Taller problemas finales

Contenidos

1 Taller final ENUNCIADOS

Se trata de resolver los siguientes problemas y cuestiones en un fichero Rmd y su salida en un informe en html, word o pdf.

1.1 Problema 1: Contraste de parámetros de dos muestras.

Queremos comparar los tiempos de realización de un test entre estudiantes de dos grados G1 y G2, y determinar si es verdad que los estudiantes de G1 emplean menos tiempo que los de G2. No conocemos σ_1 y σ_2. Disponemos de dos muestras independientes de cuestionarios realizados por estudiantes de cada grado, $n_1 = n_2 = 50$.

Los datos están en https://github.com/joanby/estadistica-inferencial/, en la carpeta `datasets` en dos ficheros `grado1.txt` y `grado2.txt`.

Para bajarlos utilizad la dirección del los ficheros `raw` que se muestran en el siguiente código

```
G1=read.csv(
  "https://raw.githubusercontent.com/joanby/estadistica-inferencial/master/datasets/grado1.txt",
            header=TRUE)$x
G2=read.csv(
  "https://raw.githubusercontent.com/joanby/estadistica-inferencial/master/datasets/grado2.txt",
  header=TRUE)$x

n1=length(na.omit(G1))
n2=length(na.omit(G2))
media.muestra1=mean(G1,na.rm=TRUE)
media.muestra2=mean(G2,na.rm=TRUE)
desv.tip.muestra1=sd(G1,na.rm=TRUE)
desv.tip.muestra2=sd(G2,na.rm=TRUE)
```

Calculamos las medias y las desviaciones típicas muestrales de los tiempos empleados para cada muestra. Los datos obtenidos se resumen en la siguiente tabla:

$$\begin{array}{llll} n_1 & = 50, & n_2 & = 50 \\ \overline{x}_1 & = 9.7592926, & \overline{x}_2 & = 11.4660825 \\ \tilde{s}_1 & = 1.1501225, & \tilde{s}_2 & = 1.5642932 \end{array}$$

Se pide:

1. Comentad brevemente el código de R explicando que hace cada instrucción.
2. Contrastad si hay evidencia de que las notas medias son distintas entre los dos grupos. En dos casos considerando las varianzas desconocidas pero iguales o desconocidas pero distintas. Tenéis que hacer el contraste de forma manual y con funciones de `R` y resolver el contrate con el p-valor.
3. Calculad e interpretar los intervalos de confianza para la diferencia de medias asociados a los dos test anteriores.
4. Comprobad con el test de Fisher. Tenéis que resolver el test de Fisher con `R` o de forma manual con ayudados para los p-valores con alguan hoja de cálculo. Decidir utilizando el p-valor.

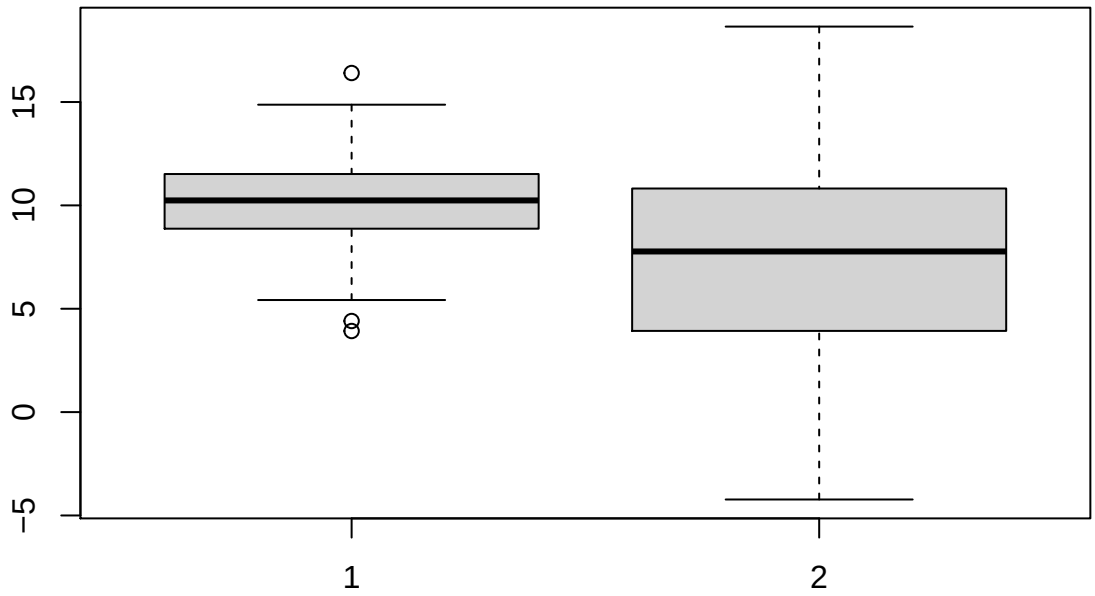
1.2 Problema 2 : Contraste dos muestras

Simulamos dos muestras con las funciones siguientes

```
set.seed(2020)
x1=rnorm(100,mean = 10,sd=2)
x2=rnorm(100,mean = 8,sd=4)
```

Dibujamos estos gráficos

```
boxplot(x1,x2)
```




```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

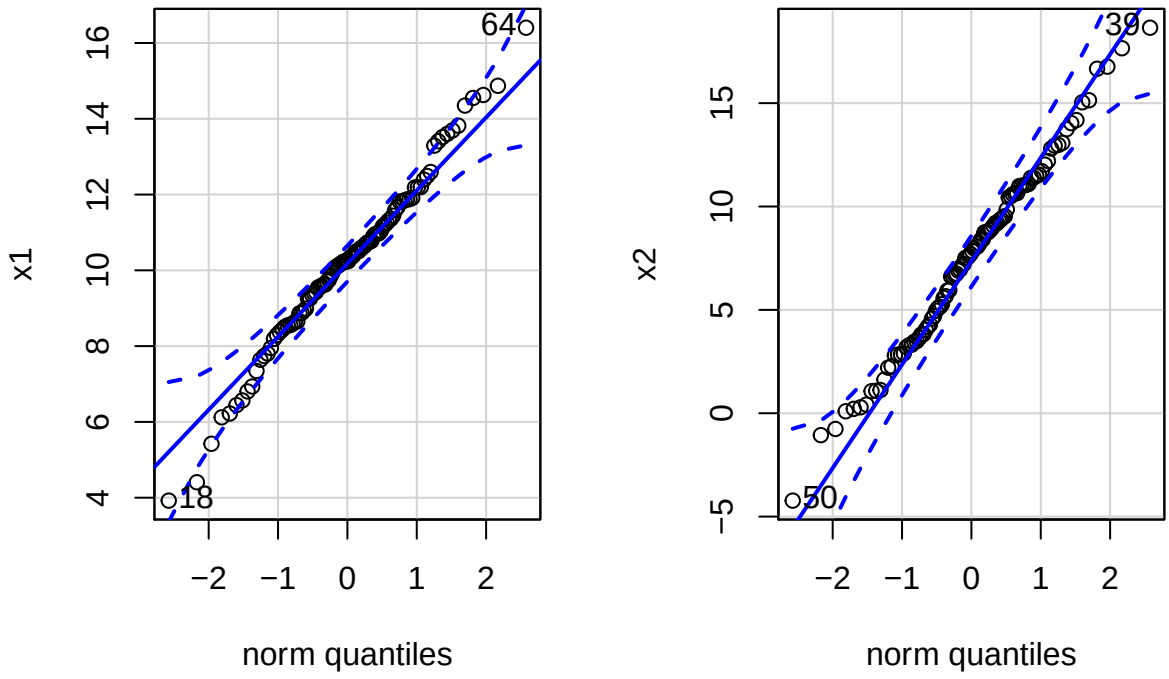
```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
par(mfrow=c(1,2))
qqPlot(x1)
```

```
## [1] 18 64
```

```
qqPlot(x2)
```




```
## [1] 50 39
```

```
par(mfrow=c(1,1))
```

Realizamos algunos contrastes de hipótesis de igual de medias entre ambas muestras

```
t.test(x1,x2,var.equal = TRUE,alternative = "greater")
```

```
##
##  Two Sample t-test
##
## data:  x1 and x2
## t = 5.3009, df = 198, p-value = 0.0000001531
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
```

```
##  1.844757      Inf
## sample estimates:
## mean of x mean of y
## 10.217784  7.537402
```

```
t.test(x1,x2,var.equal = FALSE,alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  x1 and x2
## t = 5.3009, df = 144.56, p-value = 0.0000004221
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.680966 3.679797
## sample estimates:
## mean of x mean of y
## 10.217784  7.537402
```

```
t.test(x1,x2,var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  x1 and x2
## t = 5.3009, df = 198, p-value = 0.0000003061
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.683238 3.677526
## sample estimates:
## mean of x mean of y
## 10.217784  7.537402
```

Se pide

1. ¿Cuál es la distribución y los parámetros de las muestras generadas?
2. ¿Qué muestran y cuál es la interpretación de los gráficos?
3. ¿Qué test contrasta si hay evidencia a favor de que las medias poblacionales de las notas en cada grupo sean distintas? Di qué código de los anteriores resuelve este test.
4. Para el test del apartado anterior dad las hipótesis nula y alternativa y redactar la conclusión del contraste.

1.3 Problema 3 : ANOVA Comparación de las tasas de interés para la compra de coches entre seis ciudades.

Consideremos el `data set newcar` accesible desde https://www.itl.nist.gov/div898/education/anova/newcar.dat de *Hoaglin, D., Mosteller, F., and Tukey, J. (1991). Fundamentals of Exploratory Analysis of Variance. Wiley, New York, page 71.*

Este data set contiene dos columnas:

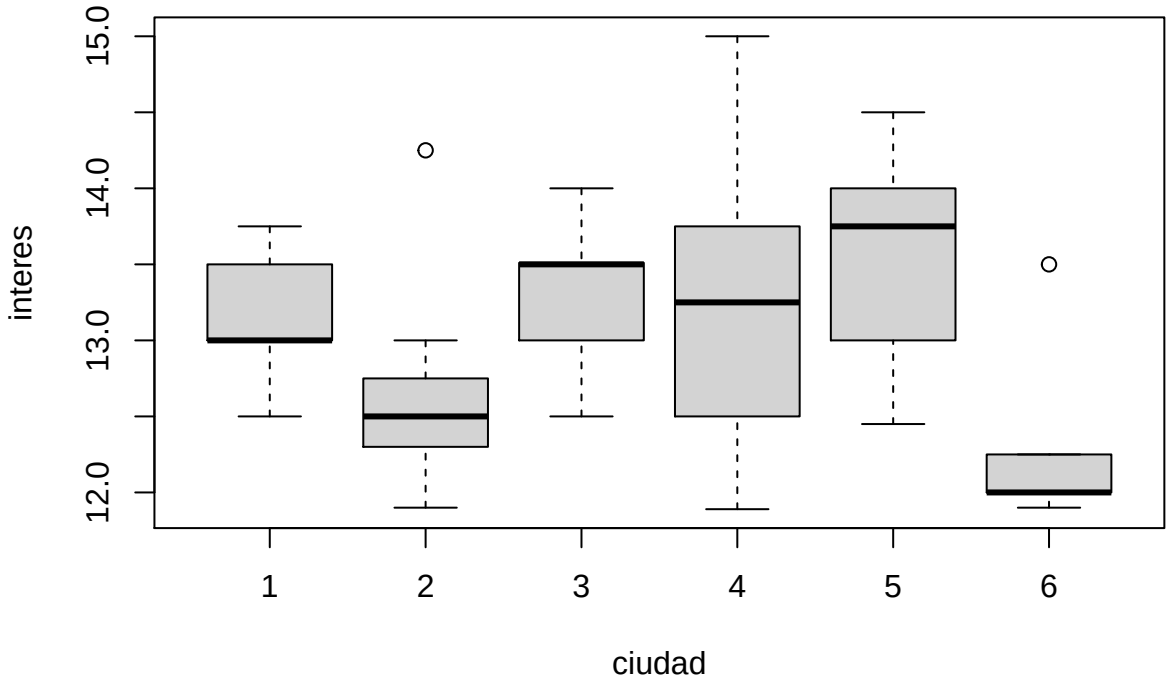
- Rate (interés): tasa de interés en la compra de coches a crédito
- City (ciudad) : la ciudad en la que se observó la tasa de interés para distintos concesionarios (codificada a enteros). Tenemos observaciones de 6 ciudades.

```
datos_interes=read.table(
  "https://www.itl.nist.gov/div898/education/anova/newcar.dat",
  skip=25)
```

```r
# salto las 25 primeras líneas del fichero,son un preámbulo que explica los datos.
names(datos_interes)=c("interes","ciudad")
str(datos_interes)
```

```
## 'data.frame':    54 obs. of  2 variables:
##  $ interes: num  13.8 13.8 13.5 13.5 13 ...
##  $ ciudad : int  1 1 1 1 1 1 1 1 1 2 ...
```

```r
boxplot(interes~ciudad,data=datos_interes)
```



Se pide:

1. Comentad el código y el diagrama de caja.
2. Se trata de contrastar si hay evidencia de que la tasas medias de interés por ciudades son distintas. Definid el ANOVA que contrasta esta hipótesis y especificar qué condiciones deben cumplir las muestras para poder aplicar el ANOVA.

3. Comprobad las condiciones del ANOVA con un test KS y un test de Levene (con código de `R`). Justificad las conclusiones.

4. Realizad el contraste de ANOVA (se cumplan las condiciones o no) y redactar adecuadamente la conclusión. Tenéis que hacedlo con funciones de `R`.

5. Se acepte o no la igualdad de medias realizar las comparaciones dos a dos con ajustando los p-valor tanto por Bonferroni como por Holm al nivel de significación $\alpha = 0.1$. Redactad las conclusiones que se obtienen de las mismas.

1.4 Problema 4: Cuestiones cortas

- Cuestión 1: Supongamos que conocemos el p-valor de un contraste. Para que valores de nivel de significación α RECHAZAMOS la hipótesis nula.
- Cuestión 2: Hemos realizado un ANOVA de un factor con 3 niveles, y hemos obtenido un p-valor de 0.001. Suponiendo que las poblaciones satisfacen las condiciones para que el ANOVA tenga sentido, ¿podemos afirmar con un nivel de significación $\alpha = 0.05$ que las medias de los tres niveles son diferentes dos a dos? Justificad la respuesta.

1.5 Problema 5: Contraste de proporciones de dos muestras independientes.

Queremos comparar las proporciones de aciertos de dos redes neuronales que detectan tipos si una foto con un móvil de una avispa es una avispa velutina o asiática. Esta avispa en una especie invasora y peligrosa por el veneno de su picadura. Para ello disponemos de una muestra de 1000 imágenes de insectos etiquetadas como avispa velutina y no velutina.

Em el github del curso os tenéis que descargar de la carpeta de datos los dicheros "algoritmo1.csv" y "algoritmo2.csv". Cada uno está en fichero los aciertos están codificados con 1 y los fallos con 0.

Se pide:

1. Cargad los datos los datos y calcular el tamaño de las muestras y la proporción de aciertos de cada muestra.
2. Contrastad si hay evidencia de que las las proporciones de aciertos del algoritmo 1 son mayores que las del algoritmo 2. Definid bien las hipótesis y las condiciones del contraste. Tenéis que hacer el contraste con funciones de `R` y resolver el contrate con el p-valor.
3. Calculad e interpretar los intervalos de confianza para la diferencia de proporciones asociados al test anterior, con funciones de R.