Tema 2: Tablas de Contingencia

Introducción

Una *tabla de contingencia* es una de las formas más comunes de resumir datos categóricos. En general, el interés se centra en estudiar si existe alguna asociación entre una variable denominada *fila* y otra variable denominada *columna* y se calcula la intensidad de dicha asociación.

De manera formal, se consideran X e Y dos variables categóricas con I y J categorías respectivamente. Una observación puede venir clasificada en una de las posibles $I \times J$ categorías que existen.

Cuando las casillas de la tabla contienen las frecuencias observadas, la tabla se denomina **tabla de contingencia**, término que fue introducido por Pearson en 1904.

Una tabla de contingencia (o tabla de clasificación cruzada), con I filas y J columnas se denomina una tabla $I \times J$.

Ejemplo

Por ejemplo, se considera la distribución conjunta de dos variables y la correspondiente tabla de contingencia en una muestra de pacientes de un hospital. Se tiene la siguiente tabla donde se consideran el riesgo de ataque al corazón respecto a la toma de aspirinas:

- $X \equiv$ Se toma aspirina o placebo ($I = 2$).
- $Y \equiv$ Se sufre ataque cardíaco o no ($J = 3$).

	Mortal	**No mortal**	**No ataque**	**Totales**
Placebo	18	171	10845	11034
Aspirina	5	99	10933	11037

Como resumen de la información que presenta la tabla, de los 11034 enfermos que tomaron un placebo, 18 tuvieron un ataque al corazón, mientras que de los 11037 que tomaron aspirina, 5 tuvieron ataques al corazón.

La distribución conjunta de dos variables categóricas determina su relación. Esta distribución también determina las distribuciones marginales y condicionales.

Distribución conjunta

La distribución conjunta viene dada por

$$\pi_{ij} = P\left(X = i, Y = j\right)$$

con $i = 1, \ldots, I$ y $j = 1, , \ldots, J$.

Es la probabilidad de (X, Y) en la casilla de la fila i y la columna j.

Distribución marginal

Las distribuciones marginales son

$$\pi_{i+} = P\left(X = i\right) = \sum_{j=1}^{J} P\left(X = i, Y = j\right) = \sum_{j=1}^{J} \pi_{ij}$$

$$\pi_{+j} = P\left(Y = j\right) = \sum_{i=1}^{I} P\left(X = i, Y = j\right) = \sum_{i=1}^{I} \pi_{ij}$$

es decir, el símbolo + indica la suma de las casillas correspondientes a un índice dado.

Se cumple siempre que

$$\sum_{j} \pi_{+j} = \sum_{i} \pi_{i+} = \sum_{i} \sum_{j} \pi_{ij} = 1$$

Distribución condicional

En muchas ocasiones en las tablas de contingencia, como en el ejemplo anterior, una de las variables, digamos Y, es una variable respuesta y la otra variable X es una variable explicativa o predictora. En esta situación no tiene sentido hablar de distribución conjunta.

Cuando se considera una categoría fija de X, entonces Y tiene una distribución de probabilidad que se expresa como una probabilidad condicionada.

Así, se puede estudiar el cambio de esta distribución cuando van cambiando los valores de X.

Distribución condicionada de Y respecto de X

$$P\left(Y=j|X=i\right)=\pi_{j|i}=\frac{\pi_{ij}}{\pi_{i+}}$$

Se tiene que

$$\sum_j \pi_{j|i}=1$$

y el vector de probabilidades $\left(\pi_{1|i},\ldots,\pi_{J|i}\right)$ forman la distribución condicionada de Y en la categoría i de X.

La mayor parte de los estudios se centran en la comparación de las distribuciones condicionadas de Y para varios niveles de las variables explicativas X.

Independencia y Homogeneidad

Cuando las variables que se consideran son de tipo respuesta, se pueden usar distribuciones conjuntas o bien distribuciones condicionales para describir la asociación entre ellas.

Dos variables son **independientes** si

$$\pi_{ij}=\pi_{i+}\cdot\pi_{+j}$$

lo cual implica que la distribución condicionada es igual a la marginal:

$$\pi_{j|i}=\pi_{+j}$$

para $j=1,\ldots,J$, dado que

$$\pi_{j|i}=\frac{\pi_{ij}}{\pi_{i+}}$$

para todo i y j.

Si X e Y son variables respuesta entonces se habla de *independencia*

Si Y es variable respuesta y X es variable explicativa entonces se habla de *homogeneidad*.

Ejemplo con SAS

Muchas veces, los datos categóricos se presentan en forma de tablas de contingencia como la anterior. Supongamos, por ejemplo:

Tratamiento	*Favorable*	*Desfavorable*
Placebo	16	48
Test	40	20

En SAS el modo de introducir esta tabla sería:

```
/* OPTIONS nodate ls=75 formchar='|----|+|---+=|-/\<>*'; */
OPTIONS nodate ls=75;

/* Para SAS University */
ODS rtf file='/folders/myfolders/resultado.rtf' style=minimal
    startpage=no;
DATA respira;
INPUT treat $ outcome $ count;
DATALINES;
placebo f 16
placebo u 48
test    f 40
test    u 20
;

PROC freq;
weight count;
tables treat*outcome;
RUN;
ODS rtf close;
```

Procedimiento FREQ

Frecuencia Porcentaje Pct fila Pct col	Tabla de treat por outcome		
	outcome		
treat	f	u	Total
placebo	16 12.90 25.00 28.57	48 38.71 75.00 70.59	64 51.61
test	40 32.26 66.67 71.43	20 16.13 33.33 29.41	60 48.39
Total	56 45.16	68 54.84	124 100.00

Estos datos también se podrían presentar también en forma de matriz de datos, donde cada individuo esté representado por una fila con valores en distintas variables.

En ese caso el programa en SAS sería semejante al anterior, pero NO se tendría que usar el comando `weight` ya que los datos se presentarían en forma de dos columnas con las dos variables.

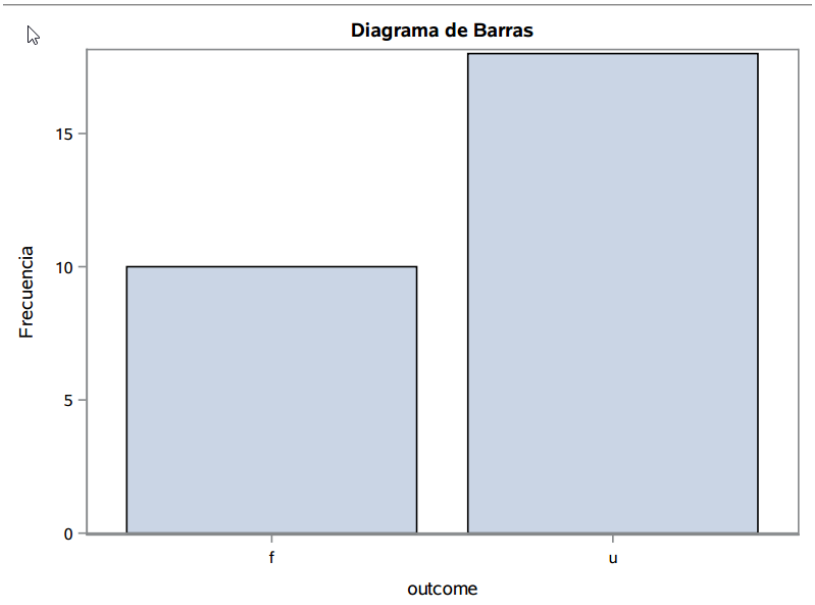
Ejemplo sobre otro tratamiento clínico con dos variables: tipo de tratamiento y tipo de respuesta. El siguiente programa funciona también con *SAS University*.

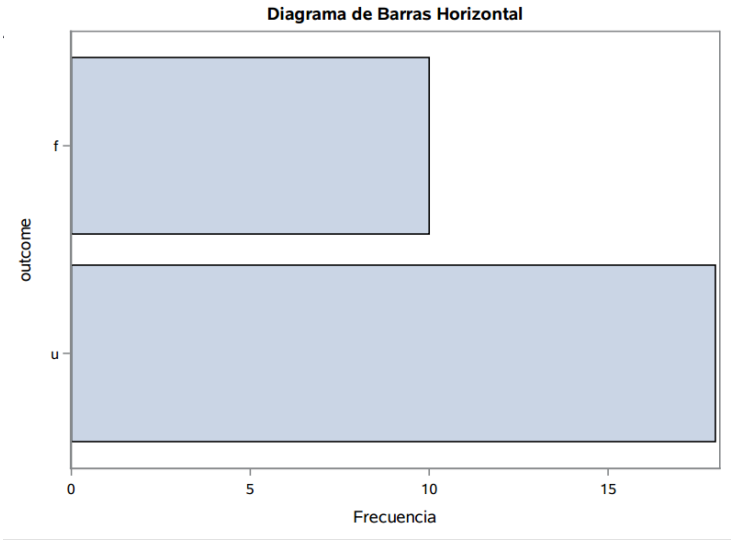
```
OPTIONS nodate ls=75;
/* Para SAS University */
ODS rtf file='/folders/myfolders/resultado.rtf' style=minimal
    startpage=no;

DATA respira;
INPUT treat $ outcome $ @@;
DATALINES;
placebo f placebo f placebo f
placebo f placebo f
placebo u placebo u placebo u
placebo u placebo u placebo u
placebo u
test f test f test f
test f test f
test u test u test u
```

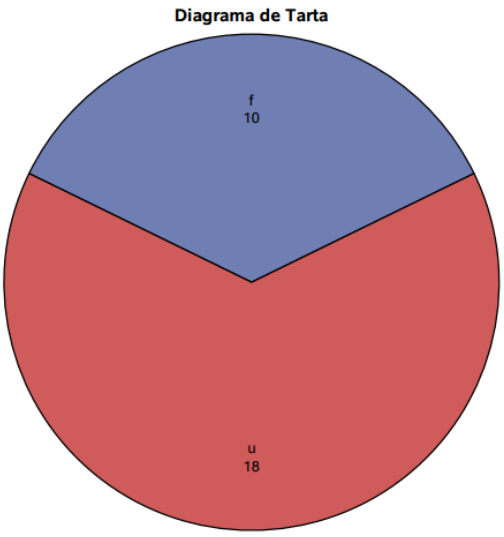
```
test u test u test u
test u test u test u
test u test u
;

PROC freq;
tables treat*outcome;
RUN;

PROC SGPLOT DATA=respira;
TITLE 'Diagrama de Barras';
      VBAR outcome;
RUN;

PROC SGPLOT DATA=respira;
TITLE 'Diagrama de Barras Horizontal';
      HBAR outcome;
RUN;
ODS rtf close;
```

Procedimiento FREQ

Frecuencia Porcentaje Pct fila Pct col	Tabla de treat por outcome		
	outcome		
treat	f	u	Total
placebo	5 17.86 41.67 50.00	7 25.00 58.33 38.89	12 42.86
test	5 17.86 31.25 50.00	11 39.29 68.75 61.11	16 57.14
Total	10 35.71	18 64.29	28 100.00







El mismo programa se puede escribir en SAS estándar:

```
/*  Opcion SAS estandar */
 OPTIONS nodate ls=75 formchar='|----|+|---+=|-/\<>*';
/* Fijo el directorio de trabajo */
 x 'cd "e:\Categoricos"';
/* Se graban los resultados en un fichero rtf o en uno pdf */
/* ODS pdf file='cosa.pdf' style=minimal startpage=no; */
   ODS rtf file='cosa.rtf' style=minimal startpage=no;

DATA respira;
INPUT treat $ outcome $ @@;
DATALINES;
placebo f placebo f placebo f
placebo f placebo f
placebo u placebo u placebo u
placebo u placebo u placebo u
placebo u
test f test f test f
test f test f
test u test u test u
test u test u test u
test u test u test u
test u test u
;

PROC freq;
tables treat*outcome;
RUN;

PROC GCHART DATA=respira;
      VBAR outcome;
RUN;

PROC GCHART DATA=respira;
      HBAR outcome/ DISCRETE;
RUN;

PROC GCHART DATA=respira;
      PIE outcome/ DISCRETE VALUE=INSIDE
            PERCENT=INSIDE SLICE=OUTSIDE;
RUN;

 ODS rtf close;
/* ODS pdf close; */
```

Cuando se consideran datos **ordinales**, es importante asegurase de que los niveles de las filas y columnas se ordenen correctamente, ya que en SAS los datos se ordenan de forma *alfanumérica* por defecto. Para ello se usa el comando `order=data`.

Por ejemplo, supongamos la siguiente tabla con tres variables:

		Mejora		
Sexo	**Tratamiento**	**Alta**	**Escasa**	**Ninguna**
Mujer	*Activo*	16	5	6
Mujer	*Placebo*	6	7	19
Hombre	*Activo*	5	2	7
Hombre	*Placebo*	1	0	10

El programa en SAS para introducir la tabla anterior es

```
OPTIONS nodate ls=75;
/* Para SAS University */
ODS rtf file='/folders/myfolders/resultado.rtf' style=minimal
    startpage=no;

DATA artritis;
INPUT genero $ tratamiento $ mejora $ recuento @@;
DATALINES;
mujer activo alta 16 mujer activo escasa 5 mujer activo
   ninguno 6 mujer placebo activo 6 mujer placebo escasa 7
   mujer placebo ninguno 19 hombre activo activo 5 hombre
   activo escasa 2 hombre activo ninguno 7 hombre placebo
   activo 1 hombre placebo escasa 0 hombre placebo ninguno 10
;
RUN;

PROC freq order=data;
weight recuento;
tables genero*tratamiento*mejora / nocol nopct;
RUN;
ODS rtf close;
```

Procedimiento FREQ

Frecuencia Pct fila	Tabla 1 de tratamiento por mejora Control para genero=mujer				
	mejora				
tratamiento	alta	escasa	ninguno	activo	Total
activo	16 59.26	5 18.52	6 22.22	0 0.00	27
placebo	0 0.00	7 21.88	19 59.38	6 18.75	32
Total	16	12	25	6	59

Frecuencia Pct fila	Tabla 2 de tratamiento por mejora Control para genero=hombre				
	mejora				
tratamiento	alta	escasa	ninguno	activo	Total
activo	0 0.00	2 14.29	7 50.00	5 35.71	14
placebo	0 0.00	0 0.00	10 90.91	1 9.09	11
Total	0	2	17	6	25

El mismo programa se puede escribir en SAS estándar incluyendo varios gráficos con `GCHART`:

```
/* Opcion SAS estandar */
 OPTIONS nodate ls=75 formchar='|----|+|---+=|-/\<>*';
/* Fijo el directorio de trabajo */
 x 'cd "cd "e:\Categoricos"';
/* Se graban los resultados en un fichero rtf o en uno pdf */
/* ODS pdf file='cosa.pdf' style=minimal startpage=no; */
 ODS rtf file='cosa.rtf' style=minimal startpage=no;

DATA artritis;
INPUT genero $ tratamiento $ mejora $ recuento @@;
DATALINES;
mujer activo activo 16 mujer activo escasa 5 mujer activo
ninguno 6 mujer placebo activo 6 mujer placebo escasa 7 mujer
placebo ninguno 19 hombre activo activo 5 hombre activo escasa
2 hombre activo ninguno 7 hombre placebo activo 1
hombre placebo escasa 0 hombre placebo ninguno 10
;
RUN;

PROC freq order=data;
weight recuento;
tables genero*tratamiento*mejora / nocol nopct;
RUN;

PROC GCHART DATA=artritis;
pie3d mejora / sumvar=recuento;
RUN;

PROC GCHART DATA=artritis;
hbar3d mejora / sumvar=recuento patternid=midpoint
   group=genero;
RUN;

 ODS rtf close;
/* ODS pdf close; */
```

Ejemplo con R

```
datos = c(16, 40, 48, 20)

tabla = cbind(expand.grid(list(Tratamiento=c("Placebo","Test"),
Situacion=c("Favor","Desfavor"))), count=datos)

# Opcion simple
ftable(xtabs(count ~ Tratamiento + Situacion, tabla))
```

```
            Situacion Favor Desfavor
Tratamiento
Placebo                16       48
Test                   40       20
```

```
# Opcion con salida de estilo SAS
library(gmodels)
CrossTable(xtabs(count ~ Tratamiento + Situacion, tabla),
expected=TRUE, format="SAS")
```

```
   Cell Contents
|-------------------------|
|                       N |
|               Expected N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  124


             | Situacion
 Tratamiento |     Favor |  Desfavor | Row Total |
-------------|-----------|-----------|-----------|
     Placebo |        16 |        48 |        64 |
             |    28.903 |    35.097 |           |
             |     5.760 |     4.744 |           |
             |     0.250 |     0.750 |     0.516 |
             |     0.286 |     0.706 |           |
             |     0.129 |     0.387 |           |
-------------|-----------|-----------|-----------|
        Test |        40 |        20 |        60 |
             |    27.097 |    32.903 |           |
             |     6.144 |     5.060 |           |
             |     0.667 |     0.333 |     0.484 |
             |     0.714 |     0.294 |           |
             |     0.323 |     0.161 |           |
-------------|-----------|-----------|-----------|
Column Total |        56 |        68 |       124 |
             |     0.452 |     0.548 |           |
-------------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  21.70868     d.f. =  1     p =  3.173515e-06

Pearson's Chi-squared test with Yates' continuity correction
------------------------------------------------------------
Chi^2 =  20.05886     d.f. =  1     p =  7.509491e-06
```

Distribución multinomial en tablas 2×2

Cuando se consideran tablas de contingencia, es habitual asumir que los recuentos de las casillas en las tablas se distribuyen como una multinomial.

En el muestreo multinomial, fijamos el tamaño total n pero no los totales de filas y columnas. Es decir, solo el tamaño muestral está fijado previamente en el experimento.

Así, si se tienen observaciones en $I \times J$ casillas, la distribución de probabilidad de los recuentos es

$$\frac{n!}{n_{11}! \cdots n_{IJ}!} \prod_i \prod_j \pi_{ij}^{n_{ij}}.$$

En otros casos, las observaciones de una variable respuesta Y aparecen de manera separada según el nivel de una variable explicativa X. En este caso se consideran los totales por filas como *fijos*.

De este modo, se simplifica la notación como $n_{i+} = n_i$ y suponemos que dado un nivel fijo de i de X, las n_i observaciones de Y son independientes entre sí y con distribución de probabilidad $\left\{\pi_{1|i}, \ldots, \pi_{J|i}\right\}$.

Los recuentos $\{n_{ij}, \quad j = 1, \ldots, J\}$ tal que $\sum_j n_{ij} = n_i$ se distribuyen como

$$\frac{n_i!}{\prod_j n_{ij}!} \prod_j \pi_{j|i}^{n_{ij}}. \tag{1}$$

Así, cuando las muestras que se toman en diferentes niveles de X son independientes, la distribución conjunta de todos los datos es el producto de distribuciones multinomiales (1) para cada nivel i de X.

Este esquema se denomina muestreo multinomial independiente o muestreo de productos de multinomiales.

Ejemplos

Consideremos el estudio del número de accidentes mortales y no mortales, con cinturón y sin cinturón.

Asumimos un muestreo multinomial: Tomamos un muestra aleatoria de 200 accidentes que tuvieron lugar el mes pasado y fijamos el tamaño total de la muestra.

Muestreo multinomial **independiente**: Tomamos una muestra de 100 accidentes donde hubo muertos y otros 100 en los que no los hubo. Fijamos los totales por columna (en este caso es un muestreo *binomial* porque hay solo dos categorías).

Comparación de proporciones en tablas 2×2

Muchos estudios se diseñan para comparar grupos basándonos en una respuesta Y binaria. Con dos grupos tenemos una tabla de contingencia 2×2.

	Exito	**Fracaso**
Grupo 1	$\pi_{1\|1}$	$\pi_{2\|1}$
Grupo 2	$\pi_{1\|2}$	$\pi_{2\|2}$

Se denota

$$\begin{aligned} \pi_{1|i} &= \pi_i \\ \pi_{2|i} &= 1 - \pi_{1|i} = 1 - \pi_i \end{aligned}$$

de modo que la tabla se puede reescribir como

	Exito	**Fracaso**
Grupo 1	π_1	$1 - \pi_1$
Grupo 2	π_2	$1 - \pi_2$

Se quiere comparar π_1 con π_2. Para ello, se puede estudiar,

(i) La diferencia de las proporciones

$$\pi_1 - \pi_2$$

(ii) El riesgo relativo

$$\frac{\pi_1}{\pi_2}$$

(iii) La razón de plausibilidad *odds*:

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

Ejemplo:

Influencia de la toma de aspirina respecto a los ataques cardíacos:

	Ataque	No ataque
Placebo	189	10845
Aspirina	104	10933

Para contrastar $H_0 : p_a = p_p$ (igual probabilidades de ataque al corazón por grupo), se puede usar el comando `prop.test`.

Para contrastar una hipótesis unilateral, $H_0 : p_a \geq p_p$ frente a $H_1 : p_a < p_p$ se hace usando la opción `alternative`.

```
x = c(104, 189) # aspirina y placebo
n = c((104+10933), (189+10845))
prop.test(x, n)
```

```
2-sample test for equality of proportions with continuity correction

data:  x out of n
X-squared = 24.429, df = 1, p-value = 7.71e-07
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.010814914 -0.004597134
sample estimates:
    prop 1     prop 2
0.00942285 0.01712887
```

```
prop.test(x, n, alt="less")
```

```
2-sample test for equality of proportions with continuity correction

data:  x out of n
X-squared = 24.429, df = 1, p-value = 3.855e-07
alternative hypothesis: less
95 percent confidence interval:
 -1.000000000 -0.005082393
sample estimates:
    prop 1     prop 2
0.00942285 0.01712887
```

Se pueden obtener las proporciones a partir del componente `estimate` que en este caso es un vector numérico de longitud 2.

Así, la diferencia de las proporciones se calcula como:

```
temp = prop.test(x,n)
names(temp$estimate) = NULL
# Diferencia de las proporciones
temp$estimate[1]-temp$estimate[2]
```

```
[1] -0.007698806
```

Se puede calcular también el riesgo relativo y la razón de odds:

```
# Riesgo relativo
temp$estimate[2]/temp$estimate[1]
```

```
[1] 1.816814
```

```
# Razon de odds
x[2]*(n[1]-x[1])/(x[1]*(n[2]-x[2]))
```

```
[1] 1.831045
```

Para programar las razones de odds en SAS, se usa:

```
/* OPTIONS nodate ls=65 formchar='|----|+|---+=|-/\<>*'; */
OPTIONS nodate ls=75;
/* Para SAS University */
ODS rtf file='/folders/myfolders/resultado.rtf' style=minimal
    startpage=no;

DATA riesgos;
INPUT ataque $ medica $ cuenta;
DATALINES;
ataque placebo 189
ataque aspirina 104
NOataque placebo 10845
NOataque aspirina 10933
;

PROC freq order=data;
weight cuenta;
```

```
tables ataque*medica / nocol;
exact or;
RUN;
ODS rtf close;
```

Se obtiene

Procedimiento FREQ

Frecuencia Porcentaje Pct fila	Tabla de ataque por medica		
	medica		
ataque	placebo	aspirina	Total
ataque	189 0.86 64.51	104 0.47 35.49	293 1.33
NOataque	10845 49.14 49.80	10933 49.54 50.20	21778 98.67
Total	11034 49.99	11037 50.01	22071 100.00

Estadísticos para la tabla de ataque por medica

Ratio de probabilidades y riesgos relativos			
Estadístico	Valor	Límites de confianza al 95%	
Ratio de probabilidad	1.8321	1.4400	2.3308
Riesgo relativo (Columna 1)	1.2953	1.1886	1.4116
Riesgo relativo (Columna 2)	0.7070	0.6056	0.8255

Ratio de probabilidad	
Ratio de probabilidad	1.8321
Lím. de confianza asintóticos	
95% Límite conf. inferior	1.4400
95% Límite conf. superior	2.3308
Límites conf. exactos	
95% Límite conf. inferior	1.4323
95% Límite conf. superior	2.3539

Tamaño de la muestra = 22071

Odds y razón de odds

Si π es la probabilidad de éxito entonces los *odds* se definen como

$$\Omega = \frac{\pi}{1-\pi}$$

o de modo equivalente

$$\pi = \frac{\Omega}{\Omega + 1}.$$

Se tiene que $\Omega > 1$ cuando un éxito es más probable que un fallo.

Por ejemplo, cuando $\pi = 0{,}75$, entonces

$$\Omega = \frac{0{,}75}{0{,}25} = 3$$

es decir un éxito es tres veces más probable que un fallo.

Si se tiene una tabla 2×2 se pueden definir los odds en la fila i:

$$\Omega_i = \frac{\pi_i}{1-\pi_i}.$$

El cociente de los odds de las dos filas se denomina *razón de odds*:

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1 / (1-\pi_1)}{\pi_2 / (1-\pi_2)}$$

y se obtiene de manera equivalente, cuando se tiene distribuciones conjuntas, π_{ij} que

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

por lo que también se denomina cociente de los *productos cruzados*.

Propiedades

- Puede ser cualquier valor positivo.
- $\theta = 1$ significa que NO hay asociación entre X e Y.
- Valores de θ alejados de 1 indican una asociación mayor.
- Se suele trabajar con $\log \theta$ ya que el valor que se obtiene es simétrico respecto a cero.
- La razón de odds no cambia cuando se intercambian filas y columnas.

Razón de odds condicionales y marginales

Las asociaciones marginales y condicionales pueden ser descritas mediante la razón de odds.

Supongamos una tabla $2 \times 2 \times K$, si denominamos μ_{ijk} a la frecuencia esperada en la celda correspondiente.

Fijamos $Z = k$, y se define la razón de odds condicional como

$$\theta_{XY(k)} = \frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}}$$

y la razón de odds marginal como

$$\theta_{XY} = \frac{\mu_{11+}\mu_{22+}}{\mu_{12+}\mu_{21+}}$$

Un valor a 1 en la razón de odds supone, o bien independencia marginal, o bien condicionada a que $Z = k$, es decir, cuando $\theta_{XY(k)} = 1$.

NOTA:

La **independencia condicional** cuando $Z = k$ es equivalente a que

$$P\left(Y = j|X = i, Z = k\right) = P\left(Y = j|Z = k\right)$$

para todo i, j.

Si se cumple para todo valor de la variable Z, entonces se dice que X e Y son condicionalmente independientes dado Z y se obtiene que

$$\pi_{ijk} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}}$$

para cualquier i, j, k.

La independencia condicional **no** implica la independencia marginal.

Ejemplo

Consideramos un ejemplo famoso en USA sobre procesamientos por asesinatos múltiples en Florida entre los años 1976 y 1987:

Raza Victima	Raza Acusado	Pena de Muerte		Porcentaje
		Si	No	Si
Blanca	*Blanco*	53	414	11.3
	Negro	11	37	22.9
Negro	*Blanco*	0	16	0
	Negro	4	139	2.8
Total	*Blanco*	53	430	11
	Negro	15	176	7.9

Esta es una tabla de contingencia $2 \times 2 \times 2$ y el ejemplo de la pena de muerte sirve para ilustrar las razones de *odds condicionales*. Se estudia el efecto de la raza del acusado (X) en el veredicto de culpabilidad (Y), tratando a la raza de la víctima (Z) como si fuera una variable control, es decir fijándola en valor dado.

```
vic.raza = c("blanca", "negra")
def.raza = vic.raza
pena.muerte = c("SI", "NO")

datalabel = list(acusado=def.raza,muerte=pena.muerte,
victima=vic.raza)
tabla = expand.grid(acusado=def.raza,muerte=pena.muerte,
victima=vic.raza)

data = c(53, 11, 414, 37, 0, 4, 16, 139)

tabla.Ori = cbind(tabla, recuento=data)
tabla = cbind(tabla, recuento=(data+0.5))
xtabs(recuento ~ acusado+muerte+victima, data=tabla.Ori)

temp = xtabs(recuento ~ acusado+muerte+victima, data=tabla)
apply(temp, 3, function(x) x[1,1]*x[2,2]/(x[2,1]*x[1,2]))
```

Se obtiene

```
, , victima = blanca

        muerte
acusado   SI  NO
  blanca  53 414
  negra   11  37

, , victima = negra

        muerte
acusado   SI  NO
  blanca   0  16
  negra    4 139
```

Se calculan la razón de odds condicionales.

```
temp = xtabs(recuento ~ acusado+muerte+victima, data=tabla)
apply(temp, 3, function(x) x[1,1]*x[2,2]/(x[2,1]*x[1,2]))
```

```
   blanca     negra
0.4208843 0.9393939
```

```
# Con el paquete vcd
library(vcd)
temp = xtabs(recuento ~ acusado+muerte+victima, data=tabla.Ori)
summary(oddsratio(temp, log=F, stratum=3))
```

```
z test of coefficients:

       Estimate Std. Error z value Pr(>|z|)
blanca  0.42088    0.15449  2.7244 0.006442 **
negra   0.93939    1.42156  0.6608 0.508728
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La función `oddsratio` añade 0.5 a cada casilla de la tabla por defecto y no hay que hacerlo *a mano*.

Asociación Homogénea

Una tabla $2 \times 2 \times K$ tiene una asociación homogénea en XY cuando

$$\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}$$

Es decir, cuando el tipo de asociación entre X e Y es el mismo para las distintas categorías de Z.

La independencia condicional entre X e Y es un caso particular donde $\theta_{XY(K)} = 1$.

Si existe una asociación XY homogénea entonces también tenemos una asociación XZ homogénea y una asociación YZ homogénea. Se dice también que no existe interacción entre las dos variables con respecto a sus efectos en la otra variable.

Cuando existe interacción, la razón de odds para cada par de variables **cambia** a lo largo de las categorías de la tercera variable.

Ejemplo

$X \equiv$ fumador (sí, no)

$Y \equiv$ cáncer de pulmón (sí, no)

Z = edad $(< 45, 45 - 65, > 65)$

Si las razones de odds observadas son

$$\begin{aligned} \theta_{XY(1)} &= 1{,}2 \\ \theta_{XY(2)} &= 3{,}9 \\ \theta_{XY(3)} &= 8{,}8 \end{aligned}$$

El efecto de fumar se acentúa conforme la edad es mayor. La edad se denomina *efecto modificador*, dado que el efecto de fumar queda modificado por la edad de las personas.

Inferencia en tablas de contingencia

Intervalos de confianza para parámetros de asociación

Intervalo para la razón de odds

El estimador que se utiliza para la razón de odds es

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

Este estimador puede ser 0 ó ∞ (si algún $n_{ij} = 0$) o no estar definido (0/0) dependiendo de los recuentos que se tengan.

Una posible opción es trabajar con el estimador *corregido*:

$$\widehat{\theta} = \frac{(n_{11} + 0{,}5)\,(n_{22} + 0{,}5)}{(n_{12} + 0{,}5)\,(n_{21} + 0{,}5)}$$

o bien con la transformación $\log(\widehat{\theta})$.

Una estimación del error estándar de $\log(\widehat{\theta})$ es

$$\widehat{\sigma}_{\log(\widehat{\theta})} = \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)^{\frac{1}{2}}$$

de modo que el correspondiente intervalo de Wald se calcula como

$$\log(\widehat{\theta}) \pm z_{\frac{\alpha}{2}} \widehat{\sigma}_{\log(\widehat{\theta})}$$

Si se toma la función exponencial (*antilogaritmo*) de los extremos, se obtiene el intervalo original correspondiente para θ.

El test es conservador ya que la probabilidad de cubrimiento es algo mayor que el nivel nominal $(1 - \alpha)$.

Intervalo de confianza para la diferencia de proporciones

Supongamos que tenemos muestras de binomiales independientes, de modo que en el grupo i tenemos $Y_i \sim \text{Bin}\,(n_i, \pi_i)$, así, el estimador es

$$\widehat{\pi}_i = \frac{Y_i}{n_i}$$

y la media y desviación estándar son

$$\begin{aligned} E\,(\widehat{\pi}_1 - \widehat{\pi}_2) &= \pi_1 - \pi_2 \\ \sigma_{\left(\widehat{\pi}_1 - \widehat{\pi}_2\right)} &= \left[\frac{\pi_1\,(1 - \pi_1)}{n_1} + \frac{\pi_2\,(1 - \pi_2)}{n_2}\right]^{1/2}, \end{aligned}$$

de modo que un estimador de la desviación estándar es

$$\widehat{\sigma}_{(\widehat{\pi}_1-\widehat{\pi}_2)} = \left[\frac{\widehat{\pi}_1\,(1-\widehat{\pi}_1)}{n_1} + \frac{\widehat{\pi}_2\,(1-\widehat{\pi}_2)}{n_2}\right]^{1/2}.$$

El intervalo de confianza de Wald es

$$\widehat{\pi}_1 - \widehat{\pi}_2 \pm z_{\frac{\alpha}{2}}\widehat{\sigma}_{(\widehat{\pi}_1-\widehat{\pi}_2)}$$

Cuando los valores de π_1 y π_2 están próximos a 0 ó 1, este intervalo tiene una probabilidad de cubrimiento menor que la teórica $(1-\alpha)$.

Intervalo de confianza para el riesgo relativo

El riesgo relativo muestral viene dado por

$$r = \frac{\widehat{\pi}_1}{\widehat{\pi}_2}$$

Se prefiere usar mejor el logaritmo ya que converge más rápido a la normal. El estimador del correspondiente error estándar es

$$\widehat{\sigma}_{\log(r)} = \left[\frac{1-\widehat{\pi}_1}{n_1\widehat{\pi}_1} + \frac{1-\widehat{\pi}_2}{n_2\widehat{\pi}_2}\right]^{1/2}$$

El intervalo de confianza de Wald para el logaritmo es

$$\log(r) \pm z_{\frac{\alpha}{2}}\widehat{\sigma}_{\log(r)}$$

Ejemplo

Por ejemplo, en la tabla de contingencia sobre el uso de la aspirina y el infarto de miocardio.

	Infarto Miocardio		
	SI	**NO**	**Total**
Placebo	28	656	684
Aspirina	18	658	656

Se crea la tabla en R:

```
medicina = c("Placebo", "Aspirina")
infarto = c("SI", "NO")
tabla = expand.grid(medicina=medicina, infarto=infarto)
cuentas = c(28, 18, 656, 658)
tabla = cbind(tabla, count=cuentas)

tablaguay = xtabs(cuentas ~ medicina + infarto, data=tabla)
```

Para calcular los intervalos, se puede usar la función de R escrita por Laura A. Thompson (*R (and S-PLUS) Manual to Accompany Agresti's Categorical Data Analysis* (2009)):

```
# Function from Laura A. Thompson
Wald.ci = function(Table, aff.response, alpha=.05){
    # Gives two-sided Wald CI's for odds ratio,
    # difference in proportions and relative risk.
    # Table is a 2x2 table of counts with rows giving
    # the treatment populations
    # aff.response is a string like "c(1,1)" giving the cell
    # of the beneficial response and the treatment category
    # alpha is significance level

    pow = function(x, a=-1) x^a
    z.alpha = qnorm(1-alpha/2)

    if(is.character(aff.response))
    where = eval(parse(text=aff.response))
    else where = aff.response
    Next = as.numeric(where==1) + 1

    # OR
    odds.ratio = Table[where[1],where[2]]*Table[Next[1],Next[2]]/
    (Table[where[1],Next[2]]*Table[Next[1],where[2]])

    se.OR = sqrt(sum(pow(Table)))
    ci.OR = exp(log(odds.ratio) + c(-1,1)*z.alpha*se.OR)

    # difference of proportions
    p1 = Table[where[1],where[2]]/(n1=Table[where[1],Next[2]] +
    Table[where[1],where[2]])

    p2=Table[Next[1],where[2]]/(n2=Table[Next[1],where[2]] +
    Table[Next[1],Next[2]])

    se.diff = sqrt(p1*(1-p1)/n1 + p2*(1-p2)/n2)
    ci.diff = (p1-p2) + c(-1,1)*z.alpha*se.diff

    # relative risk
    RR = p1/p2
    se.RR = sqrt((1-p1)/(p1*n1) + (1-p2)/(p2*n2))
    ci.RR = exp(log(RR) + c(-1,1)*z.alpha*se.RR)

    list(OR=list(odds.ratio=odds.ratio, CI=ci.OR),
    proportion.difference=list(diff=p1-p2, CI=ci.diff),
    relative.risk=list(relative.risk=RR,  CI=ci.RR))
    }
```

Así, aplicando la función anterior, se obtienen las estimaciones y los intervalos de

confianza para las razones de odds, la diferencia de proporciones y el riesgo relativo.

```
Wald.ci(tablaguay, c(1, 1))
```

Se obtiene

```
        odds.ratio
         1.560298

        odds.ratio-CI
         0.8546703 2.8485020

        proportion.difference
         0.01430845

        proportion.difference-CI
         -0.004868983  0.033485890

        relative.risk
         1.537362

        relative.risk-CI
         0.858614 2.752671
```

Como el intervalo de confianza para θ es igual a $(0{,}8546703; 2{,}8485020)$ y éste contiene a 1, entonces es razonable pensar que la verdadera razón de odds para la muerte por infarto de miocardio, es igual para los casos en que se toma aspirina o placebo.

Quizás se necesitaría tomar una muestra mayor para comprobar el posible efecto beneficioso de la aspirina, dado que con este ejemplo la razón de odds no es grande.

Contraste de independencia en tablas de doble entrada

Los contrastes de independencia se pueden aplicar tanto para muestreo multinomial (con $I \times J$ categorías) como para muestreo multinomial independiente (para las distintas filas).

En el primer caso se contrasta *independencia* y en el segundo, *homogeneidad*.

El contraste que se plantea es

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$$

para todo i, j.

$$H_1 : \pi_{ij} \neq \pi_{i+}\pi_{+j}$$

para algún i, j.

Se utiliza el contraste de Pearson.

Si es cierta H_0 entonces el número esperado de observaciones en cada casilla es

$$E\left(n_{ij}\right) = n\pi_{i+}\pi_{+j} = \mu_{ij}$$

y los estimadores de máxima verosimilitud son

$$\widehat{\mu}_{ij} = n\widehat{\pi}_{i+}\widehat{\pi}_{+j} = n\frac{n_{i+}}{n}\frac{n_{+j}}{n} = \frac{n_{i+}n_{+j}}{n}$$

Se utiliza el siguiente estadístico

$$\mathbb{X}^2 = \sum_i \sum_j \frac{\left(n_{ij} - \widehat{\mu}_{ij}\right)^2}{\widehat{\mu}_{ij}}$$

Si es cierta H_0, entonces

$$\mathbb{X}^2 \sim \chi^2_{(I-1)(J-1)}$$

es decir equivale a un test de la chi-cuadrado.

Otra alternativa es usar el test del cociente de verosimilitudes

$$G^2 = 2\sum_i \sum_j n_{ij} \log\left(\frac{n_{ij}}{\widehat{\mu}_{ij}}\right)$$

que también se distribuye como $\chi^2_{(I-1)(J-1)}$ cuando H_0 es cierta.

Se rechaza para valores grandes de $\mathbb{X}^2$. La convergencia a la distribución chi-cuadrado es mas rápida para $\mathbb{X}^2$ que para G^2.

La aproximación para $\mathbb{X}^2$ puede ser razonablemente buena si las frecuencias esperadas son mayores que 1 y la mayor parte son mayores que 5.

Cuando no se cumplen estas condiciones, se pueden utilizar métodos para muestras pequeñas.

Ejemplo

Creencias Religiosas

Educación	*Fundamentalista*	*Moderada*	*Liberal*	*Total*
< Secundaria	178	138	108	424
Secundaria	570	648	442	1660
Graduado	138	252	252	642
Total	886	1038	802	2726

```
religion.recuento = c(178, 138, 108, 570,
648, 442, 138, 252, 252)
tabla = cbind(expand.grid(list(Religiosidad = c("Fund",
"Mod", "Lib"), Grado = c("<HS", "HS o JH", "Graduado"))),
count = religion.recuento)

tablaguay = xtabs(religion.recuento ~ Religiosidad + Grado, data=tabla)
(res = chisq.test(tablaguay))

# Recuentos esperados
res$expected
```

Se obtiene

```
        Pearson's Chi-squared test

data:  tabla.array
X-squared = 69.1568, df = 4, p-value = 3.42e-14

            Grado
Religiosidad      <HS  HS o JH Graduado
        Fund 137.8078 539.5304 208.6618
        Mod  161.4497 632.0910 244.4593
        Lib  124.7425 488.3786 188.8789
```

Los contrastes y las frecuencias esperadas pueden obtenerse con el paquete `vcd`.

```
library(vcd)
assocstats(tablaguay)
```

Se obtiene

```
                    X^2 df   P(> X^2)
Likelihood Ratio 69.812  4 2.4869e-14
Pearson          69.157  4 3.4195e-14

Contingency Coeff.: 0.157
Cramer's V        : 0.113
```

Con la función `chisq.test` también se pueden hacer contrastes por simulación Montecarlo, es decir, calculando el estadístico de la chi cuadrado para todas las posibles tablas con las mismas sumas marginales por filas y columnas de la tabla original.

```
chisq.test(tablaguay, sim=T, B=2000)
```

Se obtiene

```
    Pearson's Chi-squared test with simulated p-value
    (based on 2000 replicates)

data:  tabla.array
X-squared = 69.1568, df = NA, p-value = 0.0004998
```

Ejemplo

Consideramos la siguiente tabla en relación a las primarias del partido demócrata en 2008, en una muestra de un distrito electoral.

	Candidato		
Género	*Clinton*	*Obama*	*Total*
Hombre	200	406	606
Mujer	418	418	836
Total	618	824	1442

El programa en SAS sería

```
/* OPTIONS nodate ls=65 formchar='|----|+|---+=|-/\<>*'; */
OPTIONS nodate ls=75;
/* Para SAS University */
ODS rtf file='/folders/myfolders/resultado.rtf' style=minimal
    startpage=no;

DATA primarias;
INPUT genero $ candidato $ recuento;
DATALINES;
Hombre Clinton 200
Hombre Obama 406
Mujer Clinton 418
Mujer Obama 428
;

PROC freq order=data;
weight recuento;
tables genero*candidato/expected deviation chisq relrisk;
RUN;
ODS rtf close;
```

Comentarios: En la última línea del programa se solicita la tabla de contingencia con las filas representadas por la primera variable y las columnas por la segunda variable.

Se solicitan, después,

- Las frecuencias esperadas suponiendo independencia (`expected`)
- Discrepancia entre las frecuencias observadas y las esperadas (`deviation`)
- Test de la chi-cuadrado de independencia (`chisq`)
- Razón de odds (`relrisk`)

Procedimiento FREQ

Frecuencia Esperada Desviación Porcentaje Pct fila Pct col

Tabla de genero por candidato			
	candidato		
genero	Clinton	Obama	Total
Hombre	200 257.93 -57.93 13.77 33.00 32.36	406 348.07 57.926 27.96 67.00 48.68	606 41.74
Mujer	418 360.07 57.926 28.79 49.41 67.64	428 485.93 -57.93 29.48 50.59 51.32	846 58.26
Total	618 42.56	834 57.44	1452 100.00

Estadísticos para la tabla de genero por candidato

Estadístico	DF	Valor	Prob
Chi-cuadrado	1	38.8726	<.0001
Chi-cuadrado de ratio de verosimilitud	1	39.3080	<.0001
Chi-cuadrado adj. de continuidad	1	38.2044	<.0001
Chi-cuadrado Mantel-Haenszel	1	38.8458	<.0001
Coeficiente Phi		-0.1636	
Coeficiente de contingencia		0.1615	
V de Cramer		-0.1636	

Test exacto de Fisher	
Celda (1,1) Frecuencia (F)	200
Alineado a la izquierda Pr <= F	<.0001
Alineado a la derecha Pr >= F	1.0000
Tabla de probabilidad (P)	<.0001
De dos caras Pr <= P	<.0001

Ratio de probabilidades y riesgos relativos			
Estadístico	Valor	Límites de confianza al 95%	
Ratio de probabilidad	0.5044	0.4062	0.6263
Riesgo relativo (Columna 1)	0.6680	0.5852	0.7625
Riesgo relativo (Columna 2)	1.3243	1.2140	1.4448

Tamaño de la muestra = 1452

La salida presenta la tabla de contingencia, con

- las frecuencias observadas
- las frecuencias esperadas (asumiendo independencia),
- la desviación entre las frecuencias observadas y esperadas,
- la contribución de la anterior desviación al valor global del estadístico de la chi-cuadrado.

Los restantes valores representan la proporción de casillas con respecto al número total de observaciones, la proporción relativa de las observaciones totales de la fila, y la proporción relativa respecto de las observaciones totales en cada columna.

Finalmente, aparece la prueba de independencia (incluyendo la prueba chi-cuadrado, y la prueba exacta de Fisher).

De acuerdo con los resultados, la chi-cuadrado de Pearson es igual a 38.87, con un *p-valor* menor que 0.0001, por lo que se rechaza la hipótesis nula de independencia.

El test de la razón de verosimilitudes presenta resultados similares: 39.31 y un valor de *p-valor* inferior a 0.0001.

Test de independencia con muestras pequeñas

Test exacto de Fisher para tablas 2×2

Todos los procedimientos vistos hasta ahora se basan en distribuciones asintóticas. Si tenemos muestras grandes no hay problemas, pero con muestras pequeñas es preferible usar contrastes *exactos*.

Consideramos una tabla 2×2 donde la hipótesis nula es de independencia entre las dos variables. Esto corresponde a que la razón de odds es igual a uno, $\theta = 1$.

Supongamos que los recuentos de las casillas $\{n_{ij}\}$ provienen de dos muestras aleatorias independientes, o de una única distribución multinomial definida sobre las 4 casillas de la tabla.

Se consideran todas las posibles tablas, que tienen como totales por filas y columnas los valores observados en los datos reales; entonces la distribución de los recuentos de las casillas es una *hipergeométrica*.

Ver para referencias sobre la distribución hipergeométrica:

```
http://en.wikipedia.org/wiki/Hypergeometric_distribution
```

Si condicionamos a los totales por fila y columna, solamente nos queda libre o sin fijar un recuento, por ejemplo n_{11}, de modo que éste determina los otros tres recuentos de las casillas.

t	$n_{1+} - t$	n_{1+}
$n_{+1} - t$	$n_{2+} - n_{+1} + t$	n_{2+}
n_{+1}	$n - n_{+1}$	n

Así, con la distribución hipergeométrica se calcula la probabilidad de los recuentos en términos de solo n_{11}

$$p(t) = P\left(n_{11} = t\right) = \frac{\binom{n_{1+}}{t}\binom{n_{2+}}{n_{+1} - t}}{\binom{n}{n_{+1}}}$$

donde los posibles valores son

$$m_- \leq n_{11} \leq m_+$$

con

$$\begin{aligned} m_- &= \text{máx}\left\{0, n_{1+} + n_{+1} - n\right\} \\ m_+ &= \text{mín}\left\{n_{1+}, n_{+1}\right\} \end{aligned}$$

Dados los valores totales de las marginales, las tablas que tienen mayores valores n_{11} también tienen mayores valores de la razón de odds

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

El contraste de independencia entre las dos variables se puede formular como

$$\begin{aligned} H_0 :&\quad \theta \leq 1 \\ H_1 :&\quad \theta > 1 \end{aligned}$$

El *p-valor* es igual a la probabilidad de la cola derecha de la distribución hipergeométrica de que n_{11} sea al menos tan grande como el observado, digamos t_0,

$$P\left(n_{11} \geq t_0\right),$$

de modo que se tiene una evidencia más fuerte a favor de la hipótesis alternativa H_1.

Ejemplo

Consideramos un ejemplo clásico de R. Fisher. Un colega suyo afirmaba que era capaz de distinguir en una taza de té con leche qué se había echado primero, si la leche o el té. Para comprobarlo diseñó un experimento donde se probaban 8 tazas de té. De ellas, en 4 se había echado primero la leche, y en las otras 4, primero el té.

Se trataba de adivinar en qué orden se había echado la leche y el té. Las tazas se presentaron de manera aleatoria, y se obtuvo

	Predicción		
Primer Servicio	*Leche*	*Té*	Total
Leche	3	1	4
Té	1	3	4
Total	4	4	

Evidentemente las distribuciones marginales están fijadas en 4.

Bajo H_0 "se puede distinguir el orden entre la leche y el té" se obtendría una m.a.s del total de 8 tazas. Así, La distribución nula de n_{11} es una hipergeométrica definida para todas las tablas 2×2 que tienen como marginales $(4, 4)$.

Los posibles valores para n_{11} son (0, 1, 2, 3, 4).

En la tabla observada, si se adivinan 3 de las 4 tazas donde se ha echado primero leche, la probabilidad es

$$P(3) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = \frac{16}{70} = 0{,}23$$

Así, para que se apoye la hipótesis $H_1 : \theta > 1$ el único valor extremo que se obtiene es con $n_{11} = 4$, es decir que se adivinen todas.

$$P(4) = \frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = \frac{1}{70} = 0{,}014$$

Cálculo de p-valores

Lo que aparece habitualmente programado en los paquetes informáticos es

$$p = P\left(p(n_{11}) \leq p(t_0)\right)$$

donde $p(t_0) = P(n_{11} = t_0)$.

Es decir, se suman las probabilidades de todas aquellas tablas que son menos o igual de probables que n_{11} en la tabla observada.

En el ejemplo de las tazas de té, se tiene que

$$\begin{aligned} P\left(n_{11} = 0\right) &= 0{,}014 \\ P\left(n_{11} = 1\right) &= 0{,}229 \\ P\left(n_{11} = 2\right) &= 0{,}514 \\ P\left(n_{11} = 3\right) &= 0{,}229 \\ P\left(n_{11} = 4\right) &= 0{,}014 \end{aligned}$$

Por tanto, se suman todas las probabilidades que son menores o iguales que la probabilidad

$P(3) = 0{,}229$ de la tabla observada. Se obtiene,

$$P(0) + P(1) + P(3) + P(4) = 0{,}486$$

```
# Alternativa bilateral
(fisher.test(matrix(c(3, 1, 1, 3), byrow=T, ncol=2)))
```

```
        Fisher's Exact Test for Count Data

data:  matrix(c(3, 1, 1, 3), byrow = T, ncol = 2)
p-value = 0.4857
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0.2117329 621.9337505
sample estimates:
odds ratio
  6.408309
```

```
# Alternativa unilateral
(fisher.test(matrix(c(3, 1, 1, 3), byrow=T, ncol=2),
alternative="greater"))
```

```
        Fisher's Exact Test for Count Data

data:  matrix(c(3, 1, 1, 3), byrow = T, ncol = 2)
p-value = 0.2429
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.3135693       Inf
sample estimates:
odds ratio
  6.408309
```

Categorías Ordinales

Es muy frecuente que las categorías en las tablas presenten una ordenación entre las categorías. Este hecho no lo tienen en cuenta los tests tradicionales de la χ^2.

Para tratar con variables ordinales, es mejor asignar rangos a las categorías de X ($u_1 \leq \cdots \leq u_n$) y a las categorías de Y ($v_1 \leq \cdots \leq v_n$) de modo que se conserve la ordenación original.

Si se denomina r al coeficiente de correlación entre los rangos, entonces el estadístico

$$M^2 = (n-1)r^2 \sim \chi^2_1$$

para muestras grandes (se denomina Estadístico de Mantel-Haenszel). Se rechaza la hipótesis de independencia para valores grandes de M^2.

La hipótesis alternativa es que existe una relación **lineal** entre las variables X e Y.

Ejemplo

En la siguiente tabla se presentan dos variables que son *ingresos* y *satisfacción en el trabajo* en una muestra dada.

Ambas variables son ordinales, de modo que las categorías de satisfacción laboral son *MI*: muy insatisfecho, *PI*: un poco insatisfecho, *S*: satisfecho, *MS*: muy satisfecho.

	Satisfación laboral			
Ingresos (miles $)	*MI*	*PI*	*S*	*MS*
< 15	1	3	10	6
$15 - 25$	2	3	10	7
$25 - 40$	1	6	14	12
> 40	0	1	9	11

Se programa en R:

```
income = c("<15000", "15000-25000", "25000-40000", ">40000")
jobsat = c("VD", "LD", "MS", "VS")

tabla = expand.grid(income=income, jobsat=jobsat)
data = c(1, 2, 1, 0, 3, 3, 6, 1, 10, 10, 14, 9, 6, 7, 12, 11)
tabla = cbind(tabla, count=data)
levels(tabla$income) = c(7.5, 20, 32.5, 60)
levels(tabla$jobsat) = 1:4

# Se escribe como dos columnas con observaciones individuales
res = apply(tabla[, 1:2], 2, function(x)
{as.numeric(rep(x, tabla$count))})
```

Se obtiene

```
(cor(res)[2, 1]^2)*(nrow(res) - 1)
```

```
[1] 3.807461
```

```
1 - pchisq(3.807461, 1)
```

```
[1] 0.05102474
```

O bien se puede considerar una hipótesis alternativa mas débil: la relación entre las variables ordinales es **monótona**.

En ese caso se puede considerar el coeficiente de asociación *gamma* de Goodman y Kruskal.

Se tiene que $-1 < \gamma < 1$, el caso de independencia implica que $\gamma = 0$, aunque lo contrario no es cierto.

```
Gamma2.f = function(x, pr=0.95)
{
    # Subrutina de L. Thompson
    # x is a matrix of counts.
    # You can use output of crosstabs or xtabs in R.
    # A matrix of counts can be formed from a data frame
    # by using design.table.

    # Confidence interval calculation and output from Greg Rodd

    # Check for using S-PLUS and output is from crosstabs
    if(is.null(version$language) && inherits(x, "crosstabs"))
    { oldClass(x)=NULL; attr(x, "marginals")=NULL}

    n = nrow(x)
    m = ncol(x)
    pi.c = pi.d = matrix(0,nr=n,nc=m)

    row.x = row(x)
    col.x = col(x)

    for(i in 1:(n)){
        for(j in 1:(m)){
            pi.c[i, j] = sum(x[row.x<i & col.x<j]) +
            sum(x[row.x>i & col.x>j])
            pi.d[i, j] = sum(x[row.x<i & col.x>j]) +
            sum(x[row.x>i & col.x<j])
        }
    }

    C = sum(pi.c*x)/2
    D = sum(pi.d*x)/2
    psi = 2*(D*pi.c-C*pi.d)/(C+D)^2
    sigma2 = sum(x*psi^2)-sum(x*psi)^2

    gamma = (C - D)/(C + D)
    pr2 = 1 - (1 - pr)/2
    CIa = qnorm(pr2) * sqrt(sigma2)*c(-1, 1) + gamma

    list(gamma=gamma, C=C, D=D, sigma=sqrt(sigma2),
    Level=paste(100*pr, "%", sep=""),
    CI=paste(c("[", max(CIa[1], -1),
        ", ", min(CIa[2], 1), "]"), collapse=""))
}
```

Aplicando la función anterior, se obtiene

```
temp = xtabs(formula = count ~ income + jobsat, data=tabla)
Gamma2.f(temp)
```

```
gamma
[1] 0.2211009

C
[1] 1331

D
[1] 849

sigma
[1] 0.1171628

Level
[1] "95%"

CI
[1] "[-0.00853400851071168, 0.450735843373097]"
```

Con una estimación del error estándar igual a 0.117, el intervalo de confianza al 95 % de la gamma es igual a $(-0{,}01;\ 0{,}45)$.

Así, el valor de la gamma no resulta significativamente distinto de 0.

Algunas medidas de asociación para tablas que presenta SAS

Consideramos el estadístico de la chi-cuadrado:

$$\mathbb{X}^2 = \sum_i \sum_j \frac{(n_{ij} - \widehat{\mu}_{ij})^2}{\widehat{\mu}_{ij}}$$

Coeficiente phi (φ)

Se deriva del estadístico chi-cuadrado y se define como

$$\varphi = \sqrt{\frac{\mathbb{X}^2}{n}}$$

Para tablas 2×2 se define como

$$\varphi = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1+}n_{2+}n_{+1}n_{+2}}}.$$

toma como posibles valores

$$-1 \leq \varphi \leq 1$$

V de Cramer

Para tablas 2×2 coincide con el coeficiente phi. Para tablas generales se define como

$$V = \sqrt{\frac{\mathbb{X}^2/n}{\text{mín}\,(I-1, J-1)}}$$

toma como posibles valores

$$-1 \leq V \leq 1$$

Estadístico de Mantel-Haenszel

Contrasta como hipótesis alternativa que existe una relación lineal entre las variables X e Y. Se define como

$$Q_{MH} = (n-1)r^2$$

donde r^2 es el coeficiente de correlación de Pearson calculado sobre los rangos de las dos variables.

Para más información sobre otros coeficientes se puede consultar en:

```
support.sas.com/documentation/onlinedoc/stat/131/freq.pdf
```

Ejemplo

Se presenta un estudio sobre enfermedades coronarias (Cornfield, 1962). Se consideran $n = 1329$ pacientes clasificados según el nivel de colesterol y si se han diagnosticado con enfermedad coronaria (CHD).

	0-199	**200-219**	**220-259**	**260+**	*Totales*
CHD	12	8	31	41	92
no CHD	307	246	439	245	1237
Totales	319	254	470	286	1329

```
/* OPTIONS nodate ls=75 formchar='|----|+|---+=|-/\<>*'; */
OPTIONS nodate ls=75;
/* Para SAS University */
ODS rtf file='/folders/myfolders/resultado.rtf' style=minimal
    startpage=no;

DATA chd;
INPUT CHD $ serum $ count @@;
DATALINES;
chd 0-199 12 chd 200-199 8 chd 220-259 31 chd 260+ 41
nochd 0-199 307 nochd 200-199 246 nochd 220-259 439
nochd 260+ 245
;

PROC freq;
weight count;
tables CHD*serum /chisq cmh1 scores=ridit;
RUN;
ODS rtf close;
```

Procedimiento FREQ

Frecuencia Porcentaje Pct fila Pct col	Tabla de CHD por serum				
	serum				
CHD	0-199	200-199	220-259	260+	Total
chd	12 0.90 13.04 3.76	8 0.60 8.70 3.15	31 2.33 33.70 6.60	41 3.09 44.57 14.34	92 6.92
nochd	307 23.10 24.82 96.24	246 18.51 19.89 96.85	439 33.03 35.49 93.40	245 18.43 19.81 85.66	1237 93.08
Total	319 24.00	254 19.11	470 35.36	286 21.52	1329 100.00

Estadísticos para la tabla de CHD por serum

Estadístico	DF	Valor	Prob
Chi-cuadrado	3	35.0285	<.0001
Chi-cuadrado de ratio de verosimilitud	3	31.9212	<.0001
Chi-cuadrado MH (Puntuaciones Ridit)	1	27.8381	<.0001
Coeficiente Phi		0.1623	
Coeficiente de contingencia		0.1603	
V de Cramer		0.1623	

Tamaño de la muestra = 1329

Estadísticos de sumarización para CHD por serum

Estadísticos de Cochran-Mantel-Haenszel (Basado en puntuaciones Ridit)				
Estadístico	Hipótesis alternativa	DF	Valor	Prob
1	Correlación nonzero	1	27.8381	<.0001

Tamaño total de la muestra = 1329