2. TABLAS DE CONTINGENCIA

El uso de Tablas de Contingencia permite estudiar la relación entre dos variables categóricas o criterios de clasificación.

En una Tabla, los renglones representan las categorías de un criterio y las columnas las categorías del otro.

Referencia: Everitt, B. S. 1977 “The Analysis of Contingency Tables”, Halsted Press, 128 pp.

Por ejemplo, un criterio de clasificación podría ser los ingresos anuales por familia en cierta ciudad y el otro criterio las zonas donde viven las familias de dicha ciudad.

ZONAS	INGRESOS			
	Bajo	Medio	Alto	Total
A	n_{11}	n_{12}	n_{13}	$n_{1.}$
B	n_{21}	n_{22}	n_{23}	$n_{2.}$
C	n_{31}	n_{32}	n_{33}	$n_{3.}$
D	n_{41}	n_{42}	n_{43}	$n_{4.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{..}$

Si los ingresos anuales por familia y las zonas donde viven son ***independientes***, entonces en todas las zonas de la ciudad vivirían en las mismas proporciones familias de bajos, medios y altos ingresos.

Un cuadro n_{ij} cualquiera de la tabla, contiene las frecuencias de familias clasificadas de acuerdo con los criterios de las dos categorías.

Esto significa que cada cuadro esta formado por la interacción del nivel i-ésimo de un criterio con el nivel j-ésimo del otro criterio, esto se conoce como ***celda***.

Los totales marginales para cada criterio son:

Suma de las celdas de los renglones:

$$n_{i.} = \sum_{j=1}^{c} n_{ij}$$

Suma de las celdas de las columnas:

$$n_{.j} = \sum_{i=1}^{r} n_{ij}$$

El gran total es:

Suma de las celdas de los renglones

$$n = n_{..} = \sum_{j=1}^{c}\sum_{i=1}^{r} n_{ij} = \sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{c} n_{.j}$$

Suma de las celdas de las columnas

que deberá ser igual al número total de elementos en la muestra.

Prueba χ^2 (Ji-cuadrada) para independencia

Si se denotan por O_{ij} las frecuencias observadas n_{ij} en una muestra y por E_{ij} las frecuencias esperadas.

El estadístico χ^2 de Pearson es:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

con (r-1)(c-1) grados de libertad.

Prueba χ^2 (Ji-cuadrada) para independencia

Si dos eventos son independientes, entonces la probabilidad de una intersección de eventos es el producto de sus probabilidades. Si suponemos que los niveles de un criterio son independientes de los niveles del otro criterio, entonces los valores esperados serán:

$$E_{ij} = p_{i.}p_{.j}n = \frac{n_{i.}}{n}\frac{n_{.j}}{n}n = \frac{n_{i.}n_{.j}}{n}$$

suponen cierta la hipótesis de nulidad Ho: $p_{ij} = p_{i.}p_{.j}$

Criterio de Razón de Verosimilitud

Un método alterno para obtener el estadístico χ^2, para comparar las frecuencias observadas con las esperadas bajo una hipótesis particular, es la χ^2_L de Razón de Verosimilitud; el estadístico de prueba en este caso es:

$$\chi^2_L = 2 \cdot \sum_{i=1}^{r} \sum_{j=1}^{c} O_{ij} \cdot LN\left(\frac{O_{ij}}{E_{ij}}\right)$$

que también tiene una distribución χ^2 cuando la hipótesis nula es cierta, los grados de libertad son los mismos que la χ^2 de Pearson.

Pearson vs Razón de Verosimilitud

Es posible demostrar que χ^2 es aproximadamente igual a χ^2_L para muestras grandes.

Sin embargo, algunos autores muestran que, en general, χ^2_L es preferible a χ^2 por lo que se recomienda su utilización.

Para ambas pruebas se requiere que los valores esperados sean mayores a 5 en todas las celdas, de lo contrario sus resultados no son válidos.

Análisis de residuales

Un procedimiento útil para identificar las categorías que influyen en forma significativa en los valores de la χ^2, es el análisis de los *residuales*, d_{ij} , dados por:

$$d_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$$

El estimador de la varianza de los valores d_{ij} es:

$$v_{ij} = \left(1 - \frac{n_{i.}}{n}\right) \cdot \left(1 - \frac{n_{.j}}{n}\right)$$

Análisis de residuales

Finalmente, para cada celda de la tabla de contingencia se calculan los ***residuales estandarizados***, z_{ij}, que son:

$$z_{ij} = \frac{d_{ij}}{\sqrt{v_{ij}}}$$

Cuando las variables consideradas en la tabla de contingencia son independientes, los términos presentan una distribución normal (aprox.) con media cero y varianza uno.

Los valores se contrastan con un valor de z de la distribución normal estándar para un nivel de confianza dado. Si representan una influencia en la dependencia entre las variables, se espera que sean superiores al valor de z en valor absoluto.