INTRODUCCIÓN

En esta unidad se estudiará el caso en el que más de una variable independiente puede influir en el comportamiento de la variable dependiente.

Para describir la forma de la relación que liga a estas variables se utilizarán los llamados modelos de regresión múltiple.

Por ejemplo, supongamos que la variable x_1 define el precio anual del trigo, x_2 la cantidad de fertilizantes utilizada y la variable "y", las hectáreas sembradas anualmente en una región; es posible estudiar el efecto del precio y de los fertilizantes en la producción de trigo.

5. ANÁLISIS DE REGRESIÓN LINEAL MÚLTIPLE

5.1. Análisis de correlación

A continuación, profundizaremos el tema de la correlación lineal.

El análisis de correlación mide el grado de asociación que existe entre dos variables y para ello utiliza un coeficiente denominado **coeficiente de correlación lineal**.

Ya hemos visto que el coeficiente de correlación lineal muestral se calcula de la siguiente manera:

$$r = \frac{\text{cov}\,(x, y)}{\sigma_x \,.\, \sigma_y}$$

En el numerador del coeficiente de correlación lineal aparece una medida de variabilidad conocida como covarianza.

Aunque la covarianza es difícil de interpretar, podemos decir que cuando grandes valores de x están asociados a grandes valores de y, o cuando pequeños valores de x están asociados a pequeños valores de y, la covarianza tendrá signo positivo.

En cambio, cuando altos valores de x estén asociados con bajos valores de y, y viceversa, la covarianza tomará signo negativo.

Entonces, como las desviaciones estándares de las variables siempre son positivas, el signo del coeficiente de correlación lineal depende del signo que presente la covarianza.

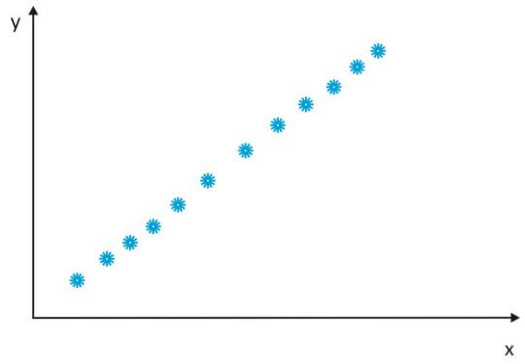
Por otra parte, la covarianza es una medida de variabilidad que se ve afectada por las unidades de medida de las variables que intervienen en su cálculo. Por ejemplo, si la variable

x está medida en metros, la covarianza adoptará un valor mayor que si está medida en centímetros.

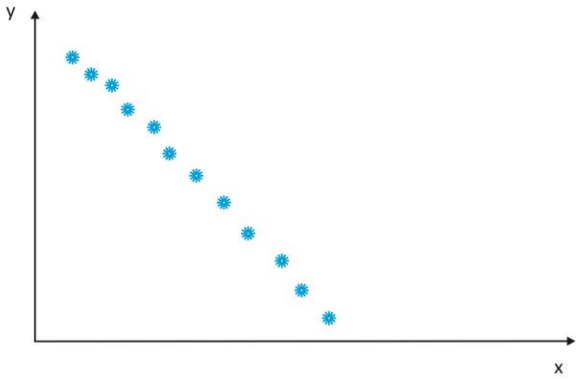
Por este motivo, se necesita construir una medida para el grado de asociación entre dos variables pero que a su vez sea independiente de su escala de medición.

Al estandarizar la covarianza dividiéndola por las correspondientes desviaciones estándares de las variables en estudio, este problema se soluciona y se obtiene así el coeficiente de correlación lineal que no depende de la escala de medición de las variables x e y.

Ya dijimos también que el coeficiente de correlación varía en el Intervalo [-1, 1]. La situación en que r = 1 refleja una relación lineal perfecta positiva entre las variables x e y. Utilizando el diagrama de dispersión, esta situación se observa del siguiente modo:

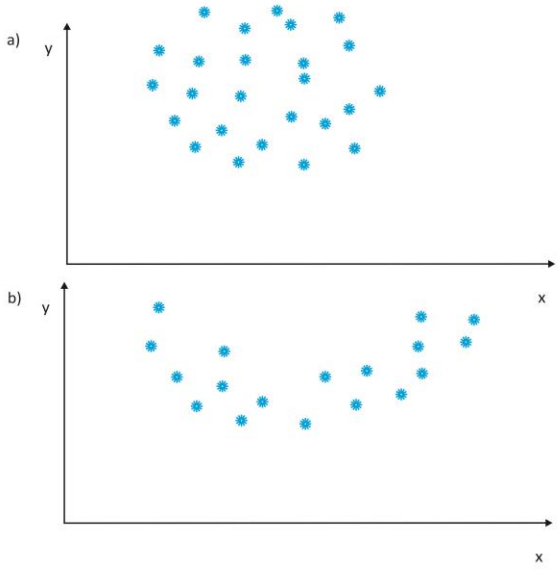


Por otra parte, cuando r = - 1 significa que existe una relación lineal perfecta negativa entre las variables en estudio. Esta situación se observa en el siguiente gráfico:

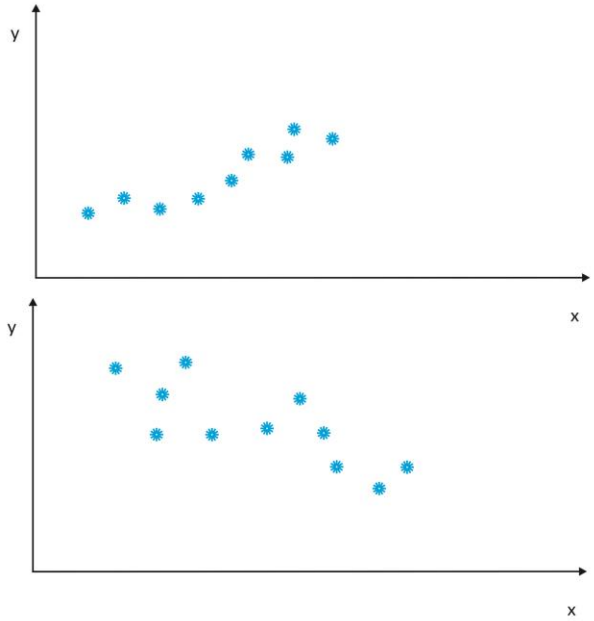


Otra situación extrema ocurre cuando r = 0. Este resultado indica tanto que no existe relación entre las variables en estudio como que la relación existente entre ellas no es lineal sino que responde, por ejemplo, a una relación del tipo parabólica.

La Figura (a) presenta la primera de estas situaciones (ausencia de relación) y la Figura (b) la segunda (relación no lineal).



En la práctica de investigación se dan generalmente situaciones Intermedias como, por ejemplo, las que se observan en las siguientes figuras:



Ahora bien, ¿cómo se puede estimar el coeficiente de correlación lineal?

Para estimar el coeficiente de correlación reemplazamos en su fórmula los parámetros poblacionales por sus correspondientes estimadores muestrales: la covarianza muestral y las desviaciones estándares muestrales de ambas variables.

Al igual que para estimar la recta de regresión poblacional, no necesitamos efectuar ningún supuesto en cuanto a la distribución de las variables en estudio. Sí deberemos hacerlo cuando efectuemos inferencias acerca del parámetro coeficiente de correlación poblacional, como veremos un poco más adelante.

El coeficiente de correlación muestral será calculado de la siguiente manera:

$$r = \frac{\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1} \cdot \frac{\sum (y_i - \bar{y})^2}{n-1}}}$$

Este coeficiente se conoce también con el nombre de **coeficiente de Pearson** pues fue uno de los primeros en estudiarlo.

Veremos ahora un ejemplo donde calcularemos el coeficiente de correlación lineal.

Ejemplo

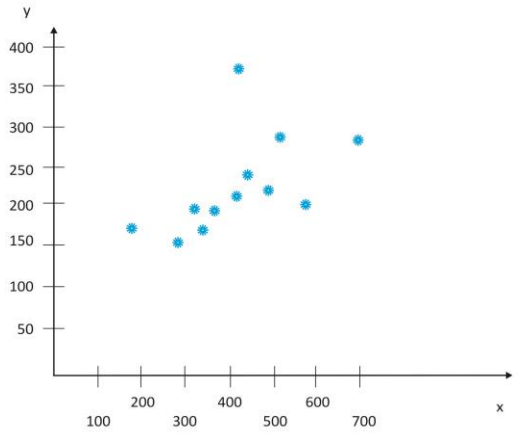
Los siguientes datos representan mediciones de resistencia al doblamiento y resistencia a la torsión de 12 raquetas de distintas marcas.

Marca	Resistencia al doblamiento (x)	Resistencia a la torsión (y)
Dunlop Maxply Fort	419	227
García 240	407	231
Bancroft Bjorn Borg	363	200
Wilson Jack Kramer	360	211
Davis Classic	257	182
Spalding Smasher III	622	304
Yonex T-7500	424	384
Prince	359	194
Wilson T-400	346	158
Yamaha YFG-30	556	225
Head Competition II	474	305
Adidas Adistar	441	235

Importante

El objetivo de estudio consiste en averiguar si las raquetas que tienen buena resistencia al doblamiento tienen también buena resistencia a la torsión.

Como primera medida construiremos el diagrama de dispersión:



Parecería que entre las dos variables existe una relación lineal positiva.
Para corroborarlo calcularemos el coeficiente de correlación lineal.
Comenzaremos con el cálculo de las medidas descriptivas de ambas variables:

$\bar{x} = 419$ $\quad \bar{y} = 238$

$S_x = 97{,}91$ $\quad S_y = 63{,}36$

Calculamos ahora la covarianza:

$$\text{cov}(x, y) = \frac{\sum x_i \, y_i - n \, \bar{x} \, \bar{y}}{n - 1}$$

Para facilitar este cálculo ordenaremos los datos en una tabla conveniente:

x_i	y_i	$x_i \cdot y_i$
419	227	95113
407	231	94017
363	200	72600
360	211	75960
257	182	46774
622	304	189088
424	384	162816
359	194	69646
346	158	54668
556	225	125100
474	305	144570
441	235	103635
5028	2856	1233987

Entonces,

$$cov\,(x, y) = \frac{1233987 - 12 \times 419 \times 238}{11} = \frac{1233987 - 1196664}{11} = 3393$$

y

$$r = \frac{3393}{97{,}91 \times 63{,}36} = \frac{3393}{6203{,}58} = 0{,}55$$

El coeficiente de correlación lineal ha dado 0,55.

Ahora bien, ¿este valor es indicador de una alta correlación entre las variables o no?

Para poder estar un poco más seguros de la decisión a adoptar, debemos apelar a un test de hipótesis con respecto al parámetro coeficiente de correlación poblacional.

Luego, la hipótesis será:

$$H_0)\ p = 0 \qquad H_1)\ p \neq 0$$

El estadístico que usaremos para probar esta hipótesis es el siguiente:

$$t = \frac{r\sqrt{n - 2}}{\sqrt{1 - r^2}}$$

Cuando x e y siguen una distribución normal bivariada, este estadístico tiene distribución t de Student con n - 2 grados de libertad.

Luego comparamos el valor del estadístico con los siguientes valores críticos obtenidos de la tabla de probabilidades de la variable t de Student:

$$t_1^+ = t_{n-2;\alpha/2} \qquad t_2^+ = t_{n-2;\,1-\alpha/2}$$

La regla de decisión será:

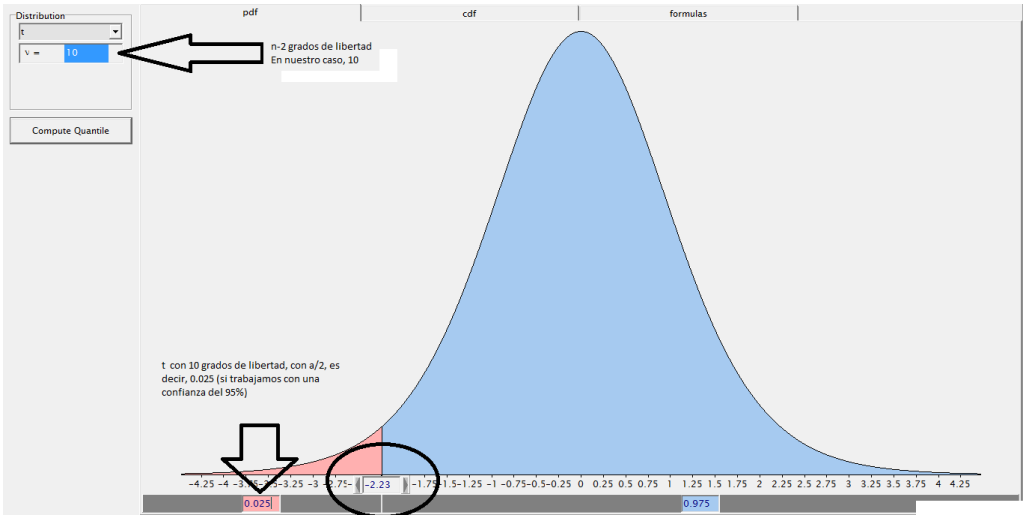
$$t = \frac{0{,}55\sqrt{12-2}}{\sqrt{1-(0{,}55)^2}} = \frac{0{,}55 \times 3{,}16}{0{,}84} = 2{,}069$$

Si $\alpha = 0{,}05$, los valores críticos son:

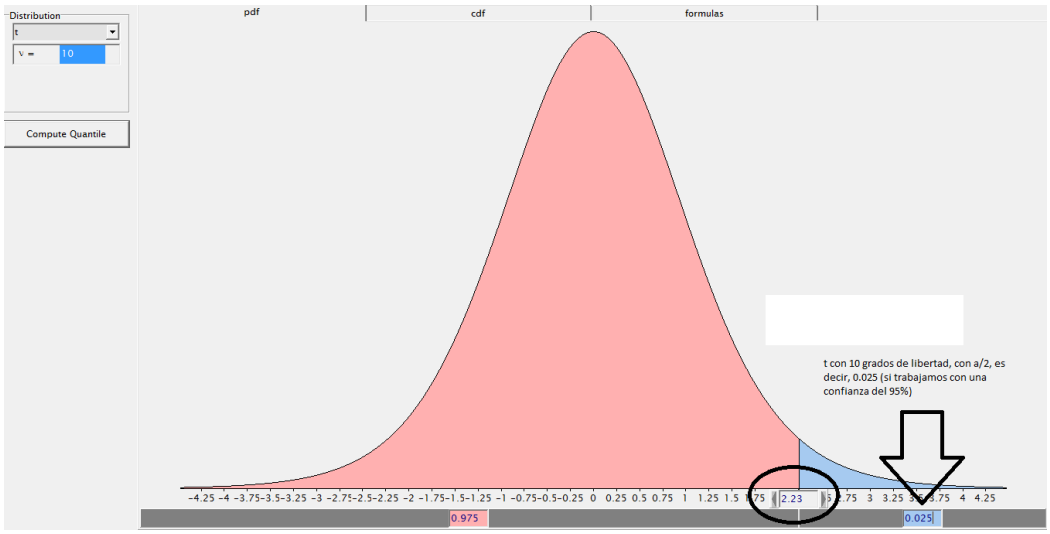
$$t_1^+ = -2{,}228 \qquad t_2^+ = 2{,}228$$

Ilustrado en PQRS , la gráfica sería:

Para el lado izquierdo:



Análogamente, en el lado derecho



Como - 2,228 < 2,069 < 2,228 se decide no rechzar H_0 . No existe evidencia estadísticamente significativa para afirmar que las variables estén correlacionadas.