# Chi cuadrada

Fernanda Sarmiento

## Introducción

Esta prueba de hipótesis se emplea al momento de trabajar con valores nominales, categóricos o cuando se trabaja con alguna clase de clasificación.

```
frec <- c(15,19)
chisq.test(frec)
```

```
    Chi-squared test for given probabilities

data:  frec
X-squared = 0.47059, df = 1, p-value = 0.4927
```

Para esta prueba se calculan los grados de libertad a partir de n-1; otro modo, para tener un resultado más crítico de chi-cuadrada se puede obtener empleando el 95%, esto se escribe de la siguiente manera.

```
qchisq(0.95,1)
```

```
[1] 3.841459
```

Los valores esperados se deben calcular a partir de:

```
chisq.test(frec)$expected
```

```
[1] 17 17
```

### Test de independencia - Fisher test

La chi-cuadrada, también conocida como bondad de ajuste, se la denomina así cuando los objetos pertenecen a una matriz de datos. Otro nombre bajo el cual se la reconoce es tabla de contingencia, esto se da cuando los datos están agrupados en filas y columnas.

A continuación, generaremos una matriz para poder evidenciar lo antes mencionado.

```
matriz <- matrix(c(4,11,10,13,3,4,6,8),nrow=2)
matriz
```

```
     [,1] [,2] [,3] [,4]
[1,]    4   10    3    6
[2,]   11   13    4    8
```

Para aplicar el test de la Chi-cuadrada se emplea la función `chisq.test(x)` de un vector x.

Al tener un p-value de 0.7328, por tal motivo no se rechaza la hipótesis. Ademas, podemos evidenciar que parte de los datos que están más alejados o el 20 % de estos son valores menores a 5. En estas circunstancias es recomendado implementar una prueba de independencia, la cual se denomina Fisher test.

Los grados de libertad para este tipo de tablas se obtienen por C-1*F-1.

```
fisher.test(matriz)
```

```
    Fisher's Exact Test for Count Data

data:  matriz
p-value = 0.7229
alternative hypothesis: two.sided
```

Podemos aumentar la precisión de la prueba si se aumenta el número de las réplicas o de las permutaciones. Esta prueba se basa en el modelo de la distribución hipergeométrica. Esta se diferencia de aplicar la `Fisdher` en que aplica una colección por continuidad para tablas.

```
fisher.test(matriz,simulate.p.value=TRUE, B=2e3)
```

```
    Fisher's Exact Test for Count Data with simulated p-value (based on
    2000 replicates)

data:  matriz
p-value = 0.7436
alternative hypothesis: two.sided
```

Para obtener el valor de la chi-cuadrara de muestras pequeñas, se puede calcular de la siguiente manera.

```
chisq.test(matriz,simulate.p.value = T)
```

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

data:  matriz
X-squared = 1.2845, df = NA, p-value = 0.7376
```

Al trabajar con una matriz de datos, hay la posibilidad de calcular las proporciones que corresponden a los datos en las columnas y en las filas.

```
prop.table(matriz)
```

```
           [,1]      [,2]       [,3]      [,4]
[1,] 0.06779661 0.1694915 0.05084746 0.1016949
[2,] 0.18644068 0.2203390 0.06779661 0.1355932
```

Esto nos ayuda a establecer la decencia de los valores en manera porcentual, además de facilitar la interpretación de los mismos.

Otras posibilidades que se tienen es presentar a través de las filas.

```
prop.table(matriz, 1)
```

```
          [,1]      [,2]      [,3]      [,4]
[1,] 0.1739130 0.4347826 0.1304348 0.2608696
[2,] 0.3055556 0.3611111 0.1111111 0.2222222
```

Así mismo se pueden proporcionar los datos con respecto a las columnas.

```
prop.table(matriz, 2)
```

```
          [,1]      [,2]      [,3]      [,4]
[1,] 0.2666667 0.4347826 0.4285714 0.4285714
[2,] 0.7333333 0.5652174 0.5714286 0.5714286
```

Matrices complejas.

A continuación aprenderemos a sintetizar o extraer parte de los datos.

```
library(MASS)
data(survey)
head(survey)
```

```
     Sex Wr.Hnd NW.Hnd W.Hnd    Fold Pulse    Clap Exer Smoke Height
M.I
1 Female   18.5   18.0 Right  R on L    92    Left Some Never 173.00
Metric
2   Male   19.5   20.5  Left  R on L   104    Left None Regul 177.80
Imperial
3   Male   18.0   13.3 Right  L on R    87 Neither None Occas     NA
<NA>
4   Male   18.8   18.9 Right  R on L    NA Neither None Never 160.00
Metric
5   Male   20.0   20.0 Right Neither    35   Right Some Never 165.00
Metric
6 Female   18.0   17.7 Right  L on R    64   Right Some Never 172.72
Imperial
     Age
1 18.250
2 17.583
3 16.917
4 20.333
5 23.667
6 21.000
```

A continuación vamos a crear una nueva matriz, en este caso vamos a usar los datos de ejercicio y smoke.

```
cuadro <- table(survey$Smoke, survey$Exer)
cuadro
```

```
        Freq None Some
  Heavy    7    1    3
  Never   87   18   84
  Occas   12    3    4
  Regul    9    1    7
```

### Test de independencia.

```
chisq.test(cuadro)
```

```
    Pearson's Chi-squared test

data:  cuadro
X-squared = 5.4885, df = 6, p-value = 0.4828
```

```
fisher.test(cuadro,simulate.p.value = TRUE)
```

```
    Fisher's Exact Test for Count Data with simulated p-value (based on
    2000 replicates)

data:  cuadro
p-value = 0.4028
alternative hypothesis: two.sided
```

A pesar de aumentar la muestra, el p-valor no cambia.

```
fisher.test(cuadro, simulate.p.value = TRUE, B=5000)
```

```
    Fisher's Exact Test for Count Data with simulated p-value (based on
    5000 replicates)

data:  cuadro
p-value = 0.4205
alternative hypothesis: two.sided
```

```
chisq.test(cuadro, simulate.p.value = T, B= 5000)
```

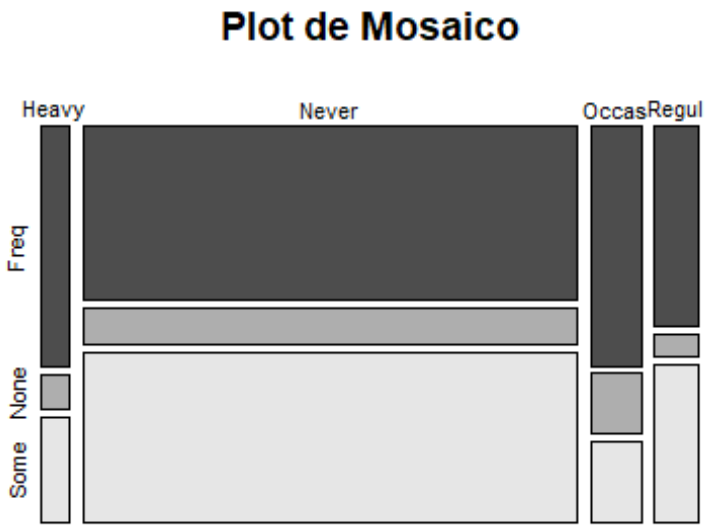
```
    Pearson's Chi-squared test with simulated p-value (based on 5000
    replicates)

data:  cuadro
X-squared = 5.4885, df = NA, p-value = 0.4859
```

## Formas Graficas Para Los Test

```
cuadro2 <- table(survey$Smoke, survey$Sex)
```

Los gráficos de tipo mosaico despliegan la información para poder examinar la relación entre dos o más variables categóricas.

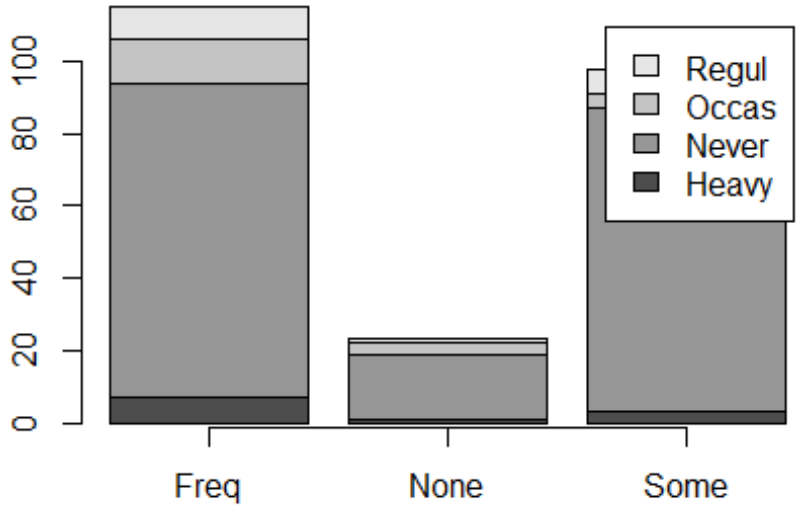
```
mosaicplot(cuadro, color=TRUE, main ="Plot de Mosaico")
```




En esta ocasión podemos evidenciar que se presenta alta frecuencia de visualizaciones en la categoría Never y las frecuencias Freq y Some, mientras que en las categorías Heavy y Regul y la frecuencia None se presentó una baja frecuencia de visualizaciones.

Posterior a la obtención de estos cuadros, se recomienda obtener las proporciones porcentuales para comparar y comprobar la coherencia de los gráficos.

```
barplot(cuadro, legend= rownames(cuadro), beside= F, axis.lty = 1)
```




El gráfico presentado anteriormente nos indica que se presenta una predominancia de la categoría Never en las barras, principalmente en las de Freq y Some, mientras que la categoría que casi no se presenta es Heavy, lo que nos indica que puede presentarse la falta de independencia en los datos.