2. DESCRIPCIÓN CONJUNTA DE VARIAS VARIABLES

Objetivo

Mostrar la relación entre dos (o más) variables discretas o categóricas.

Bibliografia recomendada

Peña y Romo (1997), Capítulo 7.

Índice

1. La tabla de doble entrada
2. La distribución conjunta de frecuencias
3. Distribuciones marginales y condicionadas
4. Ideas de independencia
5. Media y varianza condicionada.

Introducción

Puede que, exista una relación entre dos variables. Midiendo los valores de ambas variables simultaneamente, podemos intentar medir la relación.

Ejemplo 55

1. *Altura y peso. Esperamos que, por lo general, la gente más alta sea más pesada.*

2. *Número de partidos ganados y posición en la clasificación.*

3. *Número de votos y número de escaños*

4. *Tipo de vivienda y sueldo.*

5. *Provincia de nacimiento y primera idioma.*

Ejemplo 56 *La Real Academía de la Lengua Española quiere estudiar las idiomas habladas en España. Por esta razón, se pregunta a una muestra de 40 españoles su provincia de nacimiento* {**C***ataluña,* **G***alicia,* **O***tra,***P***ais Vasco* } *y su lengua materna* { **Cas***tellano,* **Cat***alan,***Eus***kera,* **Gal***lego,* **Otr***a*} *con los siguientes resultados:*

(P,Eus)	(O,Cas)	(O,Cas)	(O,Cas)	(C,Cat)	(C,Cas)
(G,Cas)	(O,Cas)	(C,Cat)	(P,Cas)	(G,Cas)	(O,Cas)
(O,Cas)	(P,Cas)	(C,Cat)	(O,Cas)	(G,Gal)	(P, Otr)
(O,Cas)	(O,Cas)	(O,Cas)	(C,Cat)	(P,Cas)	(G,Cas)
(O,Cas)	(O,Gal)	(O,Otr)	(O,Otr)	(O,Cas)	(C,Cat)
(O,Cas)	(G,Cas)	(G,Gal)	(C,Cas)	(P,Cas)	(O,Cas)
(P,Cas)	(G,Cas)	(O,Otr)	(C,Cat)		

Resumimos estos datos en una tabla de doble entrada.

La tabla de doble entrada

Construimos una tabla mostrando las frecuencias de cada combinación.

		Provincia				
		C	G	O	P	
	Cas	2	5	14	5	
	Cat	6	0	0	0	
Lengua	*Eus*	0	0	0	1	
	Gal	0	2	1	0	
	Otr	0	0	3	1	
						40

Es decir que tenemos 14 personas en la muestra que provienen de otras provincias y hablan Castellano.

Podemos convertir la tabla en una tabla de frecuencias relativas, dividiendo cada frecuencia por 40.

Tabla de frecuencias relativas

		Provincia				
		C	G	O	P	
	Cas	,050	,125	,350	,125	
	Cat	,150	,000	,000	,000	
Lengua	*Eus*	,000	,000	,000	,025	
	Gal	,000	,050	,025	,000	
	Otr	,000	,000	,075	,025	
						1

Vemos que un 12,5 % *de la gente en la muestra son del Pais Vasco y hablan Castellano.*

Supongamos que sólo nos interesan la lengua. Podemos calcular las distribución marginal.

La distribución marginal

		Provincia				
		C	G	O	P	
	Cas	,050	,125	,350	,125	,650
	Cat	,150	,000	,000	,000	,150
Lengua	*Eus*	,000	,000	,000	,025	,025
	Gal	,000	,050	,025	,000	,075
	Otr	,000	,000	,075	,025	,100
						1,000

Es decir, un 65 % *de la gente en la muestra hablan Castellano y un* 15 % *hablan Catalan.*

Igualmente, podemos añadir la distribución marginal de la región de donde proviene la gente.

		Provincia				
		C	G	O	P	
	Cas	,050	,125	,350	,125	,650
	Cat	,150	,000	,000	,000	,150
Lengua	*Eus*	,000	,000	,000	,025	,025
	Gal	,000	,050	,025	,000	,075
	Otr	,000	,000	,075	,025	,100
		,200	,175	,450	,175	1,000

Es decir que un 20 % de la gente en la muestra son Catalanes, etc.

Puede que también nos interesa la distribución de las lenguas habladas por la gente de otras provincias.

La distribución condicionada

Queremos la frecuencia de Castellanoparlantes en las otras provincias. Miramos la columna de frecuencias absolutas.

	O
Cas	14
Cat	0
Eus	0
Gal	1
Otr	3
	18

Lo podemos convertir en frecuencias condicionadas.

	O
Cas	$\frac{14}{18}$
Cat	0
Eus	0
Gal	$\frac{1}{18}$
Otr	$\frac{3}{18}$
	1

Es decir que $\frac{14}{18}$ *de la gente de otras provincias en la muestra hablan Castellano.*

Cálculo a través de la tabla de frecuencias relativas

Supongamos que queremos la distribución de las provincias de donde provienen la gente que habla Castellano.

$$
\begin{array}{cc}
 & \begin{array}{cccc} C & G & O & P \end{array} \\
\textit{Cas} & \boxed{\begin{array}{cccc} ,050 & ,125 & ,350 & ,125 \end{array}} \; ,650
\end{array}
$$

Dividimos todas las entradas por la frecuencia total (,650).

$$
\begin{array}{cc}
 & \begin{array}{cccc} C & G & O & P \end{array} \\
\textit{Cas} & \boxed{\begin{array}{cccc} ,077 & ,192 & ,538 & ,192 \end{array}} \; 1
\end{array}
$$

Casi un 54 % de la gente que hablan Castellano provienen de las otras provincias.

Fórmula General

Supongamos que medimos los valores de 2 variables X e Y con I y J distíntas categorías o valores ($x_1, \ldots, x_I$ e $y_1, \ldots, y_J$) en una muestra de n personas.

La tabla de doble entrada de frecuencias absolutas es la siguiente

$$
\begin{array}{cc|cccc|c}
 & & & & \text{Y} & & \\
 & & y_1 & y_2 & \cdots & y_J & \\
\hline
 & x_1 & n_{11} & n_{12} & \cdots & n_{1J} & n_{1\cdot} \\
 & x_2 & n_{21} & n_{22} & \cdots & n_{2J} & n_{2\cdot} \\
X & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
 & x_I & n_{I1} & n_{I2} & \cdots & n_{IJ} & n_{I\cdot} \\
\hline
 & & n_{\cdot 1} & n_{\cdot 2} & \cdots & n_{\cdot J} & n
\end{array}
$$

where n_{ij} es el número de veces que ocurre la pareja (x_i, y_j) y $n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$. Las frecuencias marginales son $n_{i\cdot} = \sum_{j=1}^{J} n_{ij}$ y $n_{\cdot j} = \sum_{i=1}^{I} n_{ij}$

La tabla de frecuencias relativas es

$$
\begin{array}{cc|cccc|c}
 & & & & \text{Y} & & \\
 & & y_1 & y_2 & \cdots & y_J & \\
\hline
 & x_1 & f_{11} = \frac{n_{11}}{n} & f_{12} = \frac{n_{12}}{n} & \cdots & f_{1J} = \frac{n_{1J}}{n} & f_{1\cdot} = \frac{n_{1.}}{n} \\
 & x_2 & f_{21} = \frac{n_{21}}{n} & f_{22} = \frac{n_{22}}{n} & \cdots & f_{2J} = \frac{n_{2J}}{n} & f_{2\cdot} = \frac{f_{2\cdot}}{n} \\
X & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
 & x_I & f_{I1} = \frac{n_{I1}}{n} & f_{I2} = \frac{n_{I2}}{n} & \cdots & f_{IJ} = \frac{f_{IJ}}{n} & f_{I\cdot} = \frac{n_{I.}}{n} \\
\hline
 & & f_{\cdot 1} = \frac{n_{\cdot 1}}{n} & f_{\cdot 2} = \frac{n_{\cdot 2}}{n} & \cdots & f_{\cdot J} = \frac{n_{\cdot J}}{n} & 1
\end{array}
$$

La distribución condicionada de Y dado $X = x_i$ es

$$f(Y = y_1|X = x_i) = \frac{f_{i1}}{f_{i\cdot}}, \quad f(Y = y_2|X = x_i) = \frac{f_{i2}}{f_{i\cdot}}, \quad \ldots, \quad f(Y = y_J|X = x_i) = \frac{f_{iJ}}{f_{i\cdot}}$$

La distribución condicionada de X dado $Y = y_j$ es

$$f(X = x_1|Y = y_j) = \frac{f_{1j}}{f_{\cdot j}}, \quad f(X = x_2|Y = y_j) = \frac{f_{2j}}{f_{\cdot j}}, \quad \ldots, \quad f(X = x_i|Y = y_j) = \frac{f_{Ij}}{f_{\cdot j}}$$