# Práctica 2 Estadística II

*Alberto Parramón Castillo*

En primer lugar cargamos los datos en la variable *goles*0809

```
load('goles0809.RData')
```

## Contrastes basados en la distribución $\chi^2$

### Ejercicio 1

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

Así tenemos como hipótesis nula: $H_0 : X \sim Uniforme$.

Guardamos los goles en casa y los goles fuera de casa en variables diferentes. Los restamos y sacamos su valor absoluto (ya que lo que nos importa es la diferencia y no el el signo), después clasificamos esos goles en una tabla:

```
golesCasa <- goles0809$casa
golesFuera <- goles0809$fuera
difGoles <- golesCasa - golesFuera
difGoles <- abs(difGoles)
difGoles <- table(difGoles)
difGoles
```

```
## difGoles
##   0   1   2   3   4   5   6
##  83 160  78  38  13   5   3
```

Agrupamos las dos ultimas columnas en una sola:

```
difGoles <- c(difGoles[1:5], sum(difGoles[6:7]))
names(difGoles)[6] <- '>4'
difGoles
```

```
##   0   1   2   3   4  >4
##  83 160  78  38  13   8
```

Por defecto la función chisq.test te calcula la diferencia de goles suponiendo una distribución uniforme :

```
chisq.test(difGoles)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  difGoles
## X-squared = 255.53, df = 5, p-value < 2.2e-16
```

Sale un p-valor muy cercano a 0, por tanto para casi cualquier nivel de significación $\alpha$ se rechaza la hipótesis nula. Rechazamos la idea de que la diferencia de goles siga una distribución uniforme.

## Ejercicio 2

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

Así tenemos como hipótesis nula: $H_0 : X \sim Poisson(lambda)$.

Al igual que antes sacamos la tabla de los goles:

```
difGoles <- golesCasa - golesFuera
difGoles <- abs(difGoles)
difGoles <- table(difGoles)
```

Ahora calculamos el EMV de $\lambda$:

```
clases = seq(0,6)
n = sum(difGoles)
lambda = sum(clases*difGoles)/n
lambda
```

```
## [1] 1.381579
```

Calculamos las probabilidades estimadas de cada clase, así como las esperanzas estimadas de cada clase:

```
prob = dpois(clases, lambda)
esp = n*prob
esp
```

```
## [1]  95.449022 131.870360  91.094656  41.951486  14.489823   4.003767
## [7]   0.921920
```

Agrupamos las clases 6 y 7 ya que valen menos de 5.

```
difGoles <- c(difGoles[1:5], sum(difGoles[6:7]))
prob <- c(prob[1:5], 1-sum(prob[1:5]))
esp <- c(esp[1:5], n-sum(esp[1:5]))
```

Obtenemos el estadístico y el p-valor, pero el p-valor que obtiene R en las hipótesis nulas compuestas no es correcto. Por ello lo calculamos con la tabla de la $\chi^2$ con $k-1-r$ grados de libertad. Como k=6 (que son las clases) y r=1 (que es la dimensión del EMV), nos queda 4:

```
t=chisq.test(difGoles, p=prob)$statistic
pvalor = 1-pchisq(t,4)
pvalor
```

```
##  X-squared
## 0.02044257
```

El p-valor es 0.02, por tanto, a veces rechazaríamos la hipótesis nula, es decir, rechazaríamos que los datos siguen una distribución de Poisson, y otras veces no. Dependerá del nivel de significación que queramos asumir, para niveles de significacion $\alpha > 0.02$ rechazaríamos la hipótesis nula.

Por ejemplo, si tenemos un nivel de significación $\alpha = 0.01$, no rechazaríamos la hipótesis nula, ya que $\alpha = 0.01$ quiere decir que queremos rechazar la hipótesis nula con una probabilidad máxima de equivocarnos del 1%, sin embargo, el análisis que hemos obtenido, nos da un $p - valor = 0.02$, eso quiere decir, que al menos tenemos que afrontar una probabilidad de equivocarnos al rechazar la hipótesis nula de un 2%.

Con nivel de significación $\alpha = 0.05$ si rechazaríamos la hipótesis nula, ya que asumimos una probabilidad máxima de equivocarnos del 5% y el p-valor nos dice que tenemos solo un 2% de probabilidades de equivocarnos.

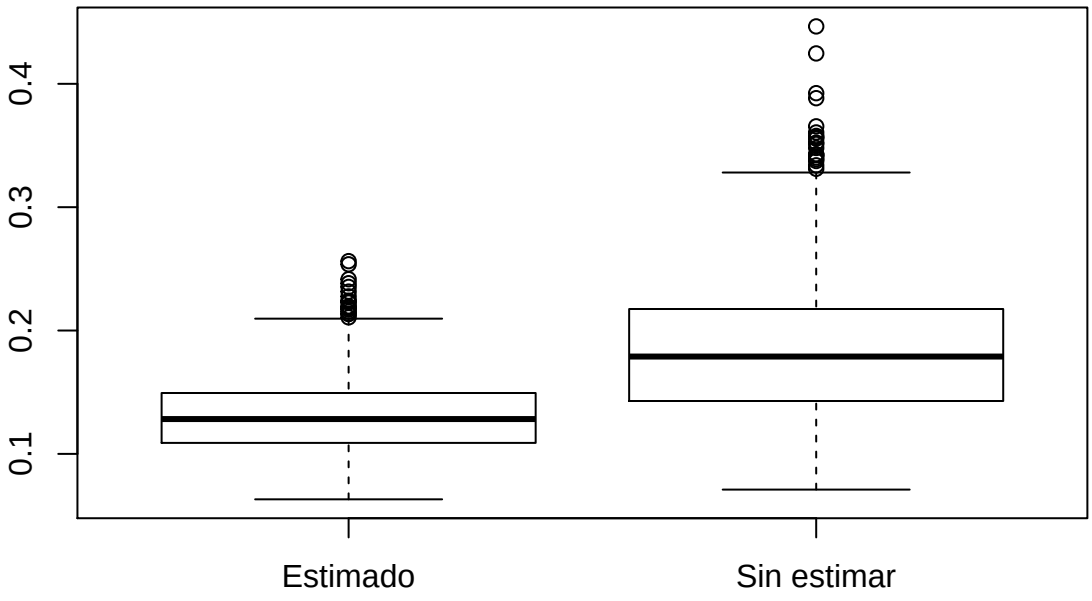
## Contraste de Kolmogorov-Smirnov

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

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

B <- 1000
n <- 20
datos <- matrix(rnorm(n*B), n)
test <- apply(datos, 2, ksest) #El 2 es para hacerlo por columnas
tnoest <- apply(datos, 2, ksnoest)

boxplot(test, tnoest, names=c("Estimado", "Sin estimar"))
```



## Ejercicio 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?

Lo que representamos en las cajas es el valor del estadístico. De media se obtienen valores más pequeños en el estimado.

Sabemos que un valor del estadístico pequeño implica un p-valor grande, y un p-valor grande implica que la probabilidad de equivocarnos si decidimos rechazar la hipótesis nula es grande. Por otro lado, sabemos que el p-valor (y por tanto el valor del estadístico) dependen de los datos de partida y de la hipótesis nula. En este caso, los datos de partida son los mismos en ambos casos (ambos proceden de generaciones aleatorias de muestras de una $N(0,1)$) por tanto la diferencia entre la primera caja y la segunda esta relacionada con la hipótesis nula ($H_0$):

En la segunda caja (Sin estimar), simplemente comparamos los datos aleatorios que generamos con una distribución $N(0,1)$, por tanto, los datos pueden haber salido un poco diferentes a esa $N(0,1)$, ya que son aleatorios; pero como provienen precisamente de una $N(0,1)$ es de esperar que se parezcan bastante a esta y que el valor del estadístico sea bastante grande, y que el p-valor sea por tanto bastante pequeño.

Mientras que en la primera caja (Estimado) estimamos la media y la desviación típica de los datos, y después suponemos como $H_0$ que los datos siguen una distribución normal de media y desviación típica las estimadas a partir de los datos. Por tanto, es natural que los datos se parezcan mucho más a esa distribución dada por $H_0$ que los de la segunda caja, y por tanto, es bastante intuitivo pensar que el p-valor va a salir bastante grande, y por tanto el valor del estadístico bastante pequeño.

En ambos casos, seguramente no rechazaríamos la hipótesis nula para valores de $\alpha$ habituales (0.01 o 0.05).

## Ejercicio 2

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

En la caja de los estimados tenemos los valores de los estadísticos más pequeños que en la caja de los contrastes sin estimar. Por tanto, los p-valores son más altos en los contrastes estimados que en los que están sin estimar. Por tanto, para un $\alpha$ en los estimados, el $\alpha$ que salga en los que están sin estimar será más pequeño.

Esto es intuitivo si volvemos a interpretar $\alpha$ como la probabilidad máxima que queremos asumir de equivocarnos al rechazar la hipótesis nula. Como hemos visto antes, la hipótesis nula es menos rechazable en de los contrastes con los datos estimados que en los que están sin estimar. Por tanto, escojo un valor $\alpha$ en los estimados, que representará una probabilidad máxima de equivocarme al rechazar $H_0$ de un $x\%$. Este $\alpha$ escogido llevará asociado un nivel crítico (valor en la tabla) en el contraste de los datos estimados, y ese mismo nivel crítico en el contraste de los parámetros sin estimar llevará asociado un valor $\alpha$ más pequeño que el anterior. Esto es razonable ya que en el contraste sin estimar la probabilidad de equivocarnos al rechazar $H_0$ es algo menor.

## Ejercicio 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 pequeña

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.

Si asumimos un $\alpha = 0.05$ es que asumimos una probabilidad máxima de equivocarnos al rechazar $H_0$ del 5%. Vamos a contrastar datos que provienen de una distribución normal, con la hipótesis nula de que siguen una distribución normal de parámetros $\mu$ y $sd$ estimados empirícamente. Por tanto, si rechazamos $H_0$ claramente nos estamos equivocando, y la prbabilidad de equivocarnos al rechazar $H_0$ viene determinada por $\alpha$. Por tanto, si $\alpha = 0.05$, lleva asociado un nivel crítico de 0.19, quiere decir que sólo nos vamos a encontrar con datos que provoquen un valor estadístico T>0.19 (es decir, entrando en la región de rechazo) en un 5% de los casos que estudiemos.

Lo comprobamos:

```
B <- 1000
n <- 20
datos <- matrix(rnorm(n*B), n)
test <- apply(datos, 2, ksest)
sum(test>0.19)/B
```

```
## [1] 0.056
```

Ahora vamos a calcular $\alpha$ sabiendo que el nivel crítico es 0.12 y las muestras son de tamaño n=40:

```
B <- 1000
n <- 40
datos <- matrix(rnorm(n*B), n)
test <- apply(datos, 2, ksest)
alpha = sum(test>0.12)/B
#Mostramos el valor de alpha:
alpha
```

```
## [1] 0.135
```

## Ejercicio 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 $H_0 \equiv N(1, 1)$. El comando rexp() puede utilizarse para generar los datos exponenciales).

Obtenemos de la tabla de Kolmogorov-Smirnov el valor para $\alpha = 0.05$: $D_{\alpha=0.05} = 0.24$.

Comprobamos que 0.24 es el nivel crítico para $\alpha = 0.05$, para ello, generamos muestras de una N(1,1) y comprobamos que la probabilidad de rechazar $H : 0$ siendo esta verdadera es de un 5%:

```
ksej4_1 <- function(datos){
  y <- ks.test(datos, pnorm, mean=1, sd=1)$statistic
  return(y)
}
```

```
B <- 10000
n <- 30
datos <- matrix(rnorm(n*B, mean=1, sd=1), n)
test <- apply(datos, 2, ksej4_1)
sum(test>0.24)/B
```

```
## [1] 0.0483
```

Vemos que nos sale aproximadamente un 5%. Ahora vamos con los que nos pide el enunciado. La potencia del contraste es ver cuántas veces se rechaza la hipótesis nula:

```
ksej4_2 <- function(datos){
  y <- ks.test(datos, pnorm, mean=1, sd=1)$statistic
  return(y)
}
```

```
B <- 10000
n <- 30
datos <- matrix(rexp(n*B), n)
test <- apply(datos, 2, ksej4_2)
sum(test>0.24)/B
```

```
## [1] 0.2886
```

Por tanto tenemos una potencia del contraste de aproximadamente un 29%

# Hoja 2 de ejercicios

## Ejercicio 9

A finales del siglo XIX el físico norteamericano Newbold descubrió que la proporción de datos que empiezan por una cifra d, p(d), en listas de datos correspondientes a muchos fenómenos naturales y demográficos es aproximadamente:

$$p(d) = log_{10}\left(\frac{d+1}{d}\right), (d = 1, 2, ..., 9)$$

Por ejemplo, $p(1) = log_{10}2 \approx 0,301030$ es la frecuencia relativa de datos que empiezan por 1. A raíz de un artículo publicado en 1938 por Benford, la fórmula anterior se conoce como ley de Benford. El fichero poblacion.RData incluye un fichero llamado poblaciones con la población total de los municipios españoles, así como su población de hombres y de mujeres. (Indicación: Puedes utilizar, si te sirven de ayuda, las funciones del fichero benford.R).

Aquí tenemos las funciones del fichero benford.R

```
#-----------------------------------------------------
#
# Una funcion para contar las frecuencias:
# Dado un vector x, esta funcion calcula la frecuencia de valores
# que empiezan por 1, 2, ..., 9
#
#-----------------------------------------------------
benford = function(x){
    n = length(x)
    proporcion = numeric(9)
    for (i in 1:9){
        proporcion[i] = sum(substr(x,1,1)==as.character(i))
    }
```

```
    return(proporcion)
}

#-----------------------------------------------------------------
# Una funcion para contar las frecuencias de los dos primeros digitos
# Dado un vector x, esta funcion calcula la tabla de frecuencias de los valores
# de los pares (i,j) donde i = 1, 2, ..., 9 y j = 0, 1, ..., 9
# (solo considera valores mayores o iguales que 10)
#
#-----------------------------------------------------------------
benford2 = function(x){
    x = x[x>=10]
    n = length(x)
    proporcion = matrix(0,9,10)
    digitos = substr(x,1,2)

    for (i in 1:9){
        for (j in 1:10){
            proporcion[i,j] = sum(digitos==paste(i,j-1,sep=''))/n
        }
    }
    colnames(proporcion) = paste(0:9)
    rownames(proporcion) = paste(1:9)
    return(proporcion)
}
```

En primer lugar cargamos el fichero benford.R

```
load('poblacion.RData')
```

A) Contrasta a nivel $\alpha = 0{,}05$ la hipótesis nula de que la población total se ajusta a la ley de Benford.

Definimos una función que nos devuelve las probabilidades de cada clase (dígito) según $H_0$, es decir, suponiendo que los dígitos siguen la distribución dada por Benford:

```
probBenford = function(){
    proporcion = numeric(9)
    for (i in 1:9){
        proporcion[i] = log10((i+1)/i)
    }
    return(proporcion)
}
```

Utilizamos el contraste de bondad de ajuste basados en la distribución $\chi^2$.

```
pobTotalFrecuencias <- benford(poblaciones$pobtotal)
prob = probBenford()
chisq.test(pobTotalFrecuencias, p=prob)
```

```
##
##  Chi-squared test for given probabilities
```

```
##
## data:  pobTotalFrecuencias
## X-squared = 13.5, df = 8, p-value = 0.09575
```

Como el p-valor es 0.095, que es mayor que 0.05, no podemos rechazar la hipótesis nula $H_0$ a nivel de significacion $\alpha = 0.05$.

B) Repite el ejercicio pero considerando sólo los municipios de más de 1000 habitantes.

```
pob1000 = poblaciones$pobtotal[poblaciones$pobtotal > 1000]
pob1000Frecuencias <- benford(pob1000)
prob = probBenford()
chisq.test(pob1000Frecuencias, p=prob)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  pob1000Frecuencias
## X-squared = 298.91, df = 8, p-value < 2.2e-16
```

Como el p-valor es 2.2e-16, que es menor que 0.05, podemos rechazar la hipótesis nula $H_0$ a nivel de significacion $\alpha = 0.05$.

C) Considera las poblaciones totales (de los municipios con 10 o más habitantes) y contrasta a nivel $\alpha =$ 0,05 la hipótesis nula de que el primer dígito es independiente del segundo.

```
n = length(poblaciones$pobtotal[poblaciones$pobtotal >= 10])
frecuencias = n*benford2(poblaciones$pobtotal)
chisq.test(frecuencias)
```

```
##
##  Pearson's Chi-squared test
##
## data:  frecuencias
## X-squared = 120.52, df = 72, p-value = 0.0002974
```

Como el p-valor es 0.0002974, que es menor que 0.05, podemos rechazar la hipótesis nula $H_0$ a nivel de significacion $\alpha = 0.05$.