# *TEMA 3: DISTRIBUCIONES BIDIMENSIONALES. CORRELACIÓN Y REGRESIÓN.*

## 1. VARIABLES ESTADÍSTICAS BIDIMENSIONALES. DISTRIBUCIONES BIDIMENSIONALES.

En esta unidad estudiaremos el comportamiento estadístico conjunto de dos características o variables estadísticas unidimensionales sobre un mismo colectivo o población. Por ejemplo:

- Horas de estudio y calificaciones en alumnos de bachillerato.
- Calificaciones en matemáticas y lengua para los mismos alumnos.
- Dinero gastado en publicidad y dinero obtenido por las ventas de cierta empresa.

**Variable estadística bidimensional** es el conjunto de pares de valores de dos caracteres o variables estadísticas unidimensionales $X$ e $Y$ sobre una misma población.

La variable estadística bidimensional se representa por el símbolo $(X, Y)$ y cada uno de los individuos de la población viene caracterizado por la pareja $(x_i, y_i)$, en el cual $x_i$ representa los datos, valores o marcas de clase $x_1, x_2, ..., x_n$ de la variable $X$; e $y_i$ representa los datos, valores o marcas de clase $y_1, y_2, ..., y_m$ de la variable $Y$.

Se denominan distribuciones bidimensionales a las tablas estadísticas bidimensionales formadas por todas las frecuencias absolutas de todos los posibles valores de la variable estadística bidimensional $(X, Y)$.

Las tablas estadísticas bidimensionales pueden ser:

a) Simples.
b) De doble entrada.

a) Las tablas estadísticas bidimensionales simples adoptan la siguiente forma:

| Variable $X$ | Variable $Y$ | Frecuencia absoluta |
|---|---|---|
| $x_1$ | $y_1$ | $f_1$ |
| $x_2$ | $y_2$ | $f_2$ |
| ⋮ | ⋮ | ⋮ |
| $x_i$ | $y_i$ | $f_i$ |
| ⋮ | ⋮ | ⋮ |
| $x_n$ | $y_m$ | $f_n$ |
| | | $\sum_i f_i = N$ |

*Ejemplo*: A cada uno de los trabajadores de una empresa se les talla y pesa. Se trata de dos variables cuantitativas.

| $X$ (tallas en m) | 1,70 | 1,70 | 1,69 | 1,68 |
|---|---|---|---|---|
| $Y$ (peso en kg) | 67 | 75 | 70 | 66 |

En este caso no aparecen las frecuencias absolutas porque habría un recluta con cada peso y talla, se podría añadir la fila correspondiente (o columna) con cada frecuencia absoluta igual a uno.

b) Las tablas estadísticas bidimensionales de doble entrada adoptan la siguiente forma:

| $X$ / $Y$ | $x_1$ | $x_2$ | $\cdots$ | $x_i$ | $\cdots$ | $x_n$ | F. absoluta de la variable $Y$ |
|---|---|---|---|---|---|---|---|
| $y_1$ | $f_{1\,1}$ | $f_{2\,1}$ | $\cdots$ | $f_{i\,1}$ | $\cdots$ | $f_{n\,1}$ | $f_{\bullet\,1}$ |
| $y_2$ | $f_{1\,2}$ | $f_{2\,2}$ | $\cdots$ | $f_{i\,2}$ | $\cdots$ | $f_{n\,2}$ | $f_{\bullet\,2}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $y_j$ | $f_{1\,j}$ | $f_{2\,j}$ | $\cdots$ | $f_{i\,j}$ | $\cdots$ | $f_{n\,j}$ | $f_{\bullet\,j}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $y_m$ | $f_{1\,m}$ | $f_{2\,m}$ | $\cdots$ | $f_{i\,m}$ | $\cdots$ | $f_{n\,m}$ | $f_{\bullet\,m}$ |
| F. absoluta de la variable $X$ | $f_{1\bullet}$ | $f_{2\bullet}$ | $\cdots$ | $f_{i\bullet}$ | $\cdots$ | $f_{n\bullet}$ | $N$ |

Denotamos por $f_{ij}$ a la frecuencia absoluta correspondiente al par $(x_i, y_j)$ y por $N$ al número total de individuos.

*Ejemplo*: Los datos obtenidos al estudiar las variables $X$ = "número de goles marcados" e $Y$ = "número de goles recibidos", en 40 partidos jugados por el equipo campeón de la liga de fútbol sala, son:

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

Elaboramos la tabla de doble entrada siguiendo estos pasos:

- Construimos una tabla con tantas columnas como valores tome $X$ y con tantas filas como valores tome $Y$ en la distribución.
  Si observamos los datos, $X$ toma los valores 1, 2, 3, 4, 5 y 6, e $Y$ toma los valores 1, 2, 3 y 4. En este caso, la tabla constará de 6 columnas y 4 filas.

- Hallamos la frecuencia absoluta de cada par de valores de la variable $(X, Y)$. Para ello contamos el número de veces que se repite ese par de valores en la distribución y lo anotamos en la casilla correspondiente.
  Así, por ejemplo, observa que el par (5, 4) aparece una sola vez; el (4, 2) aparece diez veces; y el (6, 1), ninguna.

| ***X*** / ***Y*** | **1** | **2** | **3** | **4** | **5** | **6** | **Total** |
|---|---|---|---|---|---|---|---|
| **1** | 1 | 1 | 4 | 0 | 0 | 0 | 6 |
| **2** | 0 | 2 | 3 | 10 | 0 | 0 | 15 |
| **3** | 0 | 0 | 2 | 1 | 8 | 1 | 12 |
| **4** | 0 | 0 | 0 | 1 | 1 | 5 | 7 |
| **Total** | 1 | 3 | 9 | 12 | 9 | 6 | 40 |

Fíjate en que:

- La suma de las frecuencias absolutas de una columna es la frecuencia absoluta del valor de $X$ correspondiente a esa columna.
- La suma de las frecuencias absolutas de una fila es la frecuencia absoluta del valor de $Y$ correspondiente a esa fila.

## 2. DISTRIBUCIONES MARGINALES Y DISTRIBUCIONES CONDICIONADAS

### 2.1 Distribuciones marginales.

Cuando se estudian por separado las variables unidimensionales *X* e *Y* que forman la variable bidimensional (*X*, *Y*), se habla de ***distribuciones marginales***.

La última fila y la última columna de la tabla de doble entrada contienen, respectivamente, las frecuencias absolutas de las variables *X* e *Y*, consideradas por separado. Estas frecuencias reciben el nombre de ***frecuencias marginales***. Así:

- Frecuencia absoluta marginal del valor $x_i$ de la variable *X* es el número de veces que aparece el valor $x_i$, sin tener en cuenta cual es el valor de la variable *Y*. Se representa por: $f_{i\bullet}$.

$$f_{i\bullet} = f_{i1} + f_{i2} + \cdots + f_m = \sum_{j=1}^{m} f_{ij}$$

  Estas frecuencias son las que aparecen en la última fila de la tabla de doble entrada de la distribución bidimensional.

- Frecuencia absoluta marginal del valor $y_j$ de la variable *Y* es el número de veces que aparece el valor $y_j$, sin tener en cuenta cual es el valor de la variable *X*. Se representa por: $f_{\bullet j}$.

$$f_{\bullet j} = f_{1j} + f_{2j} + \cdots + f_{nj} = \sum_{i=1}^{n} f_{ij}$$

  Estas frecuencias son las que aparecen en la última columna de la tabla de doble entrada de la distribución bidimensional.

Estas frecuencias marginales cumplen que:

$$\sum_{i=1}^{n} f_{i\bullet} = \sum_{j=1}^{m} f_{\bullet j} = \sum_{i}\sum_{j} f_{ij} = N$$

*Ejemplo*: De esta forma, la frecuencia marginal del valor 5 de la variable *X* es 9 y la frecuencia marginal del valor 4 de la variable *Y* es 7. Las distribuciones marginales completas correspondientes a las variables unidimensionales del ejemplo anterior, *X* = "número de goles marcados" e *Y* = "número de goles recibidos" son:

| Variable *X* | $f_{i\bullet}$ |
|---|---|
| 1 | 1 |
| 2 | 3 |
| 3 | 9 |
| 4 | 12 |
| 5 | 9 |
| 6 | 6 |
| | 40 = *N* |

| Variable *Y* | $f_{\bullet j}$ |
|---|---|
| 1 | 6 |
| 2 | 12 |
| 3 | 15 |
| 4 | 7 |
| | 40 = *N* |

***Ejercicio* 1**: En una clase compuesta por 30 alumnos, se ha hecho un estudio sobre el número de horas diarias de estudio *X* y el número de suspensos *Y*, obteniéndose los siguientes resultados:

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

Construye la tabla estadística bidimensional de doble entrada, y las tablas de las distribuciones marginales.

*Solución*:

| *X* / *Y* | 0 | 1 | 2 | 3 | 4 | Totales |
|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 2 | 1 | 2 | 5 |
| 1 | 0 | 0 | 8 | 2 | 0 | 10 |
| 2 | 0 | 2 | 5 | 1 | 0 | 8 |
| 3 | 0 | 2 | 0 | 0 | 0 | 2 |
| 4 | 1 | 1 | 1 | 0 | 0 | 3 |
| 5 | 2 | 0 | 0 | 0 | 0 | 2 |
| Totales | 3 | 5 | 16 | 4 | 2 | 30 |

| $x_i$ | 0 | 1 | 2 | 3 | 4 | Total |
|---|---|---|---|---|---|---|
| $f_i$ | 3 | 5 | 16 | 4 | 2 | 30 |

| $y_i$ | 0 | 1 | 2 | 3 | 4 | 5 | Total |
|---|---|---|---|---|---|---|---|
| $f_i$ | 5 | 10 | 8 | 2 | 3 | 2 | 30 |

Considerando las distribuciones marginales, como distribuciones unidimensionales es posible calcular los siguiente parámetros:

- Medias marginales:

$$\bar{x} = \frac{\sum_i x_i \, f_{i\bullet}}{N} \qquad ; \qquad \bar{y} = \frac{\sum_j y_j \, f_{\bullet j}}{N}$$

donde *N* es el número total de pares.

En una distribución bidimensional al punto $(\bar{x}, \bar{y})$ se le llama centro de gravedad de la distribución.

- Varianzas

$$\sigma_x^2 = \frac{\sum_i (x_i - \bar{x})^2 \, f_{i\bullet}}{N} = \frac{\sum_i x_i^2 \, f_{i\bullet}}{N} - \bar{x}^2 \qquad ; \qquad \sigma_y^2 = \frac{\sum_j (y_j - \bar{y})^2 \, f_{\bullet j}}{N} = \frac{\sum_j y_j^2 \, f_{\bullet j}}{N} - \bar{y}^2$$

- Desviaciones típicas

$$\sigma_x = +\sqrt{\sigma_x^2} \qquad \sigma_y = +\sqrt{\sigma_y^2}$$

Veamos un nuevo parámetro:

- Covarianza: Se llama covarianza de una variable bidimensional (*X*, *Y*) a la media aritmética de los productos de las desviaciones de cada una de las variables respecto a sus medias respectivas. Se representa por $\sigma_{xy}$.

$$\sigma_{xy} = \frac{\sum_i \sum_j (x_i - \bar{x})(y_j - \bar{y}) \, f_{ij}}{N} = \frac{\sum_i \sum_j x_i y_i \, f_{ij}}{N} - \bar{x} \cdot \bar{y}$$

A la covarianza también se le llama varianza conjunta de las variables *X* e *Y*. Más adelante veremos el significado de este parámetro, así como su interpretación según el signo.

*Ejemplo*: Para las variables $X$ = "número de goles marcados" e $Y$ = "número de goles recibidos" del ejemplo que venimos siguiendo, podemos calcular sus parámetros.

| Variable $X$ | $f_{i\bullet}$ |
|---|---|
| 1 | 1 |
| 2 | 3 |
| 3 | 9 |
| 4 | 12 |
| 5 | 9 |
| 6 | 6 |
| | $40 = N$ |

Tabla de frecuencias marginales de la variable $X$

$$\bar{x} = \frac{1\cdot 1 + 2\cdot 3 + 3\cdot 9 + 4\cdot 12 + 5\cdot 9 + 6\cdot 6}{40} = 4{,}075$$

$$\sigma_x^2 = \frac{1^2\cdot 1 + 2^2\cdot 3 + 3^2\cdot 9 + 4^2\cdot 12 + 5^2\cdot 9 + 6^2\cdot 6}{40} - (4{,}075)^2 = 1{,}57$$

| Variable $Y$ | $f_{\bullet j}$ |
|---|---|
| 1 | 6 |
| 2 | 12 |
| 3 | 15 |
| 4 | 7 |
| | $40 = N$ |

Tabla de frecuencias marginales de la variable $Y$

$$\bar{y} = \frac{1\cdot 6 + 2\cdot 12 + 3\cdot 15 + 4\cdot 7}{40} = 2{,}575$$

$$\sigma_y^2 = \frac{1^2\cdot 6 + 2^2\cdot 12 + 3^2\cdot 15 + 4^2\cdot 7}{40} - (2{,}575)^2 = 0{,}89$$

Para calcular la covarianza, podemos escribir la tabla de doble entrada como una tabla simple:

| Variable $(X, Y)$ | $f_{ij}$ |
|---|---|
| (1, 1) | 1 |
| (2, 1) | 1 |
| (2, 2) | 2 |
| (3, 1) | 4 |
| (3, 2) | 3 |
| (3, 3) | 2 |
| (4, 2) | 10 |
| (4, 3) | 1 |
| (4, 4) | 1 |
| (5, 3) | 8 |
| (5, 4) | 1 |
| (6, 3) | 1 |
| (6, 4) | 5 |
| | $40 = N$ |

$$\sigma_{xy} = \frac{1\cdot 1\cdot 1 + 2\cdot 1\cdot 1 + 2\cdot 2\cdot 2 + \cdots + 6\cdot 3\cdot 1 + 6\cdot 4\cdot 5}{40} - 4{,}075\cdot 2{,}575 = 0{,}63$$

***Ejercicio* 2**: El número de horas dedicadas al estudio de una asignatura y la calificación final obtenida en el correspondiente examen por ocho personas vienen dados en la tabla de la derecha. Halla la media y la varianza de $X$, la media y la varianza de $Y$, y la covarianza.

*Solución*: $\bar{x} = 24$; $\sigma_x^2 = 36{,}75$; $\bar{y} = 7{,}75$; $\sigma_y^2 = 1{,}31$; $\sigma_{xy} = 5{,}75$

| $X$: Horas de estudio | $Y$: Calificación del examen |
|---|---|
| 20 | 6,5 |
| 16 | 6,0 |
| 34 | 8,5 |
| 23 | 7,0 |
| 27 | 9,0 |
| 32 | 9,5 |
| 18 | 7,5 |
| 22 | 8,0 |

**2.2 Distribuciones condicionadas.**

A veces nos interesa considerar un valor particular de una de las variables; por ejemplo, el valor $x_i$ de la variable $X$. ¿Pero cómo se distribuye la variable $Y$ en el caso de que $X$ tome el valor $x_i$? Es decir, en una distribución conjunta de dos variables $X$ e $Y$, se llama distribución condicionada a $x_i$ a la distribución de la variable $Y$ correspondiente a la subpoblación formada por todos los pares que toman el valor $x_i$ en la variable $X$.

Pero, cómo se distribuye $Y$ condicionando sus valores a que $X$ sea igual a $x_i$. De la tabla de doble entrada de la distribución bidimensional podemos obtener las frecuencias absolutas de los valores de la variable $Y$ condicionados a que $X$ sea $x_i$. Dichas frecuencias serán las que aparecen en la columna correspondiente al valor $x_i$ de la variable $X$, es decir: $f_{i1}, f_{i2}, \cdots, f_m$.

Análogamente, podríamos hablar de distribuciones de frecuencias de $X$ condicionadas a que $Y$ sea igual a $y_j$.

*Ejemplo*: En el ejemplo que venimos siguiendo, la distribución de frecuencias de la variable $X$ condicionada a que $Y$ sea 3 es:

| Variable $X$ | $f\,(X / Y = 3)$ |
|---|---|
| 1 | 0 |
| 2 | 0 |
| 3 | 2 |
| 4 | 1 |
| 5 | 8 |
| 6 | 1 |

***Ejercicio* 3**: En una clase de 30 alumnos del ejercicio anterior, construye las tablas de las distribuciones de frecuencias de $Y$ condicionada a que $X = 2$, y la de $X$ condicionada a que $Y$ sea 4.

*Solución*:

| Variable $Y$ | $f\,(Y / X = 2)$ |
|---|---|
| 0 | 2 |
| 1 | 8 |
| 2 | 5 |
| 3 | 0 |
| 4 | 1 |
| 5 | 0 |

| Variable $Y$ | $f\,(X / Y = 4)$ |
|---|---|
| 0 | 1 |
| 1 | 1 |
| 2 | 1 |
| 3 | 0 |
| 4 | 0 |

Par las distribuciones de frecuencias condicionadas, también se pueden definir los parámetros estadísticos ya conocidos (media, varianza, ...), de forma análoga a como lo hicimos para las distribuciones marginales.

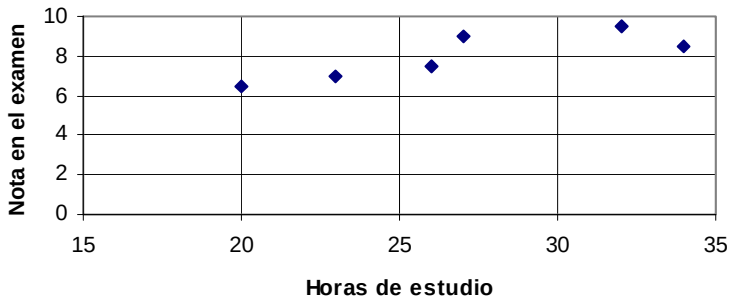
## 3. DIAGRAMAS DE DISPERSIÓN O NUBE DE PUNTOS

Podemos representar gráficamente la distribución bidimensional en un diagrama cartesiano. En el eje de abscisas representamos la variable estadística $X$ y en el eje de ordenadas la variable estadística $Y$.

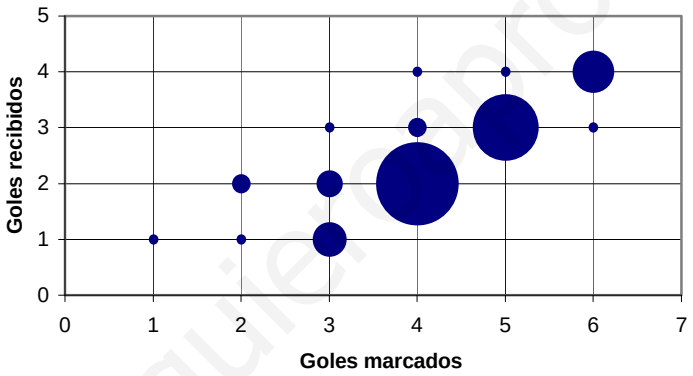
Considerando cada par de valores $(x_i, y_j)$ como las coordenadas de un punto, se consigue una gráfica denominada diagrama de dispersión o nube de puntos. Suelen dibujarse puntos de área proporcional a la frecuencia absoluta de cada par de valores que puede tomar la variable bidimensional $(X, Y)$.

*Ejemplo*: El número de horas dedicadas al estudio de una asignatura y la calificación final obtenida en el correspondiente examen por seis personas vienen dados en la tabla de la derecha. La nube de puntos o diagrama de dispersión es:

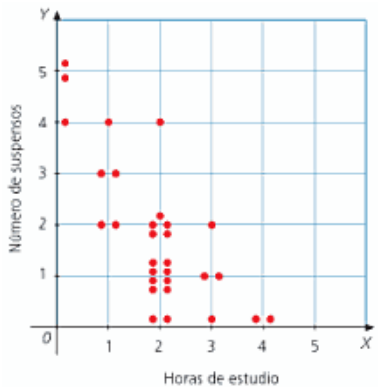
| *X*: Horas de estudio | *Y*: Calificación del examen |
|---|---|
| 20 | 6,5 |
| 26 | 7,5 |
| 34 | 8,5 |
| 23 | 7 |
| 27 | 9 |
| 32 | 9,5 |




*Ejemplo*: En el caso de que cada par de valores $(x_i, y_j)$ tenga una frecuencia mayor que uno, se suele dibujar el punto de modo que su tamaño sea proporcional a su frecuencia. Para el ejemplo anterior de *X* = "número de goles marcados" e *Y* = "número de goles recibidos" en 40 partidos, se obtiene el siguiente diagrama de dispersión:




***Ejercicio* 4**: Realiza un diagrama de dispersión para los datos del ejercicio anterior en el que se relacionaban el número de horas diarias de estudio *X* y el número de suspensos *Y*, de 30 alumnos.
*Solución*:



## 4. DEPENDENCIA O CORRELACIÓN

La etapa final de un estudio estadístico es el análisis de los datos con el fin de extraer conclusiones que puedan ser de interés. En especial, puede interesarnos estudiar si las dos variables unidimensionales que forman una variable bidimensional presentan algún tipo de relación entre ellas y cuáles son las características de esta relación.

Consideremos el siguiente ejemplo para entender mejor la relación entre variables. En una muestra de familias formadas por padre, madre y dos hijos, hemos estudiado las variables:

$X$ = estatura del padre (cm)
$Y$ = gasto anual en energía eléctrica (€)
$Z$ = consumo anual de energía eléctrica (kW · h)
$W$ = ingresos familiares anuales (€)

Los valores de $Y$ pueden determinarse exactamente a partir de los valores de $Z$ si conocemos las tarifas de la compañía eléctrica.

- Entre dos variables estadísticas existe ***dependencia funcional*** si están relacionadas de forma que sea posible determinar con exactitud los valores que toma una de ellas a partir de los que toma la otra.

Consideremos ahora las variables $W$ y $Z$. Los valores de $Z$ no pueden calcularse exactamente sólo conociendo los de $W$. Sin embargo, podemos suponer que consumirán menos energía eléctrica las familias con ingresos más modestos y, por el contrario, que consumirán más las familias con mayores recursos. Así pues, cabe esperar algún tipo de relación entre ambas variables, aunque no sea una relación exacta como en el caso anterior.

- Entre dos variables estadísticas existe ***dependencia estadística o correlación*** cuando los valores que toma una de ellas están relacionados con los valores que toma la otra, pero no de manera exacta.

Finalmente, parece razonable pensar que no existe ninguna relación entre los valores de $W$ y los de $X$.

- Dos variables estadísticas son ***independientes*** si no puede establecerse ninguna relación entre los valores que toma una de ellas y los que toma la otra.

***Ejercicio* 5**: Determina si entre los siguientes pares de variables existe dependencia funcional o estadística, o bien, si son independientes.

a) Talla de zapatos y estatura.
b) Color de cabello y profesión.
c) Radio y longitud de la circunferencia.
d) Cociente intelectual y peso.

*Solución*: a) Dependencia estadística. b) Independientes. c) Dependencia funcional. d) Independientes.

### 4.1 Interpretación gráfica de la relación entre variables.

Hemos visto que la estudiar la relación entre dos variables pueden darse tres casos: independencia, dependencia funcional y una situación intermedia a la que llamamos dependencia estadística o correlación.

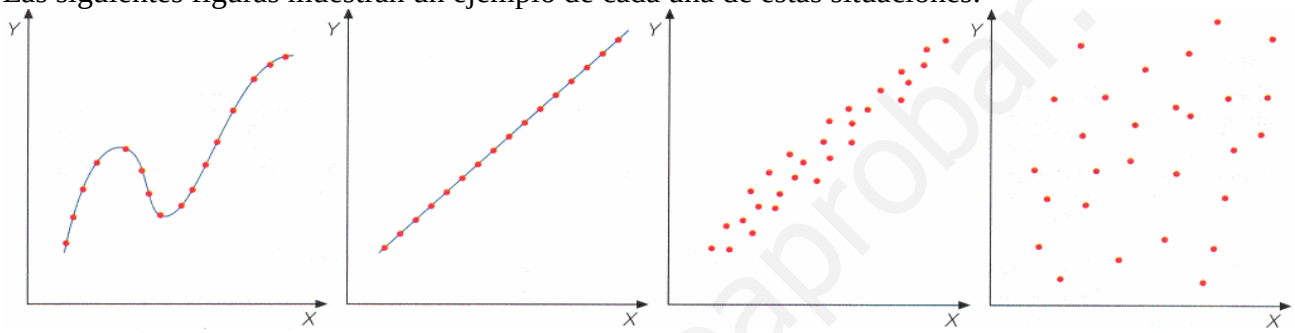
La relación existente entre dos variables queda reflejada en los diagramas de dispersión o nubes de puntos de la distribución bidimensional.

- Si los puntos de la nube se sitúan sobre una recta o una curva cuya expresión matemática podemos determinar, hablaremos de dependencia funcional entre las variables *X* e *Y*.

- Si los puntos de la nube se agrupan en torno a una posible recta, o curva, no muy definida pero reconocible, hablaremos de dependencia estadística o correlación entre las variables *X* e *Y*.

- Si los puntos de la nube no se agrupan en torno a ninguna curva, están completamente en desorden, hablaremos de independencia entre las variables *X* e *Y*.

Las siguientes figuras muestran un ejemplo de cada una de estas situaciones:




Dependencia funcional (curvilínea) | Dependencia funcional (lineal) | Dependencia estadística o correlación | Independencia

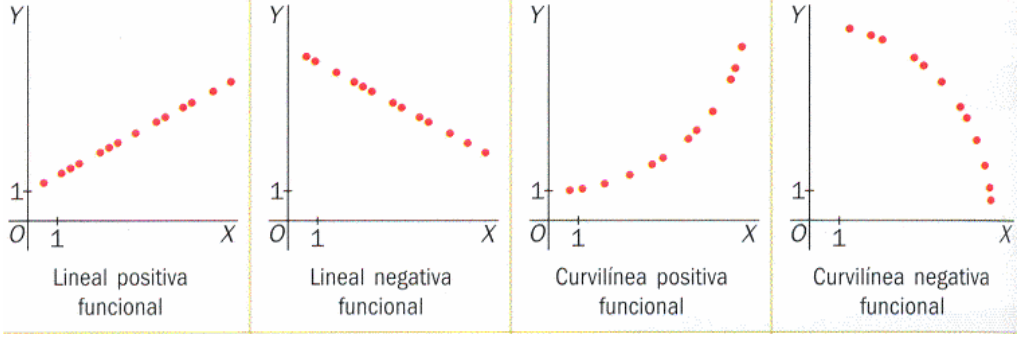
Entre los casos extremos de dependencia funcional e independencia existe una amplia gama de situaciones en que se da dependencia estadística o correlación. Por ello, al estudiar la relación entre las variables *X* e *Y*, es conveniente que tengamos en cuenta los siguientes aspectos:

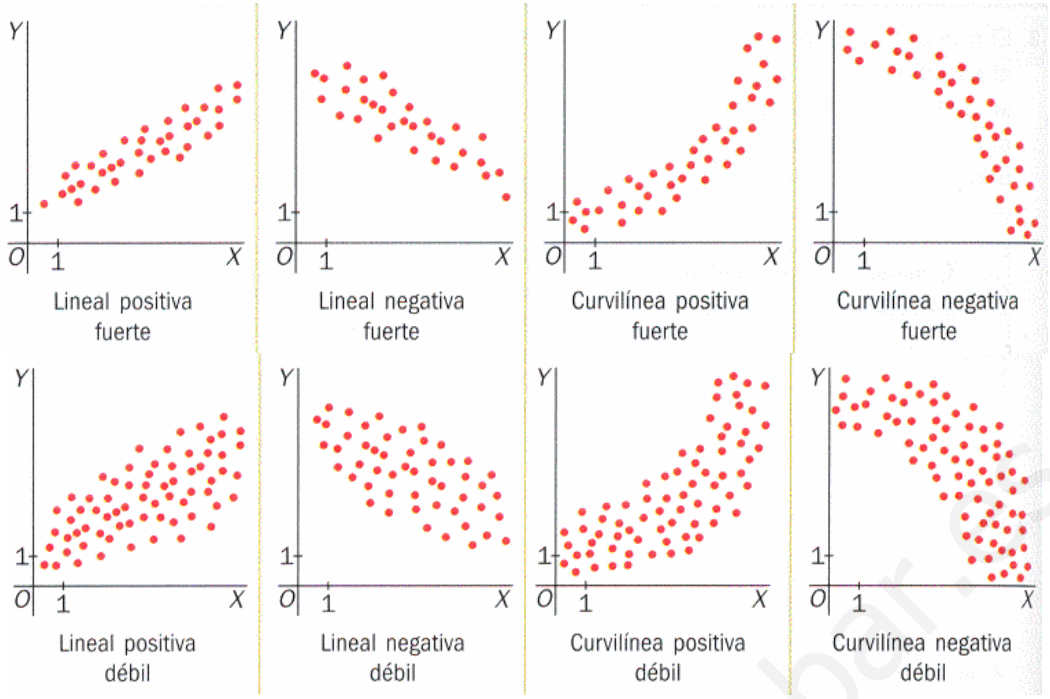
- Se dice que el grado de la correlación entre dos variables estadísticas es fuerte si la relación entre ambas se acerca a la dependencia funcional, y es débil si se acerca a la independencia.

- Entre dos variables estadísticas existe una correlación de sentido positivo cuando ambas aumentan conjuntamente, y una correlación de sentido negativo cuando una de ellas disminuye al aumentar la otra.

- Cuando los puntos del diagrama de dispersión tienden a agruparse en torno a una línea recta, decimos que existe una correlación de tipo lineal. Si los puntos se agrupan en torno a cualquier otro tipo de curva, decimos que existe una correlación de tipo curvilíneo.

En las siguientes gráficas se muestran ejemplos aclaratorios de diferentes tipos de correlación.



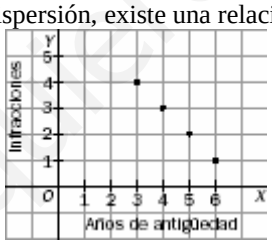



***Ejercicio* 6**: En una empresa de transportes trabajan 4 conductores. Los años de antigüedad de sus permisos de conducir y las infracciones cometidas en el último año por cada uno son los siguientes:

| | | | | |
|---|---|---|---|---|
| *X*: Años de antigüedad | 3 | 4 | 5 | 6 |
| *Y*: Infracciones | 4 | 3 | 2 | 1 |

Representa gráficamente los datos anteriores. Razona si estos muestran correlación positiva o negativa.

*Solución*: Según se aprecia en el diagrama de dispersión, existe una relación lineal negativa funcional.

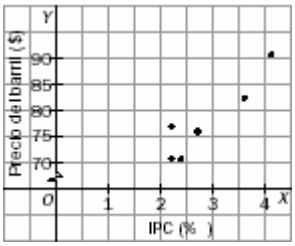



***Ejercicio* 7**: En la siguiente tabla se recoge la evolución del IPC (índice de precios al consumo) y el precio del barril de petróleo (brent) durante el segundo semestre de 2007.

| | | | | | | |
|---|---|---|---|---|---|---|
| IPC (%) | 2,4 | 2,2 | 2,2 | 2,7 | 3,6 | 4,1 |
| Precio del barril ($) | 71,54 | 77,01 | 70,73 | 76,87 | 82,50 | 90,16 |

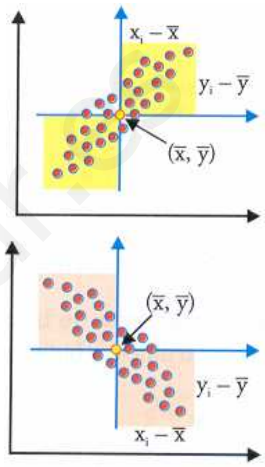
¿Se puede asegurar que la evolución del IPC está directamente relacionada con el precio del petróleo?

*Solución*: Sí, existe una correlación lineal positiva fuerte.



Como acabamos de ver, si se representan sobre unos ejes de coordenadas la nube de puntos correspondiente a la variable bidimensional, se puede apreciar de forma visual la existencia o no de relación entre las dos variables. Si la nube de puntos se condensa en torno a una recta, existe una correlación lineal entre las variables. Para muchos fenómenos y aplicaciones es de gran interés cuantificar de forma más objetiva y precisa esta correlación.

La covarianza, $\sigma_{xy}$, es un indicador numérico del grado de relación lineal que existe entre las dos variables. Además su signo nos indica el sentido de la correlación. Veámoslo. Si se calcula el centro de gravedad ( $\bar{x}$ , $\bar{y}$ ) y se toman unos ejes con el origen en este centro, se observa:




- Si ambas variables tienen una relación directa, los puntos están en el 1er y 3er cuadrantes, y por tanto los productos $(x_i - \bar{x})( y_i - \bar{y})$ mayoritariamente son positivos, con lo cual la covarianza tomará un valor positivo.

- Si la relación es inversa, los puntos están en el 2º y 4º cuadrantes, y por tanto los productos $(x_i - \bar{x})( y_i - \bar{y})$ mayoritariamente son negativos, con lo cual la covarianza tomará un valor negativo, aunque su valor absoluto sea alto.

- En el caso de que exista poca relación entre las variables, las diferencias serán aleatoriamente positivas y negativas y tenderán a compensarse, con lo cual la covarianza tendrá un valor pequeño en términos absolutos.

## 5. CORRELACIÓN LINEAL. COEFICIENTE DE CORRELACIÓN DE PEARSON

A pesar de que la covarianza es un indicador de la asociación lineal entre las dos variables, esta presenta dificultades:

- Puede verse influenciada por los puntos de la nube alejados del centro de gravedad, que distorsionan el resultado.

- Su valor depende de las unidades de medida de las variables y, en consecuencia, necesitamos un indicador que no dependa de las unidades.

Por tanto, la covarianza no indica de forma precisa la medida de la relación entre las dos variables. Para salvar estas dificultades, se define un nuevo parámetro que nos cuantifica correctamente la dependencia. Es el llamado coeficiente de correlación lineal de Pearson.

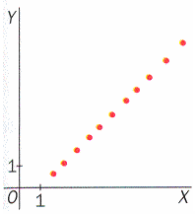
El coeficiente de correlación de Pearson se representa por *r* y es el cociente entre la covarianza y el producto de las desviaciones típicas marginales de *X* e *Y*:

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

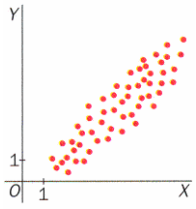
Dicho coeficiente es adimensional, es decir, no depende de las unidades utilizadas. Además el signo del coeficiente $r$ viene dado por el signo de la covarianza, ya que las desviaciones típicas son siempre positivas.

El coeficiente de correlación lineal de Pearson permite analizar el grado de aproximación de la nube de puntos a una línea recta de la siguiente manera:

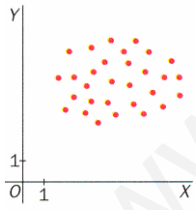
- Si $-1 < r < 0$, existe correlación lineal negativa, y será más fuerte cuanto más se aproxime $r$ a $-1$.
- Si $0 < r < 1$, existe correlación lineal positiva, y será más fuerte cuanto más se aproxime $r$ a 1.
- Si $r = 1$ ó $r = -1$, la correlación es una dependencia lineal exacta (dependencia funcional).
- Si $r = 0$, no existe correlación lineal o las variables no están correlacionadas linealmente. Esto no excluye que las variables estadísticas puedan estar relacionadas por una correlación curvilínea.



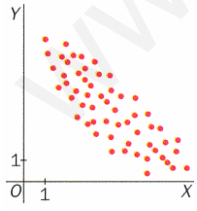

$r = 1$. Todos los valores de la variable $(X, Y)$ se encuentran sobre una recta. Las variable $X$ e $Y$ están en dependencia funcional lineal directa.



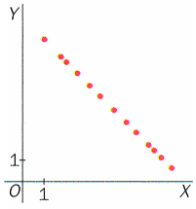

Las variables $X$ e $Y$ están en dependencia aleatoria directa, tanto más fuerte cuanto más se aproxima $r$ a 1, y más débil cuanto más se aproxima $r$ a cero.




Las variables $X$ e $Y$ son aleatoriamente independientes y $r \approx 0$. Las variables $X$ e $Y$ son incorreladas.

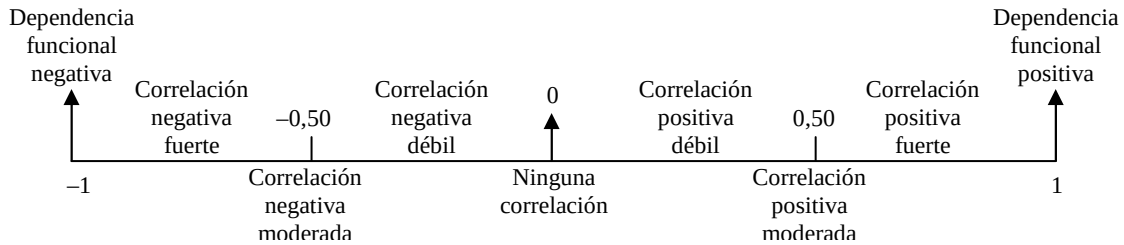



Las variables $X$ e $Y$ están en dependencia aleatoria inversa, tanto más fuerte cuanto más se aproxima $r$ a $-1$, y más débil cuanto más se aproxima $r$ a cero.




$r = -1$. Todos los valores de la variable $(X, Y)$ se encuentran sobre una recta. Las variable $X$ e $Y$ están en dependencia funcional lineal inversa.

Así, podemos resumir en el siguiente diagrama el grado de correlación lineal:




*Ejemplo*: La cotización en bolsa (en cientos de euros) de dos empresas *A* y *B*, a lo largo de 6 días de sesión son los siguientes:

| | | | | | | |
|---|---|---|---|---|---|---|
| $X$ = Empresa $A$ | 8 | 7 | 6 | 5 | 7 | 8 |
| $Y$ = Empresa $B$ | 6 | 5 | 4,5 | 4 | 4,5 | 5 |

Calcula el coeficiente de correlación de Pearson e interpreta el resultado.

En primer lugar debemos calcular las medias y desviaciones típicas de cada una de las empresas, así como la covarianza:

$$\bar{x} = \frac{41}{6} = 6{,}833 \quad ; \quad \bar{y} = \frac{29}{6} = 4{,}833$$

$$\sigma_x^2 = \frac{287}{6} - 6{,}833^2 = 1{,}143 \quad \Rightarrow \quad \sigma_x = \sqrt{1{,}143} = 1{,}069$$

$$\sigma_y^2 = \frac{142{,}5}{6} - 4{,}833^2 = 0{,}392 \quad \Rightarrow \quad \sigma_y = \sqrt{0{,}392} = 0{,}626$$

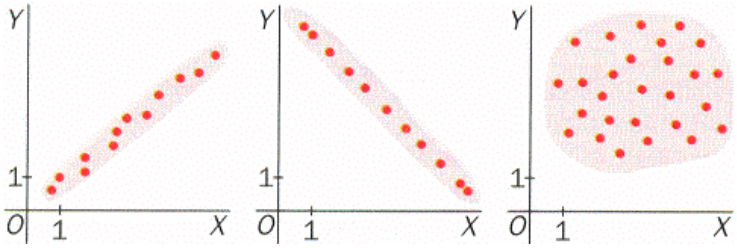
$$\sigma_{xy} = \frac{201{,}5}{6} - 6{,}833 \cdot 4{,}833 = 0{,}559$$

Así, el coeficiente de correlación de Pearson es:

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{0{,}559}{1{,}069 \cdot 0{,}626} = 0{,}835$$

El valor de este coeficiente indica una correlación lineal positiva fuerte por su proximidad a 1, lo que debe interpretarse como que ambos valores cotizan al alza o a la baja simultáneamente.

***Ejercicio* 8**: Los números 0; 0,8 y 1 son los valores absolutos del coeficiente de correlación de las distribuciones bidimensionales cuyas nubes de puntos aparecen a continuación:




Asigna a cada diagrama su coeficiente de correlación, cambiando el signo cuando sea necesario.
*Solución*: Primero: 0,8; Segundo: –1; Tercero: 0

***Ejercicio* 9**: Las puntuaciones en Matemáticas y Física de siete alumnos han sido las siguientes:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Matemáticas | 8 | 8 | 6 | 7 | 8 | 6 | 2 |
| Física | 7 | 7,5 | 5 | 7 | 7,5 | 5 | 7 |

Calcula el coeficiente de correlación de esas dos variables para los siete alumnos.

*Solución*: Medias: $\bar{x} = 6{,}43$; $\bar{y} = 6{,}57$; Varianzas: $\sigma_x^2 = 3{,}959$; $\sigma_y^2 = 1{,}031$; Covarianza: $\sigma_{xy} = 0{,}4694$; Coeficiente de correlación: $r = 0{,}232$

## 6. REGRESIÓN LINEAL

Uno de los objetivos que se persiguen, al estudiar conjuntamente dos variables *X* e *Y*, es encontrar alguna manera de predecir los valores de una de ellas conocidos los de la otra. En este sentido, es lógico pensar que, si hay una curva en torno a la cual se agrupan los puntos de un diagrama de dispersión, ésta ha de dar una aproximación de los valores reales.

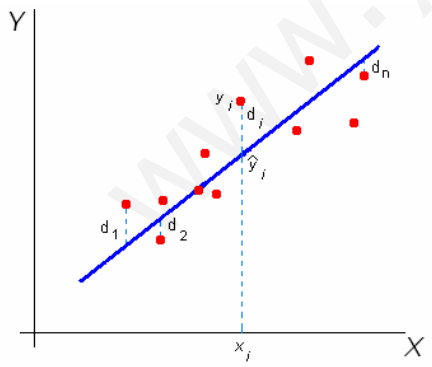
Al análisis que pretende determinar la curva que mejor aproxima un diagrama de dispersión se le llama **regresión**. En este curso estudiaremos el caso de la regresión lineal, es decir, la determinación de la recta que mejor aproxima una nube de puntos.

Es fácil hallar una recta que se ajuste aproximadamente a una distribución. Basta con dibujar la que a simple vista nos parezca más representativa de la nube de puntos. Sin embargo, éste es un método subjetivo. Para evitar este problema se considera algún criterio que permita determinar objetivamente la recta que se ajusta mejor a la distribución. Estas rectas se determinan haciendo que se cumplan las siguientes condiciones:

a) Tienen que pasar por el centro de gravedad ($\bar{x}$, $\bar{y}$).

b) Las sumas de los cuadrados de las distancias, $\sum d_i^2$, debe ser mínima, siendo $d_i = y_i - \hat{y}_i$, donde $y_i$ es el valor de la ordenada de cada punto de la nube e $\hat{y}_i$ es la ordenada del punto de la recta (criterio de los *mínimos cuadrados*).

Aclaremos esta segunda condición:




Sea $\hat{y} = mx + n$ la ecuación de la recta que mejor se aproxima a la nube de puntos. Al valor $x_i$ de la variable *X* le corresponde el valor $\hat{y}_i = mx_i + n$. El error cometido por la aproximación es la diferencia $d_i = y_i - \hat{y}_i$. La condición para calcular *m* y *n* es que la suma de los errores al cuadrado sea mínima. Se demuestra que dicha suma es mínima si:

$$m = \frac{\sigma_{xy}}{\sigma_x^2} \qquad \text{y} \qquad n = \bar{y} - \frac{\sigma_{xy}}{\sigma_x^2}\bar{x}$$

La ecuación de la recta de regresión de *Y* sobre *X* es:

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

Dicha recta permite predecir, para un valor *x*, el valor *y* que cabe esperar que presente un individuo de la población. Sin embargo, puesto que el resultado es sólo una predicción, el valor obtenido no será en general el valor real, sino una estimación de éste.

Análogamente, si nos interesa hacer predicciones de un valor *x* a partir de un valor *y*, deberemos intercambiar el papel de ambas variables. Consideraremos, en este caso, la recta de regresión de *X* sobre *Y*:

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

En este caso, obtendremos una estimación del valor de *x* proporcionada por la recta de regresión para un valor conocido de *y*.

### 6.1 Bondad del ajuste de la recta de regresión de *Y* sobre *X*.

El coeficiente de correlación lineal indica el grado de linealidad entre las dos variables, pero para analizar la bondad del ajuste de la recta de regresión se utiliza un parámetro nuevo llamado coeficiente de determinación.

Se llama coeficiente de determinación al cuadrado del coeficiente de correlación lineal, $r^2$. Dicho coeficiente, $r^2$, indica el porcentaje de la variación de *Y* que puede ser explicada por *X*.

Teniendo en cuenta en cuenta que $0 \le r^2 \le 1$, la interpretación de esta medida es 1a siguiente:

- Si $r^2 = 0$, significa que no existe tal relación lineal entre las variables (puede existir otro tipo de relación o no haber ninguna entre las dos variables). La recta es la que, según el criterio de los mínimos cuadrados, mejor se ajusta a los puntos de la nube; aún así, la bondad del ajuste es nula, porque *Y* no es una función lineal de *X*. La recta no proporciona ninguna información sobre e1 comportamiento de *Y* en función de *X* y no tendría sentido utilizarla para analizar la influencia de *X* sobre *Y*, ni para predecir el valor de *Y*, dado *X*.

- Si $r^2 = l$, significa que el ajuste es perfecto. No hemos cometido ningún error al realizarlo (todos los valores de los errores son nulos). La recta pasa por todos los puntos de la nube, obviamente porque éstos están alineados. La relación lineal entre las variables es exacta, y, por tanto, la rectal proporciona toda la información sobre el comportamiento de *Y* en función de *X*, para la muestra considerada.

- Si $0 < r^2 < 1$, en función de a cual de las dos situaciones anteriores nos acerquemos, hablaremos de ajuste malo o bueno. Además, como hemos dicho anteriormente, $r^2$, indica el porcentaje de la variación de *Y* que puede ser explicada por *X*. Así, un valor concreto de $r^2$ se puede interpretar en los siguientes términos: si $r^2$ =0,85 significa que la recta obtenida explica en un 85 % el comportamiento de *Y* en función de *X*. El 15 % restante de la variación de *Y* puede deberse al azar o a la influencia sobre *Y* de otras variables distintas.

Un coeficiente de determinación alto implica la relación lineal entre las variables es fuerte, pero eso no tiene por qué implicar que los cambios en una variable se expliquen por la otra variable. Así, si se considera la variable (*X*, *Y*), donde *X* indica la diferencia entre las temperaturas máxima y mínima diarias durante el mes de noviembre, e *Y*, el índice de bajas por catarro de una empresa, el cambio brusco de temperaturas puede explicar un aumento de las bajas por catarro, pero, evidentemente, las bajas por catarro no determinan que haya cambios bruscos de temperatura.

## 6.2 Valoración de las predicciones.

La recta de regresión nos permite predecir valores de una variable a partir de los de la otra. No obstante, hay que tener siempre presente que existen las siguientes limitaciones:

- Las predicciones realizadas a partir de una recta de regresión no son fiables si entre $X$ e $Y$ no hay un alto grado de correlación lineal, es decir, si $r$ no es, en valor absoluto, cercano a 1.

- Las predicciones deben hacerse con valores próximos a los pares considerados. Las estimaciones obtenidas para valores próximos al centro de gravedad de la distribución son más fiables que las obtenidas para valores muy alejados de él.

- La fiabilidad de una recta de regresión es mayor cuanto mayor sea el número de datos considerados para calcularla.

***Ejercicio* 13:** ¿Cuál sería la fiabilidad de un ajuste bidimensional con $r = 0{,}7$? ¿Y con $r = -0{,}8$? ¿Y con $r = 0{,}9$?
*Solución*: 49 %, 64 % y 81 %, respectivamente.

*Ejemplo*: Las notas obtenidas por cinco alumnos en matemáticas y música son las siguientes:

| Matemáticas ($X$) | 6 | 4 | 8 | 5 | 3,5 |
|---|---|---|---|---|---|
| Música ($Y$) | 6,5 | 4,5 | 7 | 5 | 4 |

a) Determina la recta de regresión de $Y$ sobre $X$ y represéntala.
b) Halla la nota de música de un alumno que tiene 7,5 en matemáticas.
c) Determina la recta de regresión de $X$ sobre $Y$ y represéntala.
d) Halla la nota de matemáticas de un alumno que tiene 6 en música.

Primero representamos la nube de puntos. Los datos se agrupan en torno a una recta, por tanto tiene sentido calcular la recta de regresión. Para hallar las rectas de regresión calculamos los parámetros marginales de $X$ e $Y$ y la covarianza.


$$\bar{x} = \frac{26{,}5}{5} = 5{,}3 \quad ; \quad \bar{y} = \frac{27}{5} = 5{,}4$$

$$\sigma_x^2 = \frac{153{,}25}{5} - 5{,}3^2 = 2{,}56 \quad \Rightarrow \quad \sigma_x = \sqrt{2{,}56} = 1{,}6$$

$$\sigma_y^2 = \frac{152{,}5}{5} - 5{,}4^2 = 1{,}34 \quad \Rightarrow \quad \sigma_y = \sqrt{1{,}34} = 1{,}16$$

$$\sigma_{xy} = \frac{152}{5} - 5{,}3 \cdot 5{,}4 = 1{,}78$$

a) Recta de regresión de $Y$ sobre $X$:

$$y - 5{,}4 = 0{,}7\,(x - 5{,}3) \quad \Rightarrow \quad y = 0{,}7\,x + 1{,}69$$

b) Se sustituye $x = 7{,}5$ en la ecuación obtenida:

$$y = 0{,}7 \cdot 7{,}5 + 1{,}69 = 6{,}94$$

Es decir, si un alumno obtuvo un 7,5 en matemáticas, se espera que obtenga un 6,94 en música.

c) Recta de regresión de $X$ sobre $Y$:

$$x - 5{,}3 = 1{,}33\,(y - 5{,}4) \quad \Rightarrow \quad x = 1{,}33y - 1{,}88$$

d) Se sustituye $y = 6$ en la ecuación obtenida:

$$x = 1{,}33 \cdot 6 - 1{,}88 = 6{,}1$$

Es decir, si un alumno obtuvo un 6 en música, se espera que obtenga un 6,1 en matemáticas.

***Ejercicio* 10**: En cierto país, el tipo de interés y el índice de la Bolsa en los últimos seis meses vienen dados por la siguiente tabla:

| Tipo de interés (%) | 8 | 7,5 | 7,2 | 6 | 5,5 | 5 |
|---|---|---|---|---|---|---|
| Índice | 120 | 130 | 134 | 142 | 150 | 165 |

Halla el índice previsto de la Bolsa en el séptimo mes, suponiendo que el tipo de interés en ese mes fue del 4,1 %, y analiza la fiabilidad de la predicción, según el valor del coeficiente de correlación.

*Solución*: $\bar{x} = 6{,}53$; $\sigma_x = 1{,}12$; $\bar{y} = 140{,}17$; $\sigma_y = 14{,}48$; $\sigma_{xy} = -15{,}01$; $y = -12{,}008x + 218{,}58$;

$y\,(4{,}1) = 169{,}35$ es el índice de Bolsa esperado para el siguiente mes; $r = -0{,}93 \Rightarrow$ el resultado obtenido es fiable.

***Ejercicio* 11**: Como consecuencia de un estudio estadístico realizado sobre 100 universitarios, se ha obtenido una estatura media de 155 cm, con una desviación típica de 15,5 cm. Además se obtuvo la recta de regresión $y = 80 + 1{,}5x$ (siendo $X$ el peso e $Y$ la altura). Determina el peso medio de estos 100 universitarios.

*Solución*: $\bar{x} = 50$ kg.

### 6.3 Comparación de las dos rectas de regresión.

En general, la recta de regresión de $Y$ sobre $X$ y la de $X$ sobre $Y$ no coinciden. Sin embargo, siempre se cumple que:

- Las rectas de regresión se cortan en el centro de gravedad, $(\bar{x}, \bar{y})$.
- Las pendientes de las rectas de regresión son del mismo signo y coinciden en signo con el coeficiente de correlación.
- El ángulo que forman las dos rectas de regresión varía según sea la correlación que hay entre las variables:
    - Si $|r|$ es próximo a 1, las rectas prácticamente coinciden. Coinciden exactamente cuando hay dependencia funcional entre las variables $X$ e $Y$ ($r = 1$).
    - Si $r$ es próximo a cero, es decir, la correlación es casi nula, el ángulo que forman las rectas es casi un ángulo recto. Si $X$ e $Y$ son independientes, las rectas son perpendiculares entre sí y paralelas a los ejes.
- Observa que la pendiente de la recta de regresión de $Y$ sobre $X$ es:

$$B = \frac{\sigma_{xy}}{\sigma_x^2}$$

Mientras que la pendiente de la recta de regresión de *X* sobre *Y* es:

$$B' = \frac{\sigma_{xy}}{\sigma_y^2}$$

Por tanto se cumple que el producto de las dos pendientes es el cuadrado del coeficiente de correlación lineal:

$$B \cdot B' = \frac{\sigma_{xy}}{\sigma_x^2} \cdot \frac{\sigma_{xy}}{\sigma_y^2} = \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2} = r^2$$

*Ejemplo*: Hemos calculado las rectas de regresión de *Y* sobre *X* y de *X* sobre *Y* en una distribución bidimensional, obteniendo las expresiones siguientes:

$$y = 0{,}16x - 0{,}1$$
$$x = 5{,}44y + 8{,}77$$

¿Cuál es el coeficiente de correlación de Pearson de la distribución?

Para hallar el coeficiente de Pearson utilizamos la propiedad anteriormente descrita:

$$B \cdot B' = r^2$$

El coeficiente *B* es la pendiente de la recta de regresión de *Y* sobre *X*, y el coeficiente *B*' es la pendiente de la recta de regresión de *X* sobre *Y*:

$$B = 0{,}16 \quad \text{y} \quad B' = 5{,}44 \quad \Rightarrow \quad B \cdot B' = 0{,}16 \cdot 5{,}44 = 0{,}8704 = r^2$$

Entonces:

$$r = \sqrt{8704} = 0{,}933$$

***Ejercicio* 12:** De una distribución bidimensional conocemos los siguientes resultados:

Recta de regresión de *Y* sobre *X*: $y = 8{,}7 - 0{,}76x$

Recta de regresión de *X* sobre *Y*: $y = 11{,}36 - 1{,}3x$

a) Calcula el centro de gravedad de la distribución.
b) Halla el coeficiente de correlación.

*Solución*: a) El centro de gravedad es $(\bar{x}, \bar{y}) = (4{,}93; 4{,}95)$; b) $r = 0{,}76$

## 8. RECTA DE TUKEY

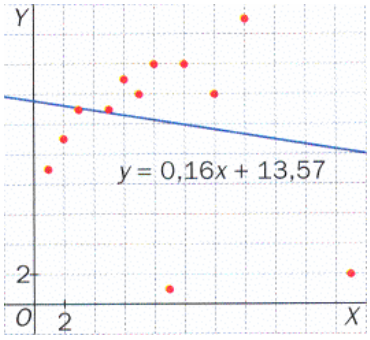
En algunos casos, la recta de regresión se ajusta muy mal a la nube de puntos, a pesar de que a simple vista los puntos parecen indicar una correlación lineal.

En estos casos se ajustan los datos a un modelo lineal mediante la recta de Tukey.

*Ejemplo*: Ajustar un modelo lineal a la distribución bidimensional dada por la siguiente tabla.

| *X* | 1 | 2 | 3 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 14 | 21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Y* | 9 | 11 | 13 | 13 | 15 | 14 | 16 | 1 | 16 | 14 | 19 | 2 |

La variable $Y$ toma dos valores, 1 y 2, que están muy alejados del resto. En la figura se ha representado la recta de regresión de $Y$ sobre $X$ y se aprecia su mal ajuste de la nube de puntos.




La recta de Tukey se calcula del siguiente modo:

1. Se ordenan los datos en orden creciente de las abcisas.

2. Se divide el conjunto ordenado de los datos en tres grupos:

$$G_1 = \{(1, 9), (2, 11), (3, 13), \{5, 13)\} \quad ; \quad G_2 = \{(6, 15), (7, 14), (8, 16), (9, 1)\}$$
$$G_3 = \{(10, 16), (12, 14), (14, 19), (21, 2)\}$$

3. Para cada grupo $G_i$ se halla el punto $P_i = (x_i, y_i)$; donde $x_i$ e $y_i$ son, respectivamente, las medianas de las abscisas y de las ordenadas del grupo $G_i$, es decir:

$$\left.\begin{array}{l}\text{Abscisas de } G_1\text{: } (1, 2, 3, 5) \Rightarrow x_1 = 2{,}5 \\ \text{Ordenadas de } G_1\text{: } (9, 11, 13, 13) \Rightarrow y_1 = 12\end{array}\right\} \Rightarrow P_1 = (2{,}5;\ 12)$$

$$\left.\begin{array}{l}\text{Abscisas de } G_2\text{: } (6, 7, 8, 9) \Rightarrow x_2 = 7{,}5 \\ \text{Ordenadas de } G_2\text{: } (1, 14, 15, 16) \Rightarrow y_2 = 14{,}5\end{array}\right\} \Rightarrow P_2 = (7{,}5;\ 14{,}5)$$

$$\left.\begin{array}{l}\text{Abscisas de } G_3\text{: } (10, 12, 14, 21) \Rightarrow x_3 = 13 \\ \text{Ordenadas de } G_3\text{: } (2, 14, 16, 19) \Rightarrow y_3 = 15\end{array}\right\} \Rightarrow P_3 = (13;\ 15)$$

4. La recta de Tukey pasa por el baricentro del triángulo $P_1$, $P_2$, $P_3$ y tiene la pendiente de la recta que pasa por $P_1$ y $P_3$.

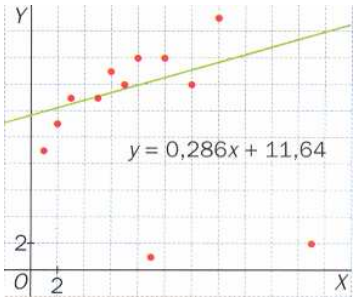
$$\text{Baricentro: } x_G = \frac{2{,}5+7{,}5+13}{3} = 7{,}67 \quad ; \quad y_G = \frac{12+14{,}5+15}{3} = 13{,}83$$

El baricentro tiene por coordenadas $G = (7{,}67;\ 13{,}83)$.

La pendiente de la recta que pasa por $P_1$ y $P_3$ es: $m = \dfrac{15-12}{13-2{,}5} = 0{,}286$

La ecuación de la recta de Tukey es:

$$y - 13{,}83 = 0{,}286\,(x - 7{,}67) \quad \Rightarrow \quad y = 0{,}286x + 11{,}64$$



*Nota*: Observa que el número de datos en este caso es $n = 12$, múltiplo de 3, y, por tanto, cada grupo está formado por 4 datos. Si el número de datos $n$ no es múltiplo de 3, puede ocurrir que:

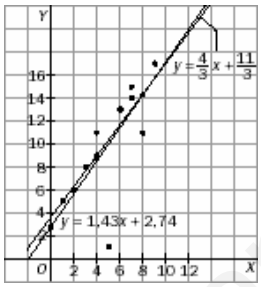
- Sea múltiplo de 3 más 1; en este caso, el grupo $G_2$ se deja con un dato más.
- Sea múltiplo de 3 más 2; en este caso, el grupo $G_2$ se deja con un dato menos.

***Ejercicio* 14**: Sea la variable bidimensional dada por la siguiente tabla.

| $X$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| $Y$ | 5 | 6 | 8 | 11 | 1 | 13 | 14 | 14 | 17 |

a) Halla la recta de Tukey.
b) Halla la recta de regresión de $Y$ sobre $X$.
c) Representa la nube de puntos y las dos rectas obtenidas.

*Solución*: a) $y = \frac{4}{3}x + \frac{11}{3}$; b) $y = 1{,}43x + 2{,}74$; c)



## *ANEXO*

## *EXTENSIONES DEL MODELO LINEAL*

El modelo de regresión lineal, estudiado para detectar dependencia lineal entre las variables, también se puede aplicar para detectar otros tipos de dependencia no lineales, como la exponencial o la potencial.

### *Dependencia exponencial*

En general, si dos variables estadísticas $X$ e $Y$ están relacionadas por un modelo exponencial:

$$y = ke^{ax}$$

al tomar logaritmos, se obtiene:

$$\ln y = ax + b$$

con $k = e^b$.

Por tanto, para encontrar dependencia exponencial en una lista de datos $(x_1, y_1)$, $(x_2, y_2)$, ..., $(x_n, y_n)$, se aplica una regresión lineal a los datos: $(x_1, \ln y_1)$, $(x_2, \ln y_2)$, ..., $(x_n, \ln y_n)$.

Si la recta de regresión es $\ln Y = a\ X + b$, entonces la dependencia exponencial en los datos originales es:

$$y = k\ e^{ax}$$

con $k = e^b$.

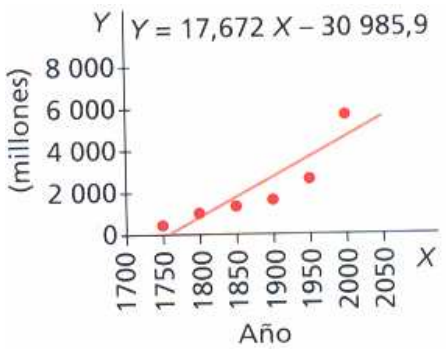
*Ejemplo*: Encuentra la dependencia exponencial de la siguiente lista de datos sobre la población mundial (en millones de habitantes):

| Año $X$ | 1750 | 1800 | 1850 | 1900 | 1950 | 1997 |
|---|---|---|---|---|---|---|
| Población $Y$ | 728 | 949 | 1171 | 1608 | 2516 | 5870 |

Si se aplica el modelo de regresión lineal a la variable bidimensional $(X, Y)$ se obtiene, tras hacer los correspondientes cálculos, la recta de regresión:

$$y = 17{,}672\ (x - 1874{,}5) + 2140{,}3 \quad \Rightarrow \quad y = 17{,}672\ x - 30985{,}9$$

con coeficiente de correlación $r = 0{,}848$.



Si se toman logaritmos en la población, se obtiene la siguiente tabla:

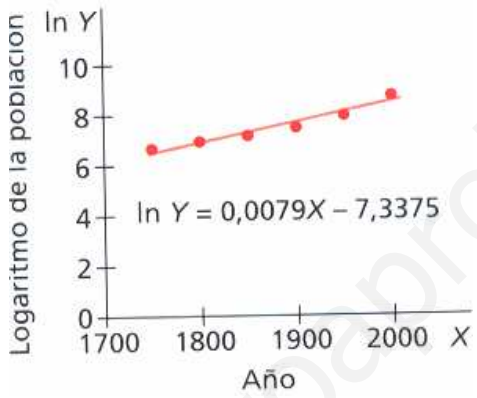
| $X$ | 1750 | 1800 | 1850 | 1900 | 1950 | 1997 |
|---|---|---|---|---|---|---|
| Ln $Y$ | 6,59 | 6,86 | 7,07 | 7,38 | 7,83 | 8,68 |

Si se aplica ahora el modelo de regresión lineal a la variable bidimensional ($X$, ln $Y$) se obtiene, tras hacer los correspondientes cálculos, la siguiente recta de regresión:

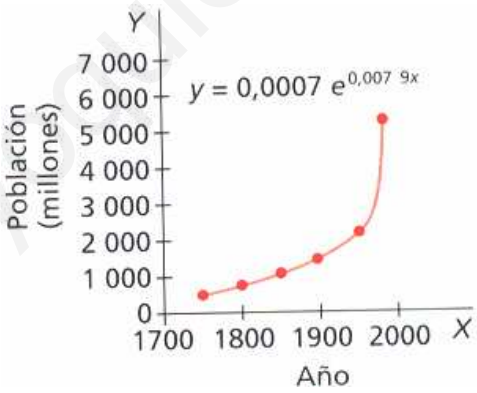
$$\ln y = 0{,}0079x - 7{,}3375$$

con coeficiente de correlación $r = 0{,}959$, de donde se obtiene la dependencia exponencial:

$$y = e^{0{,}0079x - 7{,}3375} = e^{-7{,}3375}\, e^{0{,}0079x}$$




El ajuste exponencial es mucho más fuerte que el lineal.




***Ejercicio*:** Basándote en la siguiente tabla de datos, sobre el crecimiento de la población española entre 1900 y 1981, encuentra su ley exponencial de crecimiento. ¿Cuál es la población española estimada en los años 2000 y 2010?

| Año | 1900 | 1920 | 1940 | 1960 | 1981 |
|---|---|---|---|---|---|
| Población | 18,8 | 22,0 | 26,4 | 30,8 | 37,7 |

*Solución*: $y = 1{,}493 \cdot 10^{-6}\, e^{0{,}0086x}$; $y\,(2000) = 44{,}04$; $y\,(2010) = 48{,}00$

### *Dependencia potencial*

En general, si dos variables estadísticas $X$ e $Y$ están relacionadas por un modelo potencial:

$$y = k\,x^a$$

con $k > 0$, al tomar logaritmos, se obtiene:

$$\ln y = \ln k + a \ln x$$

de donde se deduce que las variables (ln $X$, In $Y$) están relacionadas por un modelo lineal.

Para encontrar dependencia potencial en una lista de datos $(x_1, y_1)$, $(x_2, y_2)$, ..., $(x_n, y_n)$, se aplica una regresión lineal a los datos: $(\ln x_1, \ln y_1)$, $(\ln x_2, \ln y_2)$, ..., $(\ln x_n, \ln y_n)$.

Si la recta de regresión es $\ln Y = a \ln X + b$, entonces la dependencia potencial en los datos originales es:

$$y = k\,x^a$$

con $k = e^b$.

*Ejemplo*: El científico inglés Fry Richardson midi6 la costa oeste de Gran Bretaña con "reglas" de distintos tamaños, y obtuvo los siguientes datos aproximados:

| Tamaño de la regla $X$ (en km) | 1000 | 500 | 200 | 100 | 30 | 10 |
|---|---|---|---|---|---|---|
| Longitud de la costa $Y$ (en km) | 1000 | 1000 | 1200 | 1500 | 2100 | 2800 |

Calcula el coeficiente de correlaci6n para estos datos y también para los datos (ln $X$, ln $Y$). Compara los resultados. ¿Cuál es la ley que mejor describe la dependencia entre las variables $X$ e $Y$?

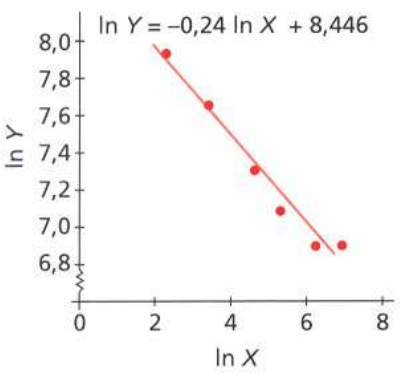
Si disminuye el tamaño de la regla, la costa aumenta de longitud, porque con reglas más pequeñas se miden más "recodos". Esto no sucedería si se estuviera midiendo la longitud de una circunferencia, que no tiene recodos. Si se hacen los correspondientes cálculos, el coeficiente de correlación para los datos ($X$, $Y$) es $r = –0{,}698$.

Si se calculan logaritmos sobre los datos, se obtiene la nueva tabla:

| ln $X$ | 6,90 | 6,21 | 5,30 | 4,60 | 3,40 | 2,30 |
|---|---|---|---|---|---|---|
| ln $Y$ | 6,90 | 6,90 | 7,09 | 7,31 | 7,65 | 7,93 |

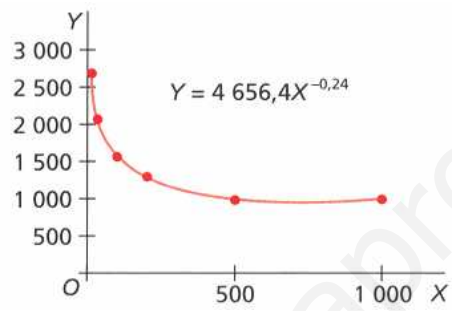
Al hacer los cálculos correspondientes, el coeficiente de correlación lineal para estos datos es $r = 0{,}985$, mucho más elevado que el anterior, y la recta de regresión es:

$$\ln y = –0{,}24 \ln x + 8{,}446$$




Esta última ley describe mucho mejor la relación entre las variables $X$ e $Y$. Al operar, se obtiene:

$$y = e^{-0,24 \ln x + 8,446} = e^{8,446}\, x^{-0,24} = 4656,4\, x^{-0,24}$$




***Ejercicio*:** La siguiente tabla de valores muestra el peso de diferentes animales y su consumo metabólico diario:

| | Peso (kg) | Tasa metabólica (kcal/día) |
|---|---|---|
| Ratón | 0,02 | 4 |
| Rata | 0,5 | 25 |
| Gato | 4 | 80 |
| Hombre | 70 | 2000 |
| Caballo | 600 | 8600 |
| Elefante | 5400 | 40000 |

Haz una regresión potencial y encuentra la ley potencial de las kilocalorías consumidas en función del peso.

*Solución*: $r = 0,994$; $\ln y = 0,7699 \ln x + 3,985167$; $y = 53,7943\, x^{0,7699}$

***Ejercicio*:** Halla el coeficiente de correlación para la regresión potencial en la siguiente lista de datos sobre el periodo orbital (en años) y su distancia media al Sol (en unidades astronómicas). ¿Cuál es la ley potencial que rige la distancia al Sol en función del periodo orbital?

| Planeta | Mercurio | Venus | Tierra | Marte | Júpiter | Saturno |
|---|---|---|---|---|---|---|
| Periodo orbital | 0,24 | 0,61 | 1,00 | 1,88 | 11,86 | 29,46 |
| Distancia al Sol | 0,39 | 0,72 | 1,00 | 1,52 | 5,20 | 9,54 |

*Solución*: $r = 0,999996$; $y = 1,002\, x^{0,665418}$

## *EJERCICIOS*

1. Construye la tabla de doble entrada correspondiente a la siguiente distribución:

(4, 1), (3, 2), (6, 0), (5, 1), (1, 5), (5, 0), (1, 6), (3, 3), (5, 1), (2, 4), (4, 2), (3, 4), (2, 4), (5, 1), (1, 7), (5, 2), (1, 6), (1, 5), (3, 3)

Construye la tabla de doble entrada de la distribución.

*Solución*:

| $Y$ \ $X$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 1 | 0 | 0 | 0 | 1 | 3 | 0 |
| 2 | 0 | 0 | 1 | 1 | 1 | 0 |
| 3 | 0 | 0 | 2 | 0 | 0 | 0 |
| 4 | 0 | 2 | 1 | 0 | 0 | 0 |
| 5 | 2 | 0 | 0 | 0 | 0 | 0 |
| 6 | 2 | 0 | 0 | 0 | 0 | 0 |
| 7 | 1 | 0 | 0 | 0 | 0 | 0 |

2. Dada la siguiente tabla de doble entrada, calcula la media y la varianza marginales de ambas variables:

| $Y$ \ $X$ | 1 | 2 | 3 |
|---|---|---|---|
| 10 | 0 | 2 | 1 |
| 20 | 3 | 0 | 4 |

*Solución*: $\bar{x} = 2{,}2$; $\sigma_x^2 = 0{,}76$ ; $\bar{y} = 17$; $\sigma_y^2 = 21$

3. Sea una variable bidimensional dada por la siguiente tabla de doble entrada:

| $Y$ \ $X$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 10 | 0 | 0 | 2 | 4 |
| 15 | 0 | 1 | 5 | 1 |
| 20 | 4 | 3 | 0 | 0 |

Calcula la media y la varianza de las variables marginales $X$ e $Y$, así como la covarianza.

*Solución*: $\bar{x} = 2{,}65$; $\sigma_x^2 = 1{,}275$ ; $\bar{y} = 15{,}25$; $\sigma_y^2 = 16{,}1875$; $\sigma_{xy} = -3{,}66$

4. En un depósito cilíndrico, la altura del agua que contiene varía conforme pasa el tiempo según la siguiente tabla:

| Tiempo (h) | 8 | 22 | 27 | 33 | 50 | 70 |
|---|---|---|---|---|---|---|
| Altura (m) | 17 | 14 | 12 | 11 | 6 | 1 |

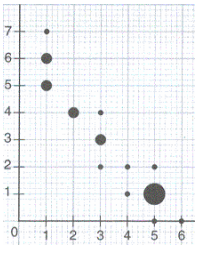
Halla:

a) Las medias y las varianzas de $X$ y de $Y$.
b) La covarianza.

*Solución*: $\bar{x} = 35$; $\sigma_x^2 = 402{,}67$ ; $\bar{y} = 10{,}17$; $\sigma_y^2 = 27{,}74$; $\sigma_{xy} = -105{,}78$

5. Representa la nube de puntos de la distribución del ejercicio 1.

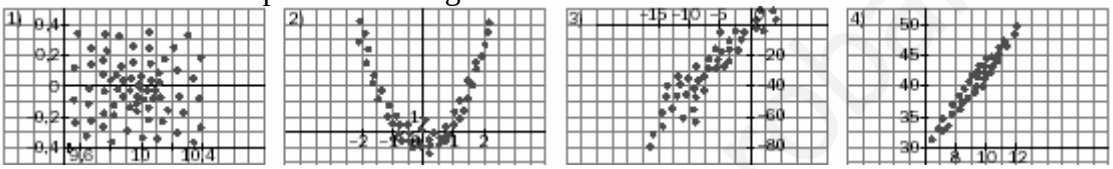
*Solución*:



6. En cada uno de los siguientes casos debes decir si, entre las dos variables que se citan, hay relación funcional o relación estadística (correlación) y, en este último caso, indicar si es positiva o negativa:
   a) En un conjunto de familias: estatura media de los padres – estatura media de los hijos.
   b) Temperatura a la que calentamos una barra de hierro – longitud alcanzada.
   c) Entre los países del mundo respecto a España: volumen de exportación – volumen de importación.
   d) Entre los países del mundo: índice de mortalidad infantil – número de médicos por cada 1000 habitantes.
   e) En las viviendas de una ciudad: kWh consumidos durante enero – coste del recibo de la luz.
   f) Número de personas que viven en cada casa – coste del recibo de la luz.
   g) Equipos de fútbol: lugar que ocupan al finalizar la liga – número de partidos perdidos.
   h) Equipos de fútbol: lugar que ocupan al finalizar la liga – número de partidos ganados.

*Solución*: a) Correlación positiva. b) Funcional. c) Correlación negativa. d) Correlación negativa. e) Funcional. f) Correlación positiva. g) Correlación positiva. h) Correlación negativa.

7. Considera las nubes de puntos de la figura.

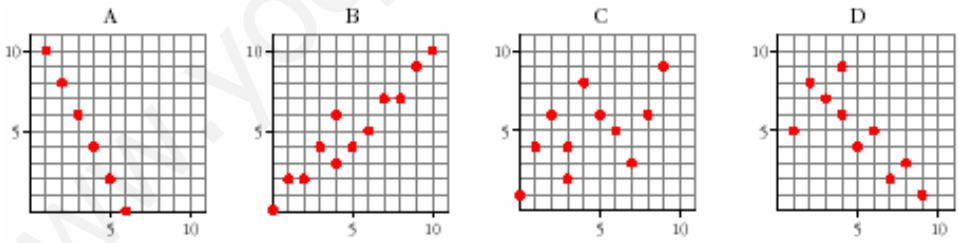



   a) Indica si hay relación de dependencia entre la variable $X$ y la variable $Y$. En caso de haberla, ¿puede considerarse esta relación lineal?
   b) Asigna a cada gráfico, si es posible, una de las siguientes rectas:

$$y = x \qquad y = 1 - 0{,}2x \qquad y = 2 + 4x$$

*Solución*: a) Hay relación entre las variables 2, 3 y 4, siendo lineal en las dos últimas.
b) La recta $y = 2 + 4x$ es la más adecuada para reflejar la relación entre las variables $X$ e $Y$ de los gráficos 3 y 4. Esta recta tiene pendiente 4 y en ambas nubes de puntos se observa que el rango de y es aproximadamente 4 veces mayor que el rango de $X$.

8. Traza, a ojo, la recta de regresión en cada una de estas distribuciones bidimensionales:




   a) ¿Cuáles de ellas tienen correlación positiva y cuáles tienen correlación negativa?
   b) Una de ellas presenta relación funcional. ¿Cuál es? ¿Cuál es la expresión analítica de la función que relaciona las dos variables?
   c) Ordena de menor a mayor las correlaciones.

*Solución*: a) B y C tienen correlación positiva; A y D, negativa; b) La A es relación funcional: $y = 12 - 2x$; d) C, D, B, A (prescindiendo del signo); A, D, C, B (considerando el signo)

9. El coeficiente de correlación de una distribución bidimensional es 0,87. Si los valores de las variables se multiplican por 10, ¿cuál será el coeficiente de correlación de esta nueva distribución?
*Solución*: El mismo, puesto que $r$ no depende de las unidades; es adimensional.

10. ¿Qué significa que en una distribución bidimensional el coeficiente de correlación sea el que se indica en cada uno de los siguientes casos?

a) $r = 1$ b) $r = 0$ c) $r = -1$ d) $r = 0{,}1$ e) $r = -0{,}75$ f) $r = 0{,}9$

*Solución*:

a) En la distribución, las variables $X$ e $Y$ están en dependencia funcional lineal directa, y todos los valores $(X, Y)$ se encuentran sobre una recta de pendiente positiva.

b) En la distribución, las variables $X$ e $Y$ son aleatoriamente independientes, y todos los valores $(X, Y)$ forman una nube de puntos sin tendencia alguna (variables incorreladas).

c) En la distribución, las variables $X$ e $Y$ están en dependencia funcional lineal inversa, y todos los valores $(X, Y)$ se encuentran sobre una recta de pendiente negativa.

d) Las variables $X$ e $Y$ están en dependencia aleatoria directa débil, y todos los valores $(X, Y)$ forman una nube de puntos ligeramente agrupada en torno a una recta de pendiente positiva.

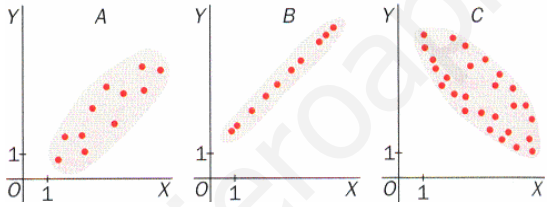
e) Las variables $X$ e $Y$ están en dependencia aleatoria inversa fuerte, y todos los valores $(X, Y)$ forman una nube de puntos medianamente agrupada en torno a una recta de pendiente negativa.

f) Las variables $X$ e $Y$ están en dependencia aleatoria directa fuerte, y todos los valores $(X, Y)$ forman una nube de puntos notablemente agrupada en torno a una recta de pendiente positiva.

11. Los coeficientes de correlación de dos conjuntos de datos estadísticos bidimensionales son $r_1 = 0{,}87$ y $r_2 = 0{,}37$. Razona en cuál de los dos conjuntos es mejor el ajuste mediante una recta de una variable en términos de la otra.

*Solución*: El ajuste será mejor en el primer conjunto ($r_1 = 0{,}87$), ya que el coeficiente de correlación es más cercano a 1, y en este caso la posible dependencia lineal de una de las variables con la otra es más fuerte.

12. Considera las siguientes nubes de puntos:



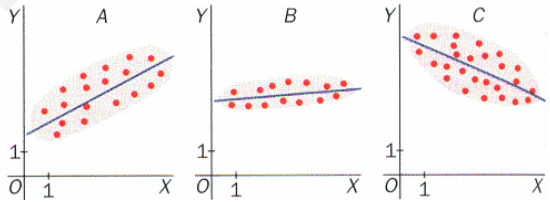

a) ¿En cual de ellas los datos se ajustarán mejor a una recta?

b) Asigna a cada una de las nubes uno de los siguientes coeficientes de correlación, fijando el signo en cada caso.

$r = \pm 0{,}99$ $r = \pm 0{,}6$ $r = \pm 0{,}8$

*Solución*: a) Se ajustará mejor a una recta la nube de puntos del apartado $B$; b) $A$: $r = 0{,}8$; $B$: $r = 0{,}99$; $C$: $r = -0{,}6$

13. En las siguientes gráficas se muestran las rectas de regresión obtenidas en tres estudios estadísticos.




a) ¿En cuál de las gráficas el coeficiente de correlación lineal será mayor?

b) Indica en qué gráficas el coeficiente de correlación lineal es negativo.

*Solución*: a) El de la gráfica $B$, ya que los puntos están más agrupados; b) El de la gráfica $C$, ya que los puntos se agrupan en torno a una recta de pendiente negativa.

14. Considera la distribución del ejercicio número 1.

a) Calcula el coeficiente de correlación lineal de Pearson.

b) ¿Es coherente el valor de este coeficiente con la correlación que se observa en el diagrama de dispersión? Justifica tu respuesta. (Ver ejercicio 5)

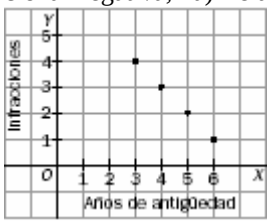
*Solución*: a) –0,945; b) Sí.

15. En una empresa de transportes trabajan 4 conductores. Los años de antigüedad de sus permisos de conducir y las infracciones cometidas en el último año por cada uno son los siguientes:

| | | | | |
|---|---|---|---|---|
| *X*: años de antigüedad | 3 | 4 | 5 | 6 |
| *Y*: infracciones | 4 | 3 | 2 | 1 |

a) Representa gráficamente los datos anteriores. Razona si estos muestran correlación positiva o negativa.
b) Calcula el coeficiente de correlación e interprétalo en términos de la situación real.

*Solución*: b) $r = -0{,}996$. Existe dependencia funcional negativa; a) Relación positiva.




16. Dada la distribución bidimensional:

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *X* | –2 | –2 | 2 | 2 | 2 | –2 | –1 | 1 | 0 | 1 | 0 | –1 |
| *Y* | 2 | 0 | 2 | –2 | 0 | –2 | 1 | –1 | –2 | 1 | 2 | –1 |

a) Dibuja el diagrama de dispersión de la distribución y describe el grado, el sentido y el tipo de correlación que observas.
b) Encuentra el valor del coeficiente de correlación lineal.
c) Relaciona los resultados obtenidos en los apartados anteriores.

*Solución*: Los puntos del diagrama de dispersión aparecen distribuidos simétricamente alrededor del origen, de manera que no se aprecia ningún tipo de agrupación de los puntos en torno a una curva reconocible. Por tanto no se observa correlación entre los valores de *X* e *Y*; b) $\sigma_{xy} = 0$; c) El hecho de que el coeficiente de correlación lineal de Pearson sea igual a 0 significa que no hay correlación lineal entre las variables *X* e *Y*, y vemos que coincide con la predicción cualitativa efectuada a partir del diagrama de dispersión.

17. Tenemos la siguiente tabla de datos de la temperatura media de varias ciudades y del gasto medio mensual en calefacción por habitante:

| | | | | | | |
|---|---|---|---|---|---|---|
| Temperatura (ºC) | 6 | 10 | 14 | 18 | 20 | 25 |
| Gasto en calefacción (€) | 50 | 45 | 25 | 15 | 10 | 2 |

Calcula el coeficiente de correlación y la recta de regresión de *Y* (gasto en calefacción) sobre *X* (temperatura media de la ciudad).

*Solución*: $r = -0{,}98$; $y = -2{,}74x + 67{,}11$

18. Las tallas y los pesos de 10 personas vienen recogidos en la siguiente tabla:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Pesos (kg) | 70 | 65 | 85 | 60 | 70 | 75 | 90 | 80 | 60 | 70 |
| Talla (cm) | 175 | 160 | 180 | 155 | 165 | 180 | 185 | 175 | 160 | 170 |

Calcula la recta de regresión de la altura sobre el peso.

*Solución*: $y = 0{,}92x + 103{,}96$

19. Cinco niñas de 2, 3, 5, 7 y 8 años de edad pesan, respectivamente, 14, 20, 32, 42 y 44 kilos.
a) Halla la ecuación de la recta de regresión de la edad sobre el peso.
b) ¿Cuál sería el peso aproximado de una niña de 6 años?
c) ¿Tendría sentido utilizar la recta de regresión hallada para estimar el peso de una adolescente de 15 años?

*Solución*: a) $x = 0{,}19y - 0{,}78$; b) $y = 5{,}15x + 4{,}65 \Rightarrow$ A una niña de 6 años le corresponde un peso de: 35,55 kg.
c) No, porque el desarrollo físico en la adolescencia difiere notablemente del que se produce en la etapa de 2 a 8 años.

20. La edad, en años, que tiene un árbol y el diámetro, en centímetros, de su tronco, medidos para un número pequeño de árboles (supongamos 10 árboles), se presentan en la siguiente tabla:

| Edad (años) | 2 | 4 | 4 | 8 | 10 | 11 | 14 | 15 | 15 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|
| Diámetro (cm) | 10 | 15 | 14 | 20 | 30 | 28 | 50 | 55 | 52 | 60 |

Calcula, utilizando la recta de regresión, el diámetro que se puede predecir para un árbol de 15 años.

*Solución*: Recta de regresión ≡ $y = 3{,}16x + 0{,}80$ ⇒ Diámetro (15 años) = 48,27 cm.

21. La siguiente tabla representa la información obtenida de 60 personas, a las que se les pesó ($X$, en kg) y se les midió ($Y$ en cm):

| $Y$ \ $X$ | [50, 60) | [60, 70) | [70, 80) | [80, 100) |
|---|---|---|---|---|
| [155, 165) | 8 | 4 | 2 | 0 |
| [165, 175) | 5 | 5 | 5 | 9 |
| [175, 185) | 1 | 2 | 8 | 10 |

a) Calcula la covarianza e interpreta su valor.
b) Calcula el coeficiente de correlación.
c) Calcula la recta de regresión adecuada y utilízala para predecir el peso de una persona de 2 metros de altura.

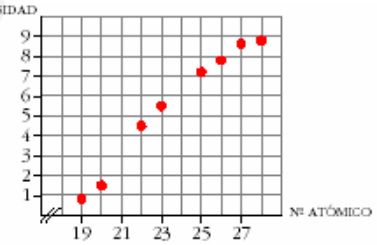
*Solución*: a) $\sigma_{xy} = 54{,}65$. Indica una dependencia directa. b) $r = 0{,}53$; c) $x = 0{,}944y - 88{,}33$ ⇒ $P$ (2 m) = 100,47 kg

22. La siguiente tabla relaciona el número atómico de varios metales de la misma fila en el sistema periódico (periodo 4), con su densidad. Representa la nube de puntos, calcula el coeficiente de correlación y halla la ecuación de la recta de regresión. A partir de ella, estima la densidad del cromo (Cr), cuyo número atómico es 24. Haz otro tanto con la del escandio (Sc), de número atómico 21.

| Elemento | K | Ca | Ti | V | Mn | Fe | Co | Ni |
|---|---|---|---|---|---|---|---|---|
| Nº Atómico | 19 | 20 | 22 | 23 | 25 | 26 | 27 | 28 |
| Densidad (g/cm$^3$) | 0,86 | 1,54 | 4,5 | 5,6 | 7,11 | 7,88 | 8,7 | 8,8 |

*Solución*: $r = 0{,}98$; $y = -16{,}5 + 0{,}93x$; $y\,(24) = 5{,}86$; $y\,(21) = 3{,}06$

Las densidades del Cr y del Sc son, aproximadamente, 5,86 y 3,01. (Los valores reales de estas densidades son 7,1 y 2,9)




23. Para una variable bidimensional se conoce $r = -0{,}5$, $\sigma_x = 2$ y $\sigma_y = 3$. Razona si alguna de las siguientes rectas de regresión de $Y$ sobre $X$ corresponde a estos datos:

a) $y = -x + 2$ b) $y = 0{,}5x - 1$ c) $3x + 4y - 4 = 0$

*Solución*: La recta buscada es la c).

24. Se sabe que entre el consumo de papel y el número de litros de agua por metro cuadrado que se recogen en una ciudad no existe relación. Responde razonadamente a las siguientes cuestiones.

a) ¿Cuál es el valor de la covarianza de estas variables?
b) ¿Cuánto vale el coeficiente de correlación lineal?
c) ¿Qué ecuaciones tienen las dos rectas de regresión y cuál es su posición en el plano?

*Solución*: a) $\sigma_{xy} = 0$; b) $r = 0$; c) Las ecuaciones de las rectas de regresión son: $y = \bar{y}$ , $x = \bar{x}$. Por tanto, son paralelas a los ejes y, en consecuencia, perpendiculares.

25. En una distribución bidimensional, la recta de regresión de $Y$ sobre $X$ es $y = \overline{y}$, siendo $\overline{y}$ la media de la distribución $Y$. ¿Cuál es la recta de regresión de $X$ sobre $Y$? ¿Existe dependencia lineal entre $Y$ y $X$? Razona las respuestas.

*Solución*: Si la recta de regresión de $Y$ sobre $X$ es $y = \overline{y}$, la recta de regresión de $X$ sobre $Y$ será $x = \overline{x}$. En estos casos no existe ningún tipo dependencia entre las variables $X$ e $Y$; por tanto, están incorreladas.

26. Se han recogido datos de temperatura y de presión en distintas ciudades:

| Temperatura (ºC) | 50 | 100 | 70 | 60 | 120 | 180 | 200 | 250 | 30 | 90 |
|---|---|---|---|---|---|---|---|---|---|---|
| Presión (mm) | 5 | 2 | 2,5 | 3,75 | 4 | 1 | 1,25 | 0,75 | 7 | 3 |

a) Calcular las rectas de regresión y el coeficiente de correlación lineal.
b) Estima la presión que habría para una temperatura de 23 ºC.
c) Estima la temperatura si la presión fuese de 830 mm.

*Solución*: a) $r = 0{,}98$; $y = 6{,}5x + 675{,}5$ ; $x = 0{,}5y - 381{,}995$; b) $y\,(23) = 825$ mm; c) $x\,(830) = 23{,}67$ ºC

27. El número de licencias de caza, en miles, y el número de votantes a un determinado partido en 6 comunidades autónomas, en decenas de miles, está expresado en la siguiente tabla:

| Nº de licencias ($X$) | 103 | 26 | 3 | 7 | 26 | 5 |
|---|---|---|---|---|---|---|
| Nº de votantes ($Y$) | 206 | 26 | 27 | 14 | 24 | 12 |

Determinar:

a) Coeficiente de correlación, interpretando su valor.
b) En el caso de que exista correlación: si en una determinada comunidad existen 50 decenas de millar de votantes, ¿cuántas licencias de caza, en miles, se puede estimar que existen?

*Solución*: a) $r = 0{,}97$, luego la correlación lineal entre las variables $X$ e $Y$ es positiva y fuerte. b) La recta de regresión de licencias sobre votantes es: $x = 0{,}485y + 3{,}3$, por lo que si en una comunidad autónoma tenemos 50 decenas de millar de votantes, $y = 50$, el número de licencias, en miles, será 27,55.

28. Hemos obtenido 10 medidas de las variables $X$ e $Y$ correspondientes a una distribución bidimensional. A partir de esos datos, conocemos:

$$\sum x_i = 200 \qquad \sum y_i = 50 \qquad r = -0{,}75$$

a) Una de las siguientes rectas es la de regresión de $Y$ sobre $X$. Di cuál de ellas es, justificadamente:

I) $y = -4{,}5 + 2{,}5x$ II) $y = 35 - 1{,}5x$
III) $y = 9 - 0{,}7x$ IV) $y = -200 + 50x$

b) Halla la recta de regresión de $X$ sobre $Y$.

*Solución*: a) La recta de regresión pasa por $(\overline{x}, \overline{y}) = (20, 5)$. Además, el signo de $r$ coincide con el signo de la pendiente de la recta de regresión; luego es la II): $y = 35 - 1{,}5x$; b) $x = 21{,}875 - 0{,}375y$

29. Halla el punto medio de una distribución tal que la recta de regresión de $Y$ sobre $X$ es $5x - 4y = -13$ y la de $X$ sobre $Y$ es $3x - 2y = -5$. ¿Cuál es el valor del coeficiente de Pearson?

*Solución*: $(\overline{x}, \overline{y}) = (3, 7)$; $r = 0{,}913$

30. Las rectas de regresión de $Y$ sobre $X$ y de $X$ sobre $Y$ en una distribución bidimensional, son las siguientes:

$$y = 0{,}91x - 5{,}88$$
$$x = 0{,}85y + 13{,}24$$

¿Cuál es el coeficiente de correlación de Pearson de la distribución?

*Solución*: $r = 0{,}879$

31. De una distribución bidimensional se sabe que $\bar{x} = 1$ e $\bar{y} = 1$. Explica por qué la recta $y = -3x + 5$ no puede ser la recta de regresión de $Y$ sobre $X$ correspondiente a esta distribución.
*Solución*: La recta de regresión ha de pasar por el centro de gravedad. Sin embargo la recta dada no cumple esta condición.

32. ¿Pueden ser las expresiones siguientes las dos rectas de regresión correspondientes a una misma distribución bidimensional?

$$2x + 3y = 29 \qquad 4x - 5y = -19$$

Justifica tu respuesta.
*Solución*: No pues las rectas de regresión han de tener pendiente de igual signo.

33. Razona si es posible encontrar alguna distribución bidimensional tal que:
La recta de regresión de $Y$ sobre $X$ sea: $y = 2x + 1$
La recta de regresión de $X$ sobre $Y$ sea: $x = 3y + 4$
*Solución*: No, pues $r^2 = 6 \Rightarrow r = \sqrt{6} > 1$, lo cual es imposible por ser $-1 \le r \le 1$

34. La siguiente tabla da los datos obtenidos para una variable bidimensional.

| $X$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| $Y$ | 14 | 4 | 18 | 16 | 13 | 18 | 15 | 10 | 11 |

a) Halla la recta de regresión de $Y$ sobre $X$.
b) Calcula la recta de Tukey.
c) Representa la nube de puntos y las dos rectas obtenidas.

*Solución*: a) $y = 0{,}034x + 13{,}05$; b) $y = -\frac{1}{2}x + \frac{97}{6}$; c)