Variables estadísticas bidimensionales

En ocasiones los datos se presentan en pares (x, y). Es decir que al recoger los datos no solo tenemos una variable, sino dos y hay una correspondencia entre los valores de una y los de la otra.

Pensemos en el siguente ejemplo:

0.0.1 Ejemplo

Una empresa observa que parece haber una relación fuerte entre las ventas en enero y las ventas en febrero. Para ello se recogen los datos de 9 años y se anotan en una tabla las ventas de enero y febrero de cada año en miles de euros.

ventas enero	ventas febrero
142.74	69.06
146.58	70.62
149.01	72.03
151.72	73.48
154.12	74.89
158.23	76.48
160.19	77.85
165.46	79.54
168.82	81.05

Para analizar estos datos, podríamos trabajar con cada variable por separado (calculando medias, varianzas, haciendo gráficas...), pero se perdería la relación entre ambas. Siguendo este procedimiento sería dificil observar por ejemplo como a mayores ventas en enero corresponden mayores ventas en febrero.

Para ello trabajamos con las dos variables juntas, a través de la **Covarianza** y el **coeficiente de correlación de pearson**. Mediante ellos intentaremos capturar la relación de dependencia entre ambas variables, particularmente la *dependencia lineal,*, que ocurre cuando una de las variables puede ser aproximadas a partir de la otra mediante una recta.

1 Covarianza

La **covarianza** es un valor que indica el grado de variación conjunta de dos variables aleatorias respecto a sus medias. Es el dato básico para determinar si existe una dependencia entre ambas variables y además es el dato necesario para estimar otros parámetros básicos, como el coeficiente de correlación lineal.

Supongamos que tenemos unos datos

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), ... (x_N, y_N)$$

La covarianza se denota por S_{xy} y se define como

$$S_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$$

$$= \frac{1}{N}((x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + ... + (x_N - \overline{x})(y_N - \overline{y}))$$

donde $\overline{x}$ denota la media de la primera variable (x), e $\overline{y}$ denota la media de la segunda (y)

1.1 Ejemplo

Para entender cómo calcularlo, usaremos el ejemplo anterior. Primero calculamos la media de ambas variables. En este caso como todos los datos son distintos no hay frecuencias así que para calcular las medias basta sumar y dividir entre el número de datos.

Obtenemos $\overline{x} \cong 155.207$, $\overline{y} \cong 77.337$. Añadimos una columna calculando las multiplicaciones $(x_i - \overline{x})(y_i - \overline{y})$

x_i	y_i	$(x_i - \overline{x})(y_i - \overline{y})$
142.74	69.06	80.389
146.58	70.62	21.981
149.01	72.03	16.658
151.72	73.48	4.142
154.12	74.89	-1.916
158.23	76.48	-1.564
160.19	77.85	2.502
165.46	79.54	27.908
168.82	81.05	114.372

y para calcular la covarianza basta sumar esta tercera columna y dividir entre el número de datos

$$S_{xy} \cong 264.474/9 \cong 29.386$$

1.2 El gráfico nubes de puntos

Una manera de visualizar la relación o dependencia entre las dos variables es dibujar cada punto (x_i, y_i) en el plano.

En el ejemplo anterior, el gráfico sería el siguente.

2 Coeficiente de correlación de Pearson

El coeficiente de **correlación de Pearson** es una medida de dependencia lineal entre dos variables estadísticas cuantitativas. A diferencia de la covarianza, la correlación de Pearson es independiente de la escala de medida de las variables.

Se define como

$$\rho_{XY} = \frac{S_{xy}}{S_X S_Y}$$

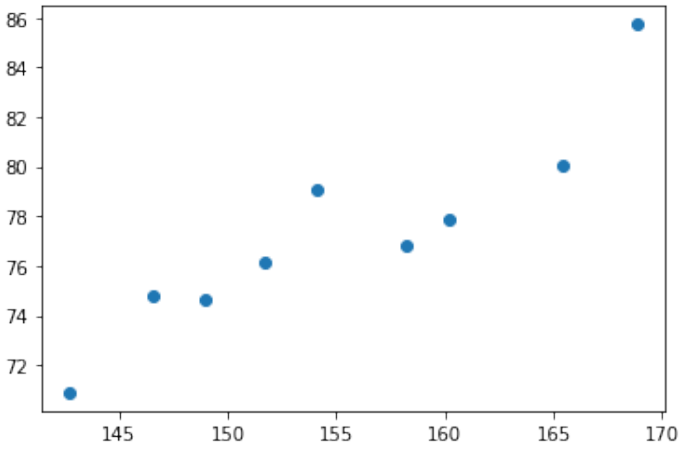

Figure 1: nube de puntos

donde S_{xy} de nota la covarianza, S_X denota la desviación típica de la primera variable y S_Y la desviación típica de la segunda.

en el ejemplo anterior, si calculamos además la desviación típica de X e Y obtenemos

$$S_x \cong 8.205$$
$$S_y \cong 3.920$$

luego

$$\rho_{XY} \cong \frac{29.386}{8.205 \cdot 3.920} \cong 0.913$$

Deducimos de aquí que dado que el coeficiente de correlación de Pearson es cercano a 1 existe una dependencia lineal directa entre las ventas de enero y febrero

2.1 Interpretación del coeficiente de correlación de Pearson

- Si $\rho_{XY} > 0$ hay dependencia lineal directa (positiva), es decir, a grandes valores de X corresponden grandes valores de Y.

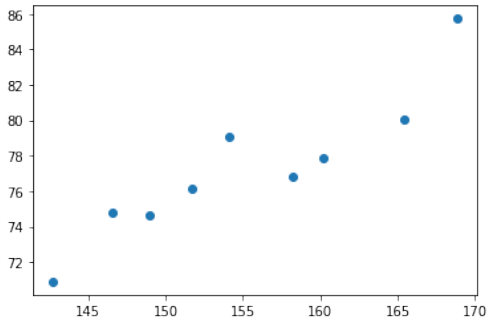

Figure 2: nube de puntos correlación positiva

- Si $\rho_{XY} = 0$ se interpreta como la no existencia de una relación lineal entre las dos variables.

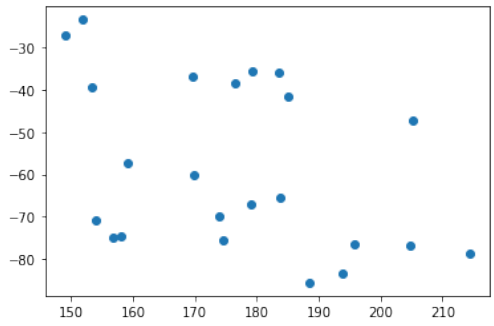

Figure 3: nube de puntos correlación 0

- Si $\rho_{XY} < 0$ hay dependencia lineal inversa o negativa, es decir, a grandes valores de X corresponden pequeños valores de Y.

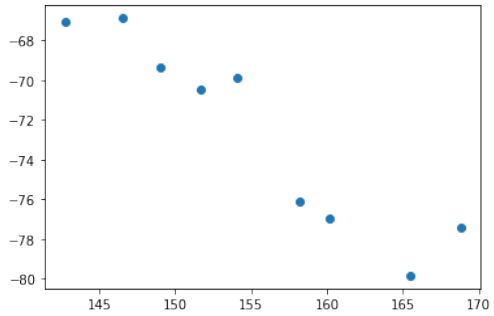

Figure 4: nube de puntos correlación negativa

3 Ejercicios

1. Los siguientes datos corresponden a los porcentajes de mortalidad obtenidos a dosis crecientes de un insecticida. Los resultados fueron los siguientes:

Dosis	Mortalidad
0	5
1	7
5	10
10	16
15	17
20	25
25	26
30	30

Calcula la covarianza y el coeficiente de correlación de Pearson e interprétalo