Tema 12. DISTRIBUCIONES BIDIMENSIONALES **Resumen**

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 probabilidad alta. Si la correlación es débil, la estimación de una variable a partir de la otra es poco fiable.

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.

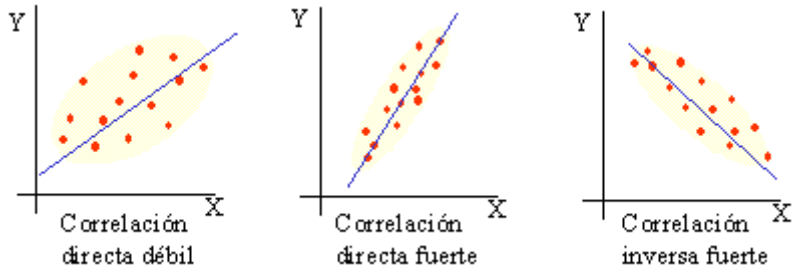
c) Las variables número de zapato y gasto en calefacción no están correlacionadas.

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: $\overline{x} = \frac{\sum x_i}{n}$; $\overline{y} = \frac{\sum y_i}{n}$

El punto $(\overline{x}, \overline{y})$ 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.

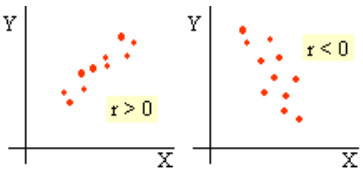
El coeficiente de correlación lineal

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) Su 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.



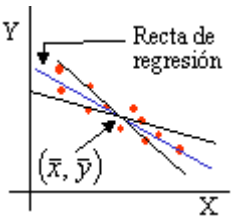
5) El cuadrado de r, r^2, indica la proporción de la variación en la variable Y que puede ser explicada por los cambios de la variable X. A r^2 se le llama coeficiente de determinación.

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 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 $b = \bar{y} - \frac{s_{xy}}{s_x^2} \cdot \bar{x}$

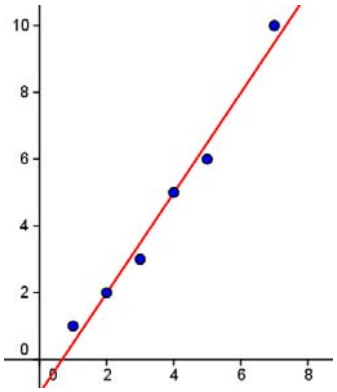
Su ecuación es: $y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x})$,

Siendo $\bar{x}$ e $\bar{y}$ las medias marginales de las variables X e Y, s_x^2 la varianza de X y s_{xy} la covarianza.

Ejemplo:
La siguiente tabla contiene las horas de asistencia a un curso de informática y las notas obtenidas por seis alumnos:

Horas de asistencia (X)	2	1	7	5	4	3
Notas (Y)	2	1	10	6	5	3

a) Representa la nube de puntos.
b) Halla el coeficiente de correlación entre X e Y, e interprétalo.
b) Halla la recta de regresión; y represéntala.
c) Si una persona asistiera seis horas al curso, ¿qué nota obtendría?
Solución:
a) La nube de puntos es la representada a la derecha.



b) Utilizando la calculadora se tienen los siguientes resultados:

$\bar{x} = 3{,}666...$ $\quad \sum x_i = 22$ $\quad \sum x_i^2 = 104$ $\quad s_x = 1{,}97203$

$\bar{y} = 4{,}5$ $\quad \sum y_i = 27$ $\quad \sum y_i^2 = 175$ $\quad s_y = 2{,}98607$

$r = 0{,}99061$. Es una correlación directa y muy fuerte.

c) La recta de regresión es: $y = 1{,}5x - 1$. Es la adjunta.
Para $x = 6$ horas, $y = 1{,}5 \cdot 6 - 1 = 8$. Obtendría un 8.

Cálculo de los parámetros anteriores "a mano"
Fórmulas:

$$\bar{x} = \frac{\sum x_i}{n};\ \bar{y} = \frac{\sum y_i}{n};\ s_x = \sqrt{\frac{\sum x_i^{\,2}}{n} - \bar{x}^2};\ s_y = \sqrt{\frac{\sum y_i^{\,2}}{n} - \bar{y}^2};\ s_{xy} = \frac{\sum x_i y_i}{n} - \overline{xy}$$

X (x_i)	Y (y_i)	x_i^2	y_i^2	$x_i \cdot y_i$
2	2	4	4	4
1	1	1	1	1
7	10	49	100	70
5	6	25	36	30
4	5	16	25	20
3	3	6	9	9
$\sum x_i = 22$	$\sum y_i = 27$	$\sum x_i^2 = 104$	$\sum y_i^2 = 175$	$\sum x_i y_i = 134$
$\bar{x} = \frac{22}{6} = 3{,}666$	$\bar{y} = \frac{27}{6} = 4{,}5$			

$$s_x = \sqrt{\frac{104}{6} - 3{,}666^2} = 1{,}972$$

$$s_y = \sqrt{\frac{175}{6} - 4{,}5^2} = 2{,}986$$

$$s_{xy} = \frac{134}{6} - \frac{22}{6} \cdot 4{,}5 = 5{,}833$$

Recta de regresión: $y - \bar{y} = \frac{s_{xy}}{s_x^{\,2}}(x - \bar{x})$

$$y - 4{,}5 = \frac{5{,}833}{1{,}972^2}\left(x - \frac{22}{6}\right) \Rightarrow y = 1{,}5x - 1$$