DISTRIBUCIONES BIDIMENSIONALE: CORELACIÓN Y REGRESIÓN LINEAL

Distribuciones bidimensionales

Se estudian a la vez dos variables aleatorias (genéricamente X e Y; sus valores serán (x_i, y_i)).

Correlación

Al estudiar distribuciones bidimensionales, el objetivo es determinar si existe relación estadística entre las dos variables consideradas; es decir, ver si los cambios en una de las variables influyen en los cambios de la otra. Cuando sucede esto, se dice que ambas variables están correlacionadas o que hay correlación entre ellas.

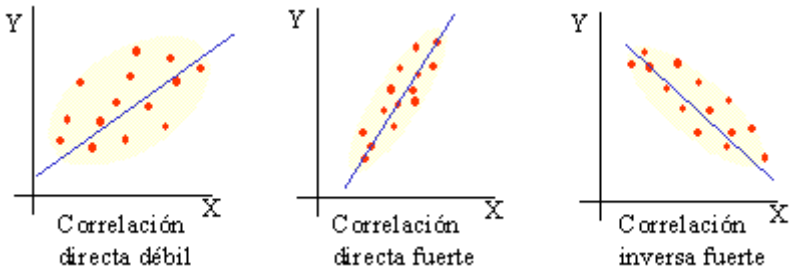
Si las variables aumentan o disminuyen conjuntamente, la correlación es directa. Si, por el contrario, al aumentar una de ellas disminuye la otra, la correlación será inversa.
Si la correlación es *fuerte*, a partir de una variable puede estimarse la otra con una fiabilidad (probabilidad) alta. Si la correlación es débil, la estimación de una variable a partir de la otra es poco fiable. (Aquí se estudiará solo la correlación lineal).

Ejemplos:
a) La correlación entre el número de zapato y la estatura de las personas es directa y fuerte.
b) Las variables temperatura ambiente y gasto de calefacción están inversamente correlacionadas: a menor temperatura más gasto en calefacción.
c) Las variables número de zapato y gasto en calefacción no están correlacionadas.

Diagramas de dispersión

El primer paso para determinar el sentido y el grado de la correlación entre dos variables consiste en representar gráficamente, en el plano cartesiano, los pares de valores conocidos. Estos gráficos, que reciben el nombre de diagramas de dispersión, permiten visualizar la posición de los datos en el plano. La forma de la nube de puntos asociada a cada diagrama permitirá establecer conjeturas sobre la correlación existente entre las variables estudiadas.
En general, dependiendo de la forma de la nube de puntos, puede asegurarse:

- Una nube de puntos alargada indica correlación lineal: los puntos se distribuyen en torno a una línea recta. La estrechez de la nube expresa que la correlación es fuerte.
- Si la recta que se ajusta a la nube tiene pendiente positiva, la **correlación** será **directa**: al crecer la variable X, lo hace también la variable Y.
- Una recta con pendiente negativa, indica que la **correlación** es **inversa**, al crecer X, disminuye Y.



La confirmación cuantitativa de estas conjeturas se deduce estudiando: 1) los parámetros estadísticos asociados a la distribución bidimensional; 2) determinando la recta de regresión.

Parámetros de una distribución bidimensional

Medias marginales para cada una de las variables X e Y. Valen: $\bar{x} = \frac{\sum x_i}{n}$; $\bar{y} = \frac{\sum y_i}{n}$

El punto $(\bar{x}, \bar{y})$ se llama centro medio de la distribución.

Varianzas y desviaciones típicas

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

Las desviaciones típicas marginales, s_x y s_y, son la raíz cuadrada de cada una de ellas.

La covarianza: La covarianza es un parámetro estadístico conjunto, pues, en su cálculo intervienen las dos variables a la vez. Se define como la media aritmética de los productos de las diferencias de los valores de cada variable respecto de su media marginal. Por tanto, vale:

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

Si $s_{xy} > 0$, la correlación es directa; si $s_{xy} < 0$, la correlación es inversa.

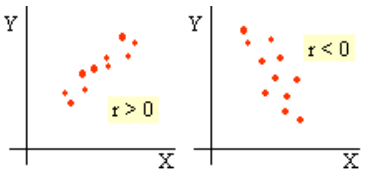
El coeficiente de correlación lineal

Da una medida de la fuerza de la correlación entre las dos variables estudiadas.

Vale: $r = \frac{s_{xy}}{s_x \cdot s_y}$. Es la razón entre la covarianza de las variables X e Y y el producto de sus desviaciones típicas marginales.

El coeficiente de correlación cumple:

1) El valor de r no cambia al hacerlo la escala de medición.
2) El signo de r es el mismo que el de la covarianza: si $r > 0$, la correlación es directa; si $r < 0$, la correlación es inversa.
3) El valor de r está entre –1 y +1: $-1 \le r \le 1$
4) Si $|r|$ toma valores cercanos a 1, la correlación es fuerte.



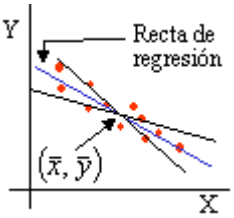
5) El cuadrado de r, r^2, indica la proporción de la variación en la variable Y que puede ser explicada por los cambios de la variable X. A r^2 se le llama coeficiente de determinación.

Ejemplo:
Si $r = 0{,}8$, el coeficiente de determinación vale $r^2 = 0{,}8^2 = 0{,}64$. Esto significa que el 64% de la variación de Y puede ser explicada a partir de la variación de X.

Recta de regresión lineal

Esta recta (de Y sobre X) permite hacer estimaciones de la variable Y a partir de la X.

La recta de regresión es la que mejor se ajusta a la nube de puntos. Es una recta ideal que asignaría a cada valor x_i de la variable X el promedio de los y_i correspondientes a x_i. En consecuencia, debe pasar por el punto $(\bar{x}, \bar{y})$, centro de gravedad de la distribución bidimensional.



La recta que mejor se ajusta a estos propósitos es la recta de regresión mínimo cuadrática, que es aquella que minimiza la suma de los cuadrados de los errores.

Si la ecuación de esta recta es $y = ax + b$, se cumple que: $a = \frac{s_{xy}}{s_x^2}$ y $b = \bar{y} - \frac{s_{xy}}{s_x^2} \cdot \bar{x}$

Su ecuación es: $y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x})$,

→ La recta de regresión de X sobre Y (que no es la misma que la de Y sobre X) permite estimar los valores de Y a partir de los de la variable X. Su ecuación es: $x-\overline{x}=\frac{s_{xy}}{s_y^{\,2}}(y-\overline{y})$.

Ejemplo:

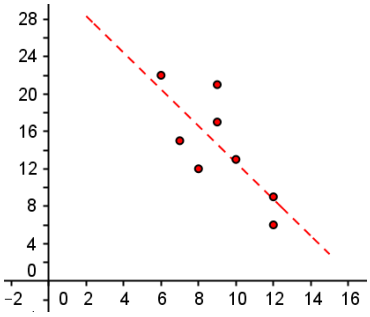
1. Ocho personas, con similar destreza en mecanografía, teclearon 20 líneas de texto en un ordenador. El tiempo empleado, en minutos, y el número de errores cometidos, fueron:

Tiempo (X)	6	7	8	9	9	10	12	12
Errores (Y)	22	15	12	17	21	13	9	6

a) Dibuja la nube de puntos asociada. ¿Qué tipo de correlación se da entre las variables estudiadas?
b) Calcula, indicando todos los pasos intermedios, el coeficiente de correlación y la recta de regresión de *Y* sobre *X*.
c)¿Cuántos errores deben esperarse para una persona que tarda 14 minutos en teclear las 20 líneas de texto?

Solución**:**

a) A partir de la lectura de los valores de la tabla se observa una correlación negativa (el número de errores tiende a disminuir al aumentar el tiempo). Esto se confirma representando los pares de valores: (6, 22), (7, 15),.., (12, 9), (12, 6).
La nube de puntos, alargada y con tendencia decreciente, sugiere una correlación lineal inversa y fuerte: los puntos se ajustan bien a una recta. Por tanto, puede deducirse que el tiempo empleado determina de alguna manera el número de errores: a menos tiempo más errores.



b) De acuerdo con las fórmulas de los parámetros hay que hacer sumas, sumas de cuadrados y sumas de productos; para ello resulta eficaz la siguiente tabla:

x_i	y_i	$x_i^{\,2}$	$y_i^{\,2}$	$x_i y_i$
6	22	36	484	132
7	15	49	225	175
8	12	64	144	96
9	17	81	289	171
9	21	81	441	189
10	13	10	169	130
12	9	144	81	108
12	6	144	36	72
$\sum x_i = 73$	$\sum y_i = 115$	$\sum x_i^{\,2} = 699$	$\sum y_i^{\,2} = 1869$	$\sum x_i y_i = 985$

Con esto:

$$\overline{x}=\frac{73}{8}=9{,}125\,;\quad \overline{y}=\frac{115}{8}=14{,}375$$

$$\sigma_x=\sqrt{\frac{699}{8}-9{,}12^2}=2{,}027\,;\ \sigma_y=\sqrt{\frac{1869}{8}-14{,}375^2}=5{,}195\,;\ \sigma_{xy}=\frac{985}{8}-9{,}125\cdot 14{,}375=-8{,}047$$

Por último, $r=\frac{-8{,}047}{2{,}027\cdot 5{,}195}=-0{,}764$.

La correlación es inversa y fuerte: si se teclea más deprisa se comenten más errores.

La recta de regresión es $y-14{,}375=\frac{-8{,}047}{2{,}027^2}(x-9{,}125)\ \Leftrightarrow\ y=-1{,}959x+32{,}251$

c) Si $x=14$, $y=-1{,}959\cdot 14+32{,}251=4{,}825$ → 5 errores.

Pequeños retos

1. En seis alumnos de bachillerato se observaron dos variables: X = "puntuación obtenida en un determinado test" e Y = "nota en un examen de filosofía". Los resultados se indican en la siguiente tabla:

Test: X	110	90	140	120	120	100
Examen: Y	6	5	9	7	8	6

a) Halla la recta de regresión de Y sobre X.
b) Sabiendo que un alumno obtuvo 130 puntos en el test, pero no realizó el examen de filosofía, predice, si es posible, la nota que hubiese obtenido.

Soluciones:

1. a) $\overline{x} = 13,33...$; $\overline{y} =,83...$; $s_x = 15,986$; $s_y = 1,344$; $s_{xy} = 20,956$; $r = 0,975$. (Correlación directa y fuerte). $y = 0,082x - 2,283$

Resultados con la calculadora: $r = 0,95693211$; $y = 0,080434782x - 2,282608696$.

b) Si $x = 130$, $y = 0,082 \cdot 130 - 2,283 = 8,377$.