Práctica 11. Contrastes chi chuadrado. Soluciones

Estadística (650008). Grado en biología sanitaria. UAH.

Actualizado: 2023-12-22

Ejercicio 1

En el trabajo de G. Mendel con guisantes, hay un vector de frecuencias observadas (semilla lisa, semilla rugosa):

```
observados = c(5474, 1850)
```

y unas proporciones (distribución) teóricas

```
probEsperados = c(3/4, 1/4)
```

¿Estaba Mendel en lo cierto?

Resuelto en clase de teoría: la hipótesis que se contrata es

H0: los valores observados por Mendal se corresponden con las probabilidades (frecuencias relativas) teóricas

```
chisq.test(x = observados, p = probEsperados)
```

```
    Chi-squared test for given probabilities

data:  observados
X-squared = 0.26288, df = 1, p-value = 0.6081
```

El p-valor es alto, y no hay motivo para dudar de H0.

Para calcular las frecuencias absolutas esperadas hay dos alternativas: 1. Contar todas las semillas y aplicarles las proporciones teóricas:

```
sum(observados)*probEsperados
```

```
[1] 5493 1831
```

1. Rescatar la información generada para hacer el contraste. Guarda el resultado de hacer el contraste en una variable (como hacíamos con los boxplots)

```
contraste = chisq.test(x = observados, p = probEsperados)
```

y usa el símbolo $ para acceder a los valores esperados

```
contraste$expected
```

```
[1] 5493 1831
```

Ejercicio 2

A partir de la siguiente tabla de contingencia, determina si hay asociación entre las variables “dar positivo en una prueba diagnóstico” y “estar enfermo”

	Enfermos	Sanos
Positivo	192	158
Negativo	10	9646

Se trata de un contraste de independencia, y se contrasta la hipótesis nula H0: "Las dos variables son independientes" frente a H1: "las dos variables NO son independientes".

Lo primero es construir una tabla de contingencia. Puede hacerse mediante una matriz

```
tablaObservada = matrix( c(192, 10, 158, 9646),
                         nrow= 2,
                         byrow = FALSE)
```

y, a continuación, se hace el contraste Chi cuadrado sobre la tabla/matriz construida

```
chisq.test(tablaObservada)
```

```
    Pearson's Chi-squared test with Yates' continuity correction

data:  tablaObservada
X-squared = 5091.5, df = 1, p-value < 2.2e-16
```

para rechazar H0: las variables son independientes.

Para calcular la tabla de valores esperados lo más sencillo es hacer usar la información calculada para hacer el contraste

```
contraste = chisq.test(tablaObservada)
contraste$expected
```

```
           [,1]      [,2]
[1,]   7.065761  342.9342
[2,] 194.934239 9461.0658
```

Ejercicio 3

En este fichero encontrarás un extracto de los datos de la encuesta nacional de salud de 2012. Aquí tienes las primeras lineas de la tabla

```
df = read.table(file = "Practica11_ENSE_chi2.csv", sep = ";", header = TRUE)
head(df)
```

```
  Aperitivos Refrescos ActividadFisica Fumar
1          1         3               1     1
2          4         4               1     4
3          1         4               1     1
4          3         5               1     4
5          4         4               1     1
6          2         3               4     4
```

- Considera las variables `Fumar` y `ActividadFisica`, que toman los valores

  1. Sí fuma, diariamente
  2. Sí fuma, pero no diariamente
  3. No fuma actualmente, pero ha fumado antes
  4. No fuma ni ha fumado nunca de manera habitual

  y

1. No hago ejercicio. El tiempo libre lo ocupo de forma casi completamente sedentaria (leer, ver la televisión, ir al cine, etc.)
2. Hago alguna actividad física o deportiva ocasional
3. Hago actividad física varias veces al mes
4. Hace entrenamiento deportivo o físico varias veces a la semana

respectivamente, analiza si hay relación entre las dos variables o, por contra, son independientes.

Ten en cuenta que los valores 8 y 9 en la tabla se refieren, respectivamente, a "no sabe" y "no contesta", por lo que hay que eliminarlos.

Se trata de contrastar H0: las variables on independientes.

Para el contrate chi cuadrado es necesario calcular la correspondiente tabla de contingencia

```
(tabla1 = table(df$ActividadFisica, df$Fumar))
```

```
      1    2    3    4    8    9
  1 2382  204 1672 5122    0   10
  2 1533  200 1691 4151    1    9
  3  510   92  451 1268    0    0
  4  334   58  363  940    0    0
  8    2    0    4    3    0    0
  9    1    0    0    3    0    3
```

Observa que aparecen los valores 8 y 9, que corresonden con no sabe/no contesta, y hay que eliminarlos de la muestra. Creamos una tabla con las variables de interés

```
df2 = df[   , c("ActividadFisica", "Fumar")]
```

Se eliminan los valores 8 y 9

```
df2 = df2[df$ActividadFisica %in% 1:5,  ]
df2 = df2[df2$Fumar %in% 1:5,  ]
```

Se comprueba que la eliminación ha sido correcta

```
(tabla2 = table(df2$ActividadFisica, df2$Fumar))
```

```
      1    2    3    4
  1 2382  204 1672 5122
  2 1533  200 1691 4151
  3  510   92  451 1268
  4  334   58  363  940
```

Y se procede ha hacer el contraste

```
 chisq.test(tabla2)
```

```
    Pearson's Chi-squared test

data:  tabla2
X-squared = 129.51, df = 9, p-value < 2.2e-16
```

Para concluir que las dos variables no son independientes, sino que están asociadas.

Ejercicio 4

Vamos a seguir parte del tutorial 12 de postdata. Este fichero contiene los datos del Lunar *Orbiter Laser Altimeter instrument* (LOLA) para determinar si los cráteres de la Luna están distribuidos de forma uniforme entre el hemisferio norte y el sur o si, por contra, hay más en uno de los hemisferios.

Las tres variables que aparecen en ese fichero:

```
crateres = read.table(file = "Cap09-LolaLargeLunarCraterCatalog.csv", sep = ",", header = TRUE)
colnames(crateres)
```

```
[1] "Lon"     "Lat"     "Diam_km"
```

que se refieren a la latitud, longitud (ambas en grados) y diámetro (en km) de los cráteres lunares y son todas ellas cuantitativas continuas.

La pregunta a responder se puede formular como sigue: ¿hay diferencia entre los diámetros de los cráteres de ambos hemisferios de la Luna?

La función `cut` permite categorizar las variables:

```
hemisphere = cut(crateres$Lat, breaks=c(-90, 0, 90))
head(hemisphere)
```

```
[1] (-90,0] (-90,0] (0,90]  (0,90]  (0,90]  (-90,0]
Levels: (-90,0] (0,90]
```

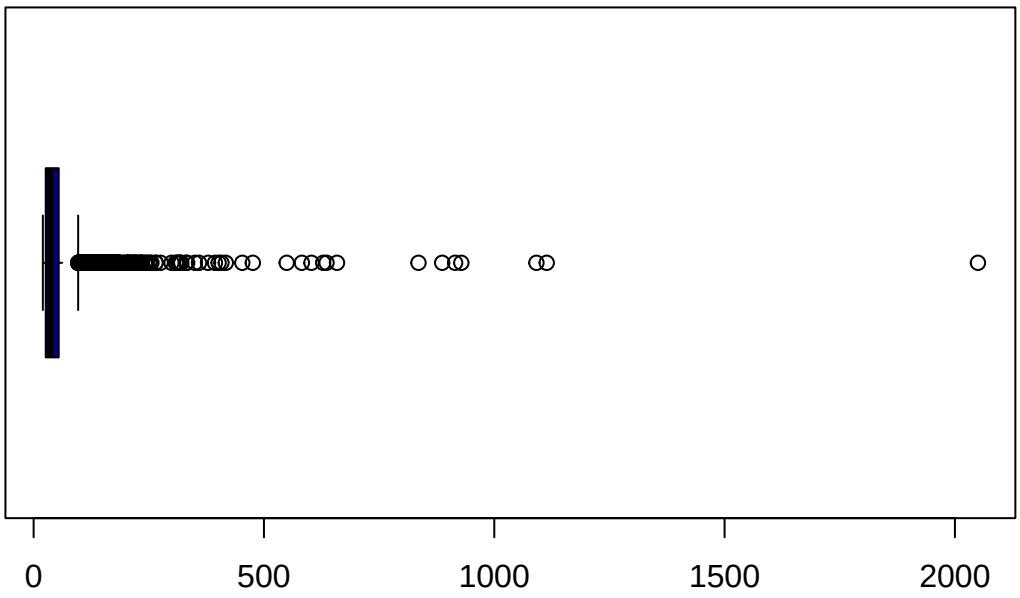
renombramos los niveles del factor

```
levels(hemisphere) = c("SUR", "NORTE")
head(hemisphere)
```

```
[1] SUR   SUR   NORTE NORTE NORTE SUR
Levels: SUR NORTE
```

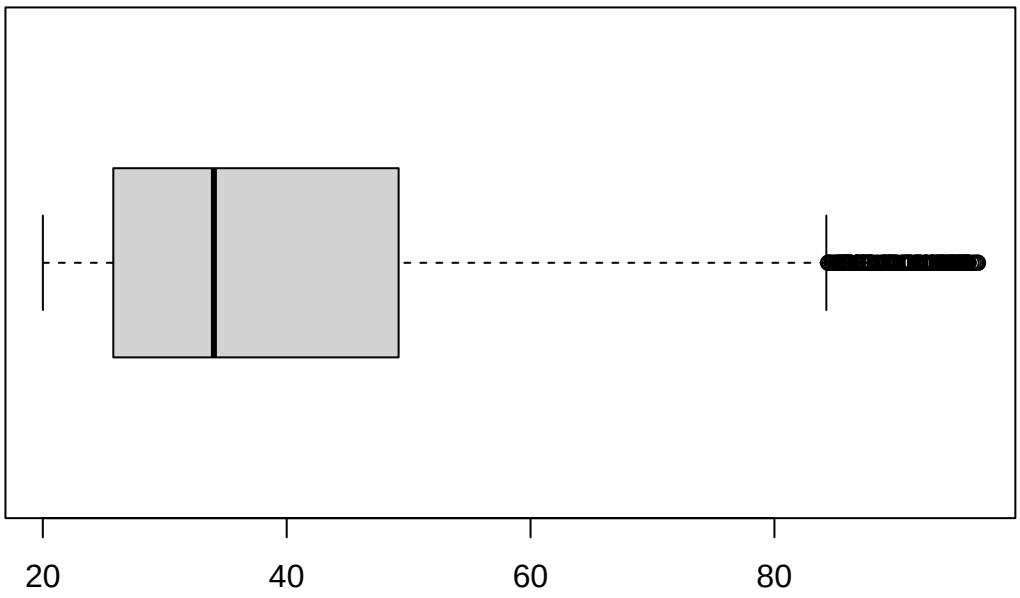
Para los diámetros de los cráteres

```
bp = boxplot(crateres$Diam_km, col = "navy", horizontal = T)
```



para decidir cómo agrupar los diámetros, observamos el boxplot si atípicos

```
eliminar = which(crateres$Diam_km %in% bp$out)
boxplot(crateres$Diam_km[-eliminar], horizontal = TRUE)
```



se pueden hacer clases de 20 - 40 - 60 - 80, mayor que 80

```
craterSize = cut(crateres$Diam_km,
breaks=c(seq(20, 80, 20), max(crateres$Diam_km)),
include.lowest=TRUE)
```

y ya podemos construir la tabla de frecuencias y hacer el contraste

```
(tabla_crateres = table(hemisphere, craterSize))
```

```
##           craterSize
## hemisphere [20,40] (40,60] (60,80] (80,2.05e+03]
##      SUR      1615     585     255           328
##      NORTE    1388     523     227           264
```

```
chisq.test(tabla_crateres)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabla_crateres
## X-squared = 1.184, df = 3, p-value = 0.7568
```

por lo que no hay evidencias para rechazar H_0.

Ejercicio 5

Sospechas que cierto caracter se hereda de forma independiente con una probabilidad de 0.2. Para comprobarlo, haces 100 experimentos, cada uno de ellos consiste en obtener 5 descedientes y contar el n'umero de ellos que

presenta dicho caracter.

El resultado es

```
table(descendientes)
```

```
descendientes
 0  1  2  3  4
26 46 23  3  2
```

¿Se cumple tu hipótesis?

De acuerdo con el enunciado, podemos considerar la variable aleatoria $X =$ "número de descendientes que hereda un caracter cuando hay 5 descendientes ys e hereda con probabilidad 0.2". Lo que nos preguntamos es

$$H_0 : X \sim B(n = 5, p = 0.2) \qquad H_1 : X \neq B(n = 5, p = 0.2)$$

La tabla que da el enunciado es la distribución observada dicha variable (número de éxitos), y hay que determinar la distribución de probabilidades teóricas. Esta no es otra que

```
(probEsperados = dbinom(0:5, size = 5, prob = 0.2))
```

```
[1] 0.32768 0.40960 0.20480 0.05120 0.00640 0.00032
```

Podemos construir también el vector de frecuencias observadas

```
frecsObservadas = c(26, 46, 23,  3,  2 )
```

y hacer el contraste chi cuadrado

```
chisq.test(x = frecsObservadas, p = probEsperados)
```

```
## Error in chisq.test(x = frecsObservadas, p = probEsperados): 'x' and 'p' must have the same number o
```

Se obtiene un error porque en ninguno de los 100 experimentos se han observado 5 éxitos; tenemos frecuencias absolutas para 0, 1, 2, 3, 4 éxitos, pero probabilidades para 0, 1, 2, 3, 4, 5 éxitos. Hay tres alternativas

1. Añadir un cero (no se obervaron 5 éxitos) al final del vector de frecuencias observadas

```
frecsObservadas = c(26, 46, 23,  3,  2, 0)
chisq.test(x = frecsObservadas, p = probEsperados)
```

```
Warning in chisq.test(x = frecsObservadas, p = probEsperados): Chi-squared
approximation may be incorrect


    Chi-squared test for given probabilities

data:  frecsObservadas
X-squared = 6.1279, df = 5, p-value = 0.294
```

procediendo así el función chisq.test funciona, pero da un aviso (warning) porque hay celdas con menos de 5 observaciones.

2. Podemos (es lo habitual si se quiere hacer un contraste chi cuadrado) agrupar las celdas con frecuencias bajas porque tiene sentido. Podemos pensar en 0, 1, 2, 3 o más éxitos. Hay que ajustar también las probabilidades:

```
frecsObservadas = c(26, 46, 23, 3+2+0)
(probEsperados = c(dbinom(0:2, size = 5, prob = 0.2), sum(dbinom(3:5, size = 5, prob = 0.2))))
```

```
[1] 0.32768 0.40960 0.20480 0.05792
```

```
chisq.test( frecsObservadas, p = probEsperados)
```

```
    Chi-squared test for given probabilities

data:  frecsObservadas
X-squared = 2.4364, df = 3, p-value = 0.4869
```

Ejercicio 6

PARA HACER EN CASA: En este fichero de datos está el genoma del bacteriófago ΦX174, primer genoma basado en ADN secuenciado (1977).

Puedes leer el contenido del fichero con

```
phiX174 = read.table(file ="phix174.txt")
```

y convertir en tipo caracter (ahora es una palabra muy larga)

```
(phiX174 = as.character(phiX174$V1))
```

```
## [1] "GAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTCGGATATTTCTGATGAGTCGAAAAATTATCTTGATAAAGCAGGAATTACTACTGCTT
```

y separar en letras individuales

```
# Separamos la secuencia en caracteres (nucleótidos)
genoma = unlist(strsplit(phiX174, split = ""))
head(genoma)
```

```
[1] "G" "A" "G" "T" "T" "T"
```

¿Dirías que los cuatro nucleotidos están igualmente representados?

Los nucleótidos estarían igualmente representados si sus frecuencias relativas fueran 1/4. Hacemos un contraste de homogeneidad

```
(chisqPhi174 = chisq.test(table(genoma), p = c(1/4, 1/4, 1/4, 1/4)))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(genoma)
## X-squared = 119.91, df = 3, p-value < 2.2e-16
```

por lo que concluimos que no están igualmente representados.

¿Qué valores habría que haber observado en caso de homogeneidad?

Hay que buscar en la información generada por `chisq.test` para hacer el contraste

```
chisqPhi174$expected
```

```
##      A      C      G      T
## 1346.5 1346.5 1346.5 1346.5
```

Nota: Si quieres hacer todo el proceso de adquisición de datos necesitarás instalar el paquete `ape`, que permite descargar el genoma del bacteriófago ΦX174, primer genoma basado en ADN secuenciado (1977). Sólo necesitas su número de acceso en la base de datos GenBank del NCBI

```
if (!requireNamespace("ape", quietly = TRUE)){install.packages("ape") }
library(ape)
```

```
Warning: package 'ape' was built under R version 4.1.3
```

```
# guardar el identificador del genoma
myID <- c("NC_001422.1")

# descargar el genoma
mySequence <- read.GenBank(access.nb = myID,    seq.names = myID,
                           species.names = TRUE,  as.character = TRUE)

# echa un vistazo
head(mySequence$NC_001422.1, 10)
```

```
 [1] "g" "a" "g" "t" "t" "t" "t" "a" "t" "c"
```

Cambiando el número de acceso, podrás descargar otros genomas allí almacenados.

Ejercicio 7

PARA HACER EN CASA: En este fichero encontrarás un extracto de los datos de la encuesta nacional de salud de 2012. Aquí tienes las primeras lineas de la tabla

```
df = read.table(file = "Practica11_ENSE_chi2.csv", sep = ";", header = TRUE)
head(df)
```

```
  Aperitivos Refrescos ActividadFisica Fumar
1          1         3               1     1
2          4         4               1     4
3          1         4               1     1
4          3         5               1     4
5          4         4               1     1
6          2         3               4     4
```

- Considera las variables `Consumo de aperitivos o comidas saladas de picar` (patatas fritas, ganchitos, galletitas saladas) y `consumo de refrescos azucarados`, cada una de las cuales puede tomar valores del 1 al 5 de cuerdo con
  1. A diario
  2. Tres o más veces a la semana, pero no a diario
  3. Una o dos veces a la semana
  4. Menos de una vez a la semana
  5. Nunca o casi nunca

  Analiza si hay relación entre las frecuencias con que se toman aperitivos y refrescos o, por contra, son independientes.

  Se trata de contrastar H0: las variables on independientes.

  Para el contrate chi cuadrado es necesario calcular la correspondiente tabla de contingencia

```
(tabla1 = table(df$Aperitivos, df$Refrescos))
```

```
   1   2   3   4   5   8   9
1  176 244 579 448 692   2   0
2   14 301 516 374 482   0   0
3   31 157 1106 753 837   0   1
4   25  82 382 1408 1113   2   0
5   80 216 984 1821 8110   5   0
8    0   1   4   3  11  31   0
9    0   0   0   0   0   0  16
```

```
Observa que aparecen los valores 8 y 9, que corresonden con no sabe/no contesta, y hay que eliminarlos de la muestr
```

```
df1 = df[   , c("Aperitivos", "Refrescos")]
```

```
Se eliminan los valores 8 y 9
```

```r
df1 = df1[df$Aperitivos %in% 1:5,  ]
df1 = df1[df1$Refrescos %in% 1:5,  ]
```

Se comprueba que la eliminación ha sido correcta

```r
(tabla1 = table(df1$Aperitivos, df1$Refrescos))
```

```
      1    2    3    4    5
  1 176  244  579  448  692
  2  14  301  516  374  482
  3  31  157 1106  753  837
  4  25   82  382 1408 1113
  5  80  216  984 1821 8110
```

Y se procede ha hacer el contraste

```r
 chisq.test(tabla1)
```

```
    Pearson's Chi-squared test

data:  tabla1
X-squared = 5832, df = 16, p-value < 2.2e-16
```

Para concluir que las dos variables no son independientes, sino que están asociadas.