Tema 62

Tablas estadísticas bidimensionales. Regresión y correlación lineal. Coeficiente de correlación. Significado y aplicaciones

62.1 Introducción

Consideremos una población de n individuos descritos simultáneamente según dos carácteres A y B. Designemos por $A_1, ..., A_i, ..., A_k$ las k modalidades del carácter A y por $B_1, ..., B_j, ..., B_p$ las p modalidades del carácter B. Sea n_{ij} el número de individuos de la población que presentan a la vez la modalidad A_i del carácter A y la modalidad B_j del carácter B. Debido a que las modalidades de A, así como las de B, son incompatibles y exhaustivas, la suma de las frecuencias absolutas n_{ij} es igual al total de la población:

$$\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij} = n$$

La tabla estadística que describe a los n individuos, es una tabla de doble entrada, donde figuran en las filas las modadlidades de A y en las columnas las modalidades de B. Se supondrá que todas las frecuencias absolutas de una misma fila (o de una misma columna) no se anulan simultáneamente. Si ocurriese, bastaría no considerar la modalidad correspondiente de A (o de B) o

bien agruparla con otra modalidad.

$A \diagdown B$	B_1	B_2	$\cdots$	B_j	$\cdots$	B_p	
A_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1p}	$n_{1\bullet}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2p}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$		$\vdots$	$\vdots$
A_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ip}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_k	n_{k1}	n_{k2}	$\cdots$	n_{kj}	$\cdots$	n_{kp}	$n_{k\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet p}$	

Se designa por un punto la totalización según el índice i ó el índice j : $n_{i\bullet}$ es el total de las frecuencias absolutas n_{ij} según j; $n_{\bullet j}$ es el total de las frecuencias absolutas n_{ij} según i, es decir:

$$n_{i\bullet} = \sum_{j=1}^{p} n_{ij} \qquad n_{\bullet j} = \sum_{i=1}^{k} n_{ij}$$

$$n_{\bullet\bullet} = \sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij} = \sum_{i=1}^{k} n_{i\bullet} = \sum_{j=1}^{p} n_{\bullet j}$$

La frecuencia absoluta $n_{i\bullet}$ es el número de individuos de la población que presentan la modalidad A_i del carácter A, independientemente de las modalidades del carácter B. Análogamente, $n_{\bullet j}$ es el número de individuos de la población que corresponden a la modalidad B_j del carácter B. Se llama frecuencia relativa de la pareja de modalidades (A_i, B_j) a la proporción de individuos que presentan simultáneamente las modalidades A_i y B_j :

$$f_{ij} = \frac{n_{ij}}{n}$$

La suma de las frecuencias totales extendida a todos los pares de modalidades posibles es igual a la unidad:

$$\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij} = 1$$

Las sumas parciales se designan igualmente por un punto en lugar del índice, que hace la función de sumatorio:

$$f_{i\bullet} = \sum_{j=1}^{p} f_{ij} = \frac{n_{i\bullet}}{n} \qquad f_{\bullet j} = \sum_{i=1}^{k} f_{ij} = \frac{n_{\bullet j}}{n}$$

$$\sum_{j=1}^{p} f_{\bullet j} = \sum_{i=1}^{k} f_{i\bullet} = 1$$

62.2 Variables estadísticas bidimensionales

Consideremos una población de n individuos descrita según dos carácteres cuantitativos X e Y, es decir, según dos variables estadísticas: el par (X, Y) se denomina variable estadística bidimensional.

Supongamos que X e Y son discretas, en el caso de que fueran continuas x_i e y_j designarían las marcas de clase:

$$X : x_1, x_2, ..., x_i, ..., x_k$$

$$Y : y_1, y_2, ..., y_j, ..., y_p$$

La tabla estadística que describe la población da la frecuencia absoluta n_{ij} de individuos que presentan a la vez el valor x_i de la variable estadística X y el valor y_j de la variable estadística Y.

$X \diagdown Y$	Y_1	Y_2	$\cdots$	Y_j	$\cdots$	Y_p	
X_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1p}	$n_{1\bullet}$
X_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2p}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$		$\vdots$	$\vdots$
X_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ip}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_k	n_{k1}	n_{k2}	$\cdots$	n_{kj}	$\cdots$	n_{kp}	$n_{k\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet p}$	

62.2.1 Distrubuciones marginales

Son distribuciones unidimensionales que estudian una variable independientemente de la otra. Sus tablas son:

X	Frec. absolutas	Frec. relativas
X_1	$n_{1\bullet}$	$f_{1\bullet}$
$\vdots$	$\vdots$	$\vdots$
X_i	$n_{i\bullet}$	$f_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$
X_k	$n_{k\bullet}$	$f_{k\bullet}$
	n	1

Y	Frec. absolutas	Frec. relativas
Y_1	$n_{\bullet 1}$	$f_{\bullet 1}$
$\vdots$	$\vdots$	$\vdots$
Y_j	$n_{\bullet j}$	$f_{\bullet j}$
$\vdots$	$\vdots$	$\vdots$
Y_p	$n_{\bullet p}$	$f_{\bullet p}$
	n	1

62.2.2 Distribuciones condicionadas

Son distribuciones unidimensionales que estudian una variable supuesto un determinado valor para la otra. Sus tablas son:

$X/_{Y=y_j}$	Frec. abs.	Frec. cond.
X_1	n_{1j}	f_1^j
$\vdots$	$\vdots$	$\vdots$
X_i	n_{ij}	f_i^j
$\vdots$	$\vdots$	$\vdots$
X_k	n_{kj}	f_k^j
	$n_{\bullet j}$	1

$Y/_{X=x_i}$	Frec. abs.	Frec. cond.
Y_1	n_{i1}	f_1^i
$\vdots$	$\vdots$	$\vdots$
Y_j	n_{ij}	f_p^i
$\vdots$	$\vdots$	$\vdots$
Y_p	n_{ip}	f_p^i
	$n_{\bullet j}$	1

62.2.3 Características marginales y condicionadas

Consideremos la columna marginal de la tabla: da las frecuencias absolutas n_i de individuos que presentan el valor x_i de X, es decir, define la variable marginal X. Las carácterísticas marginales de X se designan por:

$$\text{Media:} \qquad \overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}x_i = \sum_{i=1}^{k} f_{i\bullet}x_i$$

$$\text{Varianza:} \qquad V(X) = \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}\left(x_i - \overline{x}\right)^2 = \sum_{i=1}^{k} f_{i\bullet}\left(x_i - \overline{x}\right)^2$$

Análogamente, la variable marginal Y tiene las siguientes carácterísticas:

$$\text{Media:} \qquad \overline{y} = \sum_{j=1}^{p} f_{\bullet j}y_j$$

$$\text{Varianza:} \qquad V(Y) = \sum_{j=1}^{p} f_{\bullet j}\left(y_j - \overline{y}\right)^2$$

Consideremos ahora la j-ésima columna de la tabla estadística: describe los n_{ij} individuos que presentan el valor y_j de Y según la variable X. Define, por consiguiente, la variable condicionada $X/_{Y=y_j}$. Las carácterísticas de esta distribución condicionada son:

$$\text{Media:} \qquad \overline{x_j} = \frac{1}{n_{\bullet j}}\sum_{i=1}^{k} n_{ij}x_i = \sum_{i=1}^{k} f_i^j x_i$$

$$\text{Varianza:} \qquad V(X) = \frac{1}{n_{\bullet j}}\sum_{i=1}^{k} n_{ij}\left(x_i - \overline{x_j}\right)^2 = \sum_{i=1}^{k} f_i^j\left(x_i - \overline{x_j}\right)^2$$

Análogamente, se designan las carácterísticas de la distribución condicionada $Y\diagup_{X=x_i}$ por:

$$\text{Media:} \qquad \overline{y_i} = \sum_{j=1}^{p} f_j^i y_j$$

$$\text{Varianza:} \qquad V_i(Y) = \sum_{j=1}^{p} f_j^i \left(y_j - \overline{y_i}\right)^2$$

62.2.4 Relación de la media marginal con las condicionadas

Proposición 1 *La media marginal de X es la media ponderada de las medias condicionadas, es decir:*

$$\overline{x} = \frac{1}{n} \sum_{j=1}^{p} n_{\bullet j} \overline{x_j}$$

Demostración:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{k} n_{i\bullet} x_i \underset{(1)}{=} \frac{1}{n} \sum_{i=1}^{k} x_i \left(\sum_{j=1}^{p} n_{ij} \right) = \frac{1}{n} \sum_{j=1}^{p} \underbrace{\sum_{i=1}^{k} n_{ij} x_i}_{=\overline{x_j} n_{\bullet j}} = \frac{1}{n} \sum_{j=1}^{p} n_{\bullet j} \overline{x_j}$$

donde en (1) hemos usado que $n_{i\bullet} = \sum_{j=1}^{p} n_{ij}$. C.Q.D. □

62.2.5 Relación de la varianza marginal con las condiconadas

Proposición 2 *La varianza marginal es igual a la media ponderada de las varianzas más la varianza ponderada de las medias, es decir,*

$$V(X) = \overline{V_j(X)} + V\left(\overline{X_j}\right)$$

Demostración:

$$V(X) = \tfrac{1}{n} \sum_{i=1}^{k} n_{i\bullet} \left(x_i - \overline{x}\right)^2 \underset{(1)}{=} \tfrac{1}{n} \sum_{i=1}^{k} n_{i\bullet} \left[\left(x_i - \overline{x_j}\right) + \left(\overline{x_j} - \overline{x}\right) \right]^2 =$$

$$= \tfrac{1}{n} \sum_{i=1}^{k} n_{i\bullet} \left(x_i - \overline{x_j}\right)^2 + \tfrac{1}{n} \sum_{i=1}^{k} n_{i\bullet} \left(\overline{x_j} - \overline{x}\right)^2 + \underbrace{\frac{2}{n} \sum_{i=1}^{k} n_{i\bullet} \left(x_i - \overline{x_j}\right) \left(\overline{x_j} - \overline{x}\right)}_{:=A}$$

$$A = \tfrac{2}{n} \sum_{i=1}^{k} n_{i\bullet} \left(x_i - \overline{x_j}\right) \left(\overline{x_j} - \overline{x}\right) = \tfrac{2}{n} \sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij} \left(x_i - \overline{x_j}\right) \left(\overline{x_j} - \overline{x}\right) =$$

$$= \tfrac{2}{n} \sum_{j=1}^{p} \left(\overline{x_j} - \overline{x}\right) \sum_{i=1}^{k} n_{ij} \left(x_i - \overline{x_j}\right) \underset{(2)}{=} \tfrac{2}{n} \sum_{j=1}^{p} \left(\overline{x_j} - \overline{x}\right) 0 = 0$$

donde en (1) hemos sumado y restado $\overline{x_j}$ y en (2) hemos tenido en cuenta que

$$\sum_{i=1}^{k} n_{ij}\left(x_i-\overline{x_j}\right)=\sum_{i=1}^{k} n_{ij}x_i-\sum_{i=1}^{k} n_{ij}\overline{x_j}=n_{\bullet j}\overline{x_j}-\overline{x_j}\sum_{i=1}^{k} n_{ij}=n_{\bullet j}\overline{x_j}-\overline{x_j}n_{\bullet j}=0$$

Así:

$$V(X)=\tfrac{1}{n}\sum_{i=1}^{k} n_{i\bullet}\left(x_i-\overline{x_j}\right)^2+\tfrac{1}{n}\sum_{i=1}^{k} n_{i\bullet}\left(\overline{x_j}-\overline{x}\right)^2=\tfrac{1}{n}\sum_{j=1}^{p}\underbrace{\sum_{i=1}^{k} n_{ij}\left(x_i-\overline{x_j}\right)^2}_{=V_j(X)n_{\bullet j}}+$$

$$+\tfrac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{k} n_{ij}\left(\overline{x_j}-\overline{x}\right)^2=\sum_{j=1}^{p}\tfrac{n_{\bullet j}}{n}V_j(X)+\sum_{j=1}^{p}\tfrac{n_{\bullet j}}{n}\left(\overline{x_j}-\overline{x}\right)^2=$$

$$=\overline{V_j(X)}+V\left(\overline{X_j}\right)\qquad\qquad C.Q.D.\square$$

62.2.6 Momentos

Los momentos son valores construidos a partir de la distribución de frecuencias que resumen la información en relación a algún aspecto o propiedad de la variable.

Definición 3 *Llamamos momento no central de oreden* (r,s) *de la variable estadística bidimensional* (X,Y) *a:*

$$m_{rs}=\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}x_i^r y_j^s$$

Los primeros momentos no centrados son los siguientes:

$$m_{10}=\sum_{i=1}^{k} f_{i\bullet}x_i=\overline{x}$$

$$m_{01}=\sum_{j=1}^{p} f_{\bullet j}y_j=\overline{y}$$

$$m_{20}=\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}x_i^2=\sigma_X^2+\overline{x}^2=\sigma_X^2+m_{10}^2$$

$$m_{02}=\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}y_i^2=\sigma_Y^2+\overline{y}^2=\sigma_Y^2+m_{01}^2$$

$$m_{11}=\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}x_i y_j$$

Definición 4 *Se llama momento central (o respecto de la media) de la variable estadística bidimensional* (X,Y) *a:*

$$\mu_{rs} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\left(x_i - \overline{x}\right)^r \left(y_j - \overline{y}\right)^s$$

Los primeros momentos centrales son los siguientes:

$$\mu_{10} = 0$$

$$\mu_{01} = 0$$

$$\mu_{20} = \sigma_X^2$$

$$\mu_{02} = \sigma_Y^2$$

$$\mu_{11} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\left(x_i - \overline{x}\right)\left(y_j - \overline{y}\right)$$

Definición 5 *El momento central* μ_{11} *se llama covarianza de la variable estadística bidimensional* (X,Y) *y se representa por* $cov\,(X,Y)$ *o* $\sigma_{(X,Y)}$.

El signo de la covarianza indica el sentido en que varían conjuntamente ambas variables. Si es positivo, las dos variables por término medio varían en el mismo sentido, y si es negativo, en promedio las dos variables varían en sentido opuesto.

Relaciones entre momentos

Proposición 6 $\mu_{02} = m_{02} - m_{01}^2$ *y* $\mu_{20} = m_{20} - m_{10}^2$

Proposición 7 $\mu_{11} = m_{11} - m_{10}m_{01}$

Demostración:

$$\mu_{11} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\underbrace{\left(x_i - \overline{x}\right)\left(y_j - \overline{y}\right)}_{=x_iy_j+\overline{x}\overline{y}-\overline{x}y_j-x_i\overline{y}} =$$

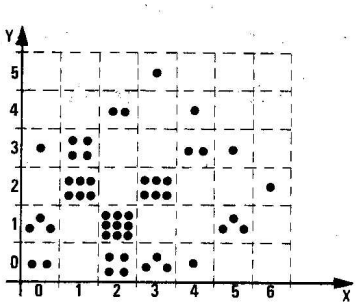
$$= \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}x_iy_j + \overline{x}\overline{y}\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij} - \overline{x}\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}y_j}_{=\overline{y}} - \overline{y}\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}x_i}_{=\overline{x}} =$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}x_iy_j + \overline{x}\overline{y} - 2\overline{x}\overline{y} = m_{11} - \overline{x}\overline{y} \qquad C.Q.D.\square$$

62.2.7 Representación gráfica

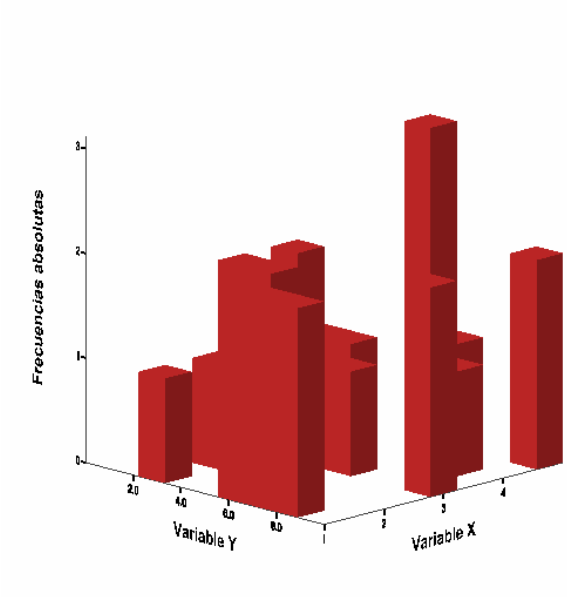
Diagrama de dispersión o nube de puntos

Consiste en unos ejes cartesianos, en los cuales van las variables. Supongamos, por comodidad, que las variables son discretas. En los puntos de confluencia de cada recta correspondiente a cada valor de la variable, se dibujan tantos puntos como frecuencia absoluta tenga el par.



Estereogramas

Consisten en una serie de barras o prismas rectangulares que tienen por altura n_i y cuya base es un punto (v.e.d.) o un rectángulo de confluencia de los intervalos (v.e.c.).



62.3 Regresión y correlación

62.3.1 Introducción

A la hora de establecer relaciones entre sucesos en un determinado campo de investigación, el investigador intenta traducirlas en estructuras manejables, haciendo uso fundamentalmente del lenguaje estadístico-matemático. Para ello, establece un conjunto de relaciones funcionales en donde un número finito de magnitudes (variables, atributos,...) $X_1, ..., X_n$ se suponen relacionadas con una variable Y a través de una determinada expresión:

$$Y = f(X_1, ..., X_n)$$

Desde esta perspectiva el problema se puede abordar con dos **enfoques**:

(1) Regresión: Determinación de la estructura de dependencia que mejor expresa el tipo de relación de la variable Y con las demás.

(2) Correlación: Estudia el grado de dependencia existente entre las variables

62.3.2 Regresión[1]

Definición 8 *Llamaremos regresión de Y sobre X, Y/X, a la función que explica la variable Y para cada valor de X.*

62.3.3 Regresión mínimo-cuadrática

Previa selección de la familia de funciones h tal que $Y = h(X)$, ajusta la "mejor" de esa familia, haciendo mínima la media de los cuadrados de los resíduos, es decir,

$$\min_{a_1,...,a_m} \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} (y_j - h(x_i))^2$$

donde $h(x_i) = h(x_i, a_1, ..., a_m)$ se ha seleccionado previamente y los $a_1, ..., a_m$ son desconocidos.

Regresión lineal

Sea (x_i, y_j, n_{ij}) la nube de puntos a la que queremos ajustar la recta $y = ax + b$. Para ello, tenemos que hacer mínima la siguiente expresión:

$$\Phi(a, b) := \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} (y_j - ax_i - b)^2$$

[1] A lo largo de todo el tema lo que se haga para Y/X se puede traducir literalmente para X/Y.

Derivamos e igualamos a cero:

$$\begin{cases} \frac{\partial}{\partial a}\Phi(a,b) = 2\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\left(y_j - ax_i - b\right)(-x_i) = 0 \\ \\ \frac{\partial}{\partial b}\Phi(a,b) = 2\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\left(y_j - ax_i - b\right)(-1) = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\left(y_j - ax_i - b\right)(-x_i) = 0 \\ \\ \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\left(y_j - ax_i - b\right)(-1) = 0 \end{cases}$$

$$\begin{cases} -\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}y_j x_i + a\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}x_i^2 + b\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}x_i = 0 \\ \\ -\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}y_j + \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}x_i + b\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij} = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}y_j x_i = a\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}x_i^2 + b\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}x_i \\ \\ \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}y_j = a\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}x_i + b\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij} \end{cases}$$

$$\begin{cases} \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}x_i y_j = a\sum_{i=1}^{k} f_{i\bullet}x_i^2 + b\sum_{i=1}^{k} f_{i\bullet}x_i \\ \\ \sum_{j=1}^{p} f_{\bullet j}y_j = a\sum_{i=1}^{k} f_{i\bullet}x_i + b \end{cases}$$

Sistema de ecuaciones normales

$$\begin{cases} \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}x_i y_j = a\sum_{i=1}^{k} f_{i\bullet}x_i^2 + b\overline{x} \qquad (*) \\ \\ \overline{y} = a\overline{x} + b \end{cases}$$

De la segunda ecuación resulta que

$$b = \overline{y} - a\overline{x}$$

y sustituyendo en $(*)$ obtenemos:

$$\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}x_iy_j = a\sum_{i=1}^{k} f_{i\bullet}x_i^2 + \overline{yx} - a\overline{x}^2$$

$$\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}x_iy_j - \overline{xy} = a\left(\sum_{i=1}^{k} f_{i\bullet}x_i^2 - \overline{x}^2\right)$$

$$\sigma_{(X,Y)} = a\sigma_X^2 \Rightarrow \begin{cases} a = \frac{\sigma_{(X,Y)}}{\sigma_X^2} \\ b = \overline{y} - \frac{\sigma_{(X,Y)}}{\sigma_X^2}\overline{x} \end{cases}$$

Sustituyendo en la ecuación de la recta obtenemos:

$$y = \frac{\sigma_{(X,Y)}}{\sigma_X^2}x + \overline{y} - \frac{\sigma_{(X,Y)}}{\sigma_X^2}\overline{x}$$

$y - \overline{y} = \frac{\sigma_{(X,Y)}}{\sigma_X^2}(x - \overline{x})$	Recta de regresión de Y/X

De forma análoga se obtiene que la recta de regresión de X/Y es:

$$x - \overline{x} = \frac{\sigma_{(X,Y)}}{\sigma_Y^2}(y - \overline{y})$$

Las dos rectas de regresión se cortan en un punto que recibe el nombre de centro de gravedad de la distribución.

Vamos a estudiar los coeficientes de la recta de regresión:
Tenemos que

$$y = \frac{\sigma_{(X,Y)}}{\sigma_X^2}x + \overline{y} - \frac{\sigma_{(X,Y)}}{\sigma_X^2}\overline{x}$$

luego

$$a = \frac{\sigma_{(X,Y)}}{\sigma_X^2}$$

que es la pendiente de la recta de regresión de Y/X. Por tanto, dependerá del signo de la covarianza el que las dos rectas sean crecientes o decrecientes.

62.3.4 Análisis de la correlación

Introducción

Llamamos correlación al grado de dependencia mutua que puede existir entre las variables según una determinada función de regresión. El objetivo de la Teoría de la Correlación es la determinación de medidas que cuantifiquen la intensidad con que las variables se relacionan según distintas funciones de regresión, por lo

tanto, lo que se va a pretender es "medir" la bondad de las funciones ajustadas a los datos dados.

Según el gráfico anterior tenemos que

$$e_{ij} = y_j - y^*$$

y por tanto, si los errores son grandes el ajuste será malo, mientras que si los errores son pequeños entonces el ajuste es bueno.

Llamamos varianza residual a la media de todos los residuos elevados al cuadrado, es decir:

$$\sigma^2_{rY} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}e^2_{ij} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\left(y_j - y^*\right)^2$$

$$\sigma^2_{rX} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\left(x_i - x^*\right)^2$$

La varianza residual puede parecer adecuada como una primera medida de la correlación en el siguiente sentido: cuanto mayor sea la varianza residual peor será el ajuste, es decir, menor será la intensidad con que las variables X e Y están relacionadas mediante la función de regresión. Si embargo, como consecuencia de lo anterior se complica mucho la comparación entre la dependencia de los diversos grupos de variables. Por lo tanto la Teoría de la Correlación va a definir como medidas de correlación unos índices adimensionales que dependan de los valores de las medidas y que armonicen el sentido entre el valor de las medidas y el sentido de la correlación.

Supongamos que es posible la siguiente descomposición:

Varianza total = Varianza residual + Varianza debida a la regresión

$$\sigma^2_Y = \sigma^2_{rY} + \sigma^2_{Y^*}$$

donde

$$\sigma^2_{Y^*} = \sum_{i=1}^{k} f_{i\bullet}\left(\overline{y_i} - \overline{y}\right)^2$$

es la varianza debida a la regresión.

Definición 9 *Llamamos razón de correlación de Y/X a la proporción de variabilidad de Y explicada por la regresión, es decir,*

$$RC_{Y/X} = \frac{\sigma^2_{Y^*}}{\sigma^2_Y} = \frac{\sum_{i=1}^{k} f_{i\bullet}\left(\overline{y_i} - \overline{y}\right)^2}{\sum_{j=1}^{p} f_{\bullet j}\left(y_j - \overline{y}\right)^2}$$

De la descomposición de la varianza resulta:

$$\sigma^2_{Y^*} = \sigma^2_Y - \sigma^2_{rY} \Leftrightarrow \frac{\sigma^2_{Y^*}}{\sigma^2_Y} = \frac{\sigma^2_Y}{\sigma^2_Y} - \frac{\sigma^2_{rY}}{\sigma^2_Y} \Leftrightarrow \frac{\sigma^2_{Y^*}}{\sigma^2_Y} = 1 - \frac{\sigma^2_{rY}}{\sigma^2_Y} \Leftrightarrow$$

$$\Leftrightarrow 1 - \frac{\sigma^2_{rY}}{\sigma^2_Y} = RC_{Y/X}$$

La razón de correlación es un índice adimensional que armoniza el sentido de su valor numérico con la intensidad de la correlación, es decir, a mayor valor de la razón de correlación mayor será el grado de intensidad con que se relacionan las variables.

Como consecuencia de lo anterior se tiene que:

$$0 \leq RC_{Y/X} \leq 1$$

Correlación y regresión lineal

Supongamos que dada una distribución bidimensional el modelo más adecuado que explica el comportamiento de una variable a partir de los valores de la otra es un modelo lineal, es decir, una ecuación del tipo $y = ax + b$, supuesta ajustada por el método de mínimos-cuadrados. Vamos a determinar el grado de asociación existente entre X e Y según esta función, esto es, vamos a cuantificar el grado en que las variables estan correladas según una recta. Por lo tanto, lo que estamos midiendo es el grado de acierto al ajustar una recta a la distribución dada.

Vamos a ver si es posible descomponer la varianza en dos términos: varianza

residual y varianza explicada por la regresión.

$$\sigma_Y^2 = \sum_{j=1}^{p} f_{\bullet j} \left(y_j - \overline{y}\right)^2 = \sum_{j=1}^{p} f_{\bullet j} \left[\left(y_j - ax_i - b\right) + \left(ax_i + b - \overline{y}\right)\right]^2 =$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} \left(y_j - ax_i - b\right)^2 + \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} \left(ax_i + b - \overline{y}\right)^2 +$$

$$\underbrace{+2 \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} \left(y_j - ax_i - b\right) \left(ax_i + b - \overline{y}\right)}_{=A} \qquad (*)$$

$$A = 2 \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} \left(y_j - ax_i - b\right) \left(ax_i + b - \overline{y}\right) =$$

$$= 2 \sum_{i=1}^{k} \left(ax_i + b - \overline{y}\right) \sum_{j=1}^{p} f_{ij} \left(y_j - ax_i - b\right) =$$

$$\underset{(1)}{=} 2 \sum_{i=1}^{k} \left(ax_i + b - a\overline{x} - b\right) \sum_{j=1}^{p} f_{ij} \left(y_j - ax_i - b\right)$$

$$\underset{(2)}{=} 2a \sum_{i=1}^{k} f_{i\bullet} \left(x_i - \overline{x}\right) \sum_{j=1}^{p} f_j^i \left(y_j - ax_i - b\right) = 0$$

donde en (1) hemos tenido en cuenta que $\overline{y} = a\overline{x} + b$ y en (2) que $f_{ij} = f_{i\bullet} f_j^i$.

Como

$$\sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} \left(y_j - ax_i - b\right)^2 = \sigma_{rY}^2$$

$$\sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} \left(ax_i + b - \overline{y}\right)^2 = \sigma_{Y^*}^2$$

sustituyendo en $(*)$ obtenemos que:

$$\sigma_Y^2 = \sigma_{rY}^2 + \sigma_{Y^*}^2$$

Definición 10 *Llamamos razón de correlación lineal a la proporción de variabilidad de Y explicada por la recta de regresión, es decir:*

$$r^2 = \frac{\sigma_{Y^*}^2}{\sigma_Y^2} = \frac{\sum_{i=1}^{k} f_{i\bullet} \left(ax_i + b - \overline{y}\right)^2}{\sum_{j=1}^{p} f_{\bullet j} \left(y_j - \overline{y}\right)^2} = 1 - \frac{\sigma_{rY}^2}{\sigma_Y^2}$$

Como consecuencia de la definición se tiene:

$$0 \leq r^2 \leq 1$$

Veamos la **interpretación** de r^2 :

Si $r^2 = 1$ entonces se dice que existe correlación lineal perfecta entre X e Y, es decir, la correlación lineal es máxima. Por tanto, la recta ajustada explica perfectamente el comportamiento de una variable por la otra. En esta situación la recta ajustada va a pasar por todos los puntos de la nube.

Si $r^2 = 0$ entonces se dice que existe correlación lineal nula entre las variables X e Y, es decir, que no existe asociación lineal entre las variables. Por lo tanto, la recta ajustada no explica en absoluto el comportamiento de una variable en relación con la otra. Ahora bien, esto no quiere decir que las variables sean independientes, sino que no están relacionadas mediante una recta.

Si $0 < r^2 < 1$ entonces existe cierto grado de correlación lineal, que será mayor cuanto más próximo esté a 1 y ello nos informará de la validez del ajuste, y cuanto más próximo esté a cero menor será la intensidad de la relación lineal de las variables, por lo tanto peor será el ajuste.

Vamos a dar una expresión analítica de la razón de correlación lineal:

$$\sigma^2_{Y^*} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\left(ax_i + b - \overline{y}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\left(ax_i + b - a\overline{x} - b\right)^2 =$$

$$= a^2 \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\left(x_i - \overline{x}\right)^2 = \left(\frac{\sigma_{X,Y}}{\sigma^2_X}\right)^2 \sigma^2_X = \frac{\sigma^2_{X,Y}}{\sigma^2_X}$$

$$r^2 = \frac{\sigma^2_{Y^*}}{\sigma^2_Y} = \frac{\frac{\sigma^2_{X,Y}}{\sigma^2_X}}{\sigma^2_Y} = \frac{\sigma^2_{X,Y}}{\sigma^2_X \sigma^2_Y}$$

Proposición 11 *r^2 es invariante por cambio de origen y de escala.*

La razón de correlación lineal también se puede expresar en función de los coeficientes de regresión:

$$\left.\begin{array}{l} y = ax + b \text{ con } a = \frac{\sigma^2_{X,Y}}{\sigma^2_X} \\ x = a'y + b' \text{ con } a' = \frac{\sigma^2_{X,Y}}{\sigma^2_Y} \end{array}\right\} \Rightarrow r^2 = aa'$$

Vamos a definir ahora el **coeficiente de correlación lineal**.

Este coeficiente va a ser una medida de correlación que además de medir el grado de asociación lineal entre las variables nos va a informar sobre el carácter de esta relación, es decir, si es positiva o negativa, entendiendo que si es positiva las variables van a crecer en el mismo sentido y si es negativa las variables

crecerán en sentidos opuestos. En la primera situación tendremos rectas de regresión crecientes y en la segunda rectas de regresión decrecientes.

$$r = \sqrt{r^2} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} = \sqrt{aa'}$$

Como **consecuencia** de su definición se verifican:
(1) El signo de este coeficiente es el signo de la covarianza.
(2) $-1 \leq r \leq 1$

Proposición 12 *El coeficiente de correlación lineal r es invariante por cambio de origen y de escala.*

Vamos a **interpretar** este coeficiente:
Si $r = 1$ se dice que existe correlación lineal perfecta positiva, es decir, la recta de regresión explica perfectamente el comportamiento de las variables (a medida que aumenta una aumenta la otra). Por tanto, la recta ajustada pasa por todos los puntos.

Proposición 13 *Si $r = 1$ las dos rectas de regresión coinciden.*

Si $r = -1$ se dice que existe correlación lineal perfecta negativa, es decir, la recta de regresión explica perfectamente el comportamiento de las variables (a medida que aumenta una disminuye la otra). Por tanto, la recta ajustada pasa por todos los puntos, y aunque la intensidad de la asociación es máxima las variables crecen en sentidos opuestos.

Proposición 14 *Si $r = -1$ las dos rectas de regresión coinciden.*

Si $r = 0$ se dice que la correlación lineal es nula, es decir, X e Y no están relacionadas mediante una recta y, por lo tanto, la recta ajustada no explica en absoluto el comportamiento de las variables.

Proposición 15 *Si $r = 0$ las rectas de regresión son perpendiculares entre sí y paralelas a los ejes de coordenadas.*

Proposición 16 *Si X e Y son independientes, entonces $r = 0$, ahora bien*

$$r = 0 \not\Rightarrow X \text{ e } Y \text{ independientes}$$

Si $0 < r < 1$ se dice que existe cierto grado de correlación lineal positiva, es decir, que las variables crecen en el mismo sentido y por lo tanto las rectas de regresión son crecientes y la intensidad de correlación será mayor cuando $r \to 1$ y más débil cuando $r \to 0$.

Si $-1 < r < 0$ se dice que existe cierto grado de correlación lineal negativa, es decir, las variables crecen en sentido opuesto, por lo tanto las rectas de

regresión son decrecientes y la intensidad de la dependencia lineal será mayor cuando $r \to -1$ y menor cuando $r \to 0$.