

* En una población de tamaño n se ha observado dos variables estadísticas X, Y , las cuales han presentado K y P modalidades distintas, respectivamente, con distribución de frecuencia conjunta $f(x_i, y_j)$, $i \in \{1, \dots, K\}$, $j \in \{1, \dots, P\}$

Llamamos a n_{ij} al número total de individuos de la población que presentan simultáneamente la modalidad x_i del carácter X y la y_j del carácter Y , denominado frecuencia absoluta del par (x_i, y_j) .

Llamamos frecuencia relativa del par (x_i, y_j) a $f_{ij} = \frac{n_{ij}}{n}$

La tabla estadística más habitual para describir a los n individuos de la población por los caracteres X e Y simultáneamente es la denominada tabla de doble entrada.

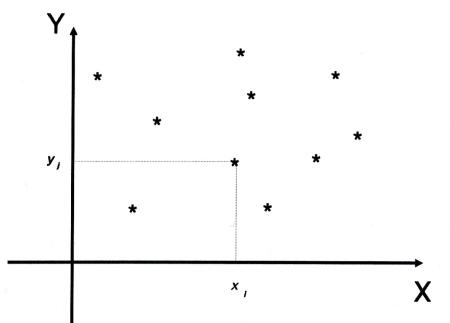
$X \backslash Y$	y_1	\dots	y_P	$n_{i \cdot}$
x_1	n_{11}	\dots	n_{1P}	$n_{1 \cdot}$
\vdots	\vdots	\ddots	\vdots	\vdots
x_K	n_{K1}	\dots	n_{KP}	$n_{K \cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	\dots	$n_{\cdot P}$	n

n₁₁ → Individuos que presentan la modalidad x_1 del carácter X
n_{.j} Individuos que presentan la modalidad y_j del carácter Y

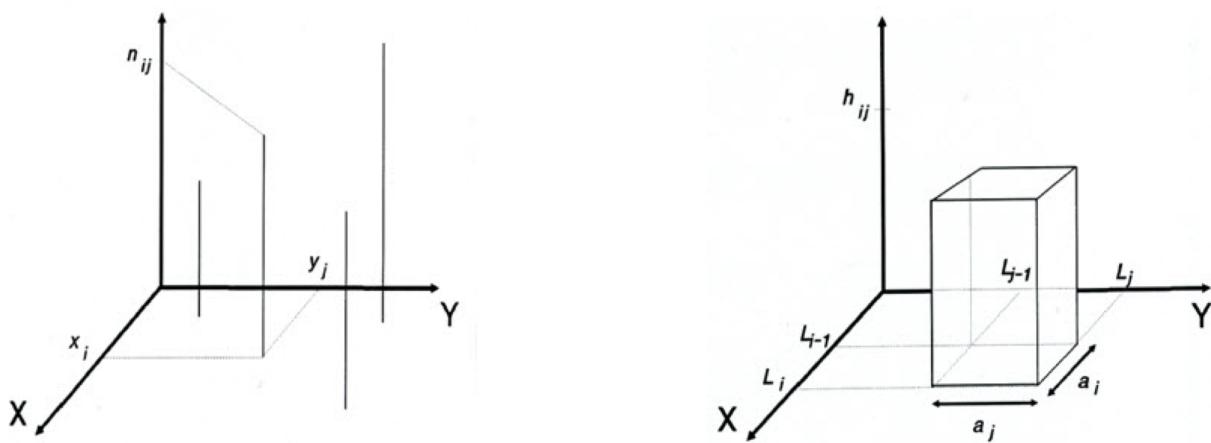
Si los caracteres son cualitativos, se llama tabla de contingencia.

Representaciones gráficas

Las dos representaciones gráficas más usadas en el caso bidimensional son la nube de puntos y el estereograma.



El diagrama de dispersión consiste en representar cada par de observaciones (x_i, y_j) por un punto o tanto. Los puntos como indica su frecuencia. Se suele usar para caracteres cuantitativos. Si el carácter es continuo se ve la marca de clase.



El estograma es un gráfico tridimensional formado por barras (si el carácter es cuantitativo discreto) o prismas (si el carácter es cuantitativo continuo)

En el primer caso la altura de cada barra será la frecuencia absoluta del par (x_i, y_j) , mientras que, en el segundo caso, la frecuencia absoluta del par se trata del volumen del prisma, teniendo en cuenta que la altura de h_{ij} es el cociente del n_{ij} del par entre el producto de las amplitudes de cada intervalo de clase.

Distribuciones marginales

Las modalidades del carácter X junto a las frecuencias $n_{i.}$ de la columna marginal de la tabla se denominan distribución marginal del carácter X , $\{x_i, n_{i.}\} \quad i=1, \dots, K$ (Análogo para el carácter Y)

Ambas son distribuciones unidimensionales, a las que es posible dar el tratamiento aparte del tema 1. Estamos representar estas distribuciones marginales en tablas de la siguiente manera:

X	$n_{i.}$	$f_{i.}$
x_1	$n_{1.}$	$f_{1.}$
\vdots	\vdots	\vdots
x_K	$n_{K.}$	$f_{K.}$
	n	1

Distribuciones condicionadas

Se llama distribución condicionada de un carácter al estudio de un carácter sobre los individuos que presentan una modalidad (o modalidades) del otro carácter.

Imaginemos el estudio del carácter X condicionado a los individuos que presentan la modalidad y_j del carácter Y . Dicho estudio corresponde al de un carácter unidimensional descrito por las frecuencias de la j -ésima columna de la tabla de doble entrada.

En el estudio de distribuciones condicionadas se representa la frecuencia relativa de la modalidad x_i del carácter X condicionado a los individuos que presentan la modalidad y_j de Y como:

$$f_{ij} = f_{ij}^d = \frac{n_{ij}}{n_{\cdot j}}$$

La tabla de una distribución condicionada será de la forma:

X	n_{ij}	f_{ij}^d
x_1	n_{1j}	f_{1j}^d
\vdots	\vdots	\vdots
x_k	n_{kj}	f_{kj}^d
	$n_{\cdot j}$	1

Al estudio de la distribución X condicionada a la modalidad y_j del carácter Y lo denotamos $X/Y = y_j$.

Vemos un ejemplo de distribución condicionada:

$X \backslash Y$	0	1	2	$n_{\cdot i}$
0	3	6	9	18
1	2	4	6	12
2	1	2	3	6
5	2	4	6	12
$n_{\cdot j}$	8	16	24	48

$X/Y = y_2$	$n_{\cdot i}$	$n_{\cdot 2}$	f_{ij}^d
0	6	16	$6/16$
1	4	16	$4/16$
2	2	16	$2/16$
5	4	16	$4/16$
	16	16	1

Dependencia e independencia estadística

Dos caracteres x e y se dirán estadísticamente dependientes cuando la variación en uno de ellos influya en la distribución del otro.

Por otro lado, se dice que el carácter X es independiente estadísticamente del carácter Y si las distribuciones de X condicionadas a cada valor y_j de Y son todas idénticas para cualquier valor de $j = 1, \dots, p$, es decir,

$$f_{i,j} = f_{i,\cdot}, \text{ no depende de } j.$$

Proposición

Si el carácter X no depende del carácter $Y \Rightarrow Y$ es independiente del carácter X .

Dependencia funcional

Se dice que el carácter X depende funcionalmente de Y si a cada modalidad y_j de Y corresponde una única modalidad posible de X con frecuencia no nula, es decir, para cualquier $j = 1, \dots, p$, la frecuencia absoluta n_{ij} es nula, excepto para un valor $i = \ell(j)$ donde $n_{ij} = n_{\cdot j}$.

Esto significa que en cada columna de la tabla de doble entrada, un término, y solo un término, es diferente de cero, aunque puede haber varios términos no nulos en una misma fila. Por ejemplo:

X	Y	y_1	y_2	y_3	y_4	y_5	
x_1		3	0	6	0	0	9
x_2		0	4	0	0	2	6
x_3		0	0	0	5	0	5
		3	4	6	5	5	20

Si hay un 0 no existe independencia funcional

Si carácter X depende funcionalmente del carácter Y , ya que a cada modalidad de Y le corresponde una sola modalidad de X con frecuencia no nula (en cada columna hay solo una frecuencia distinta de cero), sin embargo, Y no depende funcionalmente de X , ya que en cada fila no existe una única frecuencia no nula. En general, la dependencia lineal no es recíproca, solo lo será cuando la correspondencia es biunívoca (f. biyectiva).

Momentos bidimensionales

Dada una variable estadística bidimensional (x, y) , con distribución conjunta $\{(x_i, y_j); f_{ij}\}_{\substack{i=1, \dots, k \\ j=1, \dots, p}}$ se define el momento conjunto respecto al origen de orden r y s como

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

$$m_{00} = 1 \quad m_{10} = \bar{x} \quad m_{01} = \bar{y}$$

Dada una variable estadística bidimensional (x, y) , con distribución conjunta $\{(x_i, y_j); f_{ij}\}$, se define el momento conjunto central de órdenes r y s como:

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

$$M_{00} = 1 \quad M_{10} = 0 \quad M_{01} = 0 \quad \mu_{20} = \bar{x}^2$$

$$\mu_{11} = m_{11} - m_{10} m_{01} = m_{11} - \bar{x} \bar{y} \quad (\text{covarianza de } x \text{ con } y) \\ \mu_{11} = \bar{xy} = \text{Cov}(x, y)$$

Regresión

Se conoce como regresión a la determinación de la estructura de dependencia que mejor expresa el tipo de relación de una variable con las demás.

Por otro lado, está la correlación, cuyo objetivo es el estudio del grado de dependencia existente entre variables.

Existen pues, tres motivos fundamentales por los que una variable vamos a denominar dependiente o endógena está influida por otra que actúa como independiente, explicativa o exógena:

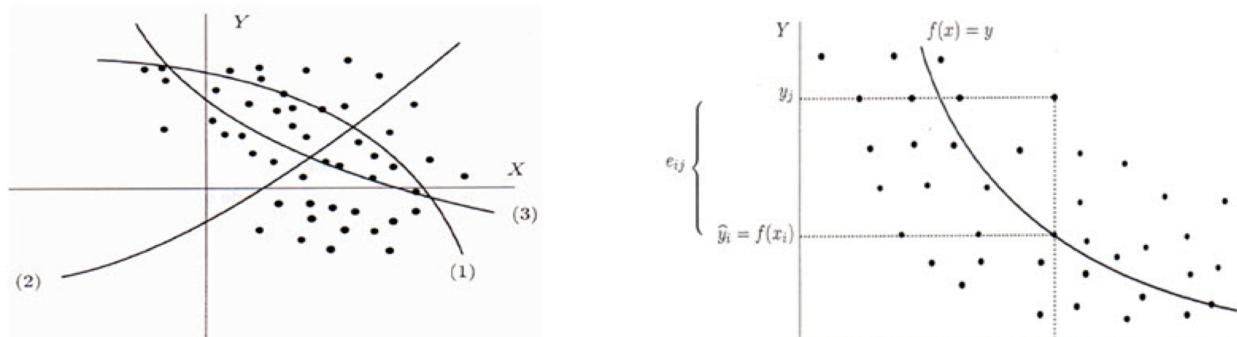
- La causalidad que ha hecho que ambas variables estén relacionadas estadísticamente.
- Una tercera variable está determinando a las que estamos estudiando.
- Existe una relación de causa-efecto.

Ajuste de funciones y datos por método de mínimos cuadrados

Si dos variables presentan una dependencia estadística, es decir, no funcional, no es posible encontrar una ecuación tal que

los valores que pueden presentar dichas variables lo satisfagan.

Ante la imposibilidad de encontrar una gráfica que pase por todos los puntos de la nube, nos centramos en buscar una función cuya gráfica más se aproxime a los datos observados.



Hacer regresión consiste, pues, en ajustar lo mejor posible una función a una serie de valores observados, encontrando una curva que, aunque no pase por todos los puntos de la nube, al menos esté lo más próxima posible a ellos. El siguiente razonamiento nos conducirá al conocido ajuste por mínimos cuadrados.

El método de mínimos cuadrados consiste en encontrar dicha función f tal que minimice la media de los cuadrados de los residuos.

$$\Psi(a_0, a_1, \dots, a_n) = \sum_{i=1}^k \sum_{j=1}^p f_{ij} (y_j - f(x_i, a_0, a_1, \dots, a_n))^2$$

Esta función, Ψ , se denomina error cuadrático medio de la función f .

El cálculo de los parámetros de la función de ajuste óptimo según el método de los mínimos cuadrados consiste en resolver el sistema siguiente, denominado sistema de ecuaciones normales:

$$\frac{\partial \Psi}{\partial a_r} = 0 \Rightarrow \sum_{i=1}^k \sum_{j=1}^p f_{ij} (y_j - f(x_i, a_0, \dots, a_n))$$

Dar

$$\frac{\partial f}{\partial a_r} = 0 \quad r = 0, 1, \dots, n$$

Dar