

# 1 Medidas de asociación

**Objetivo:** Definir medidas de asociación para contrastar la correlación o la independencia entre dos variables

## 1.1 Test de asociación lineal: Correlación y Coeficiente de Pearson

### 1.1.1 Regresión simple:

Sea el modelo de regresión simple: para  $i = 1, 2, \dots, n$ ,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

suponiendo que las desviaciones  $\varepsilon_i$  son centradas y con una varianza  $\sigma^2$ .

Queremos contrastar la hipótesis  $H_0 : \beta_1 = 0$  o sea “ $Y$  y  $X$  son incorreladas”.

El estimador de mínimos cuadrados de  $\beta_1$  es

$$\hat{\beta}_1 = \frac{\mathbf{cov}(y, x)}{s_x^2}$$

donde

$$\begin{aligned} \mathbf{cov}(y, x) &= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) \\ &= \left( \frac{1}{n} \sum_{i=1}^n y_i x_i \right) - \bar{x} \bar{y} \end{aligned}$$

es la covarianza muestral de  $(Y, X)$  y

$$\begin{aligned} s_x^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \left( \frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2 \end{aligned}$$

es la varianza muestral de  $X$ .

Una primera idea consistiría en rechazar  $H_0$  cuando  $\hat{\beta}_1$  es grande (en valor absoluto):

$$\text{Rechazamos } H_0 \text{ si } \left| \hat{\beta}_1 \right| > u$$

Sin embargo esta regla de decisión es inadecuada puesto que no tomamos en cuenta la varianza de  $\hat{\beta}_1$ ; un cambio de escala sobre  $X$  o  $Y$  afecta el valor de  $\hat{\beta}_1$  pero la relación entre  $Y$  y  $X$  no depende de la escala en la cual están expresadas ambas variables.

Una buena regla de decisión debe basarse en el estimador tipificado (bajo  $H_0$ ):

$$T = \frac{\hat{\beta}_1}{\sqrt{\widehat{\text{var}}(\hat{\beta}_1)}} = \sqrt{n-2} \frac{r(x, y)}{\sqrt{1-r^2(x, y)}}$$

donde  $\widehat{\text{var}}(\widehat{\beta}_1) = \widehat{\sigma}^2/s_x^2$  y

$$r(x, y) = \frac{\text{cov}(x, y)}{s_x s_y}$$

es el coeficiente de correlación de Pearson o coeficiente de correlación muestral de  $Y$  y  $X$ .

Vamos a rechazar  $H_0$  cuando  $|T|$  es grande o sea cuando  $|r(x, y)| > u$ .

El coeficiente de Pearson es invariante por translación y cambio de escala por lo tanto el estadístico  $T$  parece ser idóneo para contrastar  $H_0$ . Si suponemos que las desviaciones son normales, entonces conocemos la distribución de  $T$  :

$$T \sim t(n - 2)$$

Por lo tanto, la regla de decisión para un nivel  $\alpha$  será

$$\text{Rechazar } H_0 \text{ si } |T| > t_{\alpha/2}^{(n-2)}$$

o de manera equivalente, rechazar  $H_0$  si cero no pertenece al intervalo de confianza  $\left[ \widehat{\beta}_1 \pm t_{\alpha/2}^{(n-2)} \sqrt{\widehat{\text{var}}(\widehat{\beta}_1)} \right]$ .

Además, si  $T$  es positivo (rec. negativo) deducimos una relación lineal positiva (rec. negativa) significativa entre ambas variables.

### 1.1.2 Caso general

Para contrastar la correlación lineal entre dos variables  $X$  e  $Y$  cuando la distribución  $P_{(X,Y)}$  de  $(X, Y)$  pertenece a una familia paramétrica (ejemplo una normal), se suele basar la decisión en el coeficiente de Pearson

$$\begin{aligned} r(x, y) &= \frac{\mathbf{cov}(x, y)}{s_y s_x} \\ &= \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \end{aligned}$$

Bajo  $H_0$ , se conoce la distribución de  $r(x, y)$  (que depende de  $P_{(X,Y)}$ ), lo que permite deducir el umbral  $u_a$  de modo que el test

$$\text{Rechazar } H_0 \text{ si } |r(x, y)| > u_a$$

tenga un riesgo de tipo I igual  $\alpha$ .

**Ejercicio:** supongamos que la distribución conjunta de  $X$  e  $Y$  es la normal bivalente:

$$N \left( \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right)$$

Demostrar que el estimador máximo verosímil de  $\rho$  es  $\widehat{\rho} = r(x, y)$ . Deducir un test para contrastar  $H_0 : \rho = 0$ .

### 1.1.3 Test de independencia: coeficiente de Spearman y Kendall

Sabemos que si  $X$  e  $Y$  no son correladas y  $P_{(X,Y)}$  es normal entonces  $X$  e  $Y$  son independientes. Sin embargo, este resultado es falso en general. Por otra parte, si desconocemos la distribución  $P_{(X,Y)}$  no podremos conocer tampoco la distribución del coeficiente de Pearson bajo  $H_0$ , y por lo tanto no podremos construir un test basado en  $r(x, y)$ . Utilizaremos coeficientes no paramétricos (coeficientes de Spearman y Kendall) en los dos casos siguientes:

- Queremos contrastar una asociación monótona (no necesariamente lineal):  $H_0$  : “ $X$  e  $Y$  son independientes”
- Desconocemos la distribución  $P_{(X,Y)}$ .

### 1.1.4 Coeficiente de Spearman

Ordenamos los valores observados de  $Y$  por orden creciente:

- $Y_i$  está en el  $i$ -ésimo lugar
- $X_i$  está en el  $R_i$  lugar. ( $R_i$  es el rango o número de orden de  $X_i$ )

**Ejemplo:** Observamos los valores siguiente de  $(Y_i, X_i)$ ,  $i = 1, \dots, 4$ :  $(2, 8)$ ,  $(3, 5)$ ,  $(4, 2)$ ,  $(8, 4)$ . Por lo tanto

$R_i$	4	3	1	2
$X_i$	8	5	2	4
$Y_i$	2	3	4	8

El coeficiente de Spearman se define por

$$\begin{aligned} r_S &= 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^n (i - R_i)^2 \\ &= \frac{\sum_{i=1}^n (R_i - \bar{R})(i - \bar{i})}{\sqrt{\sum_{i=1}^n (R_i - \bar{R})^2} \sqrt{\sum_{i=1}^n (i - \bar{i})^2}} \end{aligned}$$

Con esta última expresión deducimos que el coeficiente de Spearman corresponde a la correlación muestral de los rangos de  $X$  e  $Y$ .

#### Comentarios:

- $r_S$  es máximo e igual a 1 si  $R_i = i$  para cualquier  $i = 1, \dots, n$ .
- $r_S$  es mínimo e igual a  $-1$  si  $R_i = n - i + 1$ , para cualquier  $i = 1, \dots, n$ .
- $r_S$  es invariante por una transformación monótona de  $X$  o  $Y$ .
- La distribución de  $r_S$  bajo  $H_0$  no depende de  $P_{(X,Y)}$

La regla de decisión basada en  $r_S$  para contrastar  $H_0$  : “ $X$  et  $Y$  son incorreladas” será por lo tanto:

$$\text{Rechazar } H_0 \text{ si } |r_S| > u_\alpha$$

donde  $u_\alpha$  se obtiene mediante la tabla de la distribución del coeficiente de Spearman.

**Ejemplo** (Calculo de la distribución de  $r_S$  en el caso  $n = 3$ ) :  $Y_1 < Y_2 < Y_3$  y tenemos 6 posible orden para los  $X_i$  :

	$r_S$
$X_1 < X_2 < X_3$	1
$X_1 < X_3 < X_2$	1/2
$X_2 < X_1 < X_3$	1/2
$X_2 < X_3 < X_1$	-1/2
$X_3 < X_1 < X_2$	-1/2
$X_3 < X_2 < X_1$	-1

Bajo  $H_0$  : “ $X$  e  $Y$  son independientes” todas las combinaciones son equiprobables, por lo tanto tenemos:

$$P(r_S = 1) = P(r_S = -1) = \frac{1}{6} \text{ y } P(r_S = \frac{1}{2}) = P(r_S = -\frac{1}{2}) = \frac{1}{3}$$

**Ejemplo:** Datos sobre  $n = 10$  alumnos para analizar la correlación entre las horas de estudio y el número de respuestas correctas en un examen

Horas de estudio ( $X$ )	5	9	17	1	2	21	3	29	7	100
nºrespuestas correctas ( $Y$ )	6	16	18	1	3	21	7	20	15	22

El grafico suggiere una asociación positiva (de tipo logaritmico) entre  $Y$  e  $X$ . Para confirmar este diagnonsis, vamos a contrastar las hiposetsis  $H_0$  : “ $X$  e  $Y$  son independientes” frente a  $H_1$  : “Existe una asociación positiva entre ambas variables” meidante el test de Spearman.

Obtenemos  $r_S = 0.976$ . Al consultar la tabla de la distribución de  $r_S$  bajo  $H_0$  para  $n = 10$ , se obtiene que  $P(r_S > 0.648|H_0) \simeq 5\%$  y  $\hat{p} = P(r_S > 0.976|H_0) < 0.001$ .

Por lo tanto, deducimos que ambas variables están fuertemente relacionadas y que esta asociación es positiva.

### 1.1.5 Coeficiente de Kendall

Suponiendo los  $Y_i$  ordenados como en el marco anterior, el coeficiente de Kendall se define por:

$$r_K = 1 - \frac{4}{n(n-1)}Q$$

donde  $Q = \sum_{i < j} \mathbf{I}\{X_i > X_j\}$  corresponde al numero de veces que el orden de los  $X_i$  no respeta el orden de los  $Y_i$ .

**Comentarios:**

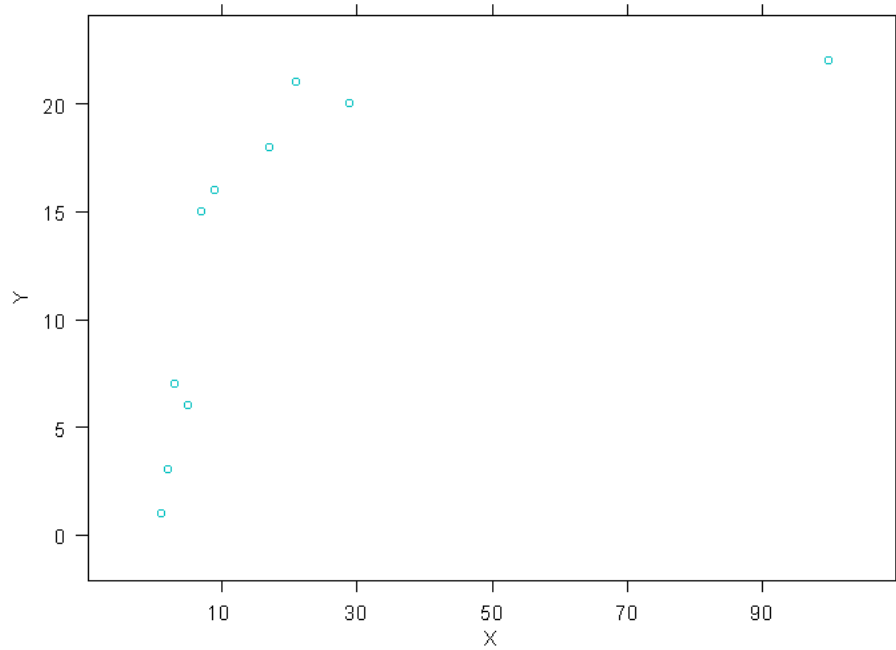


Figure 1: Horas de estudio ( $X$ ) versus n°de respuestas correctas ( $Y$ )

- $Q$  es máximo e igual a  $\frac{n(n-1)}{2}$  si los  $X_i$  están ordenado en sentido contrario del de los  $Y_i$ .
- $Q$  es mínimo e igual a 0 si los  $X_i$  e  $Y_i$  están ordenado en el mismo sentido.
- $r_K$  es invariante por una transformación monótona de  $X$  o  $Y$ .

La distribución de  $r_K$  bajo  $H_0$  no depende de  $P_{(X,Y)}$  y por lo tanto se puede construir un test para contrastar  $H_0$  cuya regla de decisión será:

$$\text{Rechazar } H_0 \text{ si } |r_K| > u_\alpha$$

donde  $u_\alpha$  verifica  $P(|r_K| > u_\alpha | H_0) = \alpha$ .