Hoja 4 (c): Tests Chi-cuadrado con R

Estadística Computacional I. Grado en Estadística

Marta Venegas Pardo

Índice

Ejercicio 1: Test de bondad de ajuste

Bondad de ajuste. Comprobar si un dado es correcto a partir del número de veces que ha salido cada lado.

```
frecu <- c(22,21,22,27,22,36)
probs <- rep(1/6,6)
```

Utilizamos el test Chi-Cuadrado, donde comparamos lo observado frente a lo esperado.

```
chisq.test(frecu, p=probs)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  frecu
## X-squared = 6.72, df = 5, p-value = 0.2423
```

Acepto que sigue ese modelo probabilístico (equiprobabilidad).

Ejercicio 2: Distribución uniforme discreta

Por defecto se compara con la unif. discreta. En el siguiente ejemplo se trata de ver si en un texto las apariciones de las letras E,T,N,R,O se distribuyen según los valores conocidos en inglés.

```r
x <-c(100,110,80,55,14)
probs <-c(29, 21, 17, 17, 16)/100
```

Si sólo me dieran los datos en lugar de la frecuencia, tendría que hacer la tabla de frecuencias y ya hacer el ejercicio.

```r
chisq.test(x,p=probs)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  x
## X-squared = 55.395, df = 4, p-value = 2.685e-11
```

Tenemos un p-valor muy extremo, existe una gran diferencia entre los valores observados y esperados. Se rechaza que la muestra siga el modelo teórico.

Ejercicio 3: Ley uniforme discreta, muestras

En la siguiente simulación se ilustra la calidad de la aproximación. Se generan M muestras de tamaño n de una ley Uinforme discreta.

```r
probabi<- c(0.03,0.25,0.45,0.27)
sum(probabi)
```

```
## [1] 1
```

```r
set.seed(12345)
n<-50 #tamaño muestral
n*probabi  #Se cumplen las condiciones
```

```
## [1]  1.5 12.5 22.5 13.5
```

```r
M<-5000
estad<- numeric(M)
```

Generar la muestra i, calcular la tabla y obtener el estadístico chi-cuadrado.

```r
for (i in 1:M) {
  x=sample(1:4,size = n, prob = probabi,rep=T)
  resultado= c(sum(x==1),
               sum(x==2),
               sum(x==3),
               sum(x==4))
```

```
  estadistico_chi=chisq.test(resultado, p=probabi)
  estad[i] =estadistico_chi$statistic
}
```

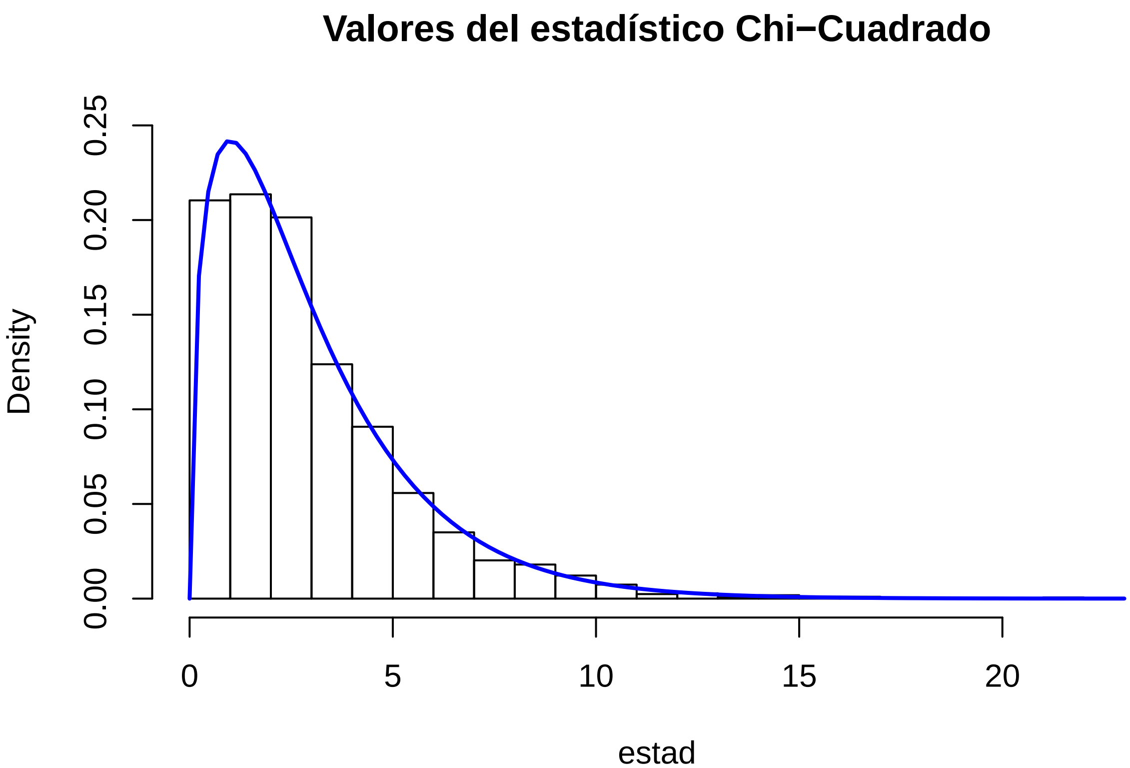
Tenemos estad con 5000 valores

```
head(estad)
```

```
## [1] 0.5244444 1.5911111 4.4651852 8.1096296 5.1940741 3.1851852
```

Histograma del estadístico y densidad de la chi-cuadrado.

```
hist(estad, breaks = 30, probability = TRUE,
     main="Valores del estadístico Chi-Cuadrado",ylim = c(0,0.25))
curve(dchisq(x,length(probabi)-1), col="blue",lwd=2,add=TRUE)
```



Ejercicio 4: Tablas de contingencia: Test

Tests de independencia en tablas de contingencia.

```
#save(TRABAJO,AUTOESTIMA,file="ej04.RData")
load("ej04.RData")
tabla<- table(TRABAJO,AUTOESTIMA)
tabla
```

```
##                        AUTOESTIMA
## TRABAJO                 Baja Media Alta
##   Actividad remunerada    90    65   91
##   Ama de Casa            101    76   42
```

Apartado a

Comprobación del p-valor y dibujar la distribución teórica, el cuantil y el valor observado.

H0 es la independencia de las muestras.

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

```
##
##  Pearson's Chi-squared test
##
## data:  tabla
## X-squared = 18.037, df = 2, p-value = 0.0001211
```

Rechazo H0, las muestras están relacionadas. Lo razonamos debido al p-valor.

```
resul$expected
```

```
##                          AUTOESTIMA
## TRABAJO                       Baja    Media     Alta
##   Actividad remunerada 101.04516 74.59355 70.36129
##   Ama de Casa           89.95484 66.40645 62.63871
```

```
resul$observed
```

```
##                          AUTOESTIMA
## TRABAJO                 Baja Media Alta
##   Actividad remunerada   90    65   91
##   Ama de Casa           101    76   42
```

Calculamos el estadístico de manera manual.

- Forma 1:

```
sum(resul$residuals^2)
```

```
## [1] 18.03737
```

- Forma 2:

```
sum((resul$observed-resul$expected)^2/(resul$expected))
```

```
## [1] 18.03737
```

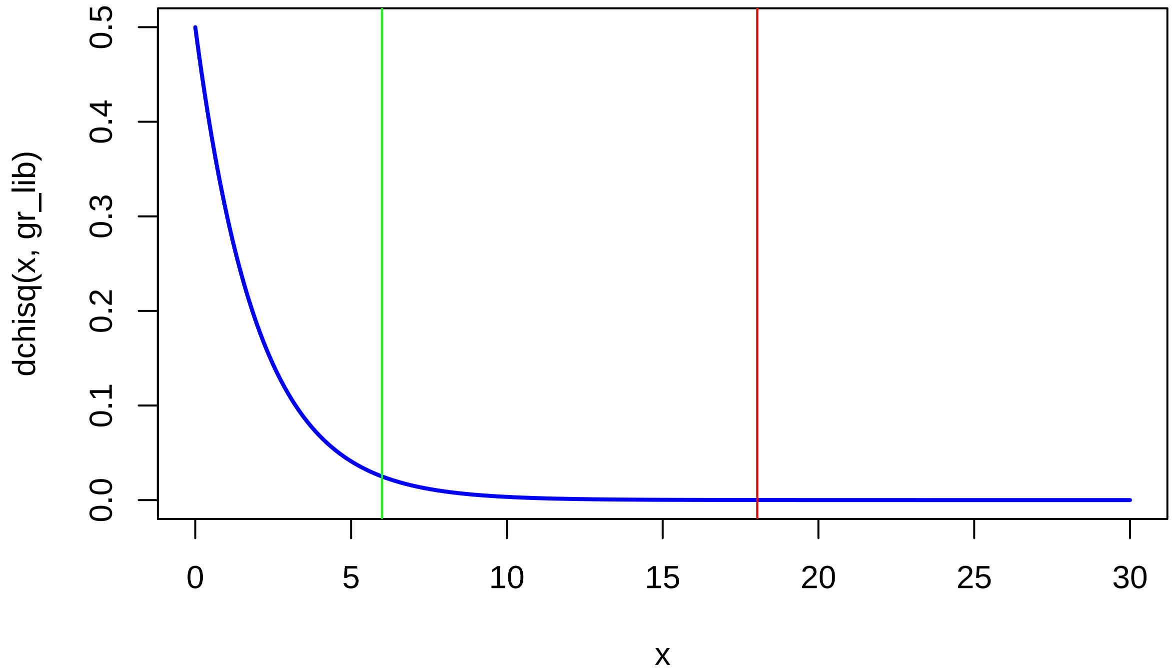
Cálculo del p-valor:

```
nr=nrow(tabla)
nc=ncol(tabla)
gr_lib=(nr-1)*(nc-1)
1-pchisq(resul$statistic,df=gr_lib)
```

```
##    X-squared
## 0.0001211255
```

Dibujar la fdd para la Chi-Cuadrado con esos grados de libertad.

```
curve(dchisq(x,gr_lib),0,30,1000,lwd=2,col="blue")
abline(v=resul$statistic,col="red")
abline(v=qchisq(0.95,gr_lib), col="green")
```



El p-valor es la probabilidad de que quede a la derecha, que como vemos es muy pequeña. La línea verde me muestra donde se encuentra el estadístico,donde empieza la región crítica.

Ejercicio 5: Librería vcd

Tests de independencia en tablas de contingencia (Uso de la librería vcd).

```
load("GSS.RData")
GSS
```

```
##        sex party count
## 1 mujeres   dem   279
## 2 hombres   dem   165
## 3 mujeres indep    73
## 4 hombres indep    47
## 5 mujeres   rep   225
## 6 hombres   rep   191
```

Tabla de frecuencias con el paquete básico.

```
tabla_GSS=xtabs(count ~ sex+party , data = GSS)
tabla_GSS
```

```
##          party
## sex       dem indep rep
##   mujeres 279    73 225
##   hombres 165    47 191
```

Test Chi_Cuadrado.

```
chisq.test(tabla_GSS)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabla_GSS
## X-squared = 7.0095, df = 2, p-value = 0.03005
```

Uso paquete vcd

Instalo el paquete vcd

```
library(vcd)
```

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

```
assocstats(tabla_GSS)
```

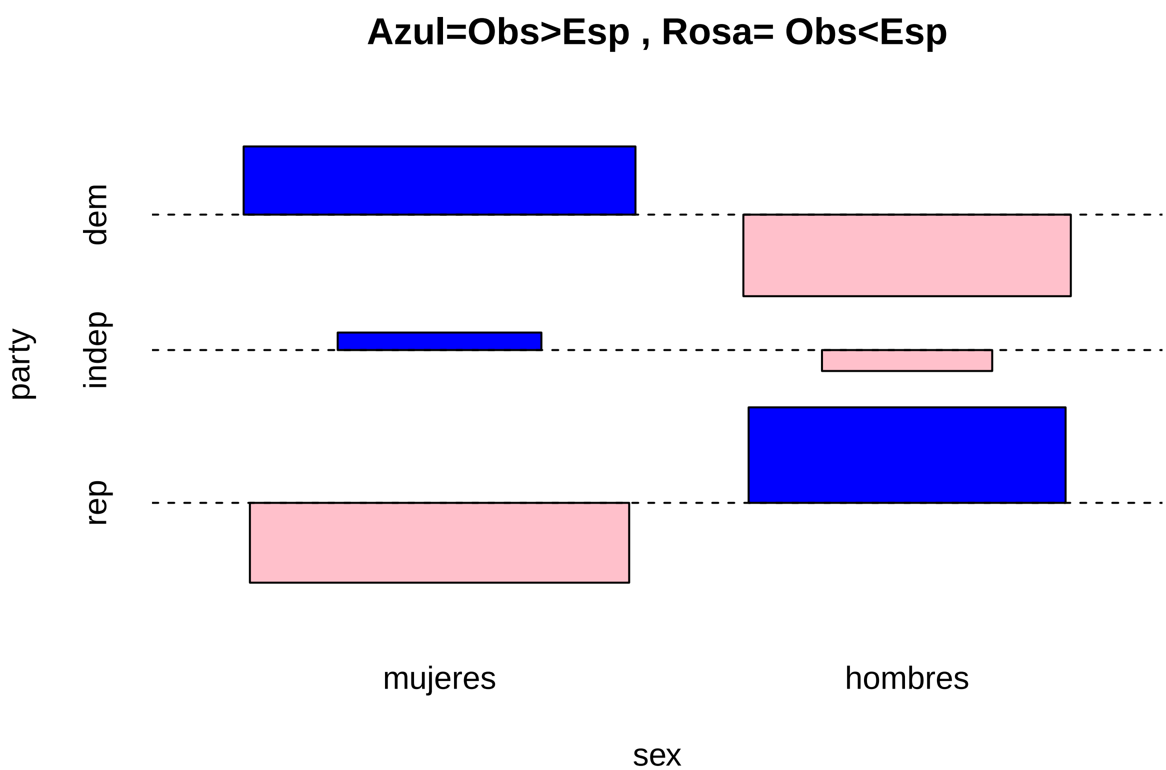
```
##                     X^2 df P(> X^2)
## Likelihood Ratio 7.0026  2 0.030158
## Pearson          7.0095  2 0.030054
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.084
## Cramer's V        : 0.085
```

```
#CrossTable(GSStab)
```

Me calcula todas las medidas de asociación. Me interesa el p-valor de pearson.

Gráfico.

```
assocplot(tabla_GSS, col=c("blue","pink"),
          main="Azul=Obs>Esp , Rosa= Obs<Esp")
```



Nos

presenta la tabla de forma gráfica, mostrando como están distribuidas las categorías.

Parece que existe un comportamiento relacionado con el sexo.

Ejercicio 6: Té

Una dama británica sostenía que era capaz de adivinar si en un té con leche se ha vertido antes el té o la leche.

Para ello se realizó un experimento donde se le pidió que lo adivinara para 8 tazas:

$$H_0 : P[diceleche = /leche] = P[dicet\acute{e}/leche]$$
$$H_1 : P[diceleche = /leche] > P[dicet\acute{e}/leche]$$

```
pruebate <-
  matrix(c(
    3,1,1,3), nrow = 2,
    dimnames = list(Predice = c("Leche", "Té"),
                    Verdad = c("Leche", "Té")))
pruebate
```

```
##        Verdad
## Predice Leche Té
##   Leche     3  1
##   Té        1  3
```

Test ChiCuadrado (a pesar de que las observadas no son mayores o iguales a 5)

```
res=chisq.test(pruebate)
```

```
## Warning in chisq.test(pruebate): Chi-squared approximation may be incorrect
```

```
res
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  pruebate
## X-squared = 0.5, df = 1, p-value = 0.4795
```

No hay razones para rechazar.

```
res$expected
```

```
##        Verdad
## Predice Leche Té
##   Leche     2  2
##   Té        2  2
```

Lo observado no se espera mucho de lo esperado.

Test exacto de Fisher

En esta situación es más apropiado aplicar el Test exacto de Fisher:

```
fisher.test(pruebate, alternative = "greater")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  pruebate
## p-value = 0.2429
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  0.3135693       Inf
## sample estimates:
## odds ratio
##   6.408309
```

Ejercicio 7: Gafas y antecedentes.

```
gafasante <-
  matrix(c(1, 8, 5, 2), nrow = 2,
         dimnames = list(Gafas = c("Sí", "No"),
                         Antecedentes = c("Sí", "No")))
gafasante
```

```
##      Antecedentes
## Gafas Sí No
##    Sí  1  5
##    No  8  2
```

Se requiere contrastar que H0: Variables categóricas independientes.

```
chisq.test(gafasante) # Muestres independientes.
```

```
## Warning in chisq.test(gafasante): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  gafasante
## X-squared = 3.8095, df = 1, p-value = 0.05096
```

Test de fisher:

```
fisher.test(gafasante)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  gafasante
## p-value = 0.03497
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.0009525702 0.9912282442
## sample estimates:
## odds ratio
## 0.06464255
```

Rechazo H0, existe un comportamiento diferente entre las variables

Ejercicio 8: Datos relacionados

Test de McNemar (datos relacionados).

Datos relacionados, por ejemplo antes-después.

Dos encuestas con un mes de separación, se pregunta a cada uno de los 1600 encuestados si aprueba o desaprueba a un gobernante.

```
datos <- matrix(c(794, 86, 150, 570), nrow = 2,
       dimnames = list("Primera encuesta" = c("Aprueba", "Desaprueba"),
                       "segunda encuesta" = c("Aprueba", "Desaprueba")))
datos
```

```
##                  segunda encuesta
## Primera encuesta Aprueba Desaprueba
##       Aprueba        794        150
##       Desaprueba      86        570
```

Muestras dependientes

```
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
```

El p-valor es muy pequeño y menor que alpha, rechazo H0: las muestras se comportan igual. Es decir, ha habido un cambio de opcinión de una encuesta a otra.