DISTRIBUCIONES BIDIMENSIONALES

Jesús García de Jalón de la Fuente

1. Introducción

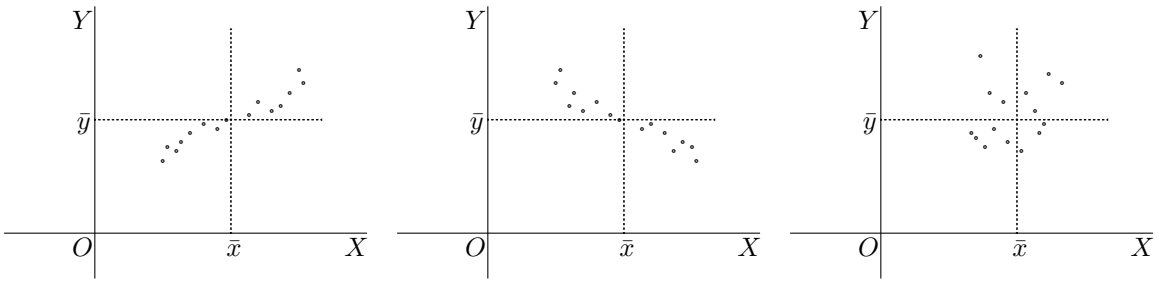
Supongamos que sobre una población se miden dos magnitudes x e y. Para cada elemento de la población se obtienen un par de valores (x, y). Considerando estos pares de números como coordenadas, podemos representar los puntos correspondientes en unos ejes de coordenadas. Se obtiene un diagrama que se llama **nube de puntos**. La cuestión que se va a plantear es si las dos variables están correlacionadas, esto es, si se puede afirmar que al aumentar una aumenta la otra (**correlación positiva**) o que al aumentar la primera disminuye la segunda (**correlación negativa**).

Lo que se diga sobre las variables, forzosamente tendrá un carácter probabilístico, es decir, si las dos variables están correlacionadas positivamente, quiere decir que si aumenta una, *es probable* que aumente la otra (pero no necesariamente) y veremos una manera de evaluar esa probabilidad.

La situación es diferente cuando existe una dependencia funcional entre las variables. En este caso, si la función es creciente, cuando una de las variables aumenta, *necesariamente* aumenta la otra.

El hecho de que dos variables estén correlacionadas no implica que exista una relación de causa-efecto entre ellas. Para que esto suceda es necesario que, además de variar conjuntamente, una de ellas sea anterior en el tiempo a la otra, debe existir una relación temporal entre ellas, aspecto éste que no estamos considerando.

Como hemos dicho, si para cada elemento de la población se representa sobre unos ejes de coordenadas los valores de las dos variables, se obtiene un diagrama que se llama **nube de puntos**:



En las figuras aparecen diagramas de puntos correspondientes a variables con correlación positiva, con correlación negativa y variables con corrrelación muy débil.

2. Correlación lineal

Supongamos que sobre una población se han medido dos magnitudes x e y y se han obtenido para ellas medias $\bar{x}$ y $\bar{y}$, y desviaciones típicas σ_x y σ_y.

El concepto clave para estudiar la correlación entre las variables es el de covarianza que se define de la siguiente forma:

$$\sigma_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{N} = \overline{xy} - \bar{x}\bar{y}$$

Como se ve, la covarianza puede calcularse (al igual que sucedía con la varianza) de dos formas diferentes, calculando la media de los productos de las diferencias con las medias de las dos variables, o como diferencia entre la media de los productos de las dos variables y el producto de las medias.

Si las dos variables están correlacionadas positivamente, la covarianza será positiva y si están correlacionadas negativamente será negativa.

Esto se entiende si representamos la nube de puntos y dibujamos unos ejes centrados en las medias (en el punto $(\bar{x}, \bar{y})$). Si la correlación es positiva, la mayor parte de los puntos debe estar en el primer y tercer cuadrante respecto a estos ejes. En este caso los dos factores del producto tienen el mismo signo, los productos son positivos y también lo será la covarianza. Lo contrario sucede si la correlación es negativa.

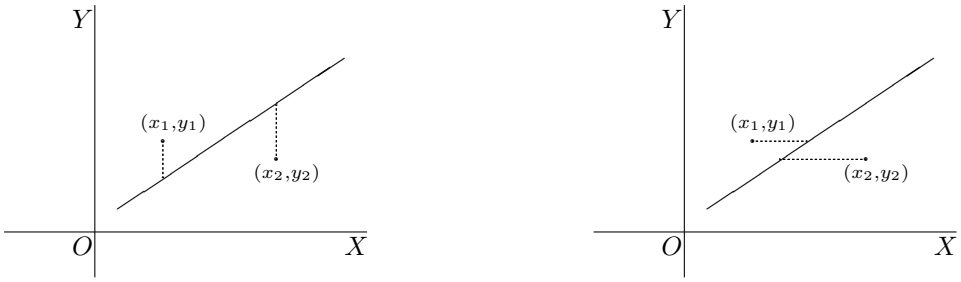
La covarianza permite saber si la correlación es positiva o negativa pero no permite saber si es fuerte o débil pues su valor depende de la unidades utilizadas. Por ello se utiliza el coeficiente de correlación:

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

que toma valores comprendidos entre -1 y $+1$. La correlación es positiva y fuerte si el valor de r es próximo a $+1$; es negativa y fuerte si el valor de r es próximo a -1. Si r es práximo a cero, la correlación es debil.

3. Recta de regresión

La recta de regresión es la recta que mejor se ajusta a la nube de puntos. Lo lógico sería tomar elegir la recta de tal forma que la suma de distancias desde los puntos a la recta fuese mínima. La obtención de esta recta es complicado matemáticamente por lo que se usan la **recta de regresión de y sobre x** y de x **sobre** y en las que se minimiza la suma de las distancias de los puntos a la recta, tomadas según paralelas a los ejes.



Las dos rectas de regresión de y sobre x y de x sobre y pasan por el punto $(\bar{x}, \bar{y})$ y se diferencian en la pendiente. Son estas:

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

Si la correlación es fuerte ambas rectas son casi coincidentes y son iguales si el coeficiente de correlación es $+1$ o -1. Si la correlación es débil, el ángulo que forman las dos rectas es grande. En el caso extremo de que el coeficiente de correlación sea 0 ambas rectas son perpendiculares, una paralela al eje de abscisas y otra paralela al eje de ordenadas.