Práctica 1: Estimación puntual e Intervalos de confianza

Módulo de Modelos Lineales.
Máster de Bioestadística, Universitat de València.

Miguel A. Martinez-Beneito

Tareas

1. La media y la mediana son dos estimadores de tendencia central en distribuciones, ampliamente conocidos y utilizados. En esta tarea nos vamos a plantear su comparación como estimadores de la media de una distribución Normal. Para ello vamos a hacer uso de procedimientos de tipo empírico más que de razonamientos teóricos. Así:

- Genera una muestra de valores de una distribución Normal standard de tamaño 50, para dicha muestra calcula su media y su mediana.
- Repite el procedimiento anterior 100 veces para 100 muestras distintas.
- A partir de todas las medias y medianas que has calculado en los pasos anteriores, calcula el Error Cuadrático Medio de ambos estimadores y comparalos ¿Qué estimador consideras más adecuado a tenor de los resultados que has obtenido?
- Por último, repito todo el proceso anterior, para una distribución t con 1 grado de libertad y valora si tus conclusiones cambian en función de la distribución de la que provienen los datos.

```r
set.seed(1)

# Generación de una muestra
n <- 50
muestra <- rnorm(n)

# Generación de 100 réplicas del proceso anterior
replicas <- 100
resul <- matrix(nrow = replicas, ncol = 2)
resul[1, ] <- c(mean(muestra), median(muestra))
for (i in 2:replicas) {
  muestra <- rnorm(n)
  resul[i, ] <- c(mean(muestra), median(muestra))
}

# MSE de las medias y medianas (theta=0 y el valor esperado de las MSE se estima como la
# media de las muestras)
c(mean(resul[, 1]^2), mean(resul[, 2]^2))
```

```
## [1] 0.01695109 0.02714257
```

```r
# Las medias parecen tener menor MSE, por tanto paraecen más aconsejables, como estimador
# del valor esperado en poblaciones Normales

# Repetimos el proceso anterior para distribuciones t con 1 grado de libertad:
muestra <- rt(n, df = 1)
resul <- matrix(nrow = replicas, ncol = 2)
```

```r
resul[1, ] <- c(mean(muestra), median(muestra))
for (i in 2:replicas) {
  muestra <- rt(n, df = 1)
  resul[i, ] <- c(mean(muestra), median(muestra))
}
c(mean(resul[, 1]^2), mean(resul[, 2]^2))
```

```
## [1] 33.85413774  0.05197988
```

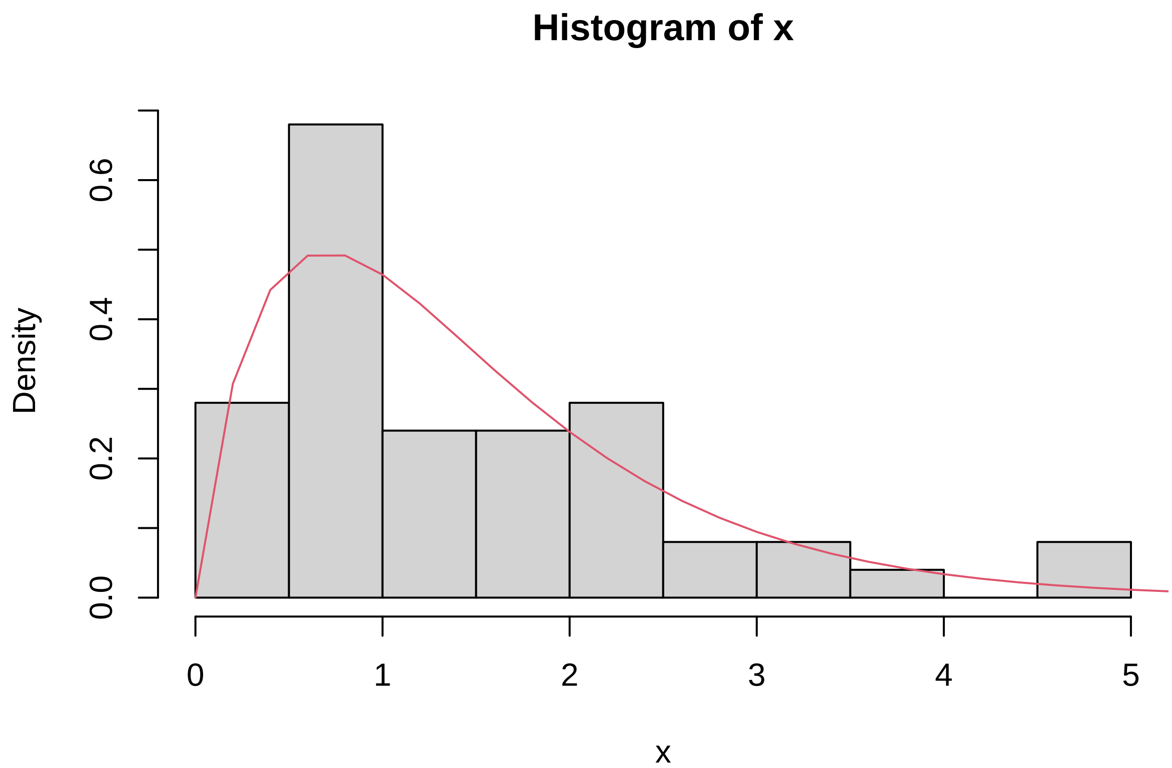
```r
# En este caso los MSE de las medias son bastante superiores a los de las medianas, por
# tanto las medianas son estimadores de tendencia central bastante mas aconsejables para
# poblaciones t con un grado de libertad.
```

2. Supongamos que disponemos de la siguiente muestra de valores: `set.seed(1); x <- exp(rnorm(50))`, todos ellos valores positivos en la recta real. Para este conjunto de datos, nos planteamos ajustarles una distribución $Gamma(\alpha, \beta)$, adecuada para este tipo de datos con valores positivos. Halla, haciendo uso de R, los estimadores MLE de α y β y representa un histograma de la muestra de valores `x`, con la distribución Gamma que hayas estimado superpuesta. Haciendo uso de la aproximación Normal de los MLE calcula un intervalo de confianza al 95% para el parámetro α de la distribución que acabas de calcular.

```r
set.seed(1)
x <- exp(rnorm(50))
minusLL <- function(alpha, beta) {
  -sum(dgamma(x, alpha, beta, log = TRUE))
}
# MLEs
estimadores <- stats4::mle(minusLL, start = list(alpha = 1, beta = 1))
estimadores
```

```
##
## Call:
## stats4::mle(minuslogl = minusLL, start = list(alpha = 1, beta = 1))
##
## Coefficients:
##    alpha     beta
## 1.896701 1.288481
```

```r
# Representación
hist(x, freq = FALSE)
lines((0:50) / 5, dgamma((0:50) / 5, estimadores@coef[1], estimadores@coef[2]), col = 2)
```




```
# ICs
sds <- sqrt(diag(estimadores@vcov))
# Extremos inferiores
estimadores@coef - 1.96 * sds
```

```
##     alpha      beta
## 1.2085722 0.7539025
```

```
# Extremos superiores
estimadores@coef + 1.96 * sds
```

```
##    alpha     beta
## 2.584830 1.823059
```

3. Reproduce por ti mismo el ejemplo de la página 17 del Tema 1 de la asignatura. Comprueba que los resultados que obtienes en cuanto a la proporción de veces que los intervalos de confianza contienen el valor 0 son similares a los de los apuntes.

```
# Cálculo de un intervalo de confianza
set.seed(1)
x <- rnorm(100)
IC.Inf <- mean(x) - 1.96 / 10
IC.Sup <- mean(x) + 1.96 / 10
c(IC.Inf, IC.Sup)
```

```
## [1] -0.08711263  0.30488737
```

```
# Replicamos el proceso anterior 1000 veces y valoramos el número de veces que los
# intervalos no contienen el valor 0
```

```r
fuera <- 0
for (i in 1:1000) {
  set.seed(i)
  x <- rnorm(100)
  IC.Inf <- mean(x) - 1.96 / 10
  IC.Sup <- mean(x) + 1.96 / 10
  if (IC.Inf > 0 | IC.Sup < 0) {
    fuera <- fuera + 1
  }
}
fuera
```

```
## [1] 42
```

4. Utiliza la función `t.test` de `R` para valorar si encuentras diferencias en las medias de las poblaciones de las que provienen las siguientes 2 muestras: `set.seed(1);x<-rnorm(10)` e `y<-rnorm(10,1)`. Eleva el tamaño muestral de ambas muestras a 20 y 30 para valorar como cambian tus conclusiones.

```r
# n<-10
set.seed(1)
n <- 10
x <- rnorm(n)
y <- rnorm(n, 1)
t.test(x, y)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -2.6669, df = 16.469, p-value = 0.01658
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.0022169 -0.2310675
## sample estimates:
## mean of x mean of y
## 0.1322028 1.2488450
```

```r
# n<-20
n <- 20
x <- rnorm(n)
y <- rnorm(n, 1)
t.test(x, y)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -4.3058, df = 37.797, p-value = 0.0001137
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.6838157 -0.6067209
## sample estimates:
##    mean of x    mean of y
## -0.006471519  1.138796773
```

```r
# n<-30
n <- 30
```

```r
x <- rnorm(n)
y <- rnorm(n, 1)
t.test(x, y)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -4.2128, df = 57.588, p-value = 8.981e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.4797304 -0.5263791
## sample estimates:
## mean of x mean of y
##  0.110278  1.113333
```

```r
# En los 3 casos encontramos que el intervalo de confianza para la diferencia de las medias
# no contiene el 0, por lo que concluimos diferencias significativas entre las medias de
# ambas poblaciones. En cualquier caso las diferencias parecen ser más evidenes conforme
# aumenta el tamaño de las muestras.
```