ANÁLISIS DE DATOS CUALITATIVOS

José Vicéns Otero
Eva Medina Moral

Enero 2005

1. CONSTRUCCIÓN DE UNA TABLA DE CONTINGENCIA

Para analizar la relación de dependencia o independencia entre dos variables cualitativas nominales o factores, es necesario estudiar su distribución conjunta o tabla de contingencia.

La tabla de contingencia es una tabla de doble entrada, donde en cada casilla figurará el número de casos o individuos que poseen un nivel de uno de los factores o características analizadas y otro nivel del otro factor analizado.

		SEXO		
		HOMBRE	**MUJER**	**MARGINAL**
FUMA	**SI**	n_{11}	n_{12}	$n_{1.}$
	NO	n_{21}	n_{22}	$n_{2.}$
	MARGINAL	$n_{.1}$	$n_{.2}$	N

donde

n_{ij} = número de observaciones que tienen el atributo i y j
$n_{i.}$ = número de individuos que tienen el atributo i (marginal i)
$n_{.j}$ = número de individuos que tienen el atributo j (marginal j)

La tabla de contingencia se define por el número de atributos o variables que se analizan conjuntamente y el número de modalidades o niveles de los mismos. El ejemplo propuesto es una tabla de contingencia 2x2, ya que tiene dos atributos (FUMA Y SEXO) y cada uno de ellos tiene dos niveles. Si quisiéramos analizar conjuntamente tres variables nominales, como por ejemplo, Fumar, Sexo y Edad, y esta última variable tuviera tres niveles (<20 años, de 20 a 40 años, >40 años), obtendríamos tres tablas como la anterior, una para cada modalidad de edad y la tabla de contingencia tendría una dimensión $3\times2\times2$.

Las tablas de contingencia tienen dos objetivos fundamentales:

1) Organizar la información contenida en un experimento cuando ésta es de carácter bidimensional, es decir, cuando está referida a dos factores (variables cualitativas).

		SEXO		
		HOMBRE	**MUJER**	**MARGINAL**
FUMA	**SI**	65	58	123
	NO	43	67	110
	MARGINAL	108	125	233

En esta tabla se puede observar en primer lugar que de los 233 individuos de los que se tiene información 108 son hombres y 125 son mujeres. Asimismo se sabe que 123 de ellos fuman y 110 no. La tabla de contingencia nos permite tener información cruzada sobre ambas variables: de los 108 hombres, 65 fuman y 43 no, mientras que en el caso de las mujeres, 58 fuman y 67 no.

2) A partir de la tabla de contingencia se puede además analizar si existe alguna relación de dependencia o independencia entre los niveles de las variables cualitativas objeto de estudio. El hecho de que dos variables sean independiente significa que los valores de una de ellas no están influidos por la modalidad o nivel que adopte la otra.

2. CONTRASTACIÓN ESTADÍSTICA DE LA RELACIÓN DE DEPENDENCIA PARA VARIABLES CUALITATIVAS

Para identificar relaciones de dependencia entre variables cualitativas se utiliza un contraste estadístico basado en el estadístico χ^2 (Chi-cuadrado), cuyo cálculo nos permitirá afirmar con un nivel de confianza estadístico determinado si los niveles de una variable cualitativa influyen en los niveles de la otra variable nominal analizada. Siguiendo con el ejemplo propuesto, el cálculo de la Chi-cuadrado nos permitiría saber si el sexo de una persona es un factor determinante en que dicha persona fume o no fume.

¿Cómo podemos determinar si existe una relación de dependencia o independencia entre las variables analizadas?

Dos variables son independientes si:

a) las frecuencias relativas condicionadas son iguales a las frecuencias relativas marginales, es decir:

$$f(A_1/B_1)=\frac{n_{11}}{n_{1\bullet}}=f(A_1/B_2)=\frac{n_{12}}{n_{1\bullet}}=....=f(A_1/B_j)=\frac{n_{1j}}{n_{1\bullet}}=\frac{n_{1\bullet}}{N}$$

$$f(A_2/B_1)=\frac{n_{21}}{n_{2\bullet}}=f(A_2/B_2)=\frac{n_{22}}{n_{2\bullet}}=....=f(A_2/B_j)=\frac{n_{2j}}{n_{2\bullet}}=\frac{n_{2\bullet}}{N}$$

$$f(A_i/B_j)=\frac{n_{ij}}{n_{i\bullet}}=f_{ij}=\frac{n_{i\bullet}}{N}$$

Frecuencias relativas marginales:

$$f(B_j/A_i)=\frac{n_{ij}}{n_{\bullet j}}=f_{ji}=\frac{n_{\bullet j}}{N}$$

b) O bien si se cumple que la frecuencia relativa conjunta es igual al producto de las frecuencias relativas marginales:

$$f(A_i\cap B_j)=\frac{n_{ij}}{N}=\frac{n_{i\bullet}}{N}x\frac{n_{\bullet j}}{N}$$

De esta forma, comparando las frecuencias teóricas esperadas en caso de independencia entre los factores con las frecuencias observadas en la muestra, podremos concluir si existe una relación de dependencia o independencia entre los factores o atributos analizados.

Según la notación de la tabla inicial, y utilizando el concepto frecuentalista de probabilidad, podemos estimar la probabilidad de que se de un suceso determinado a partir de sus frecuencias relativas:

$$P_{ij}=\frac{n_{ij}}{N};\qquad P_{i\bullet}=\frac{n_{i\bullet}}{N};\qquad P_{\bullet j}=\frac{n_{\bullet j}}{N}$$

De esta forma, si las variables son independientes

$$\hat{P}_{ij}=\frac{E_{ij}}{N}=\frac{n_{i\bullet}}{N}\times\frac{n_{\bullet j}}{N}$$

donde E_{ij} sería el número de casos o frecuencia absoluta esperada o teórica en condiciones de independencia. Por lo tanto podremos calcular las frecuencias esperadas:

$$E_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{N}$$

En lugar de los E_{ij}, habremos observado los n_{ij}. Tendremos tantos valores E_{ij} y n_{ij} como celdas de la matriz, concluyendo que si hay poca diferencia entre estos valores los atributos serán independientes, no pudiéndose afirmar lo mismo en caso contrario. Supuesto que el atributo A tiene n filas y el atributo B, k columnas, la tabla será de orden n×k. Pearson planteó la utilización del estadístico χ^2 para analizar la independencia, definido por:

$$\hat{c}^2 = \frac{\sum_{i=1}^{h}\sum_{j=1}^{k}(n_{ij} - E_{ij})^2}{E_{ij}}$$

La hipótesis nula a contrastar será la de independencia entre los factores, siendo la hipótesis alternativa la de dependencia entre los factores.

El valor de $\hat{\chi}^2$ calculado se compara con el valor tabulado de una χ^2 para un nivel de confianza determinado y (n-1) (k-1) grados de libertad. Si el valor calculado es mayor que el valor de tablas de una $\hat{c}^2_{(n-1)(k-1)}$, significará que las diferencias entre las frecuencias observadas y las frecuencias teóricas o esperadas son muy elevadas y por tanto diremos con un determinado nivel de confianza que existe dependencia entre los factores o atributos analizados.

Resumiendo:

$\hat{\chi}^2 > \hat{c}^2_{(n-1)(k-1)} \Rightarrow$ Rechazar hipótesis nula (dependencia entre las variables)

$\hat{\chi}^2 < \hat{c}^2_{(n-1)(k-1)} \Rightarrow$ Aceptar hipótesis nula (independencia entre las variables)

Veámoslo con el mismo ejemplo anterior:

		SEXO		
		HOMBRE	**MUJER**	**MARGINAL**
FUMA	**SI**	65	58	123
	NO	43	67	110
	MARGINAL	108	125	233

Frecuencias relativas marginales:

P (ser hombre) = 108/ 233 = 46.4%
P (ser mujer) = 125/ 233 = 53.6%
P (fumar) = 123/ 233 = 52.8%
P (no fumar) = 110/ 233 = 47.2%

Frecuencias relativas conjuntas:

P (hombre y fumar) = 65/ 233 = 27.9%
P (hombre y no fumar) = 43/ 233 = 18.5%
P (mujer y fumar) = 58/ 233 = 24.9%
P (mujer y no fumar) = 67/ 233 = 28.8%

Frecuencias relativas teóricas esperadas en caso de independencia:

E (hombre y fumar) = 46.4% x 52.8% = 24.5%
E (hombre y no fumar) = 46.4% x 47.2% = 21.9%
E (mujer y fumar) = 53.6% x 52.8% = 28.3%
E (mujer y no fumar) = 53.6% x 47.2% = 25.3%

Frecuencias absolutas teóricas esperadas en caso de independencia:

E (hombre y fumar) = 123 * 108 /233 = 57
E (hombre y no fumar) = 108*110/233 = 51
E (mujer y fumar) = 123*125/233 = 66
E (mujer y no fumar) = 125*110/233 = 59

Valor de la Chi-cuadrado:

$$\hat{c}^2 = \frac{\sum_{i=1}^{h}\sum_{j=1}^{k}(n_{ij} - E_{ij})^2}{E_{ij}} = \frac{(65-57)^2}{57} + \frac{(58-66)^2}{66} + \frac{(43-51)^2}{51} + \frac{(67-59)^2}{59} = 4{,}42$$

Dado que el valor calculado de la $\hat{\chi}^2$ para un nivel de confianza del 95% (5% nivel de significación) es mayor que el valor de tablas, se rechaza la hipótesis nula de independencia entre los factores, aceptando por tanto que el sexo de una persona influye en que ésta sea fumadora o no.

Cuando utilicemos el SPSS nos dará el nivel de significación, es decir la probabilidad de rechazar la hipótesis nula siendo cierta y por tanto la probabilidad de equivocarnos si rechazamos la hipótesis nula. Si esta probabilidad es muy pequeña (<0,05), rechazaremos la hipótesis nula y en consecuencia diremos que los atributos son dependientes. Por el contrario, si el nivel de significación fuera superior a 0,05, la probabilidad de equivocarnos si concluyéramos que los factores son dependientes sería muy alta, y por tanto cabría esperar que nos equivocaríamos en nuestra conclusión, y por tanto aceptaremos la hipótesis nula de independencia.

El problema de la $\hat{\chi}^2$ es que está influenciada por el tamaño muestral, es decir, que a mayor número de casos analizados (a mayor N), el valor de la $\hat{\chi}^2$ tiende a aumentar, por lo que cuanto mayor sea la muestra más fácil será que rechacemos la hipótesis nula de independencia, cuando a lo mejor podrían no ser independientes.

Otro aspecto a tener en cuenta a la hora de realizar este contraste, es que para que el contraste sea estadísticamente válido en cada celda de la tabla deberá existir un mínimo de 5 observaciones. Si no fuera así deberemos agregar filas o columnas, siempre y cuando el tipo de información lo permita.

3. PASOS A SEGUIR A TRAVÉS DE SPSS

Vamos a realizar el mismo análisis a través de SPSS.

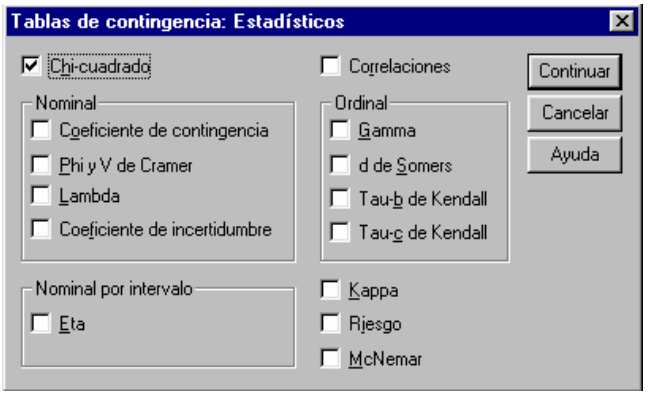
Para analizar si existe una relación de dependencia o no entre estas dos variables a través de SPSS tendremos que seleccionar en el menú “Analizar” la opción “Estadísticos” y dentro de esta la opción “Tablas de contingencia”. El cuadro de dialogo que aparece es el siguiente:

En esta pantalla debemos seleccionar las variables para las cuales queremos realizar el análisis. Normalmente introduciremos la variable dependiente (en este caso la variable "fumar") en el apartado correspondiente a filas y la variable independiente ("sexo") en la casilla correspondiente a columnas.

A continuación debemos seleccionar algunas opciones a través de bs distintos botones para que la salida de SPSS sea más completa:

ESTADÍSTICOS:

En esta pantalla será imprescindible señalar la opción Chi-cuadrado, ya que este es el estadístico que nos va a permitir contrastar la relación de dependencia o independencia entre las dos variables objeto de estudio.



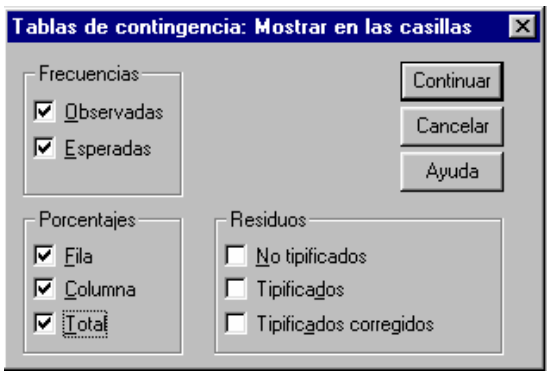
La opción de estadísticos también nos permite calcular distintas medidas de asociación para el caso en el que el valor de la Chi-cuadrado indique que existe una relación de

dependencia entre las variables. Es decir, la Chi-cuadrado permite contrastar la hipótesis de independencia, pero en el caso de que se rechace dicha hipótesis no dice nada sobre la fuerza de asociación entre las variables estudiadas debido a que su valor está afectado por el número de casos incorporados en la muestra.

Las medidas de asociación distinguen entre que las variables a analizar sean nominales u ordinales. Así, las medidas de asociación nominales sólo informan del grado de asociación existente pero no de la dirección de esa asociación. Sus valores son siempre positivos de manera que un resultado próximo a cero indica un bajo nivel de asociación, mientras que un resultado próximo a 1 indica un elevado nivel de asociación.

Por su parte, las medidas de asociación ordinales aportan información sobre la dirección de la relación, pudiendo tomar tanto valores positivos como negativos. Así, un resultado positivo indica una relación directa entre las variables analizadas, es decir, valores altos de una variable se corresponden con valores altos de la otra y valores bajos de una con valores también bajos en la otra. Sin embargo un resultado negativo representa una relación inversa entre ambas variables, es decir, valores altos en una variables se corresponden con valores bajos en la otra y viceversa.

CASILLAS:



A través de la opción "casillas" especificaremos qué valores son los que queremos que aparezcan en cada una de las celdas de la tabla de contingencia. Así podremos seleccionar las frecuencias observadas y esperadas, los porcentajes que suponen cada una de las celdas sobre la fila correspondiente, columna y sobre el total. Esta información resultará muy útil a la hora de describir las características de la muestra considerada.

En esta ventana también es posible seleccionar información sobre los residuos del ajuste. Los residuos son calculados como la diferencia entre la frecuencia observada y esperada en cada casilla, y resultan especialmente útiles para interpretar las relaciones que se observan en la tabla, especialmente cuando el número de alternativas de respuesta en alguna de las variables o en ambas es superior a 2.

En concreto, los residuos tipificados corregidos, que se distribuyen como una normal con media cero y desviación típica uno, indican que la diferencia entre la frecuencia observada y esperada es elevada, cuando su valor es superior a 1,96 en valor absoluto para un nivel de confianza del 95%. Así, un residuo tipificado corregido mayor a 1,96

en valor absoluto en una casilla indica que hay más casos, si es positivo, o menos, si es negativo, de los que debería haber en esa casilla si las variables fueran independientes, mientras que un valor comprendido entre +/-1,96 indica que la diferencia entre la frecuencia observada y la esperada es pequeña por lo que las variables en esa casilla tienen un comportamiento de independencia.

FORMATO:

Esta opción permite modificar el orden de las filas (ascendente o descendente).