Contrastes no paramétricos

Esta práctica tiene dos partes, en la primera se utilizan los contrastes basados en la distribución χ^2 para responder algunas preguntas sobre la distribución del numero de goles en un partido de fútbol. En la segunda se investigan algunas de las propiedades del contraste de Kolmogorov-Smirnov.

Contrastes basados en la distribución χ^2

El fichero **goles0809.RData** contiene el dataframe **goles0809** con los datos de los goles de la liga española de fútbol del año 2008-09. El objetivo de esta sección es contrastar si las distribuciones de los goles conseguidos en casa y fuera son iguales. Para ello, vamos a llevar a cabo un contraste χ^2 de homogeneidad.

Tenemos que situar el fichero con los datos en el directorio de trabajo. Para cargarlo en la memoria de trabajo y ver los primeros datos, escribimos el código siguiente:

```
load('goles0809.Rdata')
head(goles0809)
```

```
##   casa fuera
## 1    2     1
## 2    1     1
## 3    0     2
## 4    1     0
## 5    0     1
## 6    0     3
```

Vemos que hay dos variables con los goles conseguidos por el equipo local y el visitante en cada partido. Lo primero que tenemos que calcular son las frecuencias observadas de la dos variables:

```
obscasa <- table(goles0809$casa)
obsfuera <- table(goles0809$fuera)
obscasa
```

```
##
##   0   1   2   3   4   5   6   7
##  75 123  90  58  22   7   4   1
```

```
obsfuera
```

```
##
##   0   1   2   3   4   5   6
## 117 135  73  38  13   3   1
```

Como las frecuencias para 5 o más goles son muy bajas, las vamos a agrupar:

```
obscasa <- c(obscasa[1:5], sum(obscasa[6:8]))
obsfuera <- c(obsfuera[1:5], sum(obsfuera[6:7]))
```

Antes de poder aplicar el comando que hace los cálculos del contraste, tenemos que formar la matriz con todas las frecuencias. También damos nombre a la columna que contine los datos agrupados.

```
obs <- rbind(obscasa, obsfuera)
colnames(obs)[6] <- '>4'
obs
```

```
##            0   1  2  3  4 >4
## obscasa   75 123 90 58 22 12
## obsfuera 117 135 73 38 13  4
```

Finalmente llevamos a cabo el contraste de homogeneidad:

```
chisq.test(obs)
```

```
##
##  Pearson's Chi-squared test
##
## data:  obs
## X-squared = 21.9996, df = 5, p-value = 0.0005237
```

Se observa que el p-valor es aproximadamente 0, por lo que se rechaza que la distribución es la misma para los niveles de significación habituales. Por defecto se obtiene la información básica. En realidad lo que calcula R es una lista cuya estructura se puede ver de la forma siguiente:

```
resultado <- chisq.test(obs)
ls.str(resultado)
```

```
## data.name :  chr "obs"
## expected :  num [1:2, 1:6] 96 96 129 129 81.5 81.5 48 48 17.5 17.5 ...
## method :  chr "Pearson's Chi-squared test"
## observed :  int [1:2, 1:6] 75 117 123 135 90 73 58 38 22 13 ...
## p.value :  num 0.000524
## parameter :  Named int 5
## residuals :  num [1:2, 1:6] -2.143 2.143 -0.528 0.528 0.942 ...
## statistic :  Named num 22
## stdres :  num [1:2, 1:6] -3.506 3.506 -0.919 0.919 1.502 ...
```

Si solo nos interesa calcular el estadístico de contraste o necesitamos la matriz de frecuencias esperadas podemos calcularlos de la siguiente forma:

```
resultado$statistic
```

```
## X-squared
##   21.9996
```

```
resultado$expected
```

```
##           0   1    2  3    4 >4
## obscasa  96 129 81.5 48 17.5  8
## obsfuera 96 129 81.5 48 17.5  8
```

Ejercicios

1. Contrasta si la diferencia de goles entre los dos equipos que juegan cada partido sigue una distribución uniforme.
2. Contrasta si la diferencia de goles entre los dos equipos que juegan cada partido sigue una distribución de Poisson.

Contraste de Kolmogorov-Smirnov

En primer lugar vamos a escribir una función (**ksnoest**) que calcula el estadístico de Kolmogorov-Smirnov cuando estamos interesados en contrastar si nuestros datos siguen una distribución normal estándar:

```
ksnoest <- function(datos){
    y <- ks.test(datos,pnorm)$statistic
    return(y)
}
```

Para comprobar si funciona podemos generar una muestra normal de tamaño 20 y aplicar después la función:

```
ksnoest(rnorm(20))
```

```
##         D
## 0.1949476
```

Supongamos que queremos contrastar la hipótesis nula de que los datos son normales (con valores arbitrarios de la media y la desviación típica). Una posibilidad **(que, como veremos, no funciona)** es estimar los parámetros de la normal y comparar la función de distribución emppírica F_n con la función de distribución de una variable $\mathrm{N}(\hat{\mu}, \hat{\sigma}^2)$. La siguiente función calcula el correspondiente estadístico de Kolmogorov-Smirnov:

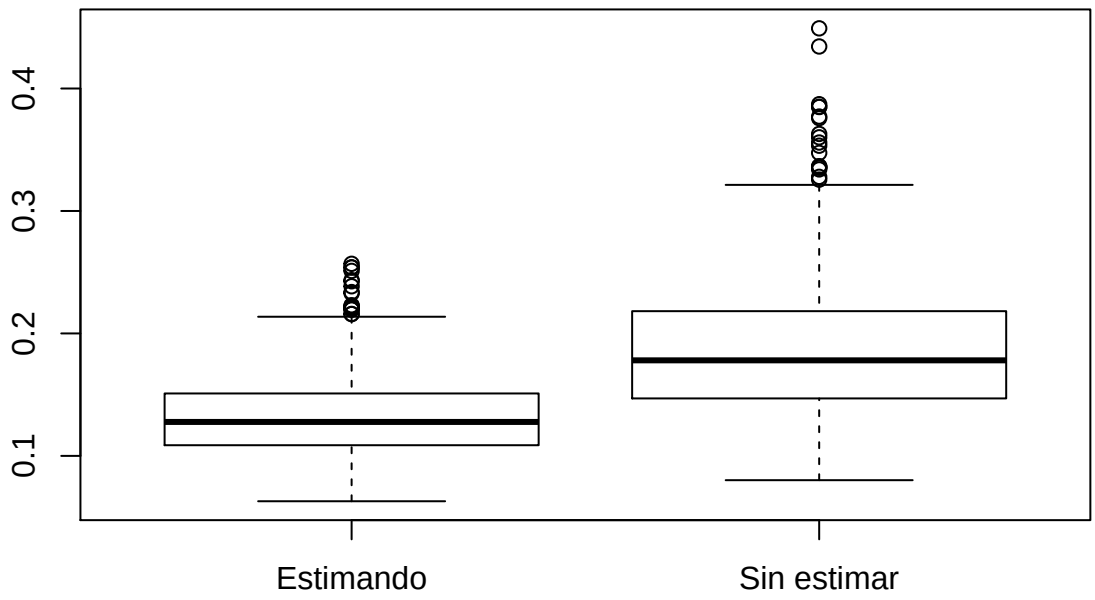
```
ksest <- function(datos){
    mu <- mean(datos)
    stdev <- sd(datos)
    y <- ks.test(datos, pnorm, mean=mu, sd=stdev)$statistic
    return(y)
}
```

Para ver la diferencia entre las dos situaciones, generamos 1000 muestras de tamaño 20 y calculamos ambos estadísticos para cada una de ellas:

```
B <- 1000
n <- 20
datos <- matrix(rnorm(n*B), n)
test <- apply(datos, 2, ksest)
tnoest <- apply(datos, 2, ksnoest)
```

Mediante diagramas de cajas, podemos comparar las distribuciones del estadístico en cada caso:

```
boxplot(test,tnoest, names=c('Estimando','Sin estimar'))
```



Ejercicios

1. Claramente las distribuciones de **test** y de **tnoest** son diferentes, por lo que no podemos usar las mismas tablas para hacer el contraste en las dos situaciones. ¿En cuál de los dos casos se obtienen en media valores menores? ¿Podrías dar una razón intuitiva?

2. Imagina que estimamos los parámetros y usamos las tablas de la distribución del estadístico de Kolmogorov-Smirnov para hacer el contraste a nivel α. El verdadero nivel de significación, ¿es mayor o menor que α?

3. Para resolver el problema se ha estudiado la distribución en el caso de muestras normales con parámetros estimados. Es lo que se conoce como contraste de normalidad de Kolmogorov-Smirnov-Lilliefors (KSL) (véase, por ejemplo, Peña (2001), pag. 471 y Tabla 9). Según la tabla del estadístico KSL, el nivel crítico para $\alpha = 0.05$ y $n = 20$ es 0.190. Esto significa que el porcentaje de valores test} mayores que 0.19 en nuestra simulación debe ser aproximadamente del 5%. Compruébalo haciendo `sum(test > 0.19)/B}`. Haz una peque~na simulación similar a la anterior para aproximar el nivel de significación del contraste KSL cuando se utiliza un valor crítico 0.12 para muestras de tamaño 40.

4. Genera $B = 10000$ muestras de tamaño $n = 30$ de una distribución exponencial de media 1 y utilízalas para determinar en este caso la potencia aproximada del test de Kolmogorov-Smirnov con $\alpha = 0.05$ para $H_0 : X \equiv \mathrm{N}(1, 1)$. (El comando `rexp()` puede utilizarse para generar los datos exponenciales).