Test estadísticos para variables cualitativas. Test exacto de Fisher, chi-cuadrado de Pearson, McNemar y Q-Cochran

Joaquín Amat Rodrigo

Enero 2016

Índice

Introducción

Los contrastes de hipótesis para variables cualitativas se realizan mediante test de frecuencia o proporciones. Dentro de esta categoría existen distintos tipos de test, la utilización de uno u otro depende de qué tipo de información se quiera obtener:

- Test de distribución esperada o *"goodness of fit"*: Se emplean para comparar la distribución observada frente a una distribución esperada o teórica.
- Test de diferencia de frecuencias o test de independencia: Se emplean para estudiar si la frecuencia de observaciones es significativamente distinta entre dos o más grupos.

En los test de "goodness of fit" solo hay una variable asociada a cada observación, mientras que en los test de independencia hay dos variables asociadas a cada observación. También se emplean distintos test dependiendo del tipo de datos (independientes o dependientes) con los que se vaya a trabajar. Las siguientes tablas muestran algunos de los más empleados.

Datos independientes

Tipo de test	Distribución esperada	Comparación de grupos
Exacto	Test binomial exacto / Test multinomial exacto	Fisher's exact
Aproximado	Chi-square goodness of fit / G-test goodness of fit	Chi-square test of independence

Datos pareados

Tipo de tabla	test
Tablas 2x2	McNemar
Tablas 2Xk	Q de Cochran

Test exactos y test aproximados

Los test exactos calculan la probabilidad de obtener los resultados observados de forma directa generando todos los posibles escenarios y calculando la proporción en los que se cumple la condición estudiada (son test de permutaciones). Los test aproximados calculan primero un estadístico y luego emplean la distribución teórica de dicho estadístico para obtener la probabilidad de que adquiera valores iguales o más extremos.

Existe bastante controversia en cuanto a si se deben de utilizar test exactos o aproximados. En la era pre-computacional, los test exactos se complicaban mucho cuando el tamaño total de muestras aumentaba, sin embargo, por medio de la computación esta barrera se ha eliminado. Los test exactos son más precisos cuando el tamaño total de observaciones es bajo o alguno de los grupos tiene pocas observaciones, una vez alcanzado un número alto de observaciones las diferencias son mínimas. En el libro *Handbook of Biological Statistics John H. McDonald* se recomienda utilizar test exactos cuando el número total de observaciones es menor a 1000 o cuando, aunque el número total sea mayor a 1000, haya algún grupo cuyo número de eventos esperados sea pequeño (normalmente menor que 5). En el caso de aplicar test aproximados sobre tamaños pequeños se suelen emplear correcciones, las más frecuentes son la *corrección de continuidad de Yate* o *la corrección de William*.

Se puede considerar a los test basados el estadístico χ^2 como una generalización del contraste de proporciones basado en la aproximación a la normal (*Z-test de una proporción* y *Z-test de dos proporciones*) cuando hay 2 o más variables cualitativas o alguna de ellas tiene 2 o más niveles. En aquellos casos en los que ambos test se pueden aplicar, el resultado de un *Z-test* y un *test* χ^2 es equivalente. Esto es debido a que en la distribución chi-cuadrado con 1 grado de libertad el estadístico χ^2 es igual al estadístico Z de una distribución normal, elevado al cuadrado.

Test exacto de Fisher

La prueba de Fisher es el test exacto utilizado cuando se quiere estudiar si existe asociación entre dos variables cualitativas, es decir, si las proporciones de una variable son diferentes dependiendo del valor que adquiera la otra variable. En la gran mayoría de casos, el test de Fisher se aplica para comparar dos variables categóricas con dos niveles cada una (tabla 2x2). Es posible utilizarlo con tablas 2xK niveles pero los requerimientos de cálculo son altos.

El test de Fisher es más preciso que sus equivalentes aproximados (*test chi-square de independencia* o *G–test de independencia*) cuando el número de eventos esperado por nivel es pequeño. Se recomienda utilizarlo siempre que sea posible (tiempo de computación) aunque para observaciones totales >1000 los resultados de los test aproximados son muy parecidos.

Es importante tener en cuenta que el test de Fisher está diseñado para situaciones en las que las frecuencias marginales de filas y columnas (los totales de cada fila y columna) son fijas, se conocen de antemano. Esta condición es relevante en los experimentos biológicos ya que no es común poder cumplirla. Si esta condición no se satisface el test de Fisher deja de ser exacto, por lo general pasando a ser más conservativo. *En varios artículos se menciona que el test de Barnard es más potente que el de Fisher cuando las frecuencias marginales no son fijas. También parece ser que aunque el test deja de ser exacto no significa que no se pueda aplicar.*

Ejemplo de experimentos con y sin frecuencias marginales fijas:

Frecuencias marginales fijas: Supóngase que se quiere saber si la preferencia que tienen dos especies de pájaros (estorninos y gorriones) para refugiarse en casetas artificiales es diferente dependiendo del material de fabricación (madera o metal). Para ellos se disponen en una pajarera 5 casetas de metal y 5 de madera y se sueltan en el interior de la jaula 4 gorriones y 6 estorninos. En este experimento se sabe que las frecuencias marginales van a ser 5, 5, 4, 6 lo que no se sabe es como se van a distribuir las observaciones dentro de la tabla.

..	metal	madera	total
gorrión	?	?	4
estornino	?	?	6
total	5	5	10

Frecuencias marginales no fijas: Supóngase que se quiere determinar si un fármaco acelera la cicatrización. Para ello se selecciona a 50 pacientes que se reparten aleatoriamente en dos grupos iguales (tratamiento y placebo), tras una semana de tratamiento se determina si la cicatrización ha finalizado (si / no). En este caso las frecuencias marginales de los tratamientos son fijas, 25 para cada grupo, sin embargo no se sabe cuántos en cada grupo van a haber cicatrizado o no, por lo que las frecuencias marginales del resultado de cicatrización no son fijas.

..	cicatrizado	no cicatrizado	total
placebo	?	?	25
tratamiento	?	?	25
total	?	?	50

Hipótesis

H_o: Las variables son independientes por lo que una variable no varía entre los distintos niveles de la otra variable.

H_a: Las variables son dependientes, una variable varía entre los distintos niveles de la otra variable.

Condiciones del test

Independencia: las observaciones de la muestra deben ser independientes unas de otras.

- Muestreo aleatorio.
- Tamaño de la muestra < 10% población.
- Cada observación contribuye únicamente a uno de los niveles.

Las frecuencias marginales de columnas y filas tienen que ser fijas. Si esta condición no se cumple, el test de Fisher deja de ser exacto.

Cálculo del *p-value*

El test exacto de Fisher se basa en la distribución hipergeométrica, que permite calcular la probabilidad exacta de obtener una determinada distribución de eventos dentro de una tabla. Supóngase la siguiente tabla de contingencia:

..	**nivel-A1**	**nivel-A2**	**total**
nivel-B1	a	b	a+b
nivel-B2	c	d	c+d
total	a+c	b+d	n= a+b+c+d

Si las frecuencias marginales son fijas (conocidas), sabiendo el valor de una celda se puede calcular el valor de las demás. La probabilidad de que *a* adquiera un determinado valor (dentro de las limitaciones impuestas por las frecuencias marginales) se corresponde con la fórmula de la distribución hipergeométrica:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{a!\ b!\ c!\ d!\ n!}$$

El test de Fisher calcula las probabilidades de todas las posibles tablas y suma las de aquellas tablas que tengan probabilidades menores o iguales que la tabla observada, generando así el *p-value* de dos colas.

Fuerza de asociación (tamaño del efecto)

Dado que el test de Fisher contrasta si las variables están relacionadas, al tamaño del efecto se le conoce como fuerza de asociación. Existen múltiples medidas de asociación, entre las que destacan *phi* o *Cramer's V*. Los límites empleados para su clasificación son:

- pequeño: 0.1
- mediano: 0.3
- grande: 0.5

En R se pueden calcular mediante la función `assocstats()` del paquete *vcd*.

Ejemplo

Se quiere estudiar si la reacción alérgica a un compuesto y una determinada mutación en un gen están relacionados. Para ello se realiza un test alérgico sobre un grupo de individuos seleccionados al azar y se genotipa el estado del gen de interés ¿Existe un diferencia significativa en la incidencia de la mutación entre los alérgicos y no alérgicos?

```
datos <- data.frame(sujeto = c("No alérgico", "No alérgico", "No alérgico",
    "No alérgico", "alérgico", "No alérgico", "No alérgico", "alérgico",
    "alérgico", "No alérgico", "alérgico", "alérgico", "alérgico", "alérgico",
    "alérgico", "No alérgico", "No alérgico", "No alérgico", "No alérgico",
    "alérgico", "alérgico", "alérgico", "alérgico", "No alérgico", "alérgico",
    "No alérgico", "No alérgico", "alérgico", "alérgico", "alérgico"),
    mutacion = c(FALSE, FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, TRUE,
        TRUE, TRUE, TRUE, TRUE, TRUE, FALSE, FALSE, TRUE, FALSE, TRUE, FALSE,
        TRUE, FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, TRUE, FALSE, TRUE))
head(datos)
```

```
##        sujeto mutacion
## 1 No alérgico    FALSE
## 2 No alérgico    FALSE
## 3 No alérgico    FALSE
## 4 No alérgico    FALSE
## 5    alérgico     TRUE
## 6 No alérgico    FALSE
```

El test de Fisher trabaja con frecuencia de eventos, por lo tanto con tablas de contingencia en las que se sumariza el número de eventos de cada tipo.

```
tabla <- table(datos$sujeto, datos$mutacion, dnn = c("Sujeto", "Estado gen"))
tabla
```

```
##              Estado gen
## Sujeto        FALSE TRUE
##   alérgico        6   10
##   No alérgico    11    3
```

Test de Fisher

```
fisher.test(x = tabla, alternative = "two.sided")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tabla
## p-value = 0.03293
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.02195148 1.03427479
## sample estimates:
## odds ratio
##  0.1749975
```

Fuerza de asociación

```
library(vcd)
```

```
## Loading required package: grid
```

```
assocstats(x = tabla)
```

```
##                     X^2 df P(> X^2)
## Likelihood Ratio 5.3356  1 0.020894
## Pearson          5.1293  1 0.023525
##
## Phi-Coefficient   : 0.413
## Contingency Coeff.: 0.382
## Cramer's V        : 0.413
```

En este ejemplo no se satisface la condición de frecuencias marginales fijas y por lo tanto el test de Fisher no es exacto. Aun así, hay evidencias para rechazar la H_0 y considerar que las dos variables sí están relacionadas. El tamaño de la fuerza de asociación (tamaño de efecto) cuantificado por *phi* o *Cramer's V* es mediano.

χ^2 de Pearson (test de independencia)

El test χ^2 *de independencia*, también conocido como χ^2 *de Pearson* se emplea para estudiar si existe asociación entre dos variables categóricas, es decir, si las proporciones de una variable son diferentes dependiendo del valor que adquiera la otra variable, cuando los datos son independientes. Se trata por lo tanto de una expansión del *Z-test para dos proporciones* cuando una de las variables estudiadas tiene dos o más niveles. Cuando ambas variables tienen dos niveles (tabla 2x2) ambos test χ^2 *goodness of fit* y *Z-test para dos proporción* son equivalentes.

Es el test aproximado equivalente a su versión exacta *test de Fisher*. Debido a los requerimientos de cálculo del test de Fisher, cuando hay muchas observaciones o muchos niveles, se emplea el test χ^2 de independencia. Es importante tener en cuenta que cuando el número de observaciones esperadas para alguno de los niveles es igual o menor a 5 la aproximación por el test χ^2 no es buena.

El test de independencia cuantifica y sumariza cómo de distinto es el número de eventos observados en cada nivel con respecto al número esperado acorde con Ho. Esto permite identificar si la desviación total es mayor que la que cabría esperar simplemente por azar.

Hipótesis

H_0: Las variables son independientes por lo que una variable no varía entre los distintos niveles de la otra variable.

H_a: Las variables son dependientes, una variable varía entre los distintos niveles de la otra variable.

Condiciones del test

Independencia: las observaciones de la muestra deben ser independientes unas de otras.

- Muestreo aleatorio.
- Tamaño de la muestra < 10% población.
- Cada observación contribuye únicamente a uno de los niveles.

Tamaño: cada nivel debe tener al menos 5 eventos **esperados** (acorde a Ho) y el número de observaciones totales (n) >30. En caso de no cumplirse esta condición, el test pierde precisión.

En el libro *Bioestadística de Francisxa Ruis Díaz* consideran que esta regla es muy estricta y rara vez se cumple en la práctica. Proponen unas condiciones más relajadas con las que no se pierde demasiada precisión: para ningún nivel el número de eventos esperados acorde a H0 es menor de 1 y como máximo un 20% de los niveles tiene menos de 5 eventos esperados.

- En caso de no cumplirse esta condición o estar en el límite, se recurre a los test exactos o, si no es posible, a la simulación.
- Aun cuando se cumplen las condiciones, son más precisos los test exactos y por lo tanto más recomendables.

Estadístico χ^2

$$\chi^2 = \sum_{i,j} \frac{(\text{observado}_{ij} - \text{esperado}_{ij})^2}{\text{esperado}_{ij}}$$

El valor esperado de cada grupo se obtiene multiplicando las frecuencias marginales de la fila y columna en la que se encuentra la celda y dividiendo por el total de observaciones. Se suman las diferencias de todos los niveles. Elevar al cuadrado las diferencias permite hacerlas todas positivas y además magnificar aquellas más grandes.

Grados de libertad

La distribución chi-cuadrado tiene un único parámetro, los grados de libertad, que determina su forma, centro y dispersión.

$$df = (niveles\ variable\ A - 1)x(niveles\ variable\ B - 1) = (columnas - 1)x(filas - 1)$$

Cálculo de *p-value*

La distribución chi-cuadrado es siempre positiva, por lo que para calcular el *p-value* solo se tiene en cuenta la cola superior.

Fuerza de asociación (tamaño del efecto

Dado que el test contrasta si las variables están relacionadas, al tamaño del efecto se le conoce como fuerza de asociación. Existen múltiples medidas de asociación, entre las que destacan *phi* o *Cramer's V*. Los límites empleados para su clasificación son:

- pequeño: 0.1
- mediano: 0.3
- grande: 0.5

En R se pueden calcular mediante la función `assocstats()` del paquete *vcd*.

Comparaciones *post-hoc*

Cuando una de las variables tiene más de dos niveles y el test χ^2 *de independencia* resulta significativo puede ser de interés estudiar en que niveles se encuentran las diferencias significativas. Para ello hay dos opciones:

Estudiar los valores residuales de Pearson: Se trata de las diferencias entre los valores observados y los esperados. Si esta diferencia se divide por la raíz cuadrada de los esperados se obtiene lo que se conoce como residuos estandarizados que se pueden comparar como Z-factor de una normal. Para un α de 0.05, si el valor residual estandarizado de un nivel supera el 1.95, se considera significativo. En R los residuos y los residuos estandarizados se almacenan dentro del test: `chisq.test().residuals` y `chisq.test().stdres`

Dividir la tabla en tablas 2x2 que abarquen todas las combinaciones de los niveles y realizar con cada tabla un nuevo test χ^2 *de independencia* con corrección del nivel de significancia α. El número total de comparaciones (k) que se puede hacer con una tabla es de:

$$k = \frac{r!}{2!\,(r-2)!} \cdot \frac{c!}{2!\,(c-2)!} = \frac{r\,(r-1)\,c\,(c-1)}{4}.$$

Ejemplo 1

Un estudio intenta comparar si existe relación entre el estatus civil de las personas y la incidencia de la obesidad. Para ello se dispone de los siguientes datos. ¿Es significativa la relación entre ambas variables para un nivel de significancia del 5%?

Obesidad	soltero	pareja	casado
Obeso	81	103	147
No_obeso	359	326	277

Solución manual

Hipótesis

$\boldsymbol{H_0}$: Obesidad y estado civil son independientes, el % de obesos no varía entre los diferentes niveles de la variable estado civil.

$\boldsymbol{H_a}$: Obesidad y estado civil son dependientes, el % de obesos sí varía entre los diferentes niveles de la variable estado civil.

Calcular el número de eventos esperados en cada combinación de niveles siendo Ho cierta

Calcular el % de obesidad en toda la muestra: % obesidad = $\frac{331}{1293} = 0.256$

Si H_0 es cierta, la cantidad de obesos en cada nivel de la variable estatus civil será igual al número marginal de ese nivel multiplicado por el % obesidad.

Obesidad	soltero	pareja	casado
Obeso	113	110	108
No_obeso	327	319	316

Otra forma de saber los eventos esperados en cada grupo = (total fila x total columna) / total tabla.

Comprobar condiciones para χ^2-test

Independencia: las observaciones de la muestra deben ser independientes unas de otras.

- Muestreo aleatorio.
- Tamaño de la muestra < 10% población.
- Cada observación contribuye únicamente a uno de los niveles.

Tamaño: todas las posible combinaciones de niveles tienen al menos 5 eventos acorde a lo esperado según H_0.

Cálculo χ^2 y grados de libertad

$$\chi^2 = \frac{(81-113)^2}{113} + \ldots + \frac{(277-316)^2}{316} = 31.68$$

$$df = (2-1)x(3-1) = 2$$

Cálculo de p-value

p-value = `pchisq(q= 31.68, df= 2, lower.tail = FALSE)` = 1.320612510^{-7}

Solución con R

La función propia `chisq.test()` permite realizar un test de independencia (teórico o simulado) dando como argumento la tabla de observaciones.

Se generar la tabla de frecuencias con los datos.

```
fila1 <- c(81, 103, 147)
fila2 <- c(359, 326, 277)
tabla <- as.table(rbind(fila1, fila2))
dimnames(tabla) = list(Peso = c("Obeso", "No obeso"), Estado_civil = c("soltero",
    "pareja", "casado"))
tabla
```

```
##           Estado_civil
## Peso       soltero pareja casado
##   Obeso         81    103    147
##   No obeso     359    326    277
```

Se realizar el test.

```
chisq.test(x = tabla)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabla
## X-squared = 30.829, df = 2, p-value = 2.021e-07
```

Solución mediante simulación.

```
chisq.test(tabla, simulate.p.value = TRUE, B = 5000)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 5000
##  replicates)
##
## data:  tabla
## X-squared = 30.829, df = NA, p-value = 2e-04
```

Dado que el test ha resultado significativo, se quiere determinar para que niveles el número de observaciones difiere significativamente de lo esperado.

Análisis de residuos de Pearson:

```
chisq.test(x = tabla)$residuals
```

```
##           Estado_civil
## Peso          soltero     pareja     casado
##   Obeso    -2.9809729 -0.6509186  3.6914422
##   No obeso  1.7485759  0.3818151 -2.1653222
```

```
chisq.test(x = tabla)$stdres
```

```
##           Estado_civil
## Peso          soltero     pareja     casado
##   Obeso    -4.2549504 -0.9231681  5.2203206
##   No obeso  4.2549504  0.9231681 -5.2203206
```

Las mayores desviaciones respecto a los valores esperados se dan en los niveles de soltero y casado. Para tener un estudio más exacto de si hay diferencias significativas en estos niveles, se divide la tabla en tablas 2X2 y se repite el test χ^2 corrigiendo el nivel de significancia.

Comparación 2 a 2 (*a modo de ejemplo se realiza con solteros y casados, pero se haría con todos*)

```
solteros_casados <- tabla[, c(1, 3)]
solteros_casados
```

```
##           Estado_civil
## Peso       soltero casado
##   Obeso          81    147
##   No obeso      359    277
```

```
chisq.test(solteros_casados)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  solteros_casados
## X-squared = 28.56, df = 1, p-value = 9.083e-08
```

```
assocstats(solteros_casados)
```

```
##                     X^2 df   P(> X^2)
## Likelihood Ratio 29.687  1 5.0775e-08
## Pearson          29.391  1 5.9140e-08
##
## Phi-Coefficient   : 0.184
## Contingency Coeff.: 0.181
## Cramer's V        : 0.184
```

Se confirma que al menos entre los grupos casados y solteros sí hay una asociación significativa de las variables "estado civil" y "obesidad" con un tamaño de asociación *Cramer's V* pequeño. Por lo tanto se puede afirmar que el porcentaje de gente obesa está asociado al estado civil.

Test de McNemar

El test de McNemar es la alternativa a los test χ^2 *de Pearson* y al test exacto de *Fisher* cuando: los datos son pareados, se trata de tablas 2x2 y ambas variables son dicotómicas (binomiales). El test de McNemar estudia si la probabilidad de evento verdadero para una variable es igual en los dos niveles de otra variable.

Condiciones

- Se trata de datos pareados.
- Se estudian dos variables, ambas de tipo binomial (dicotómicas). Tabla 2x2.
- La suma de eventos que pasan de positivo a negativo y de negativo a positivo ha de ser > 25, de lo contrario se emplea un test binomial en el que el número de aciertos es el número de eventos que han pasado de positivos a negativos y el número total de intentos es la suma de todos los que han cambiado (de positivos a negativos y de negativos a positivos).

Hipótesis

Para entender el funcionamiento, supóngase que un grupo de sujetos pasa un test cuyo resultado es binomial (positivo y negativo) antes y después de un tratamiento. El objetivo de la prueba es determinar si el tratamiento hace cambiar los test de positivo a negativo o viceversa.

..	**Después Positivo**	**Después Negativo**	**Total**
Antes Positivo	a	b	a+b
Antes Negativo	c	d	c+d
Total	a+c	b+d	n=a+b+c+d

Si el valor de una variable es independiente de la otra, se esperaría que las proporciones de pasar negativo a positivo fuesen iguales a las de pasar de positivo a negativo. Esto significa que $p_c + p_d = p_b + p_d$ y $p_c + p_d = p_b + p_d$; lo que queda como $p_b = p_c$ (hipótesis nula).

$$H_0: p_b = p_c$$

$$H_A: p_b \neq p_c$$

Estadístico

El estadístico del test de McNemar, siempre y cuando se cumpla la condición mínima de eventos, sigue una distribución χ^2 con 1 grado de libertad.

$$\chi^2 = \frac{(b-(b+c)/2)^2}{(b+c)/2} + \frac{(c-(b+c)/2)^2}{(b+c)/2} = \frac{(b-c)^2}{b+c}$$

Si $b + c < 25$ entonces se emplea un test binomial(x= c, n = c+b, p= 0.5)

Fuerza de asociación (tamaño de efecto)

Dado que el test contrasta si las variables están relacionadas, al tamaño del efecto se le conoce como fuerza de asociación. Existen múltiples medidas de asociación, entre las que destacan *phi* o *Cramer's V*. Los límites empleados para su clasificación son:

- pequeño: 0.1
- mediano: 0.3
- grande: 0.5

En R se pueden calcular mediante la función `assocstats()` del paquete *vcd*.

Ejemplo 1

Supóngase que se quiere comprobar si un tratamiento de hipnosis es capaz de hacer que las personas contesten "Sí" con mayor frecuencia. Para ello se selecciona un grupo de individuos a los que se les realiza una pregunta cuya respuesta puede ser SI/NO antes y después de someterse al tratamiento de hipnosis.

```
datos <- data.frame(sujeto = rep(1:15, each = 2), tratamiento = c("pre", "post",
    "pre", "post", "pre", "post", "pre", "post", "pre", "post", "pre", "post",
    "pre", "post", "pre", "post", "pre", "post", "pre", "post", "pre", "post",
    "pre", "post", "pre", "post", "pre", "post", "pre", "post"), respuesta =
c("NO", "SI", "SI", "SI", "NO", "SI", "SI", "NO", "SI", "SI", "NO", "SI", "NO",
    "SI", "NO", "SI", "NO", "SI", "SI", "SI", "NO", "NO", "SI", "SI", "NO",
    "SI", "NO", "NO", "NO", "SI"))
head(datos)
```

```
##   sujeto tratamiento respuesta
## 1      1         pre        NO
## 2      1        post        SI
## 3      2         pre        SI
## 4      2        post        SI
## 5      3         pre        NO
## 6      3        post        SI
```

Hipótesis

Ho: La respuesta de los sujetos es independiente del tratamiento. La proporción de sujetos que pasan de responder Sí a No es igual a la proporción de sujetos que pasan de responder No a Sí.

Ha: El tratamiento sí influye en la respuesta por lo que la proporción de sujetos que pasan de No a Sí es diferente de la de sujetos que pasan de Sí a No.

Se genera una tabla de contingencia con los datos

En este caso se trata de una tabla de contingencia en la que cada sujeto se clasifica según si sus respuestas son iguales antes y después del tratamiento y de cómo han variado. Para generar este tipo de tablas primero se tiene que generar un *data frame* en formato *tabla ancha*.

```
library(tidyr)
datos <- spread(data = datos, key = tratamiento, value = respuesta)
head(datos)
```

```
##   sujeto post pre
## 1      1   SI  NO
## 2      2   SI  SI
## 3      3   SI  NO
## 4      4   NO  SI
## 5      5   SI  SI
## 6      6   SI  NO
```

```
tabla <- table(Pre_Tratamiento = datos$pre, Post_Tratamiento = datos$post)
tabla
```

```
##                Post_Tratamiento
## Pre_Tratamiento NO SI
##              NO  2  8
##              SI  1  4
```

Condiciones

- Los datos son pareados.
- Se estudia la relación entre dos variables, ambas dicotómicas (binomiales). Tabla 2x2
- Número total de observaciones que cambian el valor de una variable en función de la otra es < 25. Dado que no se cumple el tamaño mínimo de 25, la aproximación del estadístico χ^2 no es buena, se tiene que recurrir a un test binomial. *Para fines ilustrativos se van a realizar ambos test.*

Test de McNemar

La función `McNemar()` tiene por defecto una corrección para mejorar la precisión *del p-value* cuando el número de observaciones es pequeño. Aun así el test binomial es mejor en estos casos.

```
mcnemar.test(tabla)
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  tabla
## McNemar's chi-squared = 4, df = 1, p-value = 0.0455
```

Test binomial

```
binom.test(x = 1, n = 1 + 8, p = 0.5)
```

```
##
##  Exact binomial test
##
## data:  1 and 1 + 8
## number of successes = 1, number of trials = 9, p-value = 0.03906
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.002809137 0.482496515
## sample estimates:
## probability of success
##              0.1111111
```

Conclusión

Existen evidencias significativas para rechazar Ho en favor de que sí existe relación entre el tratamiento y la respuesta de los sujetos.

Ejemplo 2

Se quiere estudiar cómo impacta en la intención de voto de un grupo de personas una determinada noticia sobre la vida personal de un candidato político. Para ello, antes de hacer pública la noticia, se pregunta a las 1600 personas del grupo si votarían al candidato. Tras hacer pública la noticia se consulta de nuevo a las mismas personas. ¿Existen evidencias significativas de que la decisión de los miembros del grupo a cambiado debido a la información?

```
datos <- matrix(c(794, 86, 150, 570), nrow = 2, dimnames = list(Desinformados =
c("Sí", "No"), Informados = c("Sí", "No")))
datos
```

```
##              Informados
## Desinformados  Sí  No
##            Sí 794 150
##            No  86 570
```

Se trata de datos pareados, dos variables dicotómicas y el número de observaciones que han cambiado su valor en los dos niveles de la otra variable es >25. Se cumplen todas las condiciones para un test de McNemar.

```
mcnemar.test(datos)
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  datos
## McNemar's chi-squared = 16.818, df = 1, p-value = 4.115e-05
```

Existen evidencias significativas para aceptar que la noticia ha influido significativamente en la decisión de los votantes.

Test Q de Cochran

El test *Q-Cochran* es el equivalente al test de McNemar para más de dos grupos. Permite estudiar la independencia entre varias muestras pareadas, es decir, si la distribución de una variable binomial es la misma en todos los grupos. Un test *Q-Conchran* de dos grupos es equivalente a un test de McNemar.

Supóngase que se pregunta a un grupo de 17 personas si compraría ropa de una cierta marca (inicial). Luego se les muestra publicidad de esa marca y se les vuelve a hacer la misma pregunta (publicidad). Finalmente, se les enseñan los comentarios y opiniones de gente en internet sobre esa marca para repetirles la misma pregunta por última vez. El estudio quiere determinar si la publicidad o los comentarios y opiniones en internet cambian la intención de compra de la gente.

..	inicial	publicidad	internet
1	1	1	1
2	0	1	1
3	1	1	1
..	..	..	..
16	0	0	1
17	0	1	1

Para saber si hay diferencias significativas en la intención de compra según la información que se le proporcione al cliente se calcula un estadístico Q a partir de la tabla anterior.

$$Q = (k-1)\frac{k\sum_{j=1}^{k} G_j^2 - \left(\sum_{j=1}^{k} G_j\right)^2}{k\sum_{i=1}^{b} L_i - \sum_{i=1}^{b} L_i^2}$$

donde k es el número de columnas, b el de filas, G_j la suma de la columna j y L_i la suma de la fila i.

Este estadístico se aproxima a una χ^2 con $k-1$ grados de libertad. Cuando mayor sea el tamaño muestral (b en la fórmula) mejor será la aproximación.

Además de para contrastar diferencias de una misma variable dicotómica en diferentes momentos o escenarios, la prueba Q puede servir comparar varias variables dicotómicas en un mismo tiempo. Por ejemplo, para estudiar si a la hora de comprar un smartphone en un primer momento la gente prefiere un iPhone, un Samsung y u otro modelo. La hipótesis nula es que las variables tienden a coincidir para un mismo sujeto ya que no hay una preferencia antes de comparar los móviles. Para la persona que quiera un smarphone, la variable adquirirá el valor 1 (comprar) en las tres categorías, y si no quiere móvil será 0 para las tres.

Hipótesis

$\boldsymbol{H_0}$: La proporción de eventos verdaderos es la misma en los diferentes grupos.
$\boldsymbol{H_a}$: La proporción de eventos verdaderos es distinta en los diferentes grupos.

Condiciones

- La variable estudiada es binomial.
- Los datos son pareados.
- Se quieren comparar 2 o más grupos.

Comparaciones Post-Hoc

Si el test *Q-Cochran* es significativo, implica que al menos dos grupos difieren entre ellos. Para identificar cuales son se recurre a comparar dos a dos los diferentes grupos mediante el test de McNemar haciendo corrección de significancia (Bonferroni, Holm u otra).

Fuerza de asociación (tamaño del efecto)

No existe un método general de calcular la fuerza de asociación para el *Q-Cochran test*. Lo que se hace en su lugar es calcular la fuerza de asociación para cada comparación en el análisis *post hoc*, en este caso el de los test *McNemar*.

Ejemplo

Supóngase que se quiere estudiar el impacto en la intención de consumo mediante diferentes canales de publicidad. Para ello se selecciona a un grupo de personas a las que se les pregunta si comprarían una determinada marca antes de conocer nada sobre ella. A continuación se les muestran varios anuncios sobre la marca y se les vuelve a preguntar. Finalmente se les muestran comentarios y opiniones de internet sobre la marca y se les vuelve a consultar ¿Existen evidencias de que el comportamiento de los compradores haya cambiado dependiendo de la información recibida?

```
Respuesta <- c(1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1,
    0, 1, 1, 0, 1, 1, 0, 1, 1)
Sujeto <- factor(c(1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 7,
    7, 7, 8, 8, 8, 9, 9, 9, 10, 10, 10))
Canal <- factor(rep(c("inicial", "anuncio", "internet"), 10))
datos <- data.frame(Sujeto, Canal, Respuesta)
head(datos)
```

```
##   Sujeto    Canal Respuesta
## 1      1  inicial         1
## 2      1  anuncio         0
## 3      1 internet         1
## 4      2  inicial         0
## 5      2  anuncio         0
## 6      2 internet         1
```

Se trata de una variable binomial cuya distribución (proporción de 1's y 0's) se quiere estudiar en más de dos grupos. Dado que los datos están pareados se emplea un test *Q-Cochran*.

Test Q-Cochran

```
require(coin)
symmetry_test(Respuesta ~ factor(Canal) | factor(Sujeto), data = datos, teststat =
"quad")
```

```
##
##  Asymptotic General Symmetry Test
##
## data:  Respuesta by
##   factor(Canal) (anuncio, inicial, internet)
##   stratified by factor(Sujeto)
## chi-squared = 8.2222, df = 2, p-value = 0.01639
```

Existen evidencias significativas de que la intención de compra es distinta al menos entre dos grupos.

Comparaciones *Post-Hoc*

Se comparan dos a dos las 3 categorías del modo de información mediante *test de McNemar*.

```
require(coin)
require(dplyr)
require(tidyr)
# Inicial vs anuncio
inicial_anuncio <- spread(data = datos %>% filter(Canal %in% c("inicial",
"anuncio")),
    key = Canal, value = Respuesta)
tabla_inicial_anuncio <- table(inicial = inicial_anuncio$inicial, anuncio =
inicial_anuncio$anuncio)
tabla_inicial_anuncio
```

```
##        anuncio
## inicial 0 1
##       0 3 5
##       1 1 1
```

```
p_value_12 <- mcnemar.test(tabla_inicial_anuncio)
p_value_12
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  tabla_inicial_anuncio
## McNemar's chi-squared = 1.5, df = 1, p-value = 0.2207
```

```r
# Inicial vs internet
inicial_internet <- spread(data = datos %>% filter(Canal %in% c("inicial",
"internet")), key = Canal, value = Respuesta)
tabla_inicial_internet <- table(inicial = inicial_internet$inicial, internet =
inicial_internet$internet)
tabla_inicial_internet
```

```
##        internet
## inicial 0 1
##       0 1 7
##       1 0 2
```

```r
p_value_13 <- mcnemar.test(tabla_inicial_internet)
p_value_13
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  tabla_inicial_internet
## McNemar's chi-squared = 5.1429, df = 1, p-value = 0.02334
```

```r
# anuncio vs internet
anuncio_internet <- spread(data = datos %>% filter(Canal %in% c("anuncio",
"internet")), key = Canal, value = Respuesta)
tabla_anuncio_internet <- table(anuncio = anuncio_internet$anuncio, internet =
anuncio_internet$internet)
tabla_anuncio_internet
```

```
##        internet
## anuncio 0 1
##       0 0 4
##       1 1 5
```

```r
p_value_23 <- mcnemar.test(tabla_anuncio_internet)
p_value_23
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  tabla_anuncio_internet
## McNemar's chi-squared = 0.8, df = 1, p-value = 0.3711
```

Se corrige el nivel de significancia, en este caso se selecciona el método de *holm*.

```
p.adjust(c(p_value_12$p.value, p_value_13$p.value, p_value_23$p.value), method =
"holm")
```

```
## [1] 0.44134272 0.07002661 0.44134272
```

A pesar de que el test *Q-Cochran* ha resultado significativo, las comparaciones dos a dos con corrección de *holm* no encuentran ninguna diferencia significativa. Es importante mencionar que en este caso no se satisfacían las condiciones para un test de McNemar por lo que habría que utilizar para las comparaciones *post-hoc* el test binomial.