Ejercicios clases 10

Test McNemar

Joaquin Cavieres G.

Introducción

El test de McNemar generalmente es usado como alternativa al test χ^2 y también al test de Fisher en donde los datos estan pareados y trabajamos con tablas de 2 x 2 con sólo dos posibles resultados en la variable respuesta (binomiales).

Recuerde las condiciones para aplicar este test:

- Trabajamos con datos pareados
- La variable respuesta sólo tiene dos posibles resultados
- Tener una tabla de 2 x 2
- Si la suma de los eventos que pasan de un estado a otro es < 25 entonces se debe trabajar con un test binomial (`binom.test()`).

Hipótesis

Generalmente queremos determinar si el tratamiento o el efecto de algún proceso sobre un grupo de personas tiene un resultado positivo (o negativo) antes y despues de un determinado tiempo de medición. Esto sería:

	Antes positivo	Despúes negativo	Total
Antes positivo	a	b	a+b
Despúes negativo	c	d	c+d
Total	a + c	b + d	n=a+b+c+d

Para este caso esperaríamos que las proporciones de pasar negativo a positivo fuesen iguales a las de pasar de positivo a negativo. Esto significa que:

$$
\begin{aligned}
pc + pd &= pb + pd \\
pc + pd &= pb + pd
\end{aligned}
$$

Esto nos lleva a proponer que pb = pc, lo que vendría siendo nuestro H_0.

Por lo anterior entonces el contraste de hipótesis es:

$$
\begin{aligned}
H_0 : p_b &= p_c \\
H_1 : p_b &\neq p_c
\end{aligned}
$$

Estadístico para el test McNemar

El estadístico del test de McNemar sigue una distribución χ^2 con 1 grado de libertad (dado los dos grupos solamente), por tanto:

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

En el caso que si la sumatoria de b + c < 25 entonces aplicamos el test binomial como declaramos previamente. En `R` esto sería `binom.test(x= c, n = c+b, p= 0.5)`

Efecto del tamaño

Al tamaño del efecto se le conoce también como fuerza de asociación. Para el test de McNemar este está dado por las diferencias entre las marginales de las proporciones tal que:

$$
d = \pi_{1+} - \pi_{+1} = \pi_{12} - \pi_{21}
$$

y límites empleados para su clasificación son:

- Pequeño: 0.1
- Mediano: 0.3
- Grande: 0.5

Estimación en muestras con n grande

Supongamos que tratamos el número total de observaciones fuera de la diagonal como fijo, entonces:

$$
n^* = n_{12} + n_{21}
$$

Considerando la hipótesis nula propuesta anteriormente entonces esperamos observar aproximadamente la misma frecuencia de conteos para un n_{12} y un n_{21}, con distribución $Bin(n^*, 0{,}5)$. Siempre que n^* sea suficientemente grande, el estadístico de prueba es:

$$z = \frac{\frac{n_{12}}{n^*} - 0{,}5}{\sqrt{\frac{0{,}5(1-0{,}5)}{n^*}}} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}},$$

donde $0{,}5n^*$ es el número esperado de n_{12} bajo H_0. Por tanto la proporción esta determinada por n_{12}/n y la varianza es $\frac{0{,}5(1-0{,}5)}{n^*}$ bajo H_0.

Bajo H_0, z es asumido $\sim N(0,1)$ y para un $n^* \geq 25$. Alternativamente podemos comparar:

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

para χ^2. Este test es válido bajo muestreo multinomial general cuando n^* no es fijo, pero n lo es.

Observación

1) Para un n grande el efecto del tamaño es:

$$\hat{d} = \frac{n_{12}}{n} - \frac{n_{21}}{n}$$

Como $\hat{d}$ es insesgado entonces tiene una aproximación Normal con varianza:

$$\begin{aligned} V(\hat{d}) &= n^{-2} V(n_{12} - n_{21}) \\ &= n^{-2}[V(n_{12}) + V(n_{21}) - 2Cov(n_{12}, n_{21})] \\ &= n^{-1}[\pi_{12}(1-\pi_{12}) + \pi_{21}(1-\pi_{21}) + 2\pi_{12}\pi_{21}] \end{aligned}$$

Un estimador de la varianza se puede obtener mediante:

$$\hat{V}(\hat{d}) = n^{-1}\left[\frac{n_{12}}{n}(1 - \frac{n_{12}}{n}) + \frac{n_{21}}{n}(1 - \frac{n_{21}}{n}) + 2\frac{n_{12}n_{21}}{n^2}\right]$$

y un intervalo de confianza del 95 % como:

$$\hat{d} \pm 1{,}96\sqrt{\hat{V}(\hat{d})}$$

2) Proporciones y sumas de marginales

Generalmente, cuando se analizan tablas de contigencia, las proporciones marginales de las filas y/o columnas nos permiten obtener una mejor visualización de las diferencias. Mediante la función `prop.table()` podemos obtener esas proporciones facilmente (obtenemos las proporciones en vez de los conteos).

Ejemplo

```
rap =  c("Sí", "No", "Sí", "No", "No", "Sí", "No", "No", "No", "No")
rock = c("No", "No", "Sí", "Sí", "Sí", "No", "Sí", "Sí", "Sí", "Sí")
```

```
# Tabla Univariada
table(rap)
```

```
## rap
## No Sí
##  7  3
```

```
# Tabla Contingencia
table(rap, rock)
```

```
##      rock
## rap  No Sí
##   No  1  6
##   Sí  2  1
```

```
# Tabla de contengencia con proporciones marginales por filas
u = table(rap, rock)     # Asignar nombre cualquiera

prop.table(u)            # Por defecto proporciones totales.
```

```
##      rock
## rap    No  Sí
##   No 0.1 0.6
##   Sí 0.2 0.1
```

```
#Sumas marginales:
margin.table(u, 1)      #Filas
```

```
## rap
## No Sí
##  7  3
```

```r
margin.table(u, 2)      #Columnas
```

```
## rock
## No Sí
##  3  7
```

3) Prueba χ^2 para test de independencias

```r
chisq.test(u)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  u
## X-squared = 0.81633, df = 1, p-value = 0.3663
```

Ejercicio 1

Se quiere determinar si un comercial de televisión puede cambiar la opinión de las personas sobre un producto en particular. Encuestaron a 100 personas para averiguar si les gustaba o no el producto en cuestion. Luego de un tiempo (en el cual se transmite el comercial), se les consulta a las mismas 100 personas. Los resultados se encuentran en la siguiente tabla:

	Antes del comercial		
Despues del comercial	Apoya	No Apoya	Total
Apoya	30	40	70
No Apoya	12	18	30
Total	42	58	100

Solución

Para determinar si hubo una diferencia estadísticamente significativa en la proporción de personas que apoyaron el producto después de ver el video, podemos realizar la prueba de McNemar.

```
data = matrix(c(30, 12, 40, 18), nrow = 2,
       dimnames = list("Antes del comercial" = c("Apoya", "No Apoya"),
                       "Despúes del comercial" = c("Apoya", "No Apoya")))

# Vemos los datos
head(data)
```

```
##                    Despúes del comercial
## Antes del comercial Apoya No Apoya
##            Apoya       30       40
##            No Apoya    12       18
```

Aplicamos el `mcnemar.test(x, y = NULL, correct = TRUE)`. Se recomienda usar la corrección (`correct = TRUE`) de continuidad cuando algunos valores en la tabla son pequeños. Se aplica normalmente cuando cualquiera de los valores es inferior a 5.

```
# Test  de McNemar's Test con corrección
mcnemar.test(data)
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  data
## McNemar's chi-squared = 14.019, df = 1, p-value = 0.000181
```

```
# Test  de McNemar's Test sin corrección
mcnemar.test(data, correct=FALSE)
```

```
##
##  McNemar's Chi-squared test
##
## data:  data
## McNemar's chi-squared = 15.077, df = 1, p-value = 0.0001032
```

En ambos casos el el p-value $<\alpha = 0{,}05$, por lo que rechazamos H_0, por lo tanto la proporción de personas que apoyaron el producto antes y después de ver el comercial fue significativamente diferente.

Ejercicio 2

Se quiere determinar si una noticia de alto impacto puede tener influencia en la percepción de las personas sobre un determinado deportista. Para ello, antes de hacer pública la noticia, se pregunta a las 2000 personas del grupo si consideran al deportista como uno de los mejores en su disciplina. 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?

Solución

```r
data2 = matrix(c(894, 186, 250, 670),
       nrow = 2,
       dimnames = list("Antes noticia" = c("Sí", "No"),
                       "Despues noticia" = c("Sí", "No")))
data2
```

```
##              Despues noticia
## Antes noticia  Sí  No
##            Sí 894 250
##            No 186 670
```

Ya que tenemos un n grande entonces podemos hacer directamente el test de McNemar.

```r
mcnemar.test(data2)
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  data2
## McNemar's chi-squared = 9.1032, df = 1, p-value = 0.002552
```

El $p - value < \alpha(0{,}05)$ entonces rechazamos H_0. Lo anterior nos indica que la noticia ha influido significativamente en la decisión de las personas

¿Y el efecto del tamaño?

```r
#install.packages("vcd")
library(vcd)
assocstats(data2)
```

```
##                     X^2 df P(> X^2)
## Likelihood Ratio 662.32  1        0
## Pearson          627.41  1        0
##
## Phi-Coefficient   : 0.56
## Contingency Coeff.: 0.489
## Cramer's V        : 0.56
```