ANÁLISIS DE DATOS MULTIDIMENSIONALES

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) ASIGN.MATRICUL.		VARIABLE 2(X_2) EDAD		VARIABLE 3 (X_3) 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) X (matriculadas.)	$\mathbf{Y_1}$ 5	$\mathbf{Y_2}$ 6	$\mathbf{Y_3}$ 7		$\mathbf{n_{i.}}$
$\mathbf{X_1}$ 7	n_{11} 0	n_{12} 1	n_{13} 0		$n_{1.}$ 1
$\mathbf{X_2}$ 8	n_{21} 0	n_{22} 0	n_{23} 1		$n_{2.}$ 1
$\mathbf{X_3}$ 9	n_{31} 2	n_{32} 1	n_{33} 0		$n_{3.}$ 3
n.j	n.1 2	n.2 2	n.3 1		**N** 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 trata de dos atributos, con distintos niveles, hablaríamos de **tabla de contingencia**.

Cada una de las frecuencias nij 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:

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

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: $n_{.j} = \sum_i n_{ij} = n^o$ 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 variable s,(prescindiendo de la información sobre los valores de las demás variables).

En el caso bidim ensional hay dos (una pa ra 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 m arginales, asignando a cada valor de la variable considerada su frecuencia marginal.

En el caso de dim ensión mayor de dos, y s upuestos los datos en form a de base 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 bidim ensional ,se pueden consid erar adem ás otras distribuciones que nos especifiquen las observaciones que hay de cad a valor de una de las variables cuando imponemos la condición de que la otra toma un valor **determinado**. Esto supone considerar únicam ente 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 multidim ensional, con u na repre sentación de ba se de datos, establecer una condición supone realizar una selección parcial de los dato s, el resultado d e esta selección s ería la dis tribución condicionada, que en este caso puede ser uni o multidimensional, dependiendo de la condición (selección).

Independencia estadística

Dos varia bles es tadísticas 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 hacers e m ás operativa, a través de la caracterización sigu iente: Dos variables son estadísticam ente indepe ndientes cuando para todos los pares de valores se cum ple que la frecuencia relati va conjunta es igual al producto de las frecuencias relativas marginales.:

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

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

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, tendríam os $\frac{20}{112} = \frac{64}{112} \cdot \frac{35}{112}$ 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 dis tribución multidimensional estudiada tiene una dim ensión superior a 2 es posible definir indicadores (basados en los momentos) que consideren a la totalid ad de las variables, en la práctica basta con anal izar la totalidad de las variables por parejas para poder contar con toda la información indispensable para manejarse adecuadamente con una distribución multidimensional.

De esta form a, dada una distribución de frecuencias m ultidimensional (de cualq uier dimensión) nos interesará, por un lado cons ervar los ind icadores univ ariantes de cada distribución m arginal (m edias, varianzas, etc.,-- de cada variable por separado) y considerar adem ás algunos indicadores (biv ariantes), de cada pareja de variables posible.

COVARIANZA

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

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

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

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

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

en el caso de disponer de la distribución sin agregar por frecuen cias (en un listado matricial de datos donde cada registro es una observación y nº 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:

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

4. La expresión de cálculo de la covarianza es $s_{xy} = a_{1,1} - \bar{x}\cdot\bar{y}$

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

$a_{1,1} = \sum_{j=1}^{k} \sum_{i=1}^{h} X_i Y_i n_{i,j}$ si las observaciones están agregadas por frecuencias , o bien:

$a_{1,1} = \sum_{j=1}^{k} \sum_{i=1}^{h} X_i Y_i$ 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 $(X_1,X_2,X_3,...,X_n)$, llamaremos vector de medias al vector columna formado por las medias de las distribuciones marginales de cada variable por separado.

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

MATRIZ DE VARIANZAS-COVARIANZAS:

Dada una variable estadí stica n-dim ensional $(X_1,X_2,X_3,...,X_n)$, llam aremos m atriz de varianzas-covarianzas (m atriz de varian zas) (m atriz de covarianzas), a la m atriz cuadrada, n´ n, que disponga en su diagonal principal de las varianzas de cada una de las distribuciones m arginales unidim ensionales, y en los elem entos no-diagonales (i,j) de las correspondientes covarianzas entre cada dos variables S_{ij}

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

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 = $S^2_x S^2_y - (S_{xy})^2$

COEFICIENTE DE CORRELACIÓN

Para poder contar con un indicador que nos perm ita, por un lado establecer la covariación conjunta de dos vari ables , y por otro, que tenga la universalidad suficiente para poder estab lecer com paraciones entre dist intos caso s, se utiliza el coeficien te de correlación (lin eal, de Pearson).La correlac ión es, pues una m edida de covariación conjunta que nos inform a del sentido de esta y de su relevancia, que está acotada y permite la comparación entre distintos casos.

El coeficien te de co rrelación en tre dos vari ables puede definirse como la covarian za existente entre sus dos variables tipificadas y tiene por expresión de cálculo:

$$r_{x,y} = S_{u,v} = S_{\frac{X-\overline{X}}{S_x},\frac{Y-\overline{Y}}{S_y}} = \frac{S_x \cdot S_y}{S_{x,y}}$$

Interpretación:

****Si r < 0** Hay **correlación negativa** : las dos variables s e correlac ionan en sentido inverso.A valores altos de una de ellas le su elen corresponder valor bajos de la otra y viceversa.Cuánto más próxim o a -1 esté el coeficiente de correlación m ás patente será esta covariación ex trema.Si r = -1 h ablaremos de **correlación negativa perfecta** lo que supone una determ inación absoluta entre las do s 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 variab les se correlacion an en sentid o directo.A valores altos de una le corresponden valores altos de la otra e igualm ente con los valores bajos. Cuánto m ás próxim o a +1 esté el coef iciente de correlación más patente será esta covariación. Si r = 1 hablaremos de **correlación positiva perfecta** lo que supone una determ inación 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 pued e estab lecerse ningún sentido de covariación.

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

MATRIZ DE CORRELACIÓN

En el caso de estar ana lizando una distribución n-di mensional con n > 2, podemos construir la llamada matriz de correlación:

La matriz de correlación **R** es una matriz cuadrada n ´ n cosntituida por los coeficientes de correlación de cada pareja de variables; de m anera que tendrá unos en su diagonal principal, y en los elem entos no diagonales (i,j) los correspondiente s coeficientes de correlación r_{ij} La m atriz de correlación será, obviam ente, sim étrica, y conservará las propiedades de ser definida- positiva y tener un determinante no negativo, (adem ás el determinante será siem pre menor o igual que 1). Puede considerarse co mo la m atriz de varianzas entre las variables tipificadas.

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