

ANÁLISIS DE DATOS MULTIDIMENSIONALES

INTRODUCCIÓN

DISTRIBUCIÓN DE FRECUENCIAS MULTIDIMENSIONAL

DISTRIBUCIONES MARGINALES

DISTRIBUCIONES CONDICIONADAS

INDEPENDENCIA ESTADÍSTICA

ESTUDIO ANALÍTICO DE DISTRIBUCIONES MULTIDIMENSIONALES

COVARIANZA

VECTOR DE MEDIAS

MATRIZ DE VARIANZAS-COVARIANZAS

COEFICIENTE DE CORRELACIÓN

MATRIZ DE CORRELACIÓN

EJEMPLO

En muchas ocasiones interesa estudiar el comportamiento de más de una característica (2 o más) en una población. Es evidente que siempre se podrá estudiar cada característica por separado, a través de su correspondiente distribución de frecuencias y analizar su comportamiento a través de los indicadores ya estudiados de posición, dispersión, forma y concentración. Pero puede resultar necesario analizar, también el comportamiento conjunto de 2 o más de las características observadas, con el fin de dilucidar la influencia de una en otra u otras, determinar las relaciones existentes entre ellas, etc. Para ello es imprescindible estudiar conjuntamente las observaciones de todas las características (variables o atributos), a través de la **distribución de frecuencias multidimensional**.

Dada una población de N individuos, de la que se disponen observaciones de varias características (supongamos cuantitativas, variables) éstas pueden, en principio representarse a través de un listado (matricial) similar a una **base de datos** en el que en cada fila aparecerá uno de los N individuos seguido de los valores que cada variable toma para cada individuo, lo que constituye un **registro**. (Cada variable es lo que en la terminología de las bases de datos se llama **campo**):

INDIVIDUO	VARIABLE 1 (X_1)	VARIABLE 2 (X_2)	VARIABLE 3 (X_3)
	ASIGN.MATRICUL.	EDAD	ASIGN. APROBADAS

1º JUAN	X_{11}	8	X_{21}	21	X_{31}	7
2º PEDRO	X_{12}	7	X_{22}	22	X_{32}	6
3º MARÍA	X_{13}	9	X_{23}	24	X_{33}	6
4º ANA	X_{14}	9	X_{24}	20	X_{34}	5
5º LUIS	X_{15}	9	X_{25}	19	X_{35}	5

Otra forma de representar los datos multidimensionales, especialmente útil en el caso bidimensional, es agrupando los datos por frecuencias.

En el caso bidimensional, consideraríamos una tabla de doble entrada para cada una de las variables, cada fila corresponde a un valor de la primera variable (x_{1i} o bien x_i), cada columna a un valor de la segunda variable (x_{2j} o bien y_j), y en cada celda aparecería la frecuencia de cada par de valores (n_{ij}).

Y(aprobadas)	Y_1	Y_2	Y_3		$n_{i.}$
X(matricul.)	5	6	7		
X_1	n_{11}	n_{12}	n_{13}		$n_{1.}$
7	0	1	0		1
X_2	n_{21}	n_{22}	n_{23}		$n_{2.}$
8	0	0	1		1
X_3	n_{31}	n_{32}	n_{33}		$n_{3.}$
9	2	1	0		3
$n_{.j}$	$n_{.1}$	$n_{.2}$	$n_{.3}$		N
	2	2	1		5

Una tabla de este tipo recibe el nombre de **tabla de correlación**. Si, en lugar de estar representadas las observaciones de dos variables (cuantitativas), se tratara de dos atributos, con distintos niveles, hablaríamos de **tabla de contingencia**.

Cada una de las frecuencias n_{ij} que nos informa del número de individuos que toman el valor x_i para la variable x , e y_j para la variable y , recibe el **nombre de frecuencia conjunta**.

Si sumamos las frecuencias conjuntas a lo largo de una fila (i) se obtiene el número total de observaciones del valor de x , x_i , con independencia del valor que tome la otra variable:

$$n_{i.} = \sum_j n_{ij} = \text{n}^\circ \text{ de observaciones de } x_i$$

Las $n_{i.}$ se conocen como **frecuencias marginales de la variable x** .

Análogamente, si sumamos las frecuencias conjuntas a lo largo de una columna (j) se obtiene el número total de observaciones del valor de y, y_j , con independencia del valor que tome la otra variable: $n_{.j} = \sum_i n_{ij} = n^\circ$ de observaciones de y_j

Las $n_{.j}$ se conocen como **frecuencias marginales de la variable y**.

Distribuciones marginales

Las distribuciones marginales son las distribuciones unidimensionales que nos informan del número de observaciones para cada valor de una de las variables, (prescindiendo de la información sobre los valores de las demás variables).

En el caso bidimensional hay dos (una para la x y otra para la y), en el caso multidimensional hay tantas como variables.

A partir de la tabla de correlación pueden construirse las distribuciones marginales, asignando a cada valor de la variable considerada su frecuencia marginal.

En el caso de dimensión mayor de dos, y supuestos los datos en forma de base de datos matricial, habrá que considerar únicamente una de las variables (una columna) y a partir del listado de observaciones, se podrá construir la tabla de frecuencias de la distribución marginal.

Las distribuciones marginales son distribuciones de frecuencias unidimensionales como las ya estudiadas y pueden analizarse de la manera habitual (media, varianza, asimetría, curtosis, etc.).

Distribuciones condicionadas

En el caso bidimensional, se pueden considerar además otras distribuciones que nos especifiquen las observaciones que hay de cada valor de una de las variables cuando imponemos la condición de que la otra toma un valor **determinado**. Esto supone considerar únicamente una columna de la tabla de correlación (distribución de x condicionada a un valor de y) o una fila de la tabla (distribución de y condicionada a un valor de x).

En el caso multidimensional, con una representación de base de datos, establecer una condición supone realizar una selección parcial de los datos, el resultado de esta selección sería la distribución condicionada, que en este caso puede ser uni o multidimensional, dependiendo de la condición (selección).

Independencia estadística

Dos variables estadísticas son **estadísticamente independientes** cuando el comportamiento estadístico de una de ellas no se ve afectado por los valores que toma la otra; esto es cuando las relativas de las distribuciones condicionadas no se ven afectadas por la condición, y coinciden en todos los casos con las frecuencias relativas marginales.

Esta definición puede hacerse más operativa, a través de la caracterización siguiente: dos variables son estadísticamente independientes cuando para todos los pares de valores se cumple que la frecuencia relativa conjunta es igual al producto de las frecuencias relativas marginales.:

$$\text{para todo } i,j : \frac{n_{i,j}}{N} = \frac{n_{i,\cdot}}{N} \cdot \frac{n_{\cdot,j}}{N}$$

Ejemplo:

Y				
X		1	2	3
				$n_{i,\cdot}$
5		1	10	5
				16
10		2	20	10
				32
15		4	40	20
				64
$n_{\cdot,j}$		7	70	35
				112

comprobemos si se cumple

$$\frac{n_{i,j}}{N} = \frac{n_{i,\cdot}}{N} \cdot \frac{n_{\cdot,j}}{N} \quad \forall i,j$$

para el primer par 1,1 tendríamos $\frac{1}{112} = \frac{16}{112} \cdot \frac{7}{112}$ que cumple

para el segundo par 1,2 tendríamos $\frac{10}{112} = \frac{16}{112} \cdot \frac{70}{112}$ que cumple

lo comprobaríamos hasta el último.....

para el último par 3,3, tendríamos $\frac{20}{112} = \frac{64}{112} \cdot \frac{35}{112}$ que cumple ,

por tanto X e Y son estadísticamente INDEPENDIENTES

Estudio analítico de distribuciones multidimensionales: Vector de Medias, matriz de Varianzas-Covarianzas

Aunque si la distribución multidimensional estudiada tiene una dimensión superior a 2 es posible definir indicadores (basados en los momentos) que consideren a la totalidad de las variables, en la práctica basta con analizar la totalidad de las variables por parejas

para poder contar con toda la información indispensable para manejarse adecuadamente con una distribución multidimensional.

De esta forma, dada una distribución de frecuencias multidimensional (de cualquier dimensión) nos interesará, por un lado conservar los indicadores univariantes de cada distribución marginal (medias, varianzas, etc.,-- de cada variable por separado) y considerar además algunos indicadores (bivariantes), de cada pareja de variables posible.

COVARIANZA

En este sentido el indicador bivalente más importante es la **covarianza**:

Dadas dos variables estadísticas x e y definiremos la covarianza S_{xy} como:

$$S_{x,y} = \frac{\sum_{j=1}^k \sum_{i=1}^k (X_i - \bar{x})(Y_i - \bar{y})n_{i,j}}{N}$$

en el caso de disponer de la distribución agregada por frecuencias en una tabla de correlación

$$S_{x,y} = \frac{\sum_{j=1}^k \sum_{i=1}^k (X_i - \bar{x})(Y_i - \bar{y})}{N}$$

en el caso de disponer de la distribución sin agregar por frecuencias (en un listado matricial de datos donde cada registro es una observación y n° de registros= N)

Propiedades:

1. La covarianza es el momento central de orden 1,1 de la distribución bidimensional.
2. Es invariante ante los cambios de origen en cualquiera de las dos variables.
3. Sin embargo depende de los cambios de unidad .Si se cambia de unidad de medida en ambas variables la covarianza se modifica proporcionalmente a ambos cambios:

$$u = a + bx \quad v = c + dy \quad S_{uv} = b \cdot d \cdot S_{xy}$$

4. La expresión de cálculo de la covarianza es $S_{xy} = \alpha_{11} - \bar{x}\bar{y}$

donde α_{11} es el llamado **momento (ordinario) mixto** y su expresión es:

$$\alpha_{11} = \frac{1}{N} \sum_{\bar{x}} \sum_{\bar{y}} x_i y_j n_{i,j}$$

si las observaciones están agregadas por frecuencias , o bien:

$$a_{11} = \frac{1}{N} \sum_{\bar{x}_i} \sum_{\bar{y}_j} x_i y_j$$

si las observaciones no están agregadas por frecuencias

5. Si dos variables son independientes su covarianza es cero (el resultado recíproco no es necesariamente cierto).

6. La covarianza nos mide la **covariación conjunta** de dos variables: Si es positiva nos dará la información de que a valores altos de una de las variable hay una mayor **tendencia** a encontrar valores altos de la otra variable y a valores bajos de una de las variable ,correspondientemente valores bajos.En cambio si la covarianza es negativa, la covariación de ambas variables será en sentido inverso: a valores altos le corresponderán bajos, y a valores bajos, altos.Si la covarianza es cero no hay una covariación clara en ninguno de los dos sentidos.Sin embargo el hecho de que la covarianza dependa de las medidas de las variables no permite establecer comparaciones entre unos casos y otros.

VECTOR DE MEDIAS:

Dada una variable estadística n-dimensional $(X_1, X_2, X_3, \dots, X_n)$, llamaremos vector de medias al vector columna formado por las medias de las distribuciones marginales de cada variable por separado.

$$M = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \dots \\ \bar{x}_n \end{pmatrix}$$

MATRIZ DE VARIANZAS-COVARIANZAS:

Dada una variable estadística n-dimensional $(X_1, X_2, X_3, \dots, X_n)$, llamaremos matriz de varianzas-covarianzas (matriz de varianzas) (matriz de covarianzas), a la matriz cuadrada, $n \times n$, que disponga en su diagonal principal de las varianzas de cada una de las distribuciones marginales unidimensionales, y en los elementos no-diagonales (i,j) de las correspondientes covarianzas entre cada dos variables S_{ij}

$$V = \begin{pmatrix} S_1^2 & S_{12} & \dots & S_{1n} \\ S_{21} & S_2^2 & \dots & \dots \\ \dots & \dots & \dots & \dots \\ S_{n1} & S_{n2} & \dots & S_n^2 \end{pmatrix}$$

Propiedades

1. La matriz de varianzas-covarianzas es **simétrica** respecto a su diagonal principal
2. La matriz de varianzas-covarianzas es definida positiva
3. El determinante de la matriz de varianzas-covarianzas (también llamado determinante de momentos) es siempre no negativo L mayor o igual a 0
4. En el caso bidimensional tendremos:

$$\det V = L = S_x^2 S_y^2 - (S_{xy})^2$$

COEFICIENTE DE CORRELACIÓN

Para poder contar con un indicador que nos permita, por un lado establecer la covariación conjunta de dos variables, y por otro, que tenga la universalidad suficiente para poder establecer comparaciones entre distintos casos, se utiliza el coeficiente de correlación (lineal, de Pearson). La correlación es, pues una medida de covariación conjunta que nos informa del sentido de esta y de su relevancia, que está acotada y permite la comparación entre distintos casos.

El coeficiente de correlación entre dos variables puede definirse como la covarianza existente entre sus dos variables tipificadas y tiene por expresión de cálculo:

$$r_{xy} = S_{xy} = \frac{S_{\left(\frac{x-\bar{x}}{S_x} \frac{y-\bar{y}}{S_y}\right)}}{S_x S_y} = \frac{S_{xy}}{S_x S_y}$$

Interpretación:

****Si $r < 0$ Hay correlación negativa** : las dos variables se correlacionan en sentido inverso. A valores altos de una de ellas le suelen corresponder valores bajos de la otra y viceversa. Cuánto más próximo a -1 esté el coeficiente de correlación más patente será esta covariación extrema. Si $r = -1$ hablaremos de **correlación negativa perfecta** lo que supone una determinación absoluta entre las dos variables (en sentido inverso): Existe una relación funcional perfecta entre ambas (una relación lineal de pendiente negativa).

**** Si $r > 0$ Hay correlación positiva**: las dos variables se correlacionan en sentido directo. A valores altos de una le corresponden valores altos de la otra e igualmente con los valores bajos. Cuánto más próximo a +1 esté el coeficiente de correlación más patente será esta covariación. Si $r = 1$ hablaremos de **correlación positiva perfecta** lo que supone una determinación absoluta entre las dos variables (en sentido directo): Existe una relación lineal perfecta (con pendiente positiva).

**** Si $r = 0$ se dice que las variables están incorrelacionadas**: no puede establecerse ningún sentido de covariación.

Propiedad importante: **Si dos variables son independientes estarán incorrelacionadas aunque el resultado recíproco no es necesariamente cierto.**

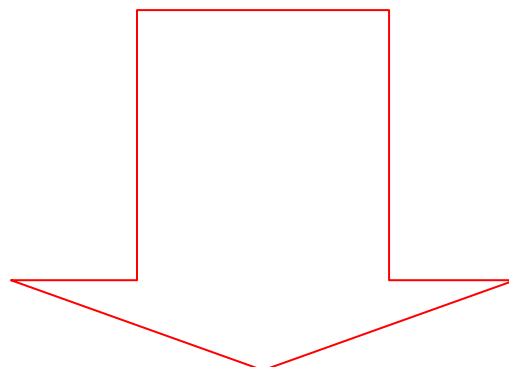
MATRIZ DE CORRELACIÓN

En el caso de estar analizando una distribución n-dimensional con $n > 2$, podemos construir la llamada matriz de correlación:

La matriz de correlación \mathbf{R} es una matriz cuadrada $n \times n$ constituida por los coeficientes de correlación de cada pareja de variables; de manera que tendrá unos en su diagonal principal, y en los elementos no diagonales (i,j) los correspondientes coeficientes de correlación r_{ij} . La matriz de correlación será, obviamente, simétrica, y conservará las propiedades de ser definida-positiva y tener un determinante no negativo, (además el determinante será siempre menor o igual que 1). Puede considerarse como la matriz de varianzas entre las variables tipificadas.

$$R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1n} \\ r_{21} & 1 & \dots & r_{2n} \\ \dots & \dots & \dots & \dots \\ r_{n1} & r_{n2} & \dots & 1 \end{pmatrix}$$

Ejemplo tabla contingencia desarrollada



Y_j	Y_1	Y_2	\dots	Y_j	\dots	Y_m	$n_{i.}$	$x_i n_{i.}$	x_i^2	$x_i^2 n_{i.}$	$\Sigma y_j n_{ij}$	$x_i \Sigma y_j n_{ij}$
X_i												
x_1	n_{11}	n_{12}		n_{1j}		n_{1m}	$n_{1.}$	$x_1 n_{1.}$	x_1^2	$x_1^2 n_{1.}$	$\Sigma y_j n_{1j}$	$x_1 \Sigma y_j n_{1j}$
x_2	n_{21}	n_{22}		n_{2j}		n_{2m}	$n_{2.}$	$x_2 n_{2.}$	x_2^2	$x_2^2 n_{2.}$	$\Sigma y_j n_{2j}$	$x_2 \Sigma y_j n_{2j}$
\vdots												
\vdots												
x_i	n_{i1}	n_{i2}		n_{ij}		n_{im}	$n_{i.}$	$x_i n_{i.}$	x_i^2	$x_i^2 n_{i.}$	$\Sigma y_j n_{ij}$	$x_i \Sigma y_j n_{ij}$
\vdots												
\vdots												
x_n	n_{n1}	n_{n2}		n_{nj}		n_{nm}	$n_{n.}$	$x_n n_{n.}$	x_n^2	$x_n^2 n_{n.}$	$\Sigma y_j n_{nj}$	$x_n \Sigma y_j n_{nj}$
n_j	$n_{.1}$	$n_{.2}$		$n_{.j}$		$n_{.m}$	N	$\Sigma x_i n_{i.}$		$\Sigma x_i^2 n_{i.}$	$\Sigma \Sigma y_j n_{ij}$	$\Sigma \Sigma x_i y_j n_{ij}$
$y_j n_{.j}$	$y_1 n_{.1}$	$y_2 n_{.2}$		$y_j n_{.j}$		$y_m n_{.m}$	$\Sigma y_j n_{.j}$					
y_j^2	y_1^2	y_2^2		y_j^2		y_m^2						
$y_j^2 n_{.j}$	$y_1^2 n_{.1}$	$y_2^2 n_{.1}$		$y_j^2 n_{.1}$		$y_m^2 n_{.1}$	$\Sigma y_j^2 n_{.j}$					
$\Sigma x_i n_{ij}$	$\Sigma x_i n_{i1}$	$\Sigma x_i n_{i2}$		$\Sigma x_i n_{ij}$		$\Sigma x_i n_{im}$	$\Sigma \Sigma x_i n_{ij}$					
$y_j \Sigma x_i n_{ij}$	$y_1 \Sigma x_i n_{i1}$	$y_2 \Sigma x_i n_{i2}$		$y_j \Sigma x_i n_{ij}$		$y_m \Sigma x_i n_{im}$	$\Sigma \Sigma x_i y_j n_{ij}$					

Ejemplo numérico

Y_j	20	22	25	27	35	n_i	$x_i n_i$	x_i^2	$x_i^2 n_i$	$\Sigma y_j n_{ij}$	$x_i \Sigma y_j n_{ij}$
X_i											
5	6	5	7	10	6	34	170	25	850	885	4425
7	2	3	2	5	2	14	98	49	686	361	2527
10	1	1	3	2	2	9	90	100	900	241	2410
11	2	2	1	1	4	10	110	121	1210	276	3036
12	10	12	10	3	2	37	444	144	5328	865	10380
15	2	3	1	2	5	13	195	225	2925	360	5400
n_j	23	26	24	23	21	117	1107		11899	2988	28178
$y_j n_j$	460	572	600	621	735	2988					
y_j^2	400	484	625	729	1225						
$y_j^2 n_j$	9200	12584	15000	16767	25725	79276					
$\Sigma x_i n_{ij}$	226	267	225	182	207	1107					
$y_j \Sigma x_i n_{ij}$	4520	5874	5625	4914	7245	28178					

	media	a_2	varianza	a_{11}	covarianza	correlacion
X	9,461538	101,7009	12,1801446			
Y	25,53846	677,5726	25,3596318			
XY				240,8376	-0,7955293	-0,0452646