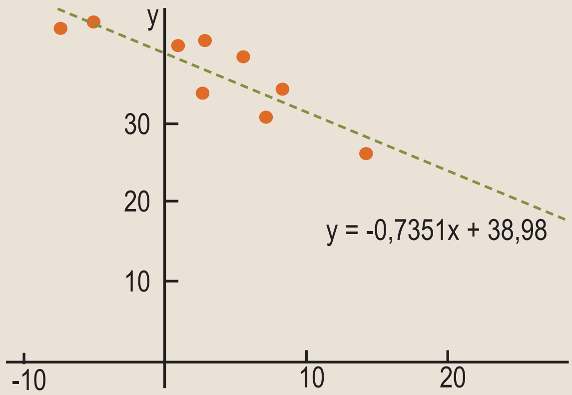
8.3 SHIFT ; 34.3 DT 5.5 SHIFT ; 38.4 DT 1 SHIFT ; 39.7 DT ... -5 SHIFT ; 42.7 DT

y con las teclas SHIFT S-SUM y 2 =, obtenemos que $\sum x_i \cdot y_i$ = 819,7

$$4^{\circ}.\ \text{Escribimos la recta de regresión: } y-36{,}57=\frac{\frac{819{,}7}{9}-3{,}28\cdot 36{,}57}{39{,}26}\cdot(x-3{,}28)$$

Haciendo operaciones, resulta la recta de regresión: $y = -0{,}735x + 38{,}98$

Gráficamente sería la recta de la figura siguiente:



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:

$$y = -0{,}735\cdot 4{,}5 + 38{,}98 = 35{,}67$$

La ciudad tendría 35,67 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.

Nota: La misma calculadora que hemos empleado en los cálculos anteriores nos da directamente la recta de regresión. Debemos ponerla primeramente en modo regresión con la tecla MODE y 3 (REG), aparece una nueva pantalla en la que elegimos 1 (Lin). Aunque puede haber otras funciones que se ajustan a la nube de puntos estudiaremos sólo la regresión lineal o el ajuste por una función lineal. Ya tenemos la máquina dispuesta para hallar la recta de regresión. Borramos, como precaución, los datos viejos de la memoria con SHIFT CLR, elegimos el número 1 (Scl) y pulsamos =, y AC para borrar la pantalla. Introducimos los datos de la variable bidimensional (*X*, *Y*) así:

8.3 , 34.3 DT 5.5 , 38.4 DT 1 , 39.7 DT ... –5 , 42.7 DT

Borramos la pantalla con AC y con las teclas SHIFT S-VAR accedemos a una pantalla en la que nos desplazamos hacia la derecha con REPLAY → hasta que aparezca:

A	*B*	*r*
1	2	3

Si pulsamos 1 = resulta 39,02253254 y si, después de borrar la pantalla y volver a la pantalla de las letras A, B y r, pulsamos 2 = aparece –0,744186777 que son la ordenada en el origen y la pendiente de la recta de regresión, $y = Bx + A$, con lo que la recta buscada resulta ser:

$$y = -0{,}744186777x + 39{,}02253254$$

Las pequeñas diferencias que se observan entre esta recta y la calculada antes estriban en que en los cálculos anteriores truncamos algunos decimales.

Actividades

5. El tiempo (en s) que tarda la sangre humana en coagular, según la temperatura (en ºC), 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.

6. Se han 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

Halla la recta de regresión.

7. Se ha anotado la potencia en caballos de vapor, la velocidad máxima que alcanzan en km/h y el peso en kg 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
Chysler N.Y.	150	188	1814
Mercedes 500	322	265	2041
BMW 750 IL	295	252	2041

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

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

8. Se han registrado las marcas olímpicas de tres especialidades de atletismo desde 1948 hasta 1992, obteniéndose la siguiente tabla:

	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 de salto de altura de las olimpiadas de Seúl (1996), Sydney (2000), Atenas (2004) y Pekín (2008).

b) Halla la recta de regresión *año olímpico – salto de longitud*. Estima las marcas de salto de longitud de las olimpiadas de Seúl (1996), Sydney (2000), Atenas (2004) y Pekín (2008).

c) Halla la recta de regresión *año olímpico – lanzamiento de disco*. Estima las marcas de lanzamiento de disco de las olimpiadas de Seúl (1996), Sydney (2000), Atenas (2004) y Pekín (2008).

9. Estima la latitud norte de una ciudad del continente americano que tuviese una temperatura media en el mes de enero de 0º C. ¿Y la latitud de una ciudad que tiene de media –10º C? Comprueba 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 Orleans	7,2	30,8
Nueva York	2,7	40,8
Boston	-5	42,7

10. Un fabricante de automóviles experimenta un tipo de frenos y registra a varias velocidades, en km/h, la distancia, en m, 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

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

Para saber más...

Hay otra recta de regresión

Si al buscar la recta de regresión en vez de la condición de que la suma de los cuadrados de las diferencias entre las ordenadas del punto y de la recta sea mínimo imponemos la condición de que la suma de los cuadrados de las diferencias entre las abscisas del punto y de la recta sea mínima, entonces se obtiene otra recta de regresión cuya ecuación es

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

y que se denomina **recta de regresión de *X* sobre *Y***. Cuando la correlación es alta, es muy parecida a la recta de regresión que hemos estudiado. Algunos autores emplean esta recta para hacer estimaciones de *X* conocido el valor de *Y*, aunque, y es importante indicarlo, cuando la correlación es alta los valores obtenidos por una u otra recta están muy próximos.

5. Concepto de correlación

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

5.1. 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, lo que 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 las 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 manejable:

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

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

La covarianza resulta ser el numerador de la pendiente de la recta de regresión y ahora veremos que es un elemento importante en el cálculo del coeficiente de correlación.

5.2. 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:

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

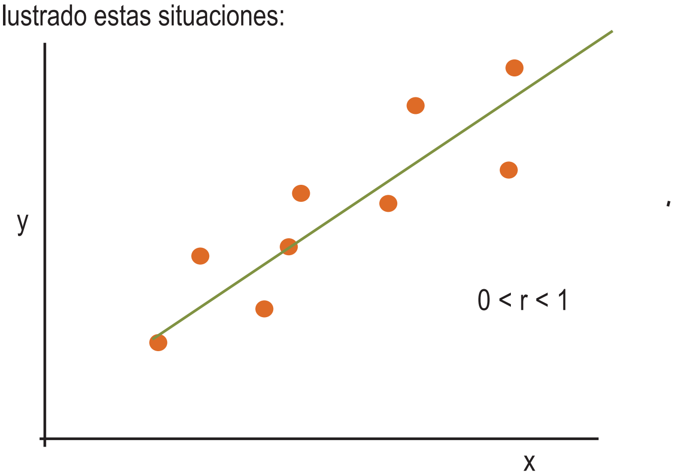
Es decir, es el cociente entre la covarianza y el producto de las desviaciones típicas de X e Y. Como la desviación típica es siempre positiva, el signo del coeficiente de correlación depende del signo de la covarianza, y podemos afirmar que:

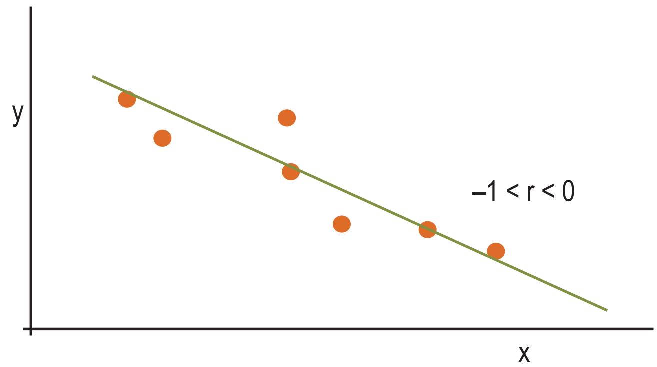
- Covarianza positiva indica correlación directa.
- Covarianza negativa indica correlación inversa.
- Covarianza nula indica que no hay correlación entre las variables.

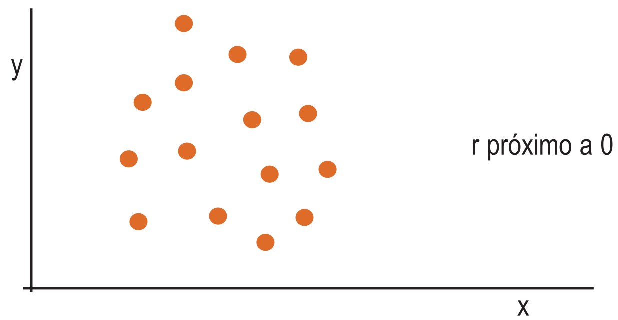
Se puede demostrar que el coeficiente de correlación es un número comprendido entre –1 y 1, y podemos encontrar las siguientes situaciones:

- 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 r está próximo a 1 y débil cuando r se aproxima a 0.
- Que $r = 0$, entonces no existe ningún tipo de relación o dependencia entre las variables.
- Que $-1 < r < 0$, entonces hay correlación inversa entre las variables. Correlación fuerte cuando r está próximo a -1 y débil cuando r está próximo 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 estas situaciones:







Resumiendo: la recta de regresión permite hacer previsiones o estimaciones, pero dichas estimaciones únicamente son fiables cuando r toma valores próximos a –1 o a 1.

Ejemplos

4. Halla 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 $s_{xy}=\dfrac{\sum x_iy_i}{n}-\bar{x}\cdot\bar{y}$ y s_x y s_y son las desviaciones típicas de X e Y.

Ya sabemos, lo hemos calculado anteriormente en el ejemplo 2, que $s_{xy}=\dfrac{\sum x_iy_i}{n}-\bar{x}\cdot\bar{y}=\dfrac{22561{,}19}{12}-23{,}5\cdot 79{,}68=$

$= 7{,}6191$

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

Introducimos de nuevo los valores de Y: 76.1 DT 77 DT 78.1 DT ... 83.5 DT

Luego, con las teclas SHIFT S -VAR y 1 =, encontramos que $\bar{y}=79{,}68$, y pulsando SHIFT S-VAR y 2 = conseguimos la desviación típica de Y: $s_y=2{,}24$. En consecuencia, $r=\dfrac{s_{xy}}{s_x\cdot s_y}=\dfrac{7{,}61}{3{,}45\cdot 2{,}24}=0{,}98.$

Este resultado indica un alto grado de correlación y las previsiones que se hagan con la recta de regresión son altamente fiables.

5. Halla 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 DT 4 DT 4 DT ... 7 DT

Con las teclas SHIFT S-VAR y 1 =, encontramos que $\bar{x}=4{,}75$, y pulsando SHIFT S-VAR y 2 =, obtenemos $s_x=1{,}58$.

2º Borramos la memoria e introducimos los datos de Lengua, variable Y: 2 DT 7 DT 4 DT ... 6 DT

Con las teclas SHIFT S-VAR y 1 =, encontramos que $\bar{y}=4{,}75$, y pulsando SHIFT S-VAR y 2 =, $s_y=2{,}12$.

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

2 SHIFT ; 2 DT 4 SHIFT ; 7 DT 4 SHIFT ; 4 DT ... 7 SHIFT ; 6 DT

y con las teclas SHIFT S-SUM y 2 =, obtenemos que $\sum x_i\cdot y_i=262$.

Ahora, $s_{xy}=\dfrac{\sum x_iy_i}{n}-\bar{x}\cdot\bar{y}=\dfrac{262}{12}-4{,}75\cdot 4{,}75=-0{,}7291$, entonces $r=\dfrac{s_{xy}}{s_x\cdot s_y}=\dfrac{-0{,}7291}{1{,}68\cdot 2{,}12}=-0{,}20$, lo que indica una correlación negativa pero muy débil.

Nota: La misma calculadora que hemos empleado en los cálculos anteriores nos da directamente el coeficiente de correlación. En primer lugar, la configuramos en modo regresión con la tecla MODE y 3 (REG); aparece una nueva pantalla en la que elegimos 1 (Lin). Ya tenemos la máquina dispuesta para hallar el coeficiente de correlación lineal. Borramos, como precaución, los datos viejos de la memoria con SHIFT CLR, elegimos el número 1 (Scl) y pulsamos =, y AC para borrar la pantalla. Introducimos los datos de la variable bidimensional (X, Y) así:

2 , 2 DT 4 , 7 DT 4 , 4 DT ... 7 , 6 DT

Borramos la pantalla con AC y con las teclas SHIFT S-VAR accedemos a una pantalla en la que nos desplazamos hacia la derecha con REPLAY → hasta que aparezca:

A B r

1 2 3

Si pulsamos 3 =, resulta −0,215995892, es decir, el coeficiente de correlación lineal:

r = −0,215995892. Hay una pequeña diferencia con el valor de r calculado antes, y que se debe al truncamiento de los decimales en los cálculos anteriores.

6. Vamos a emplear la calculadora científica para hallar la recta de regresión y el coeficiente de correlación lineal de los datos de la evaluación de 30 estudiantes donde se registró el número de asignaturas suspensas y el número de horas diarias que dedica cada uno al estudio:

(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)

y que, según vimos en el apartado 2 de este tema, se pueden ordenar en una tabla de frecuencias de doble entrada:

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

Solución:

En primer lugar, la configuramos en modo regresión con la tecla MODE y 3 (REG), aparece una nueva pantalla en la que elegimos 1 (Lin). Ya tenemos la máquina dispuesta para hallar el coeficiente de correlación lineal. Borramos, como precaución, los datos viejos de la memoria con SHIFT CLR, elegimos el número 1 (Scl) y pulsamos =, y AC para borrar la pantalla. Introducimos los datos de la variable bidimensional (X, Y) así:

0 , 2 SHIFT ; 2 DT 0 , 3 DT 0 , 4 SHIFT ; 2 DT 1 , 2 SHIFT ; 7 DT …6 , 0 DT

Borramos la pantalla con AC y con las teclas SHIFT S-VAR accedemos a una pantalla en la que nos desplazamos hacia la derecha con REPLAY → hasta que aparezca el menú:

A B r

1 2 3

Pulsando 1 = obtenemos 2,822134387, después de borrar y acceder a la misma pantalla pulsamos 2 = y nos muestra − 0,511857707; la recta de regresión es y = − 0,511857707x + 2,822134387. Por último, tras borrar y llegar a la pantalla anterior, pulsamos 3 = y nos da el coeficiente de correlación lineal − 0,774512388. Es decir, el coeficiente de correlación r = −0,774512388. Aún hay un último menú $\hat{x}$ $\hat{y}$ / 1 2 con el que podemos hacer estimaciones.

Actividades

11. Halla el coeficiente de correlación lineal de la altura 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

12. 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
Chysler N.Y.	150	188	1814
Mercedes 500	322	265	2041
BMW 750 IL	295	252	2041

13. 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

14. En el cuadro siguiente aparecen las marcas en algunas especialidades de 10 atletas de decatlón que hemos identificado con las 10 primeras letras del abecedario. 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*.

	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

Para saber más...

Sobre el nombre de recta de regresión

El nombre de recta de regresión, como se dijo en la introducción, se debe a Francis Galton (1822-1917). Éste realizó un trabajo estadístico sobre la talla de los padres y de los hijos, en el que se observa que a padres altos corresponden también hijos altos, y que padres de poca estatura engendran hijos de poca estatura. Pero, además, señalaba que la talla de los hijos tendía a regresar hacia la media de la población. Es decir, padres altos engendran hijos altos, pero menos; y padres bajos tenían hijos de poca estatura, pero más altos que ellos.

Recuerda

- ✓ Si n datos, de una característica, se distribuyen en k valores de la variable y éstos son: $x_1, x_2,...,x_k$ y $f_1, f_2,...,f_k$, las frecuencias respectivas, en donde $f_1 + f_2 +...+ f_k = n$, entonces la media aritmética se calcula con la fórmula:

$$\overline{x} = \frac{x_1 \cdot f_1 + x_2 \cdot f_2 + ... + x_k \cdot f_k}{n} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{n}.$$

- ✓ Para determinar la **varianza**, s^2, y la **desviación típica**, s, empleamos las expresiones

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 \cdot f_i}{n} - \overline{x}^2 \qquad s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{k} (x_i - \overline{x})^2 \cdot f_i}{n}} = \sqrt{\frac{\sum_{i=1}^{k} x_i^2 \cdot f_i}{n} - \overline{x}^2}$$

- ✓ La recta que mejor se ajusta a la nube de puntos que llamamos **recta de regresión** viene dada por:

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

- ✓ La **covarianza** es un parámetro que mide si la relación entre las variables es directa o inversa y se calcula por

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

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

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