Práctica 9. Análisis de Tablas de Contingencia y Tablas 2x2.

José Luis Romero Béjar, Miguel Ángel Luque y Grupo BioestadísticaR

UNIVERSIDAD
DE GRANADA

Todo el material para el conjunto de actividades de este curso ha sido elaborado y es propiedad intelectual del grupo **BioestadisticaR** formado por:
Juan de Dios Luna del Castillo,
Pedro Femia Marzo,
Miguel Ángel Montero Alonso,
Christian José Acal González,
Pedro María Carmona Sáez,
Juan Manuel Melchor Rodríguez,
José Luis Romero Béjar,
Manuela Expósito Ruíz,
Juan Antonio Villatoro García,
Juan Manuel Praena Fernández,
Miguel Ángel Luque Fernández,
Francisco Javier Arnedo Fernández.

Todos los integrantes del grupo han participado en todas las actividades, en su elección, construcción, correcciones o en su edición final, no obstante, en cada una de ellas, aparecerán uno o más nombres correspondientes a las personas que han tenido la máxima responsabilidad de su elaboración junto al grupo de **BioestadisticaR**.

Práctica 9. Análisis de Tablas de Contingencia y Tablas 2x2.

José Luis Romero Béjar (Autor) & Miguel Angel Luque Fernández (actualización)

9.1 Homogeneidad de un carácter cualitativo entre muestras independientes

En la práctica 8 anterior se abordó la comparación de dos proporciones entre muestras independientes bajo el supuesto de que sus distribuciones de probabilidad pudieran ser aproximadas por una distribución normal. Este problema se puede estudiar y generalizar a más de dos proporciones mediante el **análisis de tablas de contingencia** sin la necesidad del supuesto anterior. En esta sección se describe como contrastar la **homogeneidad** de un carácter cualitativo entre varias muestras independientes, o lo que es lo mismo, como analizar la **igualdad de proporciones** del carácter entre estas muestras.

A continuación se presentan los datos de forma adecuada para realizar este tipo de contraste de hipótesis y se describen sus distintas etapas, haciendo especial inciso en la verificación de las condiciones de validez. Finalmente se ilustra su aplicación con un ejemplo práctico resuelto de forma secuencial, en una primera resolución, y en el lenguaje R con el paquete **BioestadisticaR2** diseñado para estas prácticas, en una segunda resolución.

9.1.1 Presentación de los datos

En el supuesto de que se tengan distintas *muestras independientes* ($r \geq 2$) en las que se han recogido las frecuencias observadas de las distintas respuestas ($c \geq 2$) de un determinado *carácter cualitativo*, es conveniente reflejar toda esta información en una *tabla de doble entrada* (r x c), de la forma siguiente:

Muestra / Respuesta	Respuesta 1	Respuesta 2	...	Respuesta c	**Totales**
Muestra 1	O_{11}	O_{12}	...	O_{1c}	n_1
Muestra 2	O_{21}	O_{22}	...	O_{2c}	n_2
...	...	...	...	...	...
Muestra r	O_{r1}	O_{r2}	...	O_{rc}	n_r

donde para cada $i = 1, ..., r$ y $j = 1, ..., c$, el valor O_{ij} indica el *número de individuos observados* en la muestra i con la respuesta j del carácter bajo estudio y n_i indica el tamaño de la muestra i, $\quad \forall i = 1, ..., r$.

Como el objetivo que se persigue es analizar la homogeneidad de este carácter en sus distintas respuestas entre todas las muestras, es conveniente reflejar las respectivas *proporciones* en esta tabla.

Muestra / Respuesta	Respuesta 1	Respuesta 2	...	Respuesta c	**Totales**
Muestra 1	p_{11}	p_{12}	...	p_{1c}	1
Muestra 2	p_{21}	p_{22}	...	p_{2c}	1
...	...	...	...	...	...
Muestra r	p_{r1}	p_{r2}	...	p_{rc}	1

donde para cada $i = 1, ..., r$ y $j = 1, ..., c$, el valor $p_{ij} = \frac{O_{ij}}{n_i}$ indica la *proporción de individuos observados* en la muestra i con la respuesta j del carácter bajo estudio. Es importante destacar que las proporciones en la última columna, las de la *Respuesta c*, se pueden obtener restando al uno la suma de las proporciones en su fila, es decir que $p_{ic} = 1 - p_{i1} - p_{i2} - ... - p_{i(c-1)}, \quad \forall i = 1, ..., r$.

En esta situación el contraste de hipótesis a realizar tendría las siguientes hipótesis nula y alternativa:

$$\begin{cases} \mathcal{H}_0 : p_{1j} = p_{2j} = ... = p_{rj} \quad (\forall j = 1, ..., c-1) \\ \mathcal{H}_1 : \text{No se da alguna de las igualdades} \end{cases}$$

A modo de ejemplo, si se tiene un carácter cualitativo con dos respuestas posibles *(Si/No)* cuya homogeneidad quiere ser analizada entre dos muestras independientes, la tabla (2 x 2) adecuada para su representación sería,

Muestra / Respuesta	Respuesta 1	Respuesta 2	**Totales**
Muestra 1	p_1	$q_1 = 1 - p_1$	1
Muestra 2	p_2	$q_2 = 1 - p_2$	1

En este caso el contraste de hipótesis tendría las siguientes hipótesis nula y alternativa,

$$\begin{cases} \mathcal{H}_0 : p_1 = p_2 \\ \mathcal{H}_1 : p_1 \neq p_2 \end{cases}$$

9.1.2 Etapas para realizar el contraste de hipótesis

Para realizar este contraste de hipótesis se parte de la *tabla de frecuencias observadas*, a la que se le añade una fila y una columna con los totales de columna $C_j, \quad \forall j = 1, ..., c$ y de fila $F_i, \quad \forall i = 1, ..., r$, respectivamente.

Muestra / Respuesta	Respuesta 1	Respuesta 2	...	Respuesta c	**Totales**
Muestra 1	O_{11}	O_{12}	...	O_{1c}	F_1
Muestra 2	O_{21}	O_{22}	...	O_{2c}	F_2
...	...	...	...	...	...
Muestra r	O_{r1}	O_{r2}	...	O_{rc}	F_r
Totales	C_1	C_2	...	C_c	$T = \sum_{i=1}^{r} F_i = \sum_{j=1}^{c} C_j$

Una vez se tiene construida la tabla de frecuencias observadas con sus totales por filas y columnas, se realizan las siguientes etapas:

- **Etapa 1.** Formulación de las hipótesis.

$$\begin{cases} \mathcal{H}_0 : p_{1j} = p_{2j} = ... = p_{rj} \quad (\forall j = 1, ..., c-1) \\ \mathcal{H}_1 : \text{No se da alguna de las igualdades} \end{cases}$$

- **Etapa 2.** Obtención de cantidades esperadas bajo la hipótesis nula.

$$E_{i,j} = \frac{F_i C_j}{T} \quad \forall i = 1, ..., r; \quad j = 1, ..., c$$

Muestra / Respuesta	Respuesta 1	Respuesta 2	...	Respuesta c	**Totales**
Muestra 1	E_{11}	E_{12}	...	E_{1c}	F_1
Muestra 2	E_{21}	E_{22}	...	E_{2c}	F_2
...	...	...	...	...	...
Muestra r	E_{r1}	E_{r2}	...	E_{rc}	F_r
Totales	C_1	C_2	...	C_c	$T = \sum_{i=1}^{r} F_i = \sum_{j=1}^{c} C_j$

Es importante destacar que, si los cálculos están bien realizados, los totales de filas y columnas de ambas tablas de frecuencias observadas y esperadas coinciden.

- **Etapa 3.** Verificación de las condiciones de validez.

El test sólo es válido si se verifica que ninguna E_{ij} es inferior a 1 y que no más del 20% de las E_{ij} son inferiores o iguales a 5.

- **Etapa 4.** Obtención de la cantidad experimental.

En el supuesto de que se cumplan las condiciones de validez, el estadístico experimental que hay que calcular tiene la siguiente expresión:

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

- **Etapa 5.** Obtención del nivel de significación (p-valor).

Para obtener el nivel de significación se busca el valor experimental calculado en la etapa anterior en la tabla Chi-cuadrado con $(r-1)$ x $(c-1)$ grados de libertad.

- **Etapa 6.** Decidir por una de las hipótesis.

9.1.3 Ejemplo práctico con datos agregados

De cada uno de los 6 distritos de una ciudad se tomaron 100 individuos al azar y se encontró que había 22, 16, 15, 31, 23 y 25 hipertensos respectivamente. ¿Es igual la prevalencia de la hipertensión en todos los distritos?

Solución manual

La tabla de frecuencias observadas es:

Distrito / Hipertensión	SI	NO	**Totales**
Distrito 1	$O_{11} = 22$	$O_{12} = 78$	$F_1 = 100$
Distrito 2	$O_{21} = 16$	$O_{22} = 84$	$F_2 = 100$

Distrito / Hipertensión	SI	NO	**Totales**
Distrito 3	$O_{31} = 15$	$O_{32} = 85$	$F_3 = 100$
Distrito 4	$O_{41} = 31$	$O_{42} = 69$	$F_4 = 100$
Distrito 5	$O_{51} = 23$	$O_{52} = 77$	$F_5 = 100$
Distrito 6	$O_{61} = 25$	$O_{62} = 75$	$F_6 = 100$
Totales	$C_1 = 132$	$C_2 = 468$	$T = 600$

- **Etapa 1**. Formulación de las hipótesis.

$$\begin{cases} \mathcal{H}_0 : \text{La prevalencia de la hipertensión es la misma en todos los distritos (homogeneidad)} \\ \mathcal{H}_1 : \text{La prevalencia de la hipertensión no es la misma en todos los distritos} \end{cases}$$

o equivalentemente,

$$\begin{cases} \mathcal{H}_0 : p_1 = p_2 = p_3 = p_4 = p_5 = p_6 \\ \mathcal{H}_1 : \text{No todas las } p_i \text{ son iguales} \end{cases}$$

donde p_i se refiere a la proporción de hipertensos en el distrito i con $i = 1, 2, 3, 4, 5, 6$.

- **Etapa 2.** Obtención de cantidades esperadas.

$$E_{i,j} = \frac{F_i C_j}{T}, \quad \forall i = 1, 2, 3, 4, 5, 6; \quad j = 1, 2$$

La tabla de frecuencias esperadas es:

Distrito / Hipertensión	SI	NO	**Totales**
Distrito 1	$E_{11} = 22$	$E_{12} = 78$	$F_1 = 100$
Distrito 2	$E_{21} = 22$	$E_{22} = 78$	$F_2 = 100$
Distrito 3	$E_{31} = 22$	$E_{32} = 78$	$F_3 = 100$
Distrito 4	$E_{41} = 22$	$E_{42} = 78$	$F_4 = 100$
Distrito 5	$E_{51} = 22$	$E_{52} = 78$	$F_5 = 100$
Distrito 6	$E_{61} = 22$	$E_{62} = 78$	$F_6 = 100$
Totales	$C_1 = 132$	$C_2 = 468$	$T = 600$

- **Etapa 3.** Verificación de las condiciones de validez.

El test es válido porque se verifica que ninguna E_{ij} es inferior a 1 ni a 5.

- **Etapa 4.** Obtención de la cantidad experimental.

$$\chi^2_{exp} = \sum_{i=1}^{6} \sum_{j=1}^{2} \frac{(O_{ij})^2}{E_{ij}} - T$$

Se sustituyen las cantidades observadas y esperadas,

$$\chi^2_{exp} = \frac{(22)^2}{22} + \frac{(16)^2}{22} + \frac{(15)^2}{22} + \frac{(31)^2}{22} + \frac{(23)^2}{22} + \frac{(25)^2}{22} + \frac{(78)^2}{78} + \frac{(84)^2}{78} + \frac{(85)^2}{78} + \frac{(69)^2}{78} + \frac{(77)^2}{78} + \frac{(75)^2}{78} - 600$$

y por tanto se obtiene,

$$\chi^2_{exp} = 10.2564$$

- **Etapa 5.** Obtención del nivel de significación (p-valor).

Para obtener el nivel de significación se busca el estadístico de contraste calculado en la etapa anterior en la tabla de una Chi-cuadrado con $(6-1)$ x $(2-1) = 5$ grados de libertad. En este caso $0.05 < p < 0.1$.

- **Etapa 6.** Decidir por una de las hipótesis.

A la vista del test de hipótesis, se concluye que puede aceptarse la hipótesis de homogeneidad puesto que $p > 0.05$, si bien se aprecian indicios de significación. De modo que convendría ampliar los tamaños muestrales y volver a realizar el test.

Solución con BioestadisticaR2

Las hipótesis nula y alternativa siguen siendo,

$$\begin{cases} \mathcal{H}_0 : \text{La prevalencia de la hipertensión es la misma en todos los distritos (homogeneidad)} \\ \mathcal{H}_1 : \text{La prevalencia de la hipertensión no es la misma en todos los distritos} \end{cases}$$

o equivalentemente,

$$\begin{cases} \mathcal{H}_0 : p_1 = p_2 = p_3 = p_4 = p_5 = p_6 \\ \mathcal{H}_1 : \text{No todas las } p_i \text{ son iguales} \end{cases}$$

El paquete *BioestadisticaR2* incorpora la función **tablarxc()**, que realiza este test. Esta función puede recibir como parámetro principal un data.frame con tantas columnas como respuestas posibles del carácter cualitativo bajo estudio.

```
hipertenso_SI=c(22,16,15,31,23,25)
hipertenso_NO=c(78,84,85,69,77,75)
tabla=data.frame(hipertenso_SI,hipertenso_NO)
tablarxc(frecs=tabla,tablas="E")
```

```
##
## # Test Chi-cuadrado para tablas RxC
## # --------------------------------
##
## # Frecuencias observadas
##           hipertenso_SI   hipertenso_NO Total
##   1                  22              78   100
##   2                  16              84   100
##   3                  15              85   100
##   4                  31              69   100
##   5                  23              77   100
##   6                  25              75   100
##   Total             132             468   600
##
## # Frecuencias esperadas
##           hipertenso_SI   hipertenso_NO Total
##   1                  22              78   100
##   2                  22              78   100
##   3                  22              78   100
##   4                  22              78   100
##   5                  22              78   100
##   6                  22              78   100
##   Total             132             468   600
##
## # Test chi-cuadrado
##   Validez: Frecuencia mínima esperada = 22
```

```
##            0 frecuencias esperadas son menores a 1
##            0 son menores a 5 (el 0% de la tabla)
##    ²(5 gl) = 10.256, p = 0.068
```

La salida que proporciona esta función abarca las etapas 2 a 5 anteriores. Calcula las frecuencias esperadas (aunque no las muestra a menos que se le pidan con el parámetro *tablas="E"*), comprueba la condición de validez sobre ellas y obtiene el estadístico de contraste junto con el nivel de significación, que en este caso es $p = 0.0683$. Este nivel de significación nos indica que puede aceptarse la hipótesis nula, si bien hay indicios en su contra y sería conveniente aumentar los tamaños de muestra y realizar nuevamente el test.

Finalmente, si se está interesado en obtener la tabla de porcentajes comentada en la Sección 9.1.1, ésta se obtiene, bien por totales de filas o de columnas, con la función **tablarxc()** añadiendo el parámetro *tablas="F"* (porcentajes por filas) o *tablas="C"* (porcentajes por columnas).

```
tablarxc(frecs=tabla, tablas="F")
```

```
##
## # Test Chi-cuadrado para tablas RxC
## # --------------------------------
##
## # Frecuencias observadas
##           hipertenso_SI   hipertenso_NO Total
##   1                  22              78   100
##   2                  16              84   100
##   3                  15              85   100
##   4                  31              69   100
##   5                  23              77   100
##   6                  25              75   100
##   Total             132             468   600
##
## # Test chi-cuadrado
##   Validez: Frecuencia mínima esperada = 22
##            0 frecuencias esperadas son menores a 1
##            0 son menores a 5 (el 0% de la tabla)
##    ²(5 gl) = 10.256, p = 0.068
##
## # Porcentajes por filas
##           hipertenso_SI   hipertenso_NO Total
##   1                0.22            0.78  1.00
##   2                0.16            0.84  1.00
##   3                0.15            0.85  1.00
##   4                0.31            0.69  1.00
##   5                0.23            0.77  1.00
##   6                0.25            0.75  1.00
##   Total            0.22            0.78  1.00
```

A la vista de estos resultados, y teniendo en cuenta que hay indicios en contra de la homogeneidad en los seis distritos, se observa que los distritos 2 y 3 son los que presentan menor prevalencia de hipertensos (16% y 15% respectivamente), mientras el distrito 4 es el que presenta la mayor prevalencia (31%).

9.2 Independencia de caracteres cualitativos

En la sección anterior se ha descrito como realizar un contraste de homogeneidad de un carácter cualitativo entre distintas muestras. Sin embargo, esta no es la única posibilidad que ofrece el análisis de tablas de

contingencia. En efecto, pueden darse muchas situaciones en las que lo que interese al investigador es contrastar si dos caracteres cualitativos están o no relacionados, son independientes o no, dentro de una población. A continuación se presentan los datos de forma adecuada para realizar este tipo de contraste de hipótesis y se describen sus distintas etapas. Finalmente se plantean dos ejemplos prácticos resueltos de forma secuencial y con el lenguaje R utilizando el paquete **BioestadisticaR2**, el primero con datos agregados y el segundo con datos importados desde un fichero.

9.2.1 Presentación de los datos

En el supuesto de que se tenga *una muestra* de tamaño T en la que interesa analizar la **independencia de dos caracteres cualitativos** con $r \geq 2$ posibles respuestas, para el primero, y $c \geq 2$ posibles respuestas, para el segundo, es conveniente reflejar toda la información observada de estos caracteres en la muestra en una *tabla de doble entrada* (r x c), del tipo de la introducida en la sección 9.1.2:

Carácter 1 / Carácter 2	Respuesta 1	Respuesta 2	...	Respuesta c	**Totales**
Respuesta 1	O_{11}	O_{12}	...	O_{1c}	F_1
Respuesta 2	O_{21}	O_{22}	...	O_{2c}	F_2
...	...	...	...	...	...
Respuesta r	O_{r1}	O_{r2}	...	O_{rc}	F_r
Totales	C_1	C_2	...	C_c	$T = \sum_{i=1}^{r} F_i = \sum_{j=1}^{c} C_j$

donde para cada $i = 1, ..., r$ y $j = 1, ..., c$, el valor O_{ij} indica el *número de individuos observados* en la muestra con la respuesta i del primer carácter bajo estudio y con la respuesta j del segundo carácter analizado.

En esta situación el contraste de hipótesis a realizar tendría las siguientes hipótesis nula y alternativa:

$$\begin{cases} \mathcal{H}_0 : \text{Hay independencia entre los caracteres cualitativos} \\ \mathcal{H}_1 : \text{No hay independencia entre los caracteres cualitativos} \end{cases}$$

9.2.2 Etapas para realizar el contraste de hipótesis

Para realizar este contraste de hipótesis se realizan exactamente las mismas etapas descritas en la sección 9.1.2 con las mismas expresiones para el cálculo de frecuencias esperadas, estadístico de contraste así como para la toma de decisiones.

9.2.3 Ejemplo práctico con datos agregados

Analizada la degradación arterial mediante un análisis anatomopatológico y mediante el grado de palpación de la arteria radial, interesa estudiar si existe relación entre estos dos métodos. Para ello se ha recogido la información de 252 casos en la siguiente tabla de contingencia:

Análisis / Palpación	Baja	Media	Alta
1	20	5	5
2	60	20	10
3	45	15	25
4	20	15	12

(Los valores de análisis representan, en orden creciente, la degradación arterial: a mayor valor mayor degradación).

Solución manual

La tabla de frecuencias observadas es:

Análisis / Palpación	Baja	Media	Alta	**Totales**
1	$O_{11} = 20$	$O_{12} = 5$	$O_{13} = 5$	$F_1 = 30$
2	$O_{21} = 60$	$O_{22} = 20$	$O_{23} = 10$	$F_2 = 90$
3	$O_{31} = 45$	$O_{32} = 15$	$O_{33} = 25$	$F_3 = 85$
4	$O_{41} = 20$	$O_{42} = 15$	$O_{43} = 12$	$F_4 = 47$
Totales	$C_1 = 145$	$C_2 = 55$	$C_3 = 52$	$T = 252$

- **Etapa 1**. Formulación de las hipótesis.

$$\begin{cases} \mathcal{H}_0 : \text{El grado de palpación es independiente del análisis anatomopatológico (independencia)} \\ \mathcal{H}_1 : \text{El grado de palpación está asociado con el análisis anatomopatológico (asociación)} \end{cases}$$

- **Etapa 2.** Obtención de cantidades esperadas.

$$E_{i,j} = \frac{F_i C_j}{T}, \quad \forall i = 1, 2, 3, 4; \quad j = 1, 2, 3$$

La tabla de frecuencias esperadas es:

Análisis / Palpación	Baja	Media	Alta	**Totales**
1	$E_{11} = 17.26$	$E_{12} = 6.55$	$E_{13} = 6.19$	$F_1 = 30$
2	$E_{21} = 51.78$	$E_{22} = 19.64$	$E_{23} = 18.57$	$F_2 = 90$
3	$E_{31} = 48.91$	$E_{32} = 18.55$	$E_{33} = 17.54$	$F_3 = 85$
4	$E_{41} = 27.05$	$E_{42} = 10.26$	$E_{43} = 9.70$	$F_4 = 47$
Totales	$C_1 = 145$	$C_2 = 55$	$C_3 = 52$	$T = 252$

- **Etapa 3.** Verificación de las condiciones de validez.

El test es válido porque se verifica que ninguna E_{ij} es inferior a 1 ni a 5.

- **Etapa 4.** Obtención de la cantidad experimental.

$$\chi^2_{exp} = \sum_{i=1}^{4} \sum_{j=1}^{3} \frac{(O_{ij})^2}{E_{ij}} - T$$

Se sustituyen las cantidades observadas y esperadas,

$$\chi^2_{exp} = \frac{(20)^2}{17.26} + \frac{(60)^2}{51.78} + \frac{(45)^2}{48.91} + \frac{(20)^2}{27.05} + \frac{(5)^2}{6.55} + \frac{(20)^2}{19.64} + \frac{(15)^2}{18.55} + \frac{(15)^2}{10.26} + \frac{(5)^2}{6.19} + \frac{(10)^2}{18.57} + \frac{(25)^2}{17.54} + \frac{(12)^2}{9.70} - 252$$

y por tanto se obtiene,

$$\chi^2_{exp} = 15.033$$

- **Etapa 5.** Obtención del nivel de significación (p-valor).

Para obtener el nivel de significación se busca el estadístico de contraste calculado en la etapa anterior en la tabla de una Chi-cuadrado con $(4-1)$ x $(3-1) = 6$ grados de libertad. En este caso $p < 0.05$.

- **Etapa 6.** Decidir por una de las hipótesis.

A la vista del test de hipótesis, se concluye que se rechaza la hipótesis nula de independencia puesto que $p < 0.05$, de modo que estas técnicas están asociadas. Sería interesante responder a qué puede deberse esta significación obtenida. Una respuesta se puede encontrar analizando la tabla de porcentajes por filas y/o columnas. Se deja este análisis para la solución con el paquete *BioestadisticaR2* propuesta a continuación por una mera cuestión de eficiencia computacional.

Solución con BioestadisticaR2

Se analizan las hipótesis nula y alternativa propuestas en la solución manual anterior.

$$\begin{cases} \mathcal{H}_0: \text{El grado de palpación es independiente del análisis anatomopatológico (independencia)} \\ \mathcal{H}_1: \text{El grado de palpación está asociado con el análisis anatomopatológico (asociación)} \end{cases}$$

Tal y como se comenta en la sección 9.1.3, el paquete *BioestadisticaR2* incorpora la función **tablarxc()**, que realiza este test. Esta función recibe como parámetro principal un data.frame con tantas columnas como respuestas posibles del carácter cualitativo bajo estudio. En este caso se pide también que muestre la tabla de porcentajes por filas directamente.

```
Palp_baja=c(20,60,45,20)
Palp_media=c(5,20,15,15)
Palp_alta=c(5,10,25,12)
tabla=data.frame(Palp_baja,Palp_media,Palp_alta)
tablarxc(frecs=tabla, tablas=c("F","E"))
```

```
##
## # Test Chi-cuadrado para tablas RxC
## # --------------------------------
##
## # Frecuencias observadas
##           Palp_baja   Palp_media   Palp_alta Total
##   1              20            5           5    30
##   2              60           20          10    90
##   3              45           15          25    85
##   4              20           15          12    47
##   Total         145           55          52   252
##
## # Frecuencias esperadas
##           Palp_baja   Palp_media   Palp_alta  Total
##   1           17.26         6.55        6.19  30.00
##   2           51.79        19.64       18.57  90.00
##   3           48.91        18.55       17.54  85.00
##   4           27.04        10.26        9.70  47.00
##   Total      145.00        55.00       52.00 252.00
##
## # Test chi-cuadrado
##   Validez: Frecuencia mínima esperada = 6.19
##            0 frecuencias esperadas son menores a 1
##            0 son menores a 5 (el 0% de la tabla)
##    ²(6 gl) = 15.033, p = 0.02
##
```

```
## # Porcentajes por filas
##           Palp_baja   Palp_media   Palp_alta Total
##   1           0.667        0.167       0.167 1.000
##   2           0.667        0.222       0.111 1.000
##   3           0.529        0.176       0.294 1.000
##   4           0.426        0.319       0.255 1.000
##   Total       0.575        0.218       0.206 1.000
```

La salida que proporciona esta función abarca las etapas 2 a 5 anteriores. Calcula las frecuencias esperadas (aunque no las muestra a menos que se le pidan con el parámetro *tablas="E"*), comprueba la condición de validez sobre ellas y obtiene el estadístico de contraste junto con el nivel de significación, que en este caso es $p = 0.02$. Este nivel de significación nos indica que puede rechazarse la hipótesis nula de independencia sin ninguna duda, de modo que estos dos métodos están asociados. Del análisis de la tabla de porcentajes por filas, se observa que una *baja palpación* se da principalmente en los *análisis 1 y 2* pues son los porcentajes más elevados (66.67% en ambos casos). La *palpación media* tiene su porcentaje más elevado asociado al *análisis 4* (31.91%), mientras que la *palpación alta* tiene un porcentaje mayor para el *análisis 3* (29.41%).

9.2.4 Ejemplo práctico con datos importados desde un fichero

El fichero de datos *osteo.sav* almacena información de 27 variables distintas de 94 pacientes diagnosticados de diabetes. Para la ilustración de cómo realizar un análisis de la independencia de dos caracteres cualitativos almacenados en una base de datos como esta, en este ejemplo, se va a analizar la independencia de la presencia de retinopatía en el paciente, almacenada en la variable *retin*, con el sexo, almacenado en la variable *sexo*. En este caso solo se propone la solución con el paquete *BioestadisticaR2* puesto que el objetivo de este ejemplo es aprender a utilizar la función **tablarxc()** en esta situación.

Solución con BioestadisticaR2

Se analizan las hipótesis nula y alternativa siguientes:

$$\begin{cases} \mathcal{H}_0 : \text{La presencia de retinopatía es independiente del sexo del paciente (independencia)} \\ \mathcal{H}_1 : \text{La presencia de retinopatía está asociada al sexo (asociación)} \end{cases}$$

La función **tablarxc()** está programada para obtener la tabla de contingencia en este tipo de situaciones en las que los datos provienen de un fichero. La única diferencia con respecto a los planteamientos de ejemplos anteriores, en los que solo pasábamos al parámetro *frecs* de esta función un data.frame con la tabla de contingencia, está en que ahora han de pasarse los parámetros *fvar* con la variable que define las filas y *cvar* con la variable que define las columnas. La función **tablarxc()** admite los parámetros *fcat* y *ccat* que permiten indicar las etiquetas de cada una de las variables cualitativas enfrentadas. Esto permite una representación visual de la tabla de contingencia muy informativa. Como en ejemplos anteriores el parámetro *tablas* con el valor *"F"*, para los totales por filas, o el valor *"C"*, para los totales por columnas, o incluso ambos a la vez, va a permitir un análisis del por qué de las diferencias encontradas, a través de las convenientes tablas de porcentajes por filas o columnas.

Antes de resolver el problema planteado hay que confirmar que estamos enfrentando dos caracteres cualitativos, así como conocer sus distintas respuestas.

```
head(osteo$retin)
```

```
## <labelled<double>[6]>: Presencia de retinopatía (1=No, 2=Leve, 3=Grave)
## [1] 1 1 1 2 1 1
##
## Labels:
##  value label
##      1    No
```

```
##      2  Leve
##      3 Grave
```

```
head(osteo$sexo)
```

```
## <labelled<double>[6]>: Sexo del paciente (1=Hombre, 2=Mujer)
## [1] 1 1 1 1 2 2
##
## Labels:
##  value  label
##      1 Hombre
##      2  Mujer
```

En efecto, se confirman que son dos caracteres cualitativos que almacenan las respuestas *Hombre/Mujer* para la variable *sexo* y las respuestas *No/Leve/Grave* para la variable *retin*. Este problema se puede resolver con una sola línea gracias al uso de la función **tablarxc**.

```
tablarxc(fvar=osteo$retin,cvar=osteo$sexo,fcat=c("No","Leve","Grave"),
         ccat = c("Hombre","Mujer"),tablas=c("F","C","E"))
```

```
##
## # Test Chi-cuadrado para tablas RxC
## # --------------------------------
##
## # Frecuencias observadas
##           Hombre   Mujer Total
##   No          33      30    63
##   Leve         8      10    18
##   Grave        4       9    13
##   Total       45      49    94
##
## # Frecuencias esperadas
##           Hombre   Mujer Total
##   No       30.16   32.84 63.00
##   Leve      8.62    9.38 18.00
##   Grave     6.22    6.78 13.00
##   Total    45.00   49.00 94.00
##
## # Test chi-cuadrado
##   Validez: Frecuencia mínima esperada = 6.22
##            0 frecuencias esperadas son menores a 1
##            0 son menores a 5 (el 0% de la tabla)
##    ²(2 gl) = 2.122, p = 0.346
##
## # Porcentajes por filas
##           Hombre   Mujer Total
##   No       0.524   0.476 1.000
##   Leve     0.444   0.556 1.000
##   Grave    0.308   0.692 1.000
##   Total    0.479   0.521 1.000
##
## # Porcentajes por columnas
##           Hombre   Mujer Total
```

```
##   No       0.733   0.612 0.670
##   Leve     0.178   0.204 0.191
##   Grave    0.089   0.184 0.138
##   Total    1.000   1.000 1.000
```

La salida obtenida construye en primer lugar la tabla de contingencia con las frecuencias observadas para cada respuesta de un carácter en función de las respuestas del otro. Justifica que se verifican las condiciones de validez puesto que no hay frecuencias esperadas menores que uno, ni hay más del 20% inferiores a 5. De hecho si esto no sucediera no se realizaría el test y se informaría de ello en esta salida. Una vez verificada la condición de validez se obtiene un valor $p = 0.3461 > 0.05$ que nos lleva a no rechazar la hipótesis nula de independencia, o lo que es lo mismo: *la presencia de retinopatía es independiente del sexo.* Finalmente, si se observa la tabla de porcentajes por filas, se entiende el por qué de esta independencia. En efecto, para cada una de las respuestas posibles para la presencia de retinopatía se observan porcentajes similares tanto en hombres como en mujeres.

9.3 Enfermedad vs. Factor de riesgo: Medidas de asociación.

En secciones anteriores ha quedado clara la utilidad de las tablas de contingencia para el análisis de la independencia de un carácter cualitativo entre distintas muestras independientes, así como la independencia de dos caracteres cualitativos dentro de una misma muestra. Una aplicación directa del análisis de la independencia de dos caracteres cualitativos, cuando ambos tienen tan sólo dos posibles respuestas, es el uso de las tablas de contingencia resultantes, en este caso tablas 2 x 2, para el análisis de la independencia de un determinado factor de riesgo (FR) en relación al diagnóstico de una determinada patología o enfermedad (E).

9.3.1 Generalidades

9.3.1.1 Presentación de los datos. Se denota por E=Diagnóstico positivo para determinada enfermedad, de modo que E^c=Diagnóstico negativo para esta enfermedad. Del mismo modo se denota FR=Estar sometido a cierto factor de riesgo, y por tanto FR^c=No estar sometido a dicho factor de riesgo. Con esta notación se analizan las hipótesis nula y alternativa siguientes:

$$\begin{cases} \mathcal{H}_0 : E \text{ es independiente del } FR \\ \mathcal{H}_1 : E \text{ está asociado al } FR \end{cases}$$

y la tabla de contingencia para este problema quedaría de siguiente forma:

Enfermedad / Factor de Riesgo	FR	FR^c	**Totales**
E	O_{11}	O_{12}	F_1
E^c	O_{21}	O_{22}	F_2
Totales	C_1	C_2	T

9.3.1.2. Tipos de muestreo y tipos de estudio epidemiológico En la siguiente tabla se resumen los distintos tipos de muestreo y tipos de estudio epidemiológico que se pueden plantear en función del diseño planteado para recoger los datos, así como la dirección en la que este se plantea, de la enfermedad al factor de riesgo o del factor de riesgo a la enfermedad.

Tipo de muestreo / Tipo de estudio	Transversal	Prospectivo	Retrospectivo
Muestreo tipo I	(1)		
Muestreo tipo II		(2)	(3)

(1) Se toman T individuos al azar y se clasifican en función de E y FR.
(2) Se toman C_1 y C_2 individuos con el FR y se clasifican en base a E.
(3) Se toman F_1 y F_2 individuos con la E y se clasifican en base a FR.

9.3.1.3. Etapas para realizar el contraste de hipótesis Para realizar este contraste de hipótesis se parte de la tabla (2 x 2) de frecuencias observadas.

Enfermedad / Factor de Riesgo	FR	FR^c	**Totales**
E	O_{11}	O_{12}	F_1
E^c	O_{21}	O_{22}	F_2
Totales	C_1	C_2	T

Una vez se tiene construida esta tabla se debe de tener claro el tipo de diseño muestral así como el tipo de estudio (transversal, retrospectivo o prospectivo) que se ha planteado para realizar el análisis de independencia de la enfermedad (E) con el factor de riesgo (FR). A continuación se realizan las siguientes etapas:

- **Etapa 1**. Formulación de las hipótesis.

$$\begin{cases} \mathcal{H}_0 : E \text{ es independiente del } FR \\ \mathcal{H}_1 : E \text{ está asociado al } FR \end{cases}$$

- **Etapa 2**. Obtención de la mínima cantidad esperada de la tabla.

$$E = \frac{min(F_1, F_2) \text{ x } min(C_1, C_2)}{T}$$

- **Etapa 3**. Verificación de las condiciones de validez dependiendo del tipo de estudio y del tamaño de la muestra.

Tipo de muestreo / Tamaño de muestra	$T \leq 500$	$T > 500$
Tipo I (Transversal)	$E \geq 3.9$	$E \geq 6.2$
Tipo II (Retrospectivo o prospectivo)	$E \geq 7.7$	$E \geq 14.9$

- **Etapa 4**. Obtención de la cantidad experimental. Dependiendo del tipo de estudio se obtiene una de las siguientes cantidades experimentales:

Tipo de muestreo	Cantidad experimental
Tipo I (Transversal)	$\chi^2_{exp} = \frac{(\lvert O_{11}O_{22} - O_{12}O_{21}\rvert - 0.5)^2}{F_1F_2C_1C_2}$ x T
Tipo II (Retrospectivo o prospectivo)	$\chi^2_{exp} = \frac{(\lvert O_{11}O_{22} - O_{12}O_{21}\rvert - c)^2}{F_1F_2C_1C_2}$ x T

Observación 1: si el estudio es restrospectivo $c = 1$ si $F_1 \neq F_2$ ó $c = 2$ si $F_1 = F_2$.

Observación 2: si el estudio es prospectivo $c = 1$ si $C_1 \neq C_2$ ó $c = 2$ si $C_1 = C_2$.

- **Etapa 5**. Obtención del nivel de significación (p-valor). Para obtener el nivel de significación se busca el valor experimental calculado en la etapa anterior en la tabla Chi-cuadrado con 1 grado de libertad.
- **Etapa 6**. Decidir por una de las hipótesis.

9.3.1.4. Medidas de asociación En el supuesto de que, tras realizar el contraste de hipótesis, se rechace la hipótesis nula de independencia, existen distintas medidas de asociación, dependiendo del tipo de estudio realizado, que analizan el grado de dependencia existente entre la enfermedad E y el factor de riesgo FR.

La siguiente tabla contiene la **estimación puntual de las distintas medidas de asociación** según el tipo de estudio realizado:

Medida de asociación	Estimación	Estudios donde es válida
Diferencia de Berkson	$d = \frac{O_{11}O_{22} - O_{12}O_{21}}{C_1C_2}$	Transversales y prospectivos
Riesgo relativo	$R = \frac{O_{11}C_2}{O_{12}C_1}$	Transversales y prospectivos
Razón de producto cruzado (odds ratio)	$OR = \frac{O_{11}O_{22}}{O_{12}O_{21}}$	Transversales, retrospectivos y prospectivos

Observación: si se realiza un **estudio retrospectivo para una enfermedad rara** (prevalencia $p < 0.1$) también se puede estimar el riesgo relativo de forma puntual por medio de la razón de producto cruzado, es decir, **si el estudio es retrospectivo con** una prevalencia de la enfermedad menor al 10% (0.1%) entones la Razón del Producto Cruzado o Odds Ratio (OR) es aproximadamente igual al Riesgo Relativo (RR).

Interpretación de las distintas medias de asociación:

- La **diferencia de Berkson** (d) indica que la probabilidad de enfermar en los individuos con el FR **aumenta** en un (100 x d)% (en términos absolutos: escala aditiva) con respecto a los que no lo tienen (FR^c).
- El **riesgo relativo** (R) indica que la probabilidad de enfermar es R **veces mayor** en individuos con el FR que en los que no lo tienen (FR^c).
- La **razón de producto cruzado (odds ratio)** (O) indica que la fracción de individuos que enferman (E) frente a los que no (E^c), es O **veces mayor** en individuos con FR que los que no lo tienen (FR^c).

En las siguientes tres secciones se ilustran ejemplos para este tipo de contraste con los tres diseños introducidos: transversal, retrospectivo y prospectivo. Si bien, para los tres ejemplos se ha considerado el mismo problema pero cambiando el tipo de estudio según la forma de recoger la información, no se pretende analizar la idoneidad del mismo para analizar la relación entre la enfermedad y el factor de riesgo enfrentados.

9.3.2 Ejemplo práctico con un diseño Transversal

El trastorno por déficit de atención con o sin hiperactividad (TDA-H) está relacionado con un retraso del desarrollo neurológico en ciertas partes del cerebro que controlan, entre otras, la atención y la impulsividad.

Dada la alta incidencia de fracaso escolar asociado a este trastorno, sobre todo en la etapa de la Educación Secundaria Obligatoria (ESO), hay gran interés en identificar factores de riesgo que permitan un diagnóstico precoz. En este sentido se han encontrado evidencias relacionadas con la herencia genética. Con el objetivo de contrastar que para el diagnóstico de un niño como TDA-H hay que tener en cuenta si alguno de sus progenitores también tiene o tuvo este trastorno, se ha recogido una muestra de **200** niños de los que 93 están diagnosticados con este trastorno, teniendo 47 de ellos, al menos, a uno de sus dos progenitores con TDA-H. De los niños no diagnosticados con TDA-H tan sólo 15 de ellos tenían alguno de sus progenitores con este trastorno.

Solución manual

En primer lugar es necesario construir la tabla de contingencia con estos datos. Por comodidad se denota por E=Diagnóstico positivo por TDA-H para un niño, de modo que E^c=Diagnóstico negativo por TDA-H para un niño. Del mismo modo se denota FR=Algún progenitor fue diagnosticado por TDA-H, y por tanto FR^c=Ningún progenitor fue diagnosticado por TDA-H. La tabla de contingencia para este problema quedaría de siguiente forma:

Enfermedad / Factor de Riesgo	FR	FR^c	**Totales**
E	$O_{11} = 47$	$O_{12} = 46$	$F_1 = 93$
E^c	$O_{21} = 15$	$O_{22} = 92$	$F_2 = 107$
Totales	$C_1 = 62$	$C_2 = 138$	$T = 200$

(El color rojo indica que este es el primer número introducido en la tabla. Puede servir como regla nemotécnica para identificar que el estudio es trasnversal).

Es evidente que el tipo de diseño realizado para esta investigación es de tipo **transversal** ya que partimos de una muestra de $T = 200$ niños que clasificamos en el momento, según si están diagnosticados o no por TDA-H y según si alguno de sus progenitores también fue diagnosticado o no. Al ser un diseño transversal, en el supuesto de que se rechaze la hipótesis nula de independencia, tendrá sentido calcular e **interpretar todas las medidas de asociación** introducidas en las clases de teoría.

- **Etapa 1**. Formulación de las hipótesis. Con esta notación se analizan las hipótesis nula y alternativa siguientes:

$$\begin{cases} \mathcal{H}_0 : E \text{ es independiente del } FR \\ \mathcal{H}_1 : E \text{ está asociado al } FR \end{cases}$$

- **Etapa 2.** Obtención de la mínima cantidad esperada de la tabla.

$$E = \frac{min(F_1, F_2) \text{ x } min(C_1, C_2)}{T} = \frac{min(93, 107) \text{ x } min(62, 138)}{200} = \frac{93 \text{ x } 62}{200} = 28.83$$

- **Etapa 3.** Verificación de las condiciones de validez.

Como el estudio es transversal y el tamaño total de la muestra $T = 200 \leq 500$ es suficiente con que $E = 28.83 \geq 3.9$. Por tanto se cumple la condición de validez.

- **Etapa 4.** Obtención de la cantidad experimental.

$$\chi^2_{exp} = \frac{(|O_{11}O_{22} - O_{12}O_{21}| - 0.5)^2}{F_1F_2C_1C_2} \text{ x } T = \frac{(|47 \text{ x } 92 - 46 \text{ x } 15| - 0.5)^2}{93 \text{ x } 107 \text{ x } 62 \text{ x } 138} \text{ x } 200 = 31.0129$$

- **Etapa 5.** Obtención del nivel de significación (p-valor).

Para obtener el nivel de significación se busca el estadístico de contraste calculado en la etapa anterior en la tabla de una Chi-cuadrado con 1 grado de libertad. En este caso $p < 0.001$.

- **Etapa 6.** Decidir por una de las hipótesis.

A la vista del test de hipótesis, para este ejemplo el test exacto de Fisher proporciona un valor $p = 0.001 < 0.05$ que permite rechazar la hipótesis nula de independencia con total seguridad. Esto confirma que **el diagnóstico de un niño en edad temprana por TDA-H está relacionado con que alguno de sus progenitores también haya sido diagnosticado**

Como se ha indicado al principio de esa solución propuesta, dado que el estudio es transversal, se pueden estimar puntualmente e interpretar todas las medidas de asociación introducidas en las clases de teoría.

- *Riesgo absoluto o diferencia de Berkson*

$$d = \frac{O_{11}O_{22} - O_{12}O_{21}}{C_1C_2} = \frac{47 \text{ x } 92 - 46 \text{ x } 15}{62 \text{ x } 138} = 0.4247$$

Este valor indica que el riesgo de que un niño sea diagnosticado como TDA-H (E) aumenta un 42.27% si se tiene algún progenitor diagnosticado por este trastorno (FR) con respecto a que no lo tenga (FR^c).

- *Riesgo relativo*

$$R = \frac{O_{11}C_2}{O_{12}C_1} = \frac{47 \text{ x } 138}{46 \text{ x } 62} = 2.2742$$

  En este caso el riesgo de que un niño sea diagnosticado por TDA-H (E) es 2.2742 veces mayor si se tienen progenitores que han sido diagnosticados de este trastorno (FR) que si no se tiene a ningún progenitor diagnosticado (FR^c).

- *Odds ratio*

$$OR = \frac{O_{11}O_{22}}{O_{12}O_{21}} = \frac{47 \text{ x } 92}{46 \text{ x } 15} = 6.2667$$

  La ventaja de que un niño sea diagnosticado por TDA-H (E) frente a que no sea diagnosticado (E^c), se multiplica por 6.2667 en niños que tienen algún progenitor con este diagnóstico (FR) frente a niños que no tienen ningún progenitor diagnosticado (FR^c).

Solución con BioestadisticaR2

En primer lugar es necesario construir la tabla de contingencia con estos datos. Por comodidad se denota por E=Diagnóstico positivo por TDA-H para un niño, de modo que E^c=Diagnóstico negativo por TDA-H para un niño. Del mismo modo se denota FR=Algún progenitor fue diagnosticado por TDA-H, y por tanto FR^c=Ningún progenitor fue diagnosticado por TDA-H. Con esta notación se analizan las hipótesis nula y alternativa siguientes:

$$\begin{cases} \mathcal{H}_0 : E \text{ es independiente del } FR \\ \mathcal{H}_1 : E \text{ está asociado al } FR \end{cases}$$

y la tabla de contingencia para este problema quedaría de siguiente forma:

Enfermedad / Factor de Riesgo	FR	FR^c	**Totales**
E	$O_{11} = 47$	$O_{12} = 46$	$F_1 = 93$
E^c	$O_{21} = 15$	$O_{22} = 92$	$F_2 = 107$
Totales	$C_1 = 62$	$C_2 = 138$	$T = 200$

(El color rojo indica que este es el primero número introducido en la tabla. Puede servir como regla nemotécnica para identificar que el estudio es trasnversal).

Es evidente que el tipo de diseño realizado para esta investigación es de tipo **transversal** ya que partimos de una muestra de $T = 200$ niños que clasificamos en el momento, según si están diagnosticados o no por TDA-H y según si alguno de sus progenitores también fue diagnosticado o no. Al ser un diseño transversal, en el supuesto de que se rechaze la hipótesis nula de independencia, tendrá sentido calcular e **interpretar todas las medidas de asociación** introducidas en las clases de teoría.

Para realizar este contraste se utiliza la función **tabla2x2()** del paquete *BioestadisticaR2*. Esta función puede recibir las frecuencias observadas en la tabla de contingencia anterior en forma de un vector del tipo $(O_{11}, O_{12}, O_{21}, O_{22})$. Permite definir las etiquetas para las filas y columnas de la tabla de contingencia con los parámetros *fcat* y *ccat* comentados anteriormente. El parámetro más destacado de esta función es el parámetro *estudio* que recibe los valores **T** para un estudio transversal, **R** para un estudio retrospectivo o **P** para un estudio prospectivo. También admite el parámetro *tablas* que construye las tablas de proporciones por filas o columnas según si se le pasa el valor "F" o "C". En este ejemplo, nuestro estudio es transversal y habrá que indicar *estudio="T"*.

```
tabla2x2(o=c(47,46,15,92), fcat=c("Niños con TDA-H","Niños sin TDA-H"),
         ccat=c("Algún padre con TDA-H","Padres sin TDA-H"),estudio="T",
         tablas=c("F,C"))
```

```
##
## # Análisis de tablas 2x2
## # ----------------------
##
## # Frecuencias observadas
##                     Algún padre con TDA-H   Padres sin TDA-H Total
##   Niños con TDA-H                      47                 46    93
##   Niños sin TDA-H                      15                 92   107
##   Total                                62                138   200
##
##
## # Test Chi-cuadrado para un estudio transversal
##
##    ² = 31.013,   gl = 1,  p < 0.001, (cpc = 0.5)
##   Validez: Frecuencia mínima esperada = 28.83  > 3.9
##
##   Test exacto de Fisher (bilateral): p < 0.001
##
##   --- Otros criterios ²:
##    ² = 31.021,   gl = 1,  p < 0.001, (sin cpc)
##    ² = 29.338,   gl = 1,  p < 0.001, (cpc de Yates = 100.00)
##
## # Estimación de la prevalencia  en un estudio transversal
##   Método de Wald ajustado:
##   p=0.466; 95%-IC( )=(0.397, 0.534)
##
## # Medidas de asociación para un estudio transversal
##   [!] Las medidas de riesgo se calculan como riesgo de la categoría
##       en la 1a columna (frente a la 2a) para la categoría en la 1a
##       fila (frente a la 2a)
##
##   Riesgo absoluto (diferencia de Berkson; método de Agresti-Caffo):
##   d=0.425; 95%-IC(d)=(0.283, 0.547)
```

```
##
##   Riesgo relativo:
##   Rr=2.274; 95%-IC(Rr)=(1.714, 2.963)
##
##   Riesgo atribuible:
##   Ra=0.283; 95%-IC(Ra)= (0.167, 0.383)
##
##   Razón del producto cruzado (odds ratio):
##   OR=6.267; 95%-IC(OR)= (3.110, 11.948)
```

De la salida obtenida cabe destacar que:

- En primer lugar *muestra la tabla de contingencia con las frecuencias observadas*. Es conveniente contrastar que coincide con la que diseñamos de forma manual, en caso contrario implicaría que se ha construido mal el vector y por tanto se estarían contrastando otros datos diferentes a los considerados. En este caso todo funciona correctamente.
- En segundo lugar comprueba la *condición de validez*, que para este tipo de estudio, **transversal**, y tamaño de muestra $T \leq 500$, consiste en que la frecuencia mínima esperada exceda de 3.9. En este caso su valor es $28.83 \geq 3.9$.
- A continuación realiza el *test exacto de Fisher con corrección por continuidad (cpc), sin cpc y con la cpc de Yates*. Para este ejemplo el test exacto de Fisher proporciona un valor $p = 0.001 < 0.05$ que permite rechazar la hipótesis nula de independencia con total seguridad. Esto confirma que **el diagnóstico de un niño en edad temprana por TDA-H está relacionado con que alguno de sus progenitores también haya sido diagnosticado**.
- A continuación se muestra una estimación puntual y por intervalo de confianza al 95% de confianza para la *prevalencia de la enfermedad* según el tipo de estudio, transversal, realizado. En este ejemplo estima la **prevalencia de niños diagnosticados por TDA-H** en el $45,57\%$ de forma puntual y entre el 39.72% y 53.41% con una confianza del 95%.
- Finalmente muestra la estimación puntual y por intervalos de confianza al 95% de confianza de todas las *medidas de asociación* que se pueden obtener para este tipo de estudio. En este ejemplo, como el diseño es **transversal** tienen sentido el **riesgo absoluto (diferencia de Berkson)**, el **riesgo relativo**, **riesgo atribuible** y **razón de producto cruzado (odds ratio)**. A continuación se interpretan sus estimaciones puntuales en el contexto de este ejemplo.
  - *Riesgo absoluto o diferencia de Berkson* (**d**) = 0.4247. Este valor indica que el riesgo de que un niño sea diagnosticado como TDA-H (E) aumenta un 42.27% si se tiene algún progenitor diagnosticado por este trastorno (FR) con respecto a que no lo tenga (FR^c).
  - *Riesgo relativo* (**R**) = 2.2742. En este caso el riesgo de que un niño sea diagnosticado por TDA-H (E) es 2.2742 veces mayor si se tienen progenitores que han sido diagnosticados de este trastorno (FR) que si no se tiene a ningún progenitor diagnosticado (FR^c).
  - *Riesgo atribuible* (**Ra**) = 0.2832. No se interpreta en este ejemplo porque no es una medida introducida en las clases de teoría.
  - *Odds ratio* (**OR**) = 6.2667. La ventaja de que un niño sea diagnosticado por TDA-H (E) frente a que no sea diagnosticado (E^c), se multiplica por 6.2667 en niños que tienen algún progenitor con este diagnóstico (FR) frente a niños que no tienen ningún progenitor diagnosticado (FR^c).

Como conclusión, se ha confirmado que **la herencia genética influye en el diagnóstico temprano de un niño por TDA-H**. Además se ha dado información relevante de la **fuerza que tiene como factor de riesgo que algún progenitor haya sido diagnosticado por este trastorno**, gracias a la interpretación de las distintas medidas de asociación consideradas. Además se ha proporcionado una estimación de la **prevalencia** de este tipo de trastorno en niños gracias que se ha realizado un diseño transversal para la recogida de datos.

9.3.3 Ejemplo práctico con un diseño Retrospectivo

En este ejemplo se analiza el objetivo planteado en el ejemplo anterior desde la perspectiva de un diseño **retrospectivo**. Por tanto, se pretende contrastar que para el diagnóstico de un niño como TDA-H hay que tener en cuenta si alguno de sus progenitores también tuvo este trastorno. Para ello se ha recogido una muestra de **93** niños diagnosticados con TDA-H y **107** sin este diagnóstico. De los niños diagnosticados, 47 de ellos tienen, al menos, a uno de sus dos progenitores con TDA-H, mientras que de los niños no diagnosticados con TDA-H, tan sólo 15 de ellos tenían alguno de sus progenitores con este trastorno.

Solución con BioestadisticaR2

Del mismo modo que en el ejemplo anterior es necesario construir la tabla de contingencia con estos datos. Nuevamente, por comodidad, se denota por E=Diagnóstico positivo por TDA-H para un niño, de modo que E^c=Diagnóstico negativo por TDA-H para un niño. Del mismo modo se denota FR=Algún progenitor fue diagnosticado por TDA-H, y por tanto FR^c=Ningún progenitor fue diagnosticado por TDA-H. Con esta notación se analizan las hipótesis nula y alternativa siguientes:

$$\begin{cases} \mathcal{H}_0 : E \text{ es independiente del } FR \\ \mathcal{H}_1 : E \text{ está asociado al } FR \end{cases}$$

y la tabla de contingencia para este problema quedaría de siguiente forma:

Enfermedad / Factor de Riesgo	FR	FR^c	**Totales**
E	$O_{11} = 47$	$O_{12} = 46$	$F_1 = 93$
E^c	$O_{21} = 15$	$O_{22} = 92$	$F_2 = 107$
Totales	$C_1 = 62$	$C_2 = 138$	$T = 200$

(El color rojo indica que estos son los primeros números introducidos en la tabla. Puede servir como regla nemotécnica para identificar que el estudio es Prospectivo).

Es evidente que el tipo de diseño realizado para esta investigación es de tipo **retrospectivo** ya que partimos de dos muestras de $F_1 = 93$ niños con el trastorno (E) y $F_2 = 107$ niños sin el trastorno (E^c) que clasificamos, en cada grupo, según si alguno de sus progenitores también están diagnosticados (FR) o no (FR^c) por TDA-H. Al ser un diseño retrospectivo, en el supuesto de que se rechaze la hipótesis nula de independencia, la única medida de asociación que se puede calcular es la **razón de producto cruzado (Odss ratio)**. Si la *prevalencia de la enfermedad estudiada fuera menor de* 0.1 también se podría estimar de forma puntual el **riesgo relativo** como la Odss ratio. En este ejemplo, nuestro estudio es retrospectivo y habrá que indicar a la función **tablas2x2()** el parámetro *estudio="R"*.

```
tabla2x2(o=c(47,46,15,92), fcat=c("Niños con TDA-H","Niños sin TDA-H"),
         ccat=c("Algún padre con TDA-H","Padres sin TDA-H"),estudio="R",
         tablas=c("F,C"))
```

```
##
## # Análisis de tablas 2x2
## # ----------------------
##
## # Frecuencias observadas
##                     Algún padre con TDA-H   Padres sin TDA-H Total
##   Niños con TDA-H                      47                 46    93
##   Niños sin TDA-H                      15                 92   107
##   Total                                62                138   200
```

```
##
##
## # Test Chi-cuadrado para un estudio retrospectivo
##
##    ² = 31.004,   gl = 1,  p < 0.001, (cpc = 1)
##   Validez: Frecuencia mínima esperada = 28.83  > 7.7
##
##   Test exacto de Fisher (bilateral): p < 0.001
##
##   --- Otros criterios ²:
##    ² = 31.021,   gl = 1,  p < 0.001, (sin cpc)
##    ² = 29.338,   gl = 1,  p < 0.001, (cpc de Yates = 100.00)
##
## # Medidas de asociación para un estudio retrospectivo
##
##   Riesgo atribuible*:
##   Ra=0.637; 95%-IC(Ra)= (0.427, 0.770)
##   * La estimación de Ra para estudios retrospectivos es una aproximación
##     válida si la prevalencia de la enfermedad es baja: P(E) < 10%
##
##   Razón del producto cruzado (odds ratio):
##   OR=6.267; 95%-IC(OR)= (3.110, 11.948)
##   * La estimación para OR sirve de aproximación al riesgo relativo siempre que
##     la prevalencia de la enfermedad sea P(E) < 10%
```

De la salida obtenida cabe destacar que:

- En primer lugar *muestra la tabla de contingencia con las frecuencias observadas*. En este caso coincide con la diseñada manualmente, como debe de ser.
- En segundo lugar comprueba la *condición de validez*, que para este tipo de estudio, **retrospectivo**, y tamaño de muestra $T \leq 500$, consiste en que la frecuencia mínima esperada exceda de 7.7. En este caso su valor es $28.83 \geq 7.7$.
- A continuación realiza el *test exacto de Fisher con corrección por continuidad (cpc), sin cpc y con la cpc de Yates*. Para este ejemplo el test exacto de Fisher proporciona un valor $p = 0.001 < 0.05$ que permite rechazar la hipótesis nula de independencia con total seguridad. Esto confirma que **el diagnóstico de un niño en edad temprana por TDA-H está relacionado con que alguno de sus progenitores también haya sido diagnosticado**.
- En este caso no se puede proporcionar una estimación de la prevalencia por ser un estudio restrospectivo
- Finalmente muestra la estimación puntual y por intervalos de confianza al 95% de confianza de todas las *medidas de asociación* que se pueden obtener para este tipo de estudio. En este ejemplo, como el diseño es **retrospectivo** tienen sentido el **riesgo atribuible**, y la **razón de producto cruzado (odds ratio)**. Sus estimaciones puntuales en el contexto de este ejemplo han sido realizadas en el ejemplo de la sección 9.3.2 anterior ya que sus valores no han cambiado al trabajar con las mismas cantidades. Como la prevalencia de este trastorno no es inferior a 0.1 no se puede estimar ni interpretar el riesgo relativo a través de la odss ratio.

Como conclusión, se ha confirmado que **la herencia genética influye en el diagnóstico temprano de un niño por TDA-H**. Además se ha dado información relevante de la **fuerza que tiene como factor de riesgo que algún progenitor haya sido diagnosticado por este trastorno**, gracias a la interpretación de las distintas medidas de asociación consideradas.

9.3.4 Ejemplo práctico con un diseño Prospectivo

Para esta ilustración práctica de un estudio **prospectivo** se continua con el objetivo planteado anteriormente, pero desde la perspectiva de este diseño. En efecto, se pretende contrastar que para el diagnóstico de un niño como TDA-H hay que tener en cuenta si alguno de sus progenitores también tuvo este trastorno. Para ello se ha recogido una muestra de **62** niños cuyos progenitores fueron diagnosticados con TDA-H y **138** cuyos progenitores no tuvieron este diagnóstico. De los niños con algún progenitor diagnosticado por TDA-H, 47 de ellos también han sido diagnosticados con este trastorno, mientras que de los niños cuyos progenitores no fueron diagnosticados con TDA-H, 46 de ellos han sido diagnosticados con este trastorno.

Solución con BioestadisticaR2

Del mismo modo que en el ejemplo anterior es necesario construir la tabla de contingencia con estos datos. Nuevamente, por comodidad, se denota por E=Diagnóstico positivo por TDA-H para un niño, de modo que E^c=Diagnóstico negativo por TDA-H para un niño. Del mismo modo se denota FR=Algún progenitor fue diagnosticado por TDA-H, y por tanto FR^c=Ningún progenitor fue diagnosticado por TDA-H. Con esta notación se analizan las hipótesis nula y alternativa siguientes:

$$\begin{cases} \mathcal{H}_0 : E \text{ es independiente del } FR \\ \mathcal{H}_1 : E \text{ está asociado al } FR \end{cases}$$

y la tabla de contingencia para este problema quedaría de siguiente forma:

Enfermedad / Factor de Riesgo	FR	FR^c	**Totales**
E	$O_{11} = 47$	$O_{12} = 46$	$F_1 = 93$
E^c	$O_{21} = 15$	$O_{22} = 92$	$F_2 = 107$
Totales	$C_1 = 62$	$C_2 = 138$	$T = 200$

(El color rojo indica que estos son los primeros números introducidos en la tabla. Puede servir como regla nemotécnica para identificar que el estudio es Prospectivo).

Es evidente que el tipo de diseño realizado para esta investigación es de tipo **prospectivo** ya que partimos de dos muestras de $C_1 = 62$ niños con el factor de riesgo (FR) y $C_2 = 138$ niños sin el factor de riesgo (FR^c) que clasificamos, en cada grupo, según si están diagnosticados (E) o no (E^c) por TDA-H. Al ser un diseño prospectivo, en el supuesto de que se rechaze la hipótesis nula de independencia, tendrá sentido calcular e **interpretar todas las medidas de asociación** introducidas en las clases de teoría. En este ejemplo nuestro estudio es prospectivo y habrá que indicar a la función **tablas2x2()** el parámetro *estudio="P"*.

```
tabla2x2(o=c(47,46,15,92), fcat=c("Niños con TDA-H","Niños sin TDA-H"),
         ccat=c("Algún padre con TDA-H","Padres sin TDA-H"),estudio="P",
         tablas=c("F,C"))
```

```
##
## # Análisis de tablas 2x2
## # ----------------------
##
## # Frecuencias observadas
##                     Algún padre con TDA-H   Padres sin TDA-H Total
##   Niños con TDA-H                      47                 46    93
##   Niños sin TDA-H                      15                 92   107
##   Total                                62                138   200
##
##
```

```
## # Test Chi-cuadrado para un estudio prospectivo
##
##    ² = 31.004,   gl = 1,  p < 0.001, (cpc = 1)
##   Validez: Frecuencia mínima esperada = 28.83  > 7.7
##
##   Test exacto de Fisher (bilateral): p < 0.001
##
##   --- Otros criterios ²:
##    ² = 31.021,   gl = 1,  p < 0.001, (sin cpc)
##    ² = 29.338,   gl = 1,  p < 0.001, (cpc de Yates = 100.00)
##
## # Medidas de asociación para un estudio prospectivo
##   [!] Las medidas de riesgo se calculan como riesgo de la categoría
##       en la 1a columna (frente a la 2a) para la categoría en la 1a
##       fila (frente a la 2a)
##
##   Riesgo absoluto (diferencia de Berkson; método de Agresti-Caffo):
##   d=0.425; 95%-IC(d)=(0.283, 0.547)
##
##   Riesgo relativo:
##   Rr=2.274; 95%-IC(Rr)=(1.714, 2.963)
##
##   Razón del producto cruzado (odds ratio):
##   OR=6.267; 95%-IC(OR)= (3.110, 11.948)
```

De la salida obtenida cabe destacar que:

- En primer lugar *muestra la tabla de contingencia con las frecuencias observadas*. En este caso coincide con la diseñada manualmente, como debe de ser.
- En segundo lugar comprueba la *condición de validez*, que para este tipo de estudio, **prospectivo**, y tamaño de muestra $T \leq 500$, consiste en que la frecuencia mínima esperada exceda de 7.7. En este caso su valor es $28.83 \geq 7.7$.
- A continuación realiza el *test exacto de Fisher con corrección por continuidad (cpc), sin cpc y con la cpc de Yates*. Para este ejemplo el test exacto de Fisher proporciona un valor $p = 0.001 < 0.05$ que permite rechazar la hipótesis nula de independencia con total seguridad. Esto confirma que **el diagnóstico de un niño en edad temprana por TDA-H está relacionado con que alguno de sus progenitores también haya sido diagnosticado**.
- En este caso no se puede proporcionar una estimación de la prevalencia por ser un estudio prospectivo.
- Finalmente muestra la estimación puntual y por intervalos de confianza al 95% de confianza de todas las *medidas de asociación* que se pueden obtener para este tipo de estudio. En este ejemplo, como el diseño es **prospectivo** tienen sentido el **riesgo absoluto (diferencia de Berkson)**, el **riesgo relativo** y **razón de producto cruzado (odds ratio)**. Sus estimaciones puntuales en el contexto de este ejemplo han sido realizadas en el ejemplo de la sección 9.3.2 anterior ya que sus valores no han cambiado al trabajar con las mismas cantidades.

Como conclusión, se ha confirmado que **la herencia genética influye en el diagnóstico temprano de un niño por TDA-H**. Además se ha dado información relevante de la **fuerza que tiene como factor de riesgo que algún progenitor haya sido diagnosticado por este trastorno**, gracias a la interpretación de las distintas medidas de asociación consideradas.