

Màster universitari en Estadística i Investigació Operativa (MESIO)

Computación en Estadística y en Optimización

Test 1 con R (Grupo A)

Instrucciones:

- Bajar los ficheros `CeoGrATestR1.R` y `CeoGrATestR1.RData` de ATENEA y guardarlos en un disco local o una memoria USB.
- Cambiar el nombre del script `CeoGrATestR1.R` a `CeoGrATestR1_ApellidoNombre.R`.
- Incluir en este script todas las instrucciones necesarias para resolver los ejercicios. Posibles comentarios se pueden incluir detrás de una almohadilla (`#`).
- Entregar el script vía ATENEA o por *email* a `klaus.langohr@upc.edu` antes de las 15h.

Ejercicio 1 (5,5 puntos)

El área de trabajo `CeoGrATestR1.RData` contiene el *data frame* `decathlon` del paquete `FactoMineR` con datos de distintos atletas de decatón del año 2004. Según la página de información sobre este conjunto de datos se trata de un

“... data frame with 41 rows and 13 columns: the first ten columns correspond to the performance of the athletes for the 10 events of the decathlon. The columns 11 and 12 correspond respectively to the rank and the points obtained. The last column is a categorical variable corresponding to the sporting event (2004 Olympic Games or 2004 Decastar).”

Además, se ha añadido la variable fecha de nacimiento en la columna 14.

- a) Cargad el área de trabajo `CeoGrATestR1.RData` y cambiad los nombres de las variables del *data frame* `decathlon` a minúscula.
- b) Cread un *data frame* con nombre `olymp` que contenga solamente los datos de los Juegos Olímpicos y borrad las variables `rank` y `competition`.
- c) Ordenad el *data frame* `olymp` según el apellido de los atletas.
- d) ¿Cuántos atletas lograron saltar más de 7,5 metros en el salto de longitud (*Long jump*)?
- e) ¿Empataron algunos atletas en puntos?
- f) ¿Cuál es la marca del atleta con la mejor marca en lanzamiento de disco y qué tiempo hizo en los 1500 metros?
- g) ¿Qué día de la semana nacieron más atletas? ¿Cuántos fueron?
- h) Calculad las correlaciones entre todas las disciplinas del decatón usando el coeficiente de Pearson.
- i) ¿Entre qué disciplinas hay la máxima correlación (en valor absoluto)? ¿Cuál es el valor de esta correlación?

Ejercicio 2 (3 puntos)

Reproducid el *boxplot* de la Figura 1, que representa la distribución de las puntuaciones de los atletas en función de la competición. Además muestra la media en ambas competiciones y el nombre del atleta con la puntuación máxima. Guardad el gráfico en formato JPG.

Notas:

- Utilizad el *data frame* `decathlon` para este ejercicio.
- Utilizad colores distintos para ambas cajas.
- Los valores numéricos de la abscisa correspondientes a ambas categorías son 1 y 2.

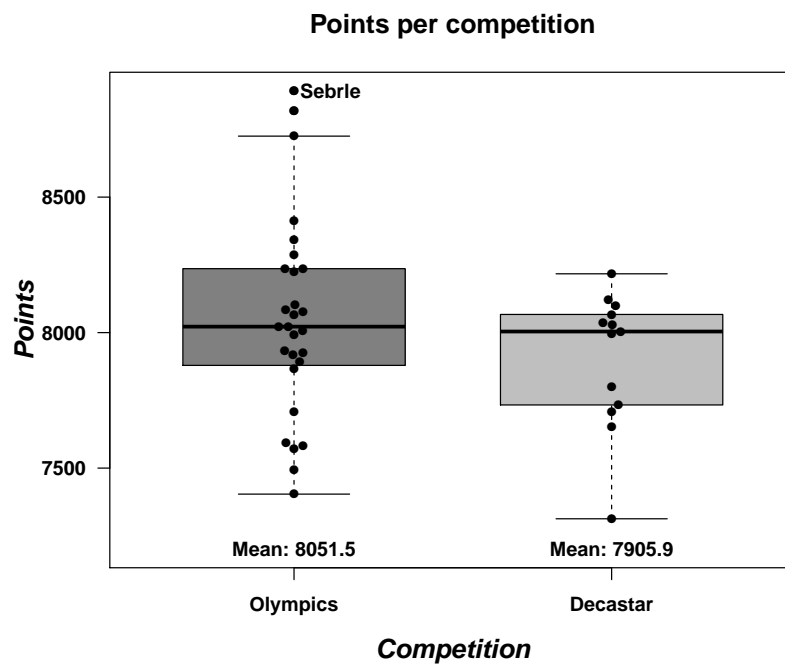


Figura 1: *Boxplot* del Ejercicio 2.

Ejercicio 3 (1,5 puntos)

El *data frame* `dfr` del área de trabajo `CeoGrATestR1.RData` contiene los valores de dos variables numéricas que han sido importados de forma errónea desde un fichero ASCII. Como resultado en R ambas variables, `x` e `y`, son factores:

```
> dfr
```

```

      x    y
1    5 1,23
2    5  8,7
3   10    5
4    2    *
5    4    *
6    7  7,5
7    4    *
8   10    2
9    6    2
10   3  3,2

```

```
> str(dfr)
```

```

'data.frame':      10 obs. of  2 variables:
 $ x: Factor w/ 7 levels "2","3","4","5",...: 4 4 7 1 3 6 3 7 5 2
 $ y: Factor w/ 7 levels "*","1,23","2",...: 2 7 5 1 1 6 1 3 3 4

```

- Convertid ambas variables en variables numéricas teniendo en cuenta que el símbolo "*" indica un valor perdido.
- Guardad el *data frame* `dfr` en un fichero RDS.