

# **Análisis de tablas de contingencia bidimensionales**

Ana María Aguilera del Pino

Abril, 2005

# Índice

1	Variables Cualitativas . . . . .	1
1.1	Clasificación de variables cualitativas . . . . .	1
2	Distribución de frecuencias observadas y de probabilidades poblacionales . . . . .	2
2.1	Independencia poblacional y muestral . . . . .	5
3	Contrastes de independencia . . . . .	7
3.1	Contraste de independencia chi-cuadrado . . . . .	7
3.2	Contraste de Independencia de Razón de Verosimilitudes	9
3.3	Test exacto de Fisher para tablas $2 \times 2$ . . . . .	9
4	Contrastes de Simetría y Homogeneidad Marginal para Tablas Cuadradas Generadas por Muestras Dependientes . . . . .	16
4.1	Test de McNemar para una tabla $2 \times 2$ . . . . .	19
4.2	Contraste de Simetría en tablas $I \times I$ : extensión de Bowker al test de McNemar . . . . .	21
4.3	Test chi-cuadrado de homogeneidad marginal: extensión de Stuart y Maxwell al test de McNemar . . . . .	22
5	Medidas de asociación . . . . .	24
5.1	Valor poblacional y su estimación muestral . . . . .	24
5.2	Cociente de ventajas . . . . .	26
5.3	Medidas basadas en el estadístico $X^2$ de Pearson . . .	28
5.4	Medida Gamma de Goodman y Kruskal para variables ordinales . . . . .	28
5.5	Medida kappa de acuerdo . . . . .	31

# 1 Variables Cualitativas

Los fenómenos aleatorios se pueden representar mediante dos tipos de variables estadísticas:

- *Cuantitativas*: aquellas cuyos posibles valores son numéricos, por ejemplo, la estatura, el peso, la edad, la temperatura, etc. Estas se clasifican a su vez en variables continuas y discretas. Las variables continuas son aquellas que pueden tomar un conjunto infinito no numerable de valores y las discretas un conjunto finito o infinito numerable. Sin embargo, en la práctica la distinción se hace entre variables que toman muchos valores o relativamente pocos porque debido a la limitación de los instrumentos de medida todas las variables se miden de forma discreta.
- *Cualitativas* (o categóricas o atributos o factores): aquellas cuyos valores son un conjunto de cualidades no numéricas a las que se le suele llamar categorías o modalidades o niveles, por ejemplo, el sexo (mujer, hombre), el color del pelo (moreno, rubio, castaño, pelirrojo), filosofía política (liberal, moderada, conservadora), estado civil (soltero, casado, divorciado, viudo), nivel de estudios (ninguno, primario, medio, universitario), etc.

Una propiedad deseable de las categorías es que sean exhaustivas (proporcionan suficientes valores para clasificar a toda la población) y mutuamente excluyentes (cada individuo se clasifica en una y solo una categoría).

Observemos que, a primera vista, la exhaustividad puede parecer muy restrictiva, por ejemplo, puede que deseemos saber que opinan los liberales y conservadores frente a la legalización del aborto. En este caso el problema se resuelve redefiniendo la población mediante eliminación de los moderados.

## 1.1 Clasificación de variables cualitativas

Hay varias formas de clasificar las variables cualitativas:

### 1. *Variables dicotómicas y politómicas*

Según el número de categorías:

- Dicotómicas: solo dos modalidades, por ejemplo, padecer una enfermedad (Si, No), Sexo (Hombre, Mujer), Resultado de una oposición (Aprobar, Suspender), en general los fenómenos de respuesta binaria, etc.

- Politómicas: más de dos categorías, por ejemplo, los fenómenos de respuesta múltiple, lugar de nacimiento, clase social, etc.

## 2. Escalas nominal, ordinal y por intervalos

Según la escala de medida de las categorías:

- Nominal: no se puede definir un orden natural entre sus categorías, por ejemplo, la raza (blanca, negra, otra), la religión (católica, judía, protestante, otra), etc.
- Ordinal: es posible establecer relaciones de orden entre las categorías lo que lleva a establecer relaciones de tipo mayor, menor, igual o preferencia entre los individuos. Por ejemplo, el rango militar (soldado, sargento, teniente, otro), la clase social (alta, media, baja), etc. Sin embargo, no se pueden evaluar distancias absolutas entre categorías. Así, podemos decir que una persona de clase alta tiene mayor poder adquisitivo que una persona de clase media pero no podemos decir exactamente cuál es la diferencia en poder adquisitivo entre ambas.
- Por intervalo: proceden de variables cuantitativas agrupadas en intervalos. Estas variables pueden ser tratadas como ordinales pero para ellas se pueden calcular, además, distancias numéricas entre dos niveles de la escala ordinal, ejemplos de este tipo son el sueldo, la edad los días del mes o el nivel de presión sanguínea.

Existen variables que pueden ser medidas en escala nominal, ordinal o cuantitativa. Por ejemplo, el tipo de educación (privado, público) es nominal, el nivel de educación (primaria, secundaria, universitaria, postgraduado) es ordinal, y el número de años de educación (0,1,2,...) es cuantitativa.

Los métodos estadísticos propios para analizar variables nominales pueden ser usados para variables ordinales pero no al revés. Lo mejor es usar métodos apropiados para cada tipo de escala.

## 2 Distribución de frecuencias observadas y de probabilidades poblacionales

Consideremos una muestra de tamaño  $n$  clasificada según dos factores A y B con I y J niveles, respectivamente. Si representamos por  $n_{ij}$  ( $i = 1, \dots, I; j = 1, \dots, J$ ) el número de individuos de la muestra que se clasifican simultáneamente en el nivel  $A_i$  de A y  $B_j$  de B, la tabla bidimensional que

		Factor B						
		$B_1$	$\dots$	$B_j$	$\dots$	$B_J$	$n_{i.}$	
Factor A	$A_1$	$n_{11}$	$\dots$	$n_{1j}$	$\dots$	$n_{1J}$	$n_{1.}$	
	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	
	$A_i$	$n_{i1}$	$\dots$	$n_{ij}$	$\dots$	$n_{iJ}$	$n_{i.}$	
	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	
	$A_I$	$n_{I1}$	$\dots$	$n_{Ij}$	$\dots$	$n_{IJ}$	$n_{I.}$	
		$n_{.j}$	$n_{.1}$	$\dots$	$n_{.j}$	$\dots$	$n_{.J}$	$n$

Tabla 1: Tabla de frecuencias observadas  $I \times J$ 

contiene en cada una de sus  $I \times J$  casillas las frecuencias observadas  $n_{ij}$  se llama *tabla de contingencia bidimensional* y su nombre es debido a Pearson (1904). También se suele llamar tabla de dos vías o tabla cruzada. A los niveles de A se le suelen llamar filas y a los de B columnas. La notación general de una tabla de contingencia es la que aparece en la Tabla 1.

A partir de la tabla de contingencia se obtienen en la siguientes forma las distribuciones de frecuencias marginales y condicionadas:

Distribuciones marginales:

*Factor A:*

$$n_{i.} = \sum_{j=1}^J n_{ij} \quad (i = 1, \dots, I)$$

*Factor B:*

$$n_{.j} = \sum_{i=1}^I n_{ij} \quad (j = 1, \dots, J)$$

verificando:

$$n = \sum_{i=1}^I \sum_{j=1}^J n_{ij} = \sum_{i=1}^I n_{i.} = \sum_{j=1}^J n_{.j}$$

Distribuciones condicionadas:

*Factor A condicionado a  $B=B_j$  ( $j = 1, \dots, J$ ):*

$$n_{i|j} = n_{ij}/n_{.j} \quad i = 1, \dots, I$$

*Factor B condicionado a  $A=A_i$  ( $i = 1, \dots, I$ ):*

$$n_{j|i} = n_{ij}/n_{i.} \quad j = 1, \dots, J$$

Tabla 2: Tabla  $I \times J$  de probabilidades poblacionales

Recordemos que cualquier individuo de la población debe clasificarse en una y solo una de las  $I \times J$  casillas de la tabla de contingencia 1. Si denotamos por  $p_{ij}$  a la probabilidad de que un individuo elegido aleatoriamente en la población se clasifique en el nivel  $A_i$  de A y en el nivel  $B_j$  de B, obtenemos la *distribución de probabilidad poblacional* de los factores A y B verificando, como cualquier distribución de probabilidad discreta, que:

Para representarla se usa la Tabla 2 que tiene la misma estructura que la tabla de contingencia

Distribuciones de probabilidad marginales:

$$p_{i.} = \sum_{j=1}^J p_{ij} \quad (i = 1, \dots, I)$$
$$p_{.j} = \sum_{i=1}^I p_{ij} \quad (j = 1, \dots, J)$$
$$\sum_{i=1}^I p_{i.} = \sum_{j=1}^J p_{.j} = 1$$

Las distribuciones marginales dan información unidimensional sobre cada variable y no dicen nada sobre la asociación entre las dos variables.

Distribuciones condicionadas:

*Factor A condicionado a  $B=B_j$  ( $j = 1, \dots, J$ ):*

$$p_{i|j} = p_{ij}/p_{.j} \quad i = 1, \dots, I$$

*Factor B condicionado a  $A=A_i$  ( $i = 1, \dots, I$ ):*

$$p_{j|i} = p_{ij}/p_{i.} \quad j = 1, \dots, J$$

## 2.1 Independencia poblacional y muestral

Dos variables cualitativas son *estadísticamente independientes* si se verifica

$$p_{j|i} = p_{.j} \quad i = 1, \dots, I. \quad (1)$$

Esto significa que las distribuciones condicionales de B son iguales que su distribución marginal. Es decir, dos variables son independientes cuando la probabilidad de clasificarse en la columna  $j$  es igual en todas las filas. Intuitivamente la independencia significa que los valores de una variable no vienen influidos por la modalidad que adopte la otra variable. De igual forma las variables son independientes si las distribuciones condicionales de A son iguales que su distribución marginal.

Además, la independencia se caracteriza en términos de la distribución conjunta de probabilidades poblacionales como sigue: A y B son independientes si y sólo si las probabilidades conjuntas son igual al producto de las probabilidades marginales, es decir:

$$p_{ij} = p_{i.}p_{.j} \quad \forall i = 1, \dots, I; j = 1, \dots, J. \quad (2)$$

A continuación veremos con un ejemplo que la relación entre dos factores puede ser de tres tipos: independencia, dependencia estadística (asociación) o independencia funcional.

**Ejemplo 1** *Supongamos que se quiere estudiar la relación entre fumar (Si, No) y padecer cáncer de pulmón (Si, No), y que se conocen las probabilidades poblacionales asociadas a sus 4 combinaciones de categorías. Consideraremos tres configuraciones distintas de probabilidades que son las que aparecen en las tablas 3, 4 y 5.*

	Fumar		
Cáncer	Si	No	$p_{i.}$
Si	0.3	0.3	0.6
No	0.2	0.2	0.4
$p_{.j}$	0.5	0.5	1

Tabla 3: Ejemplo de independencia poblacional

	Fumar		
Cáncer	Si	No	$p_{i.}$
Si	0.4	0.1	0.5
No	0.2	0.3	0.5
$p_{.j}$	0.6	0.4	1

Tabla 4: Ejemplo de dependencia estadística

	Fumar		
Cáncer	Si	No	$p_{i.}$
Si	0.6	0.0	0.6
No	0.0	0.4	0.4
$p_{.j}$	0.6	0.4	1

Tabla 5: Ejemplo de dependencia funcional



Se observa que en la Tabla 3 tener la enfermedad no está relacionado con fumar o no fumar (independencia), en la Tabla 4 la mayoría de los fumadores tienen cáncer de pulmón y la mayoría de los no fumadores no lo tienen (dependencia estadística), y en la Tabla 5 solo tienen la enfermedad los fumadores (dependencia funcional).

De forma similar se puede definir el concepto de *independencia muestral*. Para ello, asociada a una tabla de contingencia del tipo 1 se define también la *distribución muestral de proporciones* o probabilidades como  $\hat{p}_{ij} = n_{ij}/n$  que es la proporción individuos muestrales clasificados en la casilla (i,j). Entonces, la proporción de veces que un individuo de la fila  $i$  se clasifica en la columna  $j$  es  $\hat{p}_{j|i} = \hat{p}_{ij}/\hat{p}_{i.} = n_{ij}/n_{i.}$ , siendo  $\hat{p}_{i.}$  la proporción muestral de individuos en la fila  $i$ .

Las variables A y B son independientes en la muestra si se verifica:

$$\hat{p}_{j|i} = \hat{p}_{.j}$$

o equivalentemente

$$\hat{p}_{ij} = \hat{p}_{i.}\hat{p}_{.j}$$

### 3 Contrastes de independencia

Nuestro primer objetivo será contrastar la hipótesis de independencia poblacional entre dos factores basándonos en la información proporcionada por las frecuencias observadas contenidas en la tabla de contingencia.

#### 3.1 Contraste de independencia chi-cuadrado

Consideremos una tabla de contingencia  $I \times J$  generada por muestreo multinomial. Esto significa que la distribución conjunta del vector aleatorio que genera a las frecuencias observadas es una multinomial de parámetros las probabilidades poblacionales.

Supongamos que queremos contrastar la hipótesis nula de independencia

$$H_0 : p_{ij} = p_{i.}p_{.j} \quad \forall i, j$$

Sean  $\hat{m}_{ij} = n_{i.}n_{.j}/n$  los estimadores de máxima verosimilitud de las frecuencias esperadas bajo la hipótesis de independencia. Para llevar a cabo este contraste, Pearson definió el estadístico

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}},$$

que, bajo la hipótesis nula de independencia, tiene distribución de probabilidad asintótica (para tamaños muestrales grandes)  $\chi^2_{(I-1)(J-1)}$ .

Se rechaza la hipótesis de independencia al nivel  $\alpha$  cuando se verifica  $X^2_{obs} \geq \chi^2_{(I-1)(J-1);\alpha}$ .

Pearson (1990) estableció erróneamente que los grados de libertad del estadístico  $X^2$  eran  $(IJ-1)$  ya que hay en total  $IJ$  casillas en la tabla e  $(IJ-1)$  parámetros libres de la distribución multinomial de las frecuencias observadas. Fue Fisher en 1922 quién corrigió el error y estableció que los grados de libertad se obtienen como el número de parámetros libres menos el número de parámetros a estimar:  $df = IJ - 1 - (I - 1) - (J - 1) = (I - 1)(J - 1)$ .

### Expresión operativa del estadístico $X^2$ en tablas $2 \times 2$

El principal inconveniente del estadístico  $X^2$  es que su cálculo es muy laborioso porque conlleva calcular en primer lugar las frecuencias esperadas.

Se demuestra fácilmente que para una tabla  $2 \times 2$  su valor es:

$$X^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}} \quad (3)$$

### Corrección por continuidad del estadístico $X^2$

Los contrastes de independencia basados en la distribución chi-cuadrado aproximan una distribución discreta (la del estadístico del contraste que es función de la multinomial) por una continua (la distribución chi-cuadrado). Esto da lugar a un error considerable cuando el tamaño muestral no es demasiado grande. Los métodos que disminuyen este error se llaman correcciones por continuidad.

Para mejorar esta aproximación en el caso de tamaños muestrales pequeños, Yates (1934) propuso un método que se llama corrección por continuidad de Yates para el estadístico  $X^2$  en una tabla  $2 \times 2$ .

Esta corrección consiste en restar  $1/2$  a las desviaciones positivas y sumar  $1/2$  a las desviaciones negativas. Por lo tanto, el estadístico  $X^2$  corregido es de la forma:

$$X^2_c = \sum_i \sum_j \frac{(|n_{ij} - \hat{m}_{ij}| - (1/2))^2}{\hat{m}_{ij}} \quad (4)$$

que en una tabla  $2 \times 2$  es de la forma:

$$X^2_c = \frac{n(|n_{11}n_{22} - n_{12}n_{21}| - (n/2))^2}{n_{1.}n_{.1}n_{2.}n_{.2}} \quad (5)$$

Existe mucha polémica sobre la aplicación de la corrección de Yates porque hay casos prácticos en los que con el estadístico corregido se rechaza la independencia con bastante menor significación que con el test no corregido. Es decir, el p-valor corregido es bastante mayor que el no corregido y el test corregido es más conservativo. En caso de tamaño muestral grande la corrección tiene efecto pequeño sobre el valor del estadístico  $X^2$ .

### 3.2 Contraste de Independencia de Razón de Verosimilitudes

Vamos a usar ahora el test de razón de verosimilitudes para contrastar la hipótesis de independencia

$$H_0 : p_{ij} = p_{i.}p_{.j} \quad \forall i, j$$

en una tabla de contingencia  $I \times J$  generada por muestreo multinomial.

Sea el estadístico

$$G^2 = -2 \log \Lambda = 2 \sum_i \sum_j n_{ij} \log \frac{n_{ij}}{\hat{m}_{ij}} \quad (6)$$

siendo  $\hat{m}_{ij} = n_{i.}n_{.j}/n$  las frecuencias esperadas bajo la hipótesis de independencia.  $G^2$  se llama estadístico chi-cuadrado de razón de verosimilitudes y se distribuye asintóticamente como una v.a.  $\chi^2$  con  $(I-1)(J-1)$  grados de libertad.

Aplicando el test de razón de verosimilitudes se rechaza la hipótesis de independencia al nivel  $\alpha$  si se verifica

$$G_{obs}^2 \geq \chi_{(I-1)(J-1), \alpha}^2$$

### 3.3 Test exacto de Fisher para tablas $2 \times 2$

Consideremos una tabla  $2 \times 2$  con los totales marginales  $(n_{1.}, n_{2.}, n_{.1}, n_{.2})$  de ambas variables fijos.

Supongamos que queremos contrastar la hipótesis de independencia poblacional que para una tabla  $2 \times 2$  se puede expresar como:

$$H_0 : p_{1|1} = p_{1|2} = p.$$

La distribución exacta bajo  $H_0$  de cualquier distribución de frecuencias observadas cuyos totales marginales coincidan con los fijados de antemano,

es la hipergeométrica de parámetros  $(n, n_{.1}, n_{1.})$  dada por

$$\frac{\binom{n_{1.}}{n_{11}} \binom{n - n_{1.}}{n_{.1} - n_{11}}}{\binom{n}{n_{.1}}} = \frac{n_{1.}! n_{2.}! n_{.1}! n_{.2}!}{n! n_{11}! n_{12}! n_{21}! n_{22}!} \quad (7)$$

que bajo la hipótesis de independencia no depende de ningún parámetro desconocido, y expresa la distribución de las cuatro casillas de la tabla en términos del elemento  $n_{11}$  debido a que, dados los totales marginales, su valor determina las otras tres casillas de la tabla. El rango de posibles valores para  $n_{11}$  es  $\max\{0, n_{.1} - (n - n_{1.})\} \leq n_{11} \leq \min\{n_{.1}, n_{1.}\}$ .

Dado que aparecen en la fórmula de la distribución hipergeométrica un total de  $IJ+I+J+1$  factoriales, el cálculo de las probabilidades de todas las posibles tablas es complicado y laborioso. Este proceso se simplifica aplicando la fórmula de Feldman y Kingler que calcula una de estas probabilidades, por ejemplo la de la tabla observada, y las demás se obtienen a partir de ella. Si denotamos por  $p_{n_{11}}$  a la probabilidad que asigna la distribución hipergeométrica a una tabla con frecuencia  $n_{11}$  en la casilla (1,1), la expresión para la probabilidad de las demás tablas es

$$p_{n_{11}+1} = \frac{n_{12}n_{21}}{(n_{11}+1)(n_{22}+1)} p_{n_{11}}, \quad (8)$$

$$p_{n_{11}-1} = \frac{n_{11}n_{22}}{(n_{12}+1)(n_{21}+1)} p_{n_{11}}. \quad (9)$$

El paso siguiente es fijar una hipótesis alternativa y seleccionar aquellas tablas que se alejan de  $H_0$  tanto o más que la tabla observada en la dirección de la hipótesis alternativa considerada.

### Test de Fisher de dos colas

Fijemos como hipótesis alternativa

$$H_1^1 : p_{1|1} \neq p_{1|2} \quad (\text{asociación}).$$

Como bajo  $H_0$  se tiene que verificar  $n_{11} = \hat{m}_{11} = n_{1.}n_{.1}/n$ , se rechazará  $H_0$  cuando la distancia  $|d_{11}| = |n_{11} - \hat{m}_{11}|$  sea suficientemente grande. Es decir, la región de rechazo será de la forma  $|d_{11}| \geq k$  bajo la restricción de tamaño  $P[|d_{11}| \geq k] = \alpha$  (tamaño que puede no alcanzarse debido a la distribución discreta de  $n_{11}$ ).

Las tablas que se alejan de  $H_0$  tanto o más que la observada en la dirección de  $H_1^1$  (son al menos tan favorables a la hipótesis alternativa como la tabla

observada) son aquellas que verifican  $|d_{11}| \geq |d_{11Obs}|$ , definiendo  $d_{11Obs} = n_{11Obs} - \hat{m}_{11}$ . Por lo tanto, el p-valor del test es

$$p = P[n_{11} : |d_{11}| \geq |d_{11Obs}|].$$

$$\text{Si } n_{11Obs} \geq \hat{m}_{11} \text{ entonces } p = \sum_{\{n_{11}:n_{11} \notin (-n_{11Obs}+2\hat{m}_{11}, n_{11Obs})\}} p_{n_{11}}.$$

$$\text{Si } n_{11Obs} < \hat{m}_{11} \text{ entonces } p = \sum_{\{n_{11}:n_{11} \notin (n_{11Obs}, -n_{11Obs}+2\hat{m}_{11})\}} p_{n_{11}}.$$

Finalmente, el test resulta significativo para rechazar la independencia si  $p \leq \alpha$  siendo  $\alpha$  el nivel de significación previamente fijado para el contraste. Debido a que la distribución de las frecuencias observadas se obtiene condicionando sobre todos los totales marginales, el test de Fisher es conocido también como *test condicional exacto*.

A continuación nos planteamos contrastes que en caso de rechazar la hipótesis de independencia nos den información sobre el tipo de asociación.

### Test exacto de Fisher de una cola (Asociación positiva)

Fijemos como hipótesis alternativa

$$H_1^2 : p_{1|1} > p_{1|2} \quad (\text{asociación positiva}).$$

En este caso las tablas que son al menos tan favorables como la observada a la hipótesis alternativa son las que verifican  $d_{11} \geq d_{11Obs}$ , es decir,  $n_{11} - \hat{m}_{11} \geq n_{11Obs} - \hat{m}_{11}$ .

$$\text{En este caso el p-valor es } p = \sum_{\{n_{11}:n_{11} \geq n_{11Obs}\}} p_{n_{11}}.$$

Observemos que si  $n_{11Obs} < \hat{m}_{11}$  el p-valor del test es mayor que 0.5 debido a la forma de la distribución hipergeométrica. Por ello en este caso el test es siempre no significativo (se acepta la independencia).

### Test de Fisher de una cola (Asociación negativa)

Fijemos como hipótesis alternativa

$$H_1^3 : p_{1|1} < p_{1|2} \quad (\text{asociación negativa}).$$

En este caso las tablas que son al menos tan favorables como la observada a la hipótesis alternativa son las que verifican  $d_{11} \leq d_{11Obs}$ , es decir,  $(n_{11} - \hat{m}_{11}) \leq n_{11Obs} - \hat{m}_{11}$ .

$$\text{Por lo tanto el p-valor es } p = \sum_{\{n_{11}:n_{11} \leq n_{11Obs}\}} p_{n_{11}}.$$

	B	
A	Leche	Té
Leche	3	1
Té	1	3

Tabla 6: Experimento de la bebedora de té

Observemos que si  $n_{11Obs} > \hat{m}_{11}$  el p-valor del test es de nuevo mayor que 0.5 de modo que el test es no significativo (se acepta la independencia). Esta es la razón por la que programas como BMDP proporcionan sólo uno de los dos p-valores de los dos contrastes de Fisher, el de una cola a la derecha si  $n_{11Obs} < \hat{m}_{11}$  y el de una cola a la izquierda si  $n_{11Obs} > \hat{m}_{11}$ .

**Ejemplo 2** *Fisher en 1935 describió el siguiente experimento. Bebiendo té, una mujer británica exclamó: ‘Yo sería capaz de distinguir si la leche o el té fue puesto en primer lugar en la taza’. Para contrastar su afirmación, se le dieron ocho tazas de té y se le dijo que en cuatro de ellas el té había sido puesto en primer lugar. El orden de presentación de las tazas a la mujer fue aleatorio. En la Tabla 6 figuran los resultados del experimento, representando por A el ingrediente que se puso en primer lugar y por B la predicción que hizo la mujer sobre el ingrediente puesto en primer lugar. Se pretende contrastar si una señora es capaz de distinguir si la leche o el té fue puesto en primer lugar en la taza.*

Según lo expuesto anteriormente  $n_{11}$  tiene, bajo la hipótesis de independencia, distribución  $H(8,4,4)$ . Cuando alguno de los totales marginales es pequeño (en nuestro caso todos) la variable hipergeométrica  $n_{11}$  toma muy pocos valores, en este ejemplo: 0,1,2,3,4. Por lo tanto, hay 5 posibles tablas de frecuencias observadas. Las probabilidades de ocurrencia de cada una de estas tablas son las siguientes:

$n_{11}$	$p_{n_{11}}$
0	0.0143
1	0.2286
2	0.5142
3	0.2286
4	0.0143

Si nos planteamos contrastar si hay o no asociación entre la predicción de la señora y la realidad (un test de dos colas) el p-valor es:  $p = p_3 + p_0 + p_1 +$

$p_4 = 0.4858 > \alpha$  para cualquiera de los niveles de significación aceptables. Por lo tanto se acepta la independencia.

Observando con detalle este ejemplo, se ve que el contraste adecuado para decidir si la señora tiene razón tendría como hipótesis alternativa  $H_1^1$ , es decir, la probabilidad de acertar que la leche se puso en primer lugar es mayor para aquellas tazas en las que la leche se puso realmente en primer lugar.

El p-valor para este contraste es  $p = p_3 + p_4 = 0.243$  que no es significativo con los niveles usuales para rechazar la independencia. Observemos que a pesar de que acertó en 6 ocasiones de 8 se ha rechazado que la señora tuviese razón. Además, se observa que con un nivel de significación inferior a 0.0143 la hipótesis de independencia no se rechazaría nunca y con un nivel 0.05 solo se rechazaría si la tabla observada es la de valor  $n_{11} = 4$ .

Esto es debido a que cuando el tamaño muestral o alguno de los totales marginales es pequeño la distribución hipergeométrica es altamente discreta (toma muy pocos valores), el p-valor tiende a ser muy grande y el nivel de significación no se alcanza. Esto repercute en un test muy conservador (poco potente).

### Aleatorización de Tocher del test exacto de Fisher

Cuando se plantea un test de hipótesis interesa que su potencia (probabilidad de rechazar la hipótesis nula cuando es falsa, o lo que es lo mismo, probabilidad de tomar una decisión adecuada) sea lo más elevada posible para un nivel de significación fijo. En este sentido se ha demostrado que se necesitan tamaños muestrales elevados para detectar síntomas moderados de asociación con el test de Fisher.

Una solución a este problema es aleatorizar el test para alcanzar el nivel de significación y aumentar la potencia.

La aleatorización propuesta por Tocher consiste en dividir el p-valor del test en dos sumandos  $p = p_{Obs} + p_t$  definiendo  $p_t$  como la suma de las probabilidades de todas aquellas tablas que se alejan de la independencia más que la observada en el sentido marcado por la hipótesis alternativa.

El test de Fisher consiste en rechazar la independencia con probabilidad uno si  $p \leq \alpha$  (es decir para todas aquellas tablas observadas para las que su probabilidad más la de las que se alejan tanto o más de la independencia es menor o igual que  $\alpha$ ) y con probabilidad cero si  $p > \alpha$ . La aleatorización de Tocher consiste en lo siguiente:

- Rechazar la independencia con probabilidad 1 para todas aquellas tablas observadas para las que  $p_t < \alpha$  y  $p_{Obs} + p_t \leq \alpha$ .

- Rechazar la independencia con probabilidad  $(\alpha - p_t)/p_{Obs}$  para todas aquellas tablas observadas tales que  $p_t < \alpha$  y  $p_{Obs} + p_t > \alpha$ .
- Rechazar la independencia con probabilidad cero (aceptar) para todas aquellas tablas observadas para las que  $p_t \geq \alpha$ .

**Ejemplo 3** *Retomemos de nuevo el ejemplo de la bebedora de té para comprobar que se alcanza el nivel de significación mediante la aleatorización de Tocher.*

Si observamos la tabla donde figuran las probabilidades de cada una de las posibles tablas de frecuencias y fijamos nivel de significación 0.05, se tiene usando la aleatorización de Tocher que se rechaza la independencia con probabilidad 1 si la tabla observada es  $n_{11} = 4$ , con probabilidad 0.157 si la tabla observada es  $n_{11} = 3$  y con probabilidad cero para las otras tres tablas.

Como el nivel de significación es la esperanza de la probabilidad de rechazar la hipótesis nula siendo cierta, condicionada a los datos muestrales, se tiene

$$E[P[\text{Rechazar } H_0 | n_{11}]] = 1p_4 + 0.157p_3 + 0(p_2 + p_1 + p_0) = 0.05.$$

La aleatorización de Tocher del test de Fisher lleva en este caso a aceptar que la señora tenía razón con probabilidad 0.157.

En la práctica el test aleatorizado se realiza de la siguiente forma:

- Si  $p_t < \alpha$  y  $p_{Obs} + p_t \leq \alpha$  se rechaza  $H_0$ .
- Si  $p_t < \alpha$  y  $p_{Obs} + p_t > \alpha$  se calcula  $T = (\alpha - p_t)/p_{Obs}$  y se genera un número aleatorio  $s$  entre 0 y 1. Entonces, si  $s \leq T$  se rechaza  $H_0$  y si  $s > T$  se acepta.
- Si  $p_t \geq \alpha$  se acepta  $H_0$ .

### Observaciones sobre el Test exacto de Fisher

1. Aunque Tocher (1950) utilizó la aleatorización propuesta para aumentar la potencia del test de Fisher, ésta es aplicable a cualquier otro test de discreción elevada.
2. Tocher (1950) demostró que el test de Fisher con la aleatorización de Tocher es el test insesgado uniformemente más potente de los contrastes de una cola para tablas  $2 \times 2$ .



Sexo	Resultado del examen		
	Aprobado	Suspenso	Total
Hombre	2 (4)	8 (6)	10
Mujer	6 (4)	4 (6)	10
Total	8	12	20

Tabla 7: Éxito en los estudios por sexo

3. En general, en muchos ámbitos como el de la medicina y la biología se prefiere utilizar un test asintótico y su posterior corrección por continuidad en lugar de la aleatorización de la región crítica.
4. El test de Fisher se suele recomendar cuando la frecuencia mínima esperada es menor que 5, de hecho paquetes estadísticos como BMDP lo implementan solo bajo este supuesto.

**Ejemplo 4** *Para contrastar la hipótesis de que las mujeres tienen más éxito en sus estudios que los hombres, se ha tomado una muestra de 10 chicos y otra 10 de chicas que han sido examinados por un profesor que aprueba siempre al 40% de los alumnos presentados a examen. Teniendo en cuenta que sólo aprobaron 2 chicos, utilizar el test de hipótesis más adecuado para decidir si la citada hipótesis es cierta.*

Para construir la tabla de contingencia asociada a este experimento, calculamos en primer lugar el número total de aprobados que para este profesor es conocido antes de que corrija los exámenes, y corresponde al 40% de 20 que es 8. Observemos que como consecuencia no sólo están fijados por el diseño muestral los totales marginales de filas de la Tabla 7 de datos, sino también los totales marginales de columna. Por lo tanto, la Tabla 7 ha sido generada por muestreo hipergeométrico.

Como más del 20% de las frecuencias esperadas bajo la hipótesis de independencia (valores entre paréntesis en la Tabla 7) son menores que 5, se utilizará el test exacto de Fisher para el contraste de independencia.

Para contrastar la hipótesis nula de que la probabilidad de aprobar es igual en hombres que en mujeres frente a la hipótesis alternativa de que la probabilidad de aprobar es mayor para las mujeres que para los hombres, se planteará un test exacto de Fisher de una cola a la izquierda

$$H_0 : p_{1/1} = p_{1/2}$$

$$H_1^3 : p_{1/1} < p_{1/2}.$$

Para todas aquellas tablas de frecuencias con los mismos totales marginales que la tabla observada,  $n_{11}$  tiene bajo  $H_0$  distribución  $H(20, 8, 10)$  de modo que  $n_{11}$  podría tomar los valores  $\{0, 1, 2, 3, 4, 5, 6, 7, 8\}$ . Esto implica que hay nueve posibles tablas de frecuencias que responden al mismo diseño muestral que la observada.

El p-valor de este contraste exacto es la suma de las probabilidades de todas las tablas para las que se verifica  $n_{11} \geq n_{11Obs}$ . En este caso  $p = p_0 + p_1 + p_2$ .

Para calcular estas probabilidades se utiliza la fórmula de las probabilidades hipergeométricas para calcular la probabilidad de la tabla observada  $p_2$  y las demás se obtienen a partir de ella mediante la correspondiente fórmula de Feldman y Kinger. De este modo de tiene

$$\begin{aligned} p_2 &= \frac{10! \times 10! \times 8! \times 12!}{20! \times 2! \times 8! \times 6! \times 4!} = 0.07502, \\ p_1 &= \frac{2 \times 4}{9 \times 7} p_2 = 0.00953, \\ p_0 &= \frac{1 \times 3}{10 \times 8} p_1 = 0.00036. \end{aligned}$$

Como consecuencia el p-valor de este test es  $p = 0.08489 > \alpha = 0.05$  que lleva a aceptar la hipótesis de independencia con este nivel de significación. Observemos que aumentando el nivel de significación a 0.09 el test sería significativo para concluir que la probabilidad de aprobar es mayor para las mujeres que para los hombres.

A continuación vamos a aumentar la potencia del test de Fisher haciendo uso de la aleatorización propuesta por Tocher. Para ello descomponemos el p-valor en la forma  $p = p_{Obs} + p_t = p_2 + p_t$  siendo  $p_t = p_0 + p_1$ . Como  $p_t = 0.00987 < \alpha = 0.05$  y  $p > 0.05$ , la aleatorización de Tocher lleva a rechazar la independencia con probabilidad  $(0.05 - 0.00987)/0.07502 = 0.53492$ .

## 4 Contrastes de Simetría y Homogeneidad Marginal para Tablas Cuadradas Generadas por Muestras Dependientes

En esta sección analizaremos datos categóricos que proceden de muestras dependientes. Estos datos se representan mediante tablas que tienen el mismo número de filas y de columnas y reciben el nombre de *Tablas cuadradas*.

Los diseños muestrales más comunes que generan tablas de contingencia cuadradas a partir de datos dependientes son los siguientes:

- Una muestra de pares de individuos se clasifica de acuerdo a una variable respuesta cualitativa. Este tipo de estudio se llama *diseño apareada o por parejas* y a los datos que resultan se les llama datos apareados.

Los diseños apareados surgen de forma natural en muchas aplicaciones. Por ejemplo, si se desea estudiar si las profesiones se transmiten de padres a hijos, cada observación es una par representando la ocupación del padre y la del hijo. Otro ejemplo podría ser estudiar si hay diferencia de ideas políticas en un matrimonio, para lo que es necesario tomar una muestra de matrimonios.

Sin embargo, el objetivo de realizar un diseño apareado suele ser anular el efecto de factores ‘extraños’ que puedan interferir en el contraste de la independencia entre dos factores. Retomemos, como ejemplo, el estudio sobre si la vitamina C es buena para curar el resfriado común. Un diseño apareado significaría tomar una muestra de pares de individuos, de modo que los dos individuos de cada par tengan características muy similares (misma edad, sexo, estado de salud, forma física, etc). Después, en cada par, a un individuo se le da vitamina C y al otro no, observándose el número de resfriados después de cierto tiempo. De este modo las diferencias entre el grupo control y el que toma vitamina C no puede deberse a las posibles diferencias físicas entre un grupo y otro.

En cualquier diseño apareado, los individuos de cada par deben tener características similares que pueden variar de par en par. Para realizar el estudio, dentro de cada par se seleccionará un individuo al azar que se someterá al tratamiento de interés.

- Un único conjunto de individuos se clasifica según una variable de respuesta cualitativa en dos momentos de tiempo diferentes. Este tipo de estudio se llama *Diseño longitudinal* y consiste en tomar *medidas repetidas* de una variable cualitativa para un conjunto de individuos. Por ejemplo, un profesor desea estudiar como asimilan sus alumnos los conocimientos en Matemáticas que les enseña. Para ello les hace un examen cada mes clasificando el resultado como: malo, regular o bueno. Después de 9 meses de curso, tendrá 9 muestras dependientes porque los alumnos son cada mes los mismos y se tienen nueve observaciones repetidas de cada uno. Si el profesor desea contrastar si cada dos

	Nota Mes 2		
Nota Mes 1	Mala	Regular	Buena
Mala	5	5	5
Regular	2	4	6
Buena	1	2	4

Tabla 8: Nota en Matemáticas mes a mes

meses ha habido cambios significativos en el resultado de los exámenes, construirá tablas  $3 \times 3$  como la 8.

Debido a la limitación del investigador para construir muestras apareadas, muchas veces se utiliza cada elemento muestral como su propio control. Es decir, si se trata de comparar dos tratamientos médicos se aplicarían ambos sobre el mismo individuo. Esta es otra forma de obtener observaciones repetidas de una variable cualitativa que recibe el nombre de *método del control único*. Observemos que es una variante del diseño longitudinal en la que, en lugar de clasificar a los individuos en dos momentos diferentes, se les clasifica bajo dos codiciones o tratamientos diferentes. A veces el control único es desaconsejable porque el efecto del orden de aplicación de los tratamientos puede afectar al resultado del análisis.

Otra variante de datos dependientes procedentes de medidas repetidas es cuando dos profesionales clasifican, de forma independiente, a los individuos de una muestra según una variable cualitativa. Por ejemplo, cuando dos médicos diferentes diagnostican un tumor como maligno o benigno.

Las tablas que resultan de la representación de datos dependientes son cuadradas y además con las mismas categorías para filas y columnas. Para ellas el estudio de la independencia no tiene interés. Serán dos los conceptos de interés (debido a su interpretación en relación a si existen cambios significativos de la variable de interés en el tiempo o bajo los tratamientos considerados): la simetría y la homogeneidad marginal.

- Simetría:  $p_{ij} = p_{ji}$  ( $i \neq j$ )
- Homogeneidad marginal:  $p_{i.} = p_{.i}$  ( $i = 1, \dots, I$ )

En una tabla  $2 \times 2$  ambos conceptos son equivalentes. Sin embargo, en general, la simetría implica homogeneidad marginal pero el recíproco no es cierto.

### 4.1 Test de McNemar para una tabla $2 \times 2$

La estructura de una tabla  $2 \times 2$  procedente de datos dependientes binarios (clasificados según una variable dicotómica) es la siguiente:

Caso I	Caso II	
	A <sub>1</sub>	A <sub>2</sub>
A <sub>1</sub>	$n_{11}$	$n_{12}$
A <sub>2</sub>	$n_{21}$	$n_{22}$

Vamos a obtener en primer lugar un test aproximado (basado en la distribución chi-cuadrado) para contrastar la hipótesis nula:

$$H_0 : p_{1.} = p_{.1} \Leftrightarrow p_{12} = p_{21}$$

frente a la alternativa:

$$H_1 : p_{1.} \neq p_{.1} \Leftrightarrow p_{12} \neq p_{21}$$

$H_0$  significa que la probabilidad de clasificarse en la categoría  $A_1$  es la misma en los dos casos considerados. Es decir, se contrasta que no hay cambios significativos de la clasificación en los dos casos frente a la existencia de cambios. Por ello el test se suele llamar, también, test de significación de cambios.

El test es debido a McNemar (1955) y se obtiene como sigue. Bajo  $H_0$  los EMV (suponiendo muestreo multinomial para la tabla  $2 \times 2$ ) de las probabilidades poblacionales son:

$$\hat{p}_{11} = n_{11}/n; \hat{p}_{22} = n_{22}/n; \hat{p}_{12} = \hat{p}_{21} = (n_{12} + n_{21})/2n$$

Por lo tanto, los EMV de las frecuencias esperadas bajo  $H_0$  son:

$$\hat{m}_{11} = n_{11}; \hat{m}_{22} = n_{22}; \hat{m}_{12} = \hat{m}_{21} = (n_{12} + n_{21})/2$$

Entonces, sustituyendo en el estadístico  $X^2$  de bondad de ajuste se tiene la expresión:

$$X^2 = \frac{(n_{12} - n_{21})^2}{(n_{12} + n_{21})}$$

que tiene, bajo  $H_0$  distribución asintótica  $\chi^2$  con 1 grado de libertad, y es conocido como estadístico de McNemar.

El número de parámetros libres de la tabla es 3. El número de parámetros estimados es dos. Por lo tanto la diferencia es el número de grados de libertad igual a 1.

	Opinión		
	Si	No	Total
Ambulatorio	62	63	125
Centro Salud	70	55	125

Tabla 9: Opinión sobre asistencia sanitaria

Por lo tanto, el test resultará significativo para rechazar la hipótesis de independencia al nivel  $\alpha$  si se verifica:

$$X^2 \geq \chi_{1;\alpha}^2$$

Se puede comprobar fácilmente que la corrección de Yates del estadístico de McNemar es de la forma:

$$X_C^2 = \frac{(|n_{12} - n_{21}| - 1)^2}{(n_{12} + n_{21})}$$

La corrección de Yates es aconsejable cuando las frecuencias esperadas en el caso de simetría (u homogeneidad marginal) son menores que 5, es decir,  $(n_{12} + n_{21})/2 < 5$ .

**Ejemplo 5** *Se quiere saber si en los ambulatorios la asistencia sanitaria es del mismo tipo que en los centros de salud. Para ello se pregunta se eligen al azar 125 individuos y se les pregunta si están o no de acuerdo con el tipo de asistencia recibida en el centro de salud y lo mismo con la recibida en el ambulatorio.*

En principio la Tabla 9 podría representar a este estudio. Sin embargo, de acuerdo con esta tabla habríamos consultado a 250 individuos de los cuales 125 fueron preguntados sobre la asistencia sanitaria en el ambulatorio y los otros 125 sobre la asistencia en el centro de salud. El error está en que se trata de 125 individuos de los que se han tomado un par de observaciones de la misma variable (opinión) para cada uno: opinión sobre el ambulatorio y opinión sobre el centro de salud. Al ser un caso de control único le corresponde una tabla cuadrada  $2 \times 2$  (Tabla 10) donde se agrupan los individuos según la pareja de respuestas que ofrecen.

Estamos interesados en contrastar si la proporción de individuos que están de acuerdo con la asistencia recibida es igual en el ambulatorio que en el centro de salud. Para ello, como la frecuencia esperada en este caso es  $(27 + 20)/2 = 23.5 > 5$ , aplicaremos el test de McNemar.

El valor del estadístico de McNemar es  $X^2 = 0.8205 < \chi_{1;0.05}^2 = 3.8415$  que implica aceptar la hipótesis de homogeneidad marginal.

	Ambulatorio		
Centro Salud	Si	No	Total
Si	27	43	70
No	35	20	55
Total	62	63	125

Tabla 10: Datos dependientes: opinión sobre asistencia sanitaria

## 4.2 Contraste de Simetría en tablas $I \times I$ : extensión de Bowker al test de McNemar

Consideremos ahora la siguiente tabla cuadrada  $I \times I$  que contiene datos dependientes generados por un diseño longitudinal o por un diseño apareado o de control único:

	Caso II				
Caso I	$A_1$	...	$A_i$	...	$A_I$
$A_1$	$n_{11}$	...	$n_{1i}$	...	$n_{1I}$
...	...	...	...	...	...
$A_i$	$n_{i1}$	...	$n_{ii}$	...	$n_{iI}$
...	...	...	...	...	...
$A_I$	$n_{I1}$	...	$n_{Ii}$	...	$n_{II}$

Una vez estudiado el caso en que la variable cualitativa de interés es dicotómica, estudiaremos el caso general en que esta variable tiene tres o más categorías. Por ejemplo, horas de estudio en épocas de exámenes y en épocas sin exámenes, grado de visión en el ojo izquierdo y en el ojo derecho, altura de maridos y esposas, profesión de padres e hijos, ideas políticas de padres e hijos, etc.

Vamos a construir un contraste estadístico para la hipótesis nula

$$H_0 : p_{ij} = p_{ji} \quad i > j$$

frente a la alternativa

$$H_1 : p_{ij} \neq p_{ji} \quad \text{para algún par } (i,j)$$

$H_0$  significa que los cambios de categoría de un caso a otro se producen en ambas direcciones con igual probabilidad, es decir, las observaciones de las casillas situadas simétricamente respecto de la diagonal principal tienen la misma probabilidad de ocurrir.

El test es debido a Bowker (1948) y se obtiene como sigue. Recordemos que se demostró en el Capítulo 2 que los estimadores MV (suponiendo muestreo multinomial para la tabla  $I \times I$ ) de las probabilidades poblacionales son

$$\hat{p}_{ii} = n_{ii}/n \quad i = 1, \dots, I; \quad \hat{p}_{ij} = \hat{p}_{ji} = (n_{ij} + n_{ji})/2n \quad i \neq j.$$

Por lo tanto, los estimadores MV de las frecuencias esperadas bajo  $H_0$  son

$$\hat{m}_{ii} = n_{ii} \quad i = 1, \dots, I; \quad \hat{m}_{ij} = \hat{m}_{ji} = (n_{ij} + n_{ji})/2$$

Entonces, sustituyendo en el estadístico  $X^2$  de bondad de ajuste se tiene la expresión

$$X^2 = \sum_{i < j} \frac{(n_{ij} - n_{ji})^2}{(n_{ij} + n_{ji})},$$

que sigue, bajo  $H_0$ , una distribución asintótica  $\chi^2$  con  $I(I-1)/2$  grados de libertad.

El número de parámetros libres de la tabla es  $I^2 - 1$ . El número de parámetros estimados es el número de casillas de la tabla cuadrada por encima de la diagonal, más el número de casillas en la diagonal, menos uno

$$I + (I-1) + (I-2) + \dots + (I - (I-1)) - 1,$$

dado por

$$\left[ \sum_{i=1}^I I - (i-1) \right] - 1 = I^2 - \frac{I(I-1)}{2} - 1 = \frac{I(I+1)}{2} - 1.$$

Por lo tanto la diferencia es el número de grados de libertad  $I(I-1)/2$ .

Como es habitual, el test resultará significativo para rechazar la hipótesis de simetría al nivel  $\alpha$  si se verifica

$$X^2 \geq \chi_{\frac{1}{2}I(I-1); \alpha}^2.$$

### 4.3 Test chi-cuadrado de homogeneidad marginal: extensión de Stuart y Maxwell al test de McNemar

Si la hipótesis de simetría es rechazada puede ser de interés contrastar la hipótesis de homogeneidad marginal

$$H_0 : p_{i.} = p_{.i} \quad i = 1, \dots, I,$$

que constituye una extensión del test de McNemar a tablas  $(I \times I)$  desarrollada por Stuart (1955) y Maxwell (1970). El test consiste en contrastar si las diferencias entre las proporciones marginales son significativas.



El estadístico del test es de la forma

$$X^2 = \underline{d}' \mathbf{V}^{-1} \underline{d},$$

donde  $\underline{d}$  es un vector columna que contiene (I-1) cualesquiera de las diferencias  $d_i = n_{i.} - n_{.i}$  ( $i = 1, \dots, I$ ) y  $\mathbf{V}$  es la matriz de covarianzas del vector  $\underline{d}$ , cuyos elementos están definidos por

$$v_{ii} = n_{i.} + n_{.i} - 2n_{ii} \quad \text{y} \quad v_{ij} = -(n_{ij} + n_{ji}).$$

Observemos que en el vector  $\underline{d}$  se ha excluido una diferencia con el objeto de que  $\mathbf{V}$  sea no singular para que exista  $\mathbf{V}^{-1}$ . Stuart y Maxwell demuestran que el valor de  $X^2$  no cambia cualquiera que sea la diferencia que se omita.

Bajo  $H_0$ ,  $X^2$  tiene distribución asintótica  $\chi^2$  con I-1 grados de libertad. Por lo tanto, se rechazará la homogeneidad marginal al nivel  $\alpha$  si se verifica

$$X_{Obs}^2 \geq \chi_{I-1;\alpha}^2$$

Fleiss y Everitt (1971) demuestran que para una tabla  $3 \times 3$ , el estadístico  $X^2$  de homogeneidad marginal es de la forma

$$X^2 = \frac{\hat{m}_{23}d_1^2 + \hat{m}_{13}d_2^2 + \hat{m}_{12}d_3^2}{2(\hat{m}_{12}\hat{m}_{23} + \hat{m}_{12}\hat{m}_{13} + \hat{m}_{13}\hat{m}_{23})}$$

donde  $\hat{m}_{ij}$  son las frecuencias esperadas bajo la hipótesis de simetría dadas por

$$\hat{m}_{ij} = \frac{1}{2}(n_{ij} + n_{ji}).$$

**Ejemplo 6** *Existe una teoría sobre movilidad social que afirma que puede existir cierta herencia de la clase social del padre al hijo pero que, una vez que el hijo se sale de la clase social del padre, la nueva clase no tiene relación alguna con la que tenía el padre. Para contrastar esta idea se han clasificado 3497 parejas de padres e hijos según su clase social, como muestra la tabla 11.*

El valor del estadístico  $X^2$  del contraste de simetría es  $X^2 = 31.29$ . Los grados de libertad son 3 y el punto crítico al nivel  $\alpha = 0.05$  es  $\chi_{3;0.05}^2 = 7.8147$ . Por lo tanto el test es muy significativo para rechazar la hipótesis de simetría. Observemos que la mayor desviación es entre  $n_{32} = 320$  y  $n_{23} = 447$ , por lo que parece que más hijos de clase media pasan a clase baja que al contrario. Lo mismo ocurre para el número de hijos de clase alta que pasan a clase baja que es mayor que el número de hijos que cambian de clase en dirección opuesta.

	Clase Social Hijo			
Clase Social Padre	Alta	Media	Baja	Total
Alta	588	395	159	1142
Media	349	714	447	1510
Baja	114	320	411	845
Total	1051	1429	1017	3497

Tabla 11: Clase Social Padres-Hijos

	Clase Social		
	Alta	Media	Baja
Padres	0.33	0.43	0.24
Hijos	0.30	0.41	0.29

Tabla 12: Proporciones marginales por clase social

Contrastemos ahora la homogeneidad marginal con los datos de la Tabla 11. Las diferencias observadas entre los totales marginales de filas y columnas son  $d_1 = -91$ ,  $d_2 = -81$  y  $d_3 = 17$ . Una vez calculados los valores  $\hat{m}_{12} = 372$ ,  $\hat{m}_{13} = 136.5$  y  $\hat{m}_{23} = 383.5$ , se obtiene  $X^2 = 30.67$ , con 1 grado de libertad. El valor crítico es  $\chi^2_{1;0.05} = 3.8415$ , por lo que el test es muy significativo y concluye que la distribución marginal de la clase social de los padres es diferente de la de los hijos. En la Tabla 12 figuran las proporciones marginales muestrales que indican una tendencia a que la clase social de los hijos sea más baja que la de los padres.

## 5 Medidas de asociación

En caso de rechazar la independencia entre los dos factores de una tabla de contingencia, se plantea la necesidad de definir índices que describan no solo la intensidad de la asociación, sino también su dirección. El estudio de estos índices, que se conocen con el nombre genérico de *medidas de asociación*, es el objetivo principal de este capítulo.

### 5.1 Valor poblacional y su estimación muestral

El estudio pormenorizado de cada medida de asociación se hará en los cuatro pasos siguientes:

1. Definición de su valor poblacional que es una función de los parámetros poblacionales asociados al diseño muestral generador de la tabla de contingencia. En la mayoría de los casos, y salvo que se especifique lo contrario, se asumirá muestreo multinomial completo, en cuyo caso una medida de asociación es una función de las probabilidades poblacionales.

De este modo, una medida general de asociación para una tabla  $I \times J$  es una función  $f(\underline{p})$  definida del espacio de todos los posibles vectores de probabilidades poblacionales  $\underline{p} = (p_{ij})$  en un intervalo de la recta real.

2. Rango de variación e interpretación.
3. Estimación muestral, obtenida como la función que la define aplicada al vector de proporciones muestrales, que es el estimador MV del valor poblacional. Es decir, la estimación máximo-verosímil de  $f(\underline{p})$  es  $f(\hat{\underline{p}})$ , siendo  $\hat{\underline{p}}$  el vector de proporciones muestrales ( $\hat{p}_{ij} = n_{ij}/n$ ) asociadas a la tabla de contingencia.
4. Obtención de un intervalo de confianza asintótico para el valor poblacional y construcción de contrastes de independencia basados en dicha medida de asociación. Su cumplimiento conlleva la obtención de la distribución asintótica de cada una de las medidas de asociación.

Del Teorema Central del Límite se obtiene que el vector de proporciones muestrales  $\hat{\underline{p}}$  tiene distribución normal multivariante asintótica de vector de medias  $\underline{p}$  y matriz de covarianzas  $(\text{Diag}(\underline{p}) - \underline{p}\underline{p}')/n$ .

El método Delta proporciona la normalidad asintótica de una función de un vector aleatorio con distribución normal asintótica. Aplicando este método al valor muestral de una medida de asociación,  $f(\hat{\underline{p}})$ , se obtiene su distribución asintótica que es una normal de media  $f(\underline{p})$  y varianza  $\sigma^2[f(\hat{\underline{p}})] = (\underline{\Phi}'(\text{Diag}(\underline{p}) - \underline{p}\underline{p}')\underline{\Phi})/n$ , siendo  $\underline{\Phi}$  el vector de derivadas parciales de la función  $f$  respecto de las probabilidades poblacionales. Esta varianza asintótica es función de los probabilidades poblacionales, y se estimará mediante el valor  $\hat{\sigma}^2[f(\hat{\underline{p}})]$  obtenido sustituyendo en  $\sigma^2[f(\hat{\underline{p}})]$  el parámetro  $\underline{p}$  por su estimador MV  $\hat{\underline{p}}$ .

Por lo tanto, un intervalo de confianza asintótico al nivel  $(1-\alpha)$  para  $f(\underline{p})$  será de la forma

$$(f(\hat{\underline{p}}) - z_{\alpha/2}\hat{\sigma}[f(\hat{\underline{p}})], f(\hat{\underline{p}}) + z_{\alpha/2}\hat{\sigma}[f(\hat{\underline{p}})])$$

donde  $z_{\alpha/2}$  es el cuantil de orden  $1 - \alpha/2$  de la normal estándar.

## 5.2 Cociente de ventajas

El cociente de ventajas o razón de producto cruzado es una medida de asociación para una tabla  $2 \times 2$  definida por

$$\theta = \frac{p_{11}p_{22}}{p_{12}p_{21}}$$

Es una medida no negativa cuya interpretación es

1.  $\theta = 1$  sii las variables son independientes.
2.  $\theta > 1$  sii la probabilidad de clasificarse en la segunda columna en lugar de en la primera es mayor para los individuos de la segunda fila que para los de la primera fila.
3.  $\theta < 1$  sii la probabilidad de clasificarse en la segunda columna en lugar de en la primera es mayor para los individuos de la primera fila que para los de la segunda fila.

La estimación puntual de la razón de producto cruzado es

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

sumando 0.5 a cada casilla en el caso de que alguna de las frecuencias observadas sea cero.

Una transformación del cociente de probabilidades que toma valores en el intervalo  $[-1, 1]$ , lo que la hace más intuitiva, es la Q de Yule cuyo valor muestral es

$$\hat{Q} = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}$$

El valor  $Q=0$  corresponde a la independencia. Si Q es positiva los individuos de la primera fila están asociados con mayor probabilidad con los de la primera columna, y si Q es negativa los individuos de la primera fila están asociados con mayor probabilidad con los de la segunda columna.

**Ejemplo 7** *Para estudiar la asociación entre la raza y la imposición de pena de muerte a individuos culpables de homicidio se llevó a cabo un estudio sobre 326 individuos culpables de homicidio en Florida durante 1976-1977. Los datos observados para las variables ‘Veredicto de pena de muerte’ (Si,No) y ‘Raza del acusado’ (Blanca, Negra) figuran en la tabla 13.*

Raza	Pena de muerte	
	Si	No
Blanca	19	141
Negra	17	149

Tabla 13: Estudio Veredicto de pena de muerte-Raza del acusado

El cociente de ventajas muestral es

$$\hat{\theta} = \frac{19 * 149}{17 * 141} = 1.18.$$

Esta estimación puntual significa que en la muestra la probabilidad estimada de ser condenado a muerte en lugar de no serlo es 1.18 veces mayor para una persona de raza blanca que para una de raza negra.

Para poder inferir este resultado a la población vamos a construir un intervalo de confianza sobre  $\theta$  al nivel 0.95. Para ello calculamos  $\ln \hat{\theta} = 0.166$  cuyo error estándar estimado es

$$\hat{\sigma}(\ln \hat{\theta}) = \sqrt{\frac{1}{19} + \frac{1}{141} + \frac{1}{17} + \frac{1}{149}} = 0.354.$$

Por lo tanto, un intervalo de confianza al nivel 0.95 para  $\ln \theta$  es

$$0.166 \pm 1.96 * 0.354,$$

de modo que el intervalo

$$(e^{-0.527}, e^{0.860}) = (0.5902, 2.3633)$$

es un intervalo de confianza aproximado al nivel 0.95 para  $\theta$ . Dado que este intervalo contiene al valor  $\theta = 1$  de independencia, los datos no son significativos para concluir que los acusados de raza blanca y los de raza negra tienen distinta probabilidad de recibir pena de muerte.

La  $Q$  de Yule muestral se obtiene como

$$\hat{Q} = \frac{19 * 149 - 141 * 17}{19 * 149 + 141 * 17} = 0.00830.$$

Para construir un intervalo de confianza sobre  $Q$  al nivel 0.95 estimamos su error estándar asintótico  $\hat{\sigma}(\hat{Q}) = 0.17578$ . Entonces el intervalo de confianza es

$$0.00830 \pm 1.96 * 0.17578,$$

obteniendo finalmente el intervalo  $(-0.2615, 0.4275)$  que contiene al valor  $Q = 0$  de independencia.

### 5.3 Medidas basadas en el estadístico $X^2$ de Pearson

El valor de  $X^2$  no puede ser tomado directamente como medida de asociación porque su valor se hace imprevisiblemente grande cuando aumenta el tamaño muestral. Para corregir el efecto del tamaño muestral se consideran las siguientes medidas de asociación:

1. Medida phi cuadrado de Pearson:

$$\Phi^2 = \frac{X^2}{n}$$

2. Coeficiente de contingencia:

$$C^2 = \frac{X^2}{X^2 + n}$$

3. T de Tschuprow

$$T^2 = \frac{X^2}{n\sqrt{(I-1)(J-1)}}$$

4. V de Cramer

$$V^2 = \frac{X^2}{n \min\{(I-1)(J-1)\}}$$

Estas medidas toman valores entre 0 y 1, siendo preferible la V de Cramer porque alcanza el valor 1 cuando hay asociación perfecta cualquiera que sea el número de filas y columnas.

### 5.4 Medida Gamma de Goodman y Kruskal para variables ordinales

Consideremos un par de individuos o de observaciones clasificados según dos variables cualitativas ordinales. En relación al orden de estos dos individuos en cada una de las variables, el par se puede clasificar como *concordante* o *discordante*. Un par de observaciones es concordante si aquella que se ordena más alta sobre la variable A también se ordena más alta sobre B. Un par es discordante si la observación que se ordena más alta sobre A se ordena más baja sobre B. No todos los  $\binom{n}{2}$  pares de observaciones son concordantes o discordantes. Los pares que no son ni concordantes ni discordantes se llaman *ligados*. Un par de observaciones es *ligado* si ambas tienen la misma clasificación en alguna de las variables o en ambas.

Basándose en las probabilidades de concordancia y discordancia asociadas a una tabla de contingencia, se definen medidas de asociación para variables ordinales, análogas al coeficiente de correlación lineal para variables continuas, que describen el grado de monotonía de la asociación, es decir, si una variable tiende a aumentar cuando lo hace la otra o al revés.

Seleccionados al azar un par de observaciones o de individuos, la probabilidad de concordancia en la tabla es

$$P_c = 2 \sum_i \sum_j p_{ij} \left( \sum_{k>i} \sum_{l>j} p_{kl} \right)$$

y la de discordancia

$$P_d = 2 \sum_i \sum_j p_{ij} \left( \sum_{k>i} \sum_{l<j} p_{kl} \right).$$

El factor 2 se debe a que la primera observación podría estar en la casilla (i,j) y la segunda en la casilla (k,l) o viceversa.

Por otro lado el número total de pares concordantes en una tabla  $I \times J$  viene dado por:

$$C = \sum_i \sum_j n_{ij} \left( \sum_{k>i} \sum_{l>j} n_{kl} \right)$$

Análogamente, el número total de pares discordantes es:

$$D = \sum_i \sum_j n_{ij} \left( \sum_{k>i} \sum_{l<j} n_{kl} \right)$$

Se dirá que existe *asociación positiva* si  $P_c - P_d > 0$ , es decir valores bajos de A tienden a ocurrir con valores bajos de B y valores altos de A con valores altos de B. Análogamente, existe *asociación negativa* si  $P_c - P_d < 0$  lo que significa que valores altos de A tienden a ocurrir con valores bajos de B y viceversa.

La medida Gamma fue propuesta por Goodman y Kruskal en 1954 y su valor poblacional se define como:

$$\gamma = \frac{P_c - P_d}{P_c + P_d} \quad (10)$$

que es la diferencia entre las probabilidades de concordancia y discordancia para aquellos pares no ligados por alguna variable.

Su rango de variación es  $-1 \leq \gamma \leq 1$ . Si las variables son independientes, entonces,  $P_c = P_d$  lo que implica que  $\gamma = 0$ . El recíproco no es cierto.

	D	M	A	Total
EP	209	101	237	547
EM	151	126	426	703
EU	16	21	138	175
Total	376	248	801	1425

Tabla 14: Asociación opinión respecto a las drogas-nivel de estudios

La asociación positiva es tanto mayor cuanto más se aproxima  $\gamma$  a 1 y tanto menor cuanto más se aproxima a cero. A medida que se acerca a -1, aumenta la asociación negativa.

La estimación MV de  $\gamma$  es de la forma:

$$\hat{\gamma} = \frac{C - D}{C + D} \quad (11)$$

que recibe el nombre de *Gamma de Goodman y Kruskal muestral*.

**Ejemplo 8** Con el fin de estudiar la relación entre la opinión sobre la legalización de las drogas y el nivel de estudios se dispone de los datos de la Tabla 14. La opinión respecto a la legalización de las drogas se clasifica en tres categorías: D (desaprobación), M (Aprobación moderada) y A (Aprobación general). El nivel de estudios se clasifica también en tres niveles: EP (Estudios Primarios), EM (Estudios medios) y EU (Estudios universitarios). Nos proponemos estudiar la asociación entre la actitud respecto a legalizar las drogas y el nivel de estudios mediante los cocientes de ventajas locales que tienen en cuenta su naturaleza ordinal.

En primer lugar contrastaremos la independencia mediante los contrastes chi-cuadrado y de razón de verosimilitudes que pueden ser aplicados ya que la mínima frecuencia esperada bajo independencia es 30.46. Los valores observados de estos estadísticos son  $X_{Obs}^2 = 93.034$  y  $G_{Obs}^2 = 96.527$ . Para ambos sus p-valores que corresponden a la probabilidad acumulada a su derecha por una  $\chi_4^2$  son exactamente cero. Esto lleva a rechazar la hipótesis nula de independencia con cualquier nivel de significación.

Los valores estimados de las medidas de asociación basadas en  $X^2$  son

$$\hat{\phi} = 0.256; \quad \hat{C} = 0.248; \quad \hat{V} = 0.181.$$

Observemos que el grado de asociación global estimado con estas medidas es muy pequeño.



Como ambas variables son ordinales vamos a estimar la asociación mediante la medida gamma. El número de pares concordantes es el siguiente:

$$\begin{aligned} C &= 209(126 + 426 + 21 + 138) + 101(426 + 138) \\ &+ 151(21 + 138) + 126 \times 138 = 246960. \end{aligned}$$

Análogamente, el número de pares discordantes es

$$\begin{aligned} D &= 101(151 + 16) + 237(151 + 126 + 16 + 21) \\ &+ 126 \times 16 + 426(16 + 21) = 109063. \end{aligned}$$

Como consecuencia el valor muestral de la medida gamma es  $\hat{\gamma} = 0.3873$ . Su error estándar estimado  $\hat{\sigma}(\hat{\gamma}) = 0.037$ , permite construir el intervalo de confianza al nivel 0.95 dado por (0.31448, 0.45952). Observemos que el valor estimado es muy preciso dado que el intervalo de confianza es estrecho de modo que podemos concluir, en base al signo positivo de esta medida, que existe cierto grado de monotonía creciente en la asociación. Esto significa que la probabilidad de estar a favor del aborto aumenta con el nivel de estudios.

## 5.5 Medida kappa de acuerdo

Un problema de gran interés en la práctica es medir el acuerdo entre dos observadores que clasifican independientemente una muestra de individuos sobre la misma escala categórica. Por ejemplo, cuando dos psicólogos clasifican a una muestra de personas según su salud mental usando la escala: normal, trastornada o enajenada. En estos casos las clasificaciones conjuntas de ambos observadores se pueden representar en una tabla cuadrada con las mismas categorías en cada dimensión, cuya diagonal principal representa los casos en los que hay acuerdo.

Antes de pasar a definir formalmente medidas de acuerdo, es necesario distinguir entre acuerdo y asociación. Para que las clasificaciones de los dos observadores estén de acuerdo deben caer en categorías idénticas mientras que para que dos respuestas estén perfectamente asociadas solo se requiere que se pueda predecir la clasificación de un observador a partir de la del otro. Por lo tanto una tabla en la que hay asociación alta puede haber acuerdo alto o bajo. En el caso de una escala ordinal, si uno de los observadores clasifica sistemáticamente a los individuos un nivel más alto que el otro, el acuerdo es muy débil mientras que la asociación es muy fuerte.

Para medir el grado de acuerdo se estudiará la medida *kappa* introducida por Cohen en 1960, como una función de la diferencia entre la probabilidad de que los dos observadores estén de acuerdo y la probabilidad de acuerdo si sus clasificaciones conjuntas fuesen independientes.

$\kappa$	Dimensión del acuerdo
0	Pobre
0.00-0.20	Pequeño
0.21-0.40	Mediano
0.41-0.60	Moderado
0.61-0.80	Importante
0.81-1.00	Casi perfecto

Tabla 15: Reglas para evaluar el grado de acuerdo

Si denotamos por  $p_a = \sum_i p_{ii}$  a la probabilidad de acuerdo y por  $p_a^I = \sum_i p_{i.} p_{.i}$  a la probabilidad de acuerdo bajo independencia (probabilidad de acuerdo esperado o por azar), el valor poblacional de la medida kappa es

$$k = \frac{p_a - p_a^I}{1 - p_a^I}.$$

La independencia implica claramente que  $k = 0$ . El valor cero de esta medida implica que el acuerdo es igual al esperado por azar (ausencia de acuerdo). Si hay acuerdo perfecto,  $\sum_i p_{ii} = 1$ , entonces  $k = 1$ . Podría tomar valores negativos raras veces cuando el acuerdo es menor que el debido al azar.

El valor muestral de kappa es

$$\hat{k} = \frac{n \sum_i n_{ii} - \sum_i n_{i.} n_{.i}}{n^2 - \sum_i n_{i.} n_{.i}}.$$

En la práctica raramente ocurre que el acuerdo no sea mayor que el esperado por azar. Por ello será más importante construir un intervalo de confianza para medir la longitud del acuerdo que contrastar si el acuerdo es cero.

En Everitt (1992) se recogen algunas normas subjetivas pero muy útiles para interpretar el grado de acuerdo en base al valor de la medida kappa, que aparecen en la Tabla 15.

**Ejemplo 9** *Estimar el grado de acuerdo entre la clase social de padres e hijos a partir de la Tabla 11.*

El valor muestral de la medida kappa de acuerdo es

$$\hat{k} = \frac{3497(588 + 714 + 411) - (1142 \times 1051 + 1510 \times 1429 + 845 \times 1017)}{3497^2 - (1142 \times 1051 + 1510 \times 1429 + 845 \times 1017)} = 0.221,$$

con desviación estándar  $\hat{\sigma}(\hat{\kappa}) = 0.013$ .

Como consecuencia un I.C. aproximado del 95% se calcula como

$$0.221 \pm 0.013 \times 1.96,$$

y da como resultado (0.19552, 0.24648). La interpretación es que el grado de acuerdo es mediano como cabía esperar si se tiene en cuenta que se rechazó tanto la simetría como la homogeneidad marginal.



# Bibliografía

- Abad Montes, F. y M. Vargas Jiménez (2002). *Análisis de Datos para las Ciencias Sociales con SPSS*. Urbano, Granada.
- Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. Wiley, New York.
- Aguilera, A. M. (2001). *Tablas de contingencia bidimensionales*. Hespérides-La Muralla, Salamanca.
- Everitt, B. (1992). *The Analysis of Contingency Tables*. Chapman Hall, London.
- Ocaña, F., M. Valderrama, A. Aguilera y R. Gutiérrez Jáimez (1995). Repercusión económica en granada de los estudiantes universitarios foráneos. *Ars Pharmaceutica*, **36**, 59–71.
- Ruiz-Maya, L., F. J. Martín-Pliego, J. M. Montero y P. Uriz Tomé (1995). *Análisis Estadístico de Encuestas: Datos Cualitativos*. AC, Madrid.