Taller problemas finales

Contenidos

1 Taller final SOLUCIONES

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.

1.1.1 Solución

Apartado 1. El cogido R carga en las variables G1 y G2 la variables `x` de dos data frames de un servidor en github y por lo tanto hemos tenido que pasar la url del fichero original o *raw*.

Luego calcula los estadísticos básicos para realizar las siguientes preguntas. Para los tamaños muestrales n_1 y n_2 se omiten los valores `NA` antes de asignar la `length` de los arrays. También se calculan las medias y las desviaciones típicas muestrales omitiendo (si es que hay) los valores no disponibles.

Apartado 2. Denotemos por μ_1 y μ_2 las medias de las notas de los grupos 1 y 2 respectivamente. El contraste que se pide es

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

Estamos en un diseño de comparación de medias de dos grupos con dos muestras independientes de tamaño 50 que es grande. Tenemos dos casos varianzas desconocidas pero iguales y varianzas desconocidas pero distintas. Las funciones de R del contraste para estos casos son:

Varianzas iguales

```
# test para varianzas iguales
t.test(G1,G2,var.equal = TRUE,alternative = "two.sided")
```

```
##
##  Two Sample t-test
##
## data:  G1 and G2
## t = -6.2159, df = 98, p-value = 0.00000001248
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.251691 -1.161889
## sample estimates:
## mean of x mean of y
##  9.759293 11.466083
```

Varianzas distintas

```
# test para varianzas distintas
t.test(G1,G2,var.equal = FALSE,alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  G1 and G2
## t = -6.2159, df = 89.996, p-value = 0.00000001562
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -2.252298 -1.161282
## sample estimates:
## mean of x mean of y
##  9.759293 11.466083
```

El p-valor en ambos casos es muy pequeño así que la muestra no aporta evidencias rechazar la hipótesis nula las medias son iguales contra que son distintas.

Veamos el cálculo manual.

Varianzas desconocidas pero iguales, n_1 y n_2 grande

Si suponemos que $\sigma_1 = \sigma_2$, el estadístico de contraste es

$$t0 = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2}) \cdot \frac{((n_1-1)\widetilde{S}_1^2 + (n_2-1)\widetilde{S}_2^2)}{(n_1+n_2-2)}}} = \frac{9.7592926 - 11.4660825}{\sqrt{(\frac{1}{50} + \frac{1}{50}) \cdot \frac{((50-1)1.1501225^2 + (50-1)1.5642932^2)}{(50+50-2)}}}$$

```
t0=(media.muestra1-media.muestra2)/sqrt((1/n1+1/n2)*
((n1-1) *desv.tip.muestra1^2+(n2-1)*desv.tip.muestra2^2)/(n1+n2-2))
t0
```

```
## [1] -6.215931
```

operando obtenemos que $t0 = -6.215931$. y sabemos que sigue una distribución t-Student $t_{n_1+n_2-2} = t_{98}$. Para este hipótesis alternativa el p-valor es

$2 \cdot P(t_{98} > |-6.2159314|)$, lo calculamos con R

```
t0=(media.muestra1-media.muestra2)/sqrt((1/n1+1/n2)*
((n1-1) *desv.tip.muestra1^2+(n2-1)*desv.tip.muestra2^2)/(n1+n2-2))
t0
```

```
## [1] -6.215931
```

```
n1
```

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

```
n2
```

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

```
2*(1-pt(abs(t0),df=n1+n2-2)) # calculo la probabilidad del complementario
```

```
## [1] 0.00000001247958
```

```
2*pt(abs(t0),df=n1+n2-2,lower.tail = FALSE)# calcula el área la cola  superior
```

```
## [1] 0.00000001247958
```

Varianzas desconocidas pero distintas, n_1 y n_2 grande

Si suponemos que $\sigma_1 \neq \sigma_2$, el estadístico de contraste es $t0 = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\widetilde{S}_1^2}{n_1} + \frac{\widetilde{S}_2^2}{n_2}}} \sim t_f$, que, cuando $\mu_1 = \mu_2$, tiene distribución (aproximadamente, en caso de muestras grandes) t_f con

$$f = \frac{\left(\frac{\widetilde{S}_1^2}{n_1} + \frac{\widetilde{S}_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{\widetilde{S}_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{\widetilde{S}_2^2}{n_2}\right)^2}.$$

Calculamos el estadístico y los grados de libertad con R

```
t0=(media.muestra1-media.muestra2)/sqrt(desv.tip.muestra1^2/n1+desv.tip.muestra2^2/n2)
#calculo el valor dentro del floor que es el que utiliza la función t.test de R para este caso.
t0
```

```
## [1] -6.215931
```

```
f=(desv.tip.muestra1^2/n1+desv.tip.muestra2^2/n2)^2/(
  (1/(n1-1))*(desv.tip.muestra1^2/n1)^2+(1/(n2-1))*(desv.tip.muestra2^2/n2)^2)
f
```

```
## [1] 89.99613
```

El p valor es

```
# el p-valor de la función t.test de  R
2*(pt(abs(t0),f,lower.tail = FALSE))
```

```
## [1] 0.00000001562353
```

Apartado 3

Los intervalos de confianza al nivel del 95% los podemos obtener así

```
t.test(G1,G2,var.equal = TRUE,alternative = "two.sided",conf.level = 0.95)$conf.int
```

```
## [1] -2.251691 -1.161889
## attr(,"conf.level")
## [1] 0.95
```

```
t.test(G1,G2,var.equal = FALSE,alternative = "two.sided",conf.level = 0.95)$conf.int
```

```
## [1] -2.252298 -1.161282
## attr(,"conf.level")
## [1] 0.95
```

Son similares, podemos asegurar que la diferencia de medias se encuentra $-2.25 < \mu_1 - \mu_2 < -1.16$ al nivel del 95 el grupo 2 tiene una media entre 2.25 y 1.16 puntos mayor que el grupo 2 aproximadamente.

Apartado 4 El test que nos piden es el de igualdad de varianzas

$$\begin{cases} H_0: & \sigma_1^2 = \sigma_2^2 \\ H_1: & \sigma_1^2 \neq \sigma_2^2 \end{cases}.$$

El test de Fisher de igualdad de varianzas

```
var.test(G1,G2,alternative ="two.sided" )
```

```
##
##  F test to compare two variances
##
## data:  G1 and G2
## F = 0.54057, num df = 49, denom df = 49, p-value = 0.03354
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3067606 0.9525862
## sample estimates:
## ratio of variances
##            0.54057
```

Obtenemos un p-valor alto no podemos rechazar la igualdad de varianzas.

De forma manual el estadístico de este test sabemos que es

$$f_0 = \frac{\tilde{S_1}^2}{\tilde{S_2}^2} = \frac{1.3227818}{2.4470131} = 0.54057.$$

Que sigue una ley de distribución de Fisher y el p_valor es $\min\{2 \cdot P(F_{n_1-1,+n_2-1} \leq f_0), 2 \cdot P(F_{n_1-1,+n_2-1} \geq f_0)$.

que con R es

```
n1
```

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

```
n2
```

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

```
f0=desv.tip.muestra1^2/desv.tip.muestra2^2
f0
```

```
## [1] 0.54057
```

```
pvalor=min(2*pf(f0,n1-1,n2-2),2*pf(f0,n1-1,n2-2,lower.tail = FALSE))
pvalor
```

```
## [1] 0.03420609
```

Obtenemos los mismos resultados que con la función `var.test`.

El test de Levene con R tiene las mismas hipótesis que el anterior

```
library(car,quietly = TRUE)# pongo quietly para que quite avisos
```

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

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

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

```
notas=c(G1,G2)
grupo=as.factor(c(rep(1,length(G1)),rep(2,length(G1))))
leveneTest(notas~grupo)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  1.8029 0.1825
##       98
```

El p-valor obtenido es alto así que el test de Levene no aporta evidencias contra la igualdad de varianzas entre las notas de los dos grupos.

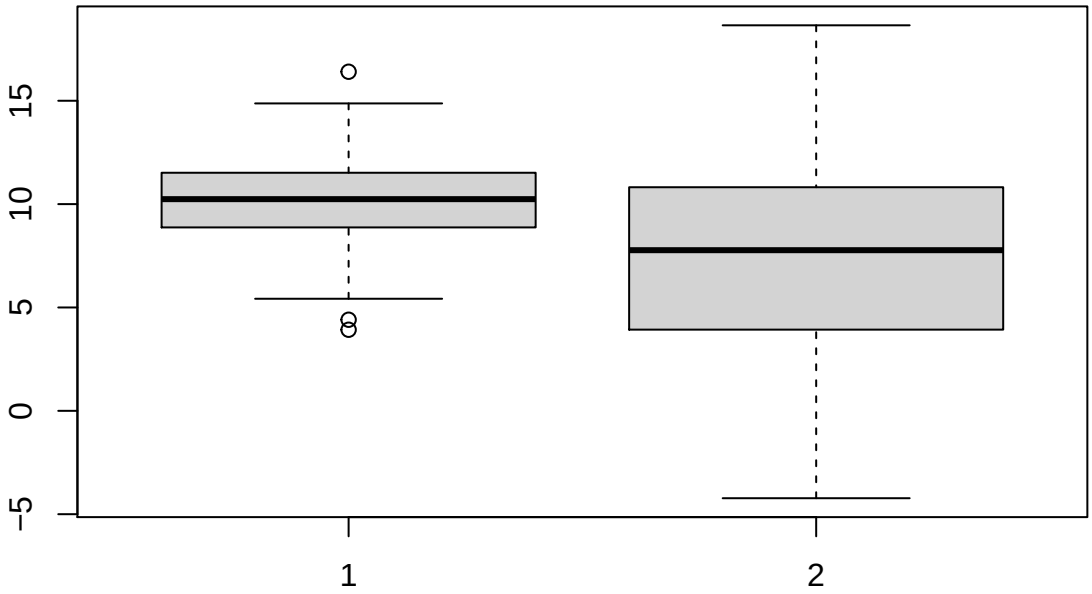
1.2 Problema 2 : Contraste dos muestras

Simulamos dos muestras con las funciones siguientes

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

Dibujamos estos gráficos

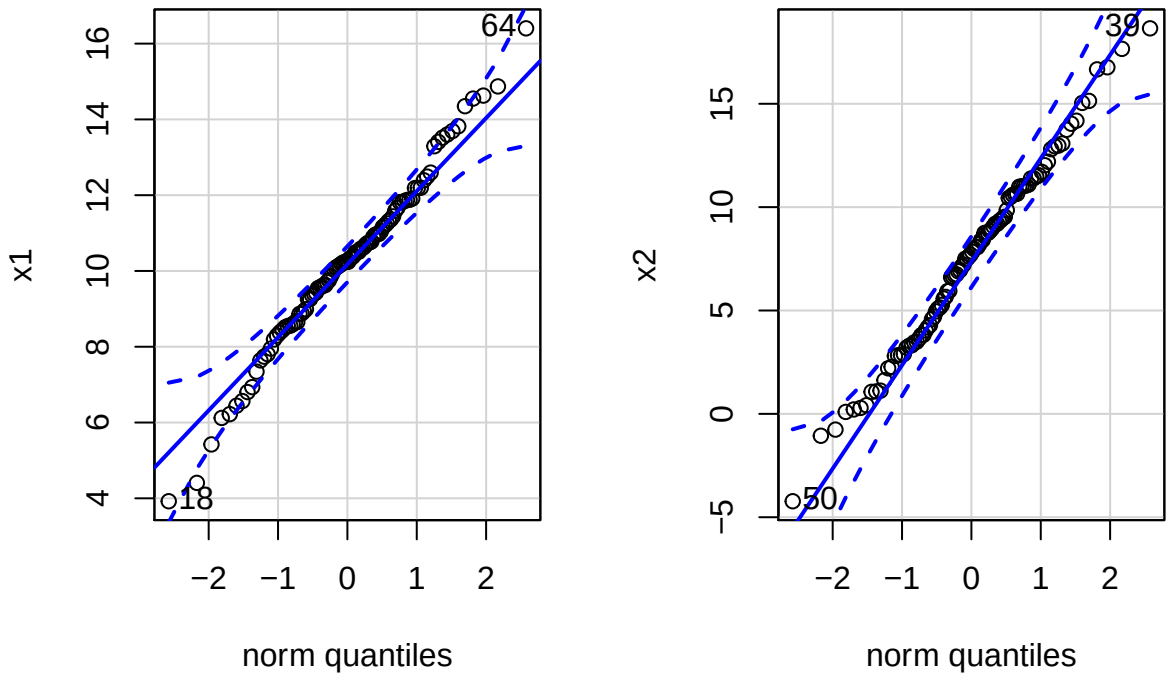
```r
boxplot(x1,x2)
```




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

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

```r
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.

1.2.1 Solución

Apartado 1

Se generan dos muestras de poblaciones normales de medias 10 y 8 y desviaciones típicas 2 y 4.

Apartado 2 El primer gráficos es un diagrama de caja (*boxplot*) que compara las distribuciones de los datos. Vemos que efectivamente la muestra 1 tiene una caja y unos bigotes más comprimidos que la muestra 2 así que la primera tiene menos varianza. Vemos que los valores medianos de la muestra 1 son más grandes que los de la muestra 2. Recordemos que la distribución normal es simétrica por lo que la media y la mediana coinciden. La muestra 1 tiene valores atípicos en la parte superior 1 y en la inferior parece que 2.

El segundo gráfico es un gráfico cuantil-cuantil o qqplot de normalidad. Compara los cuantiles muestrales con los teóricos de una normal y nos da un intervalo de confianza para esas observaciones.

Vemos que los cuantiles teóricos no difieren excesivamente de los muestrales en cada una de las muestras y que muy pocos valores se escapan de los intervalos de confianza esperados en el caso de normalidad. Así que no hay motivo para pensar que las distribuciones de ambas muestras proceden de poblaciones normales.

Apartado 3

El código es

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

El primer test contrasta para muestras independientes supuestas varianzas desconocidas pero iguales $H_0; \mu_1 = \mu_2$ contra $H_1 : \mu_1 > \mu_2$. **Así que este TEST NO ES**

El segundo test contrasta para muestras independientes supuestas varianzas desconocidas pero iguales $H_0; \mu_1 = \mu_2$ contra $H_1 : \mu_1 \neq \mu_2$. **Así que este TEST SÍ PUEDE SER** contrasta contra medias distintas para el caso de varianzas distintas.

El tercer test contrasta para muestras independientes supuestas varianzas desconocidas pero iguales $H_0; \mu_1 = \mu_2$ contra $H_1 : \mu_1 > \mu_2$ pues la opción por defecto de la función. **Así que este TEST SÍ PUEDE SER** contra medias distintas para el caso de varianzas iguales.

Apartado 4 El contrastes es

$$\begin{cases} H_0 : & \mu_1 = \mu_2 \\ H_1 : & \mu_1 \neq \mu_2 \end{cases}$$

En los dos últimos test los p-valores son muy muy pequeños así que hay evidencias en contra de la igualdad de medias entre las dos muestras. Además claramente los intervalos de confianza no contienen al cero.

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.*

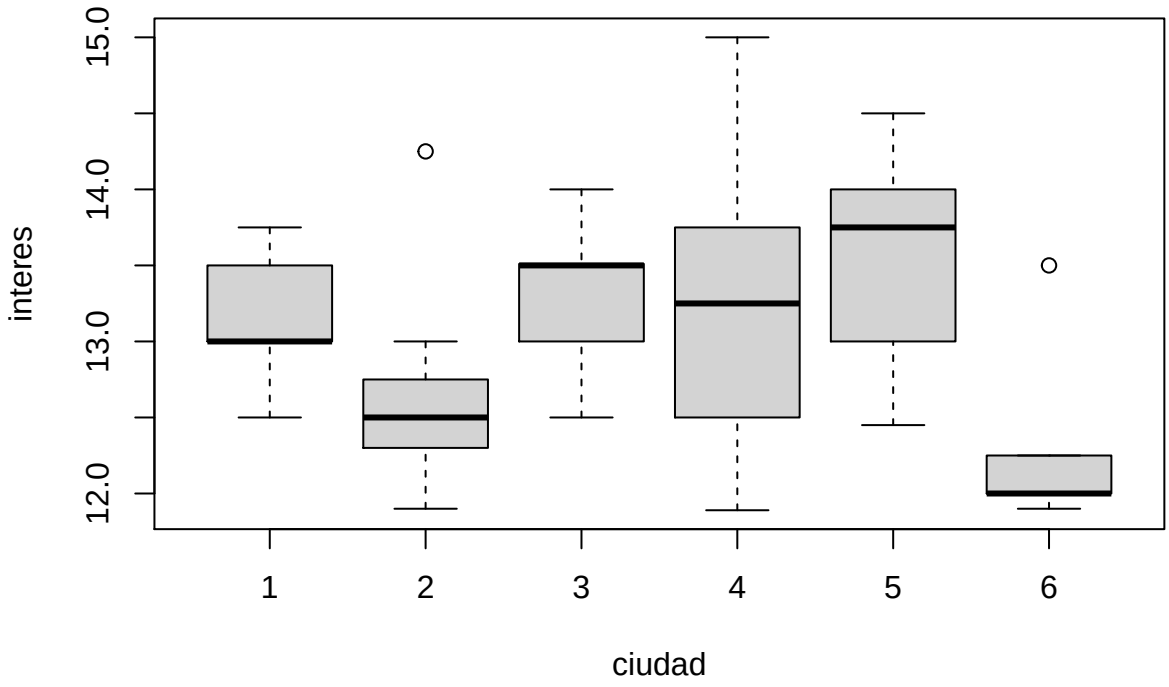
Este data set contiene dos columnas:

- 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)
# 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 ...
```

```
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.

1.3.1 Solución

Apartado 1

El código del enunciado nos carga los datos de una web, tenemos que pasar el parámetro skip=25 para que se salte las 25 primeras lineas del fichero de texto que son un preámbulo que explica los datos.

En el diagrama de caja vemos que las medias las distribuciones de la `Rate` por ciudad son muy distintas, no parecen tener ni medias ni varianzas iguales

Apartado 2

Las condiciones para realizar un ANOVA son:

- Muestras independientes y aleatorias
- Distribución normal de la Rate $N(\mu_i, \sigma_i)$ para las seis ciudades $i = 1, 2, 3, 4, 5, 6$.
- homocedasticidad; igualdad de varianzas $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = \sigma_5 = \sigma_6$.

El ANOVA que se pide es

$$\begin{cases} H_0: & \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 \\ H_1: & \text{no todas las medias son iguales.} \end{cases}$$

Apartado 3

El siguiente código realiza un test KS con corrección de Lillie para la normalidad de la variable Rate en cada una de la seis ciudades

```
library(nortest)
lillie.test(datos_interes$interes[datos_interes$ciudad==1])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos_interes$interes[datos_interes$ciudad == 1]
## D = 0.22384, p-value = 0.2163
```

```
lillie.test(datos_interes$interes[datos_interes$ciudad==2])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos_interes$interes[datos_interes$ciudad == 2]
## D = 0.22884, p-value = 0.1903
```

```
lillie.test(datos_interes$interes[datos_interes$ciudad==3])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos_interes$interes[datos_interes$ciudad == 3]
## D = 0.19145, p-value = 0.4459
```

```
lillie.test(datos_interes$interes[datos_interes$ciudad==4])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos_interes$interes[datos_interes$ciudad == 4]
## D = 0.11264, p-value = 0.9852
```

```
lillie.test(datos_interes$interes[datos_interes$ciudad==5])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos_interes$interes[datos_interes$ciudad == 5]
## D = 0.20021, p-value = 0.3743
```

```
lillie.test(datos_interes$interes[datos_interes$ciudad==6])
```

```
##
```

```
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  datos_interes$interes[datos_interes$ciudad == 6]
## D = 0.3494, p-value = 0.002236
```

No podemos rechazar la normalidad con el lillie.test en las 5 primeras ciudades, pero parece que la última está lejos de ser normal.

Ahora comprobemos que

$$\begin{cases} H_0: & \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2 = \sigma_6^2 \\ H_1: & \text{no todas las varianzas son iguales.} \end{cases}$$

con el test de Levene (o el de Bartlett)

```
library(car)
print(leveneTest(datos_interes$interes~as.factor(datos_interes$ciudad)))
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  5  1.2797 0.2882
##       48
```

El test de levene nos da un p-valor superior a 0.28 aceptamos la igualdad de varianzas

Apartado 4

Resolvemos el ANOVA con el código siguiente

```
summary(aov(datos_interes$interes~as.factor(datos_interes$ciudad)))
```

```
##                               Df Sum Sq Mean Sq F value  Pr(>F)
## as.factor(datos_interes$ciudad)  5  10.95  2.1891   4.829 0.00117 **
## Residuals                       48  21.76  0.4533
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El p-valor es muy bajo 0.00117 rechazamos la igualdad de las seis medias, al menos hay dos distintas.

Comprobamos por gusto el p-valor a partir de los datos del summary

```
Fest=2.1891/0.4533
Fest
```

```
## [1] 4.829252
```

```
1-pf(Fest,5,48)
```

```
## [1] 0.001174782
```

```
pf(Fest,5,48,lower.tail = FALSE)
```

```
## [1] 0.001174782
```

Apartado 5

Comparemos las medias dos a dos son 15 comparaciones

```
pairwise.t.test(datos_interes$interes,as.factor(datos_interes$ciudad),p.adjust.method = "holm")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
```

```
## data:  datos_interes$interes and as.factor(datos_interes$ciudad)
##
##   1      2      3      4      5
## 2 0.5781 -      -      -      -
## 3 1.0000 0.3330 -      -      -
## 4 1.0000 0.4651 1.0000 -      -
## 5 1.0000 0.0926 1.0000 1.0000 -
## 6 0.0353 1.0000 0.0148 0.0244 0.0028
##
## P value adjustment method: holm
```

Nos piden que decidamos con $\alpha = 0.1$, así que rechazaremos la igualdad de medias de todas las comparaciones con p-valor inferior a 0.1.

Tenemos que rechazar la igualdad de medias entre la ciudad 2 con la 5 y la de la ciudad 6 con las ciudades 1, 3, 4 y 5.

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.

1.4.1 Solución

Cuestion 1: Rechazamos la hipótesis nula para todos los niveles de significación α menores que el p-valor.

Cuestion 2: No, no podemos afirmar eso. Lo que sabemos después de rechazar un ANOVA es que hay al menos dos medias distintas.

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.

1.5.1 Solución

```
algoritmo1=read.table("../../../Datos/algoritmo1.csv")
algoritmo2=read.table("../../../Datos/algoritmo1.csv")
```

Proporción aciertos de cada algoritmo

```
n1=dim(algoritmo1)[1]
n1
```

```
## [1] 500
```

```
n1=length(algoritmo1$V1)
n1
```

```
## [1] 500
```

```
n2=length(algoritmo2$V1)
n2
```

```
## [1] 500
```

```
aciertos_absolutos_algoritmo1=table(algoritmo1)["1"]
aciertos_absolutos_algoritmo1
```

```
##   1
## 396
```

```
p1=prop.table(table(algoritmo1))["1"]
p1
```

```
##     1
## 0.792
```

```
aciertos_absolutos_algoritmo2=table(algoritmo2)["1"]
aciertos_absolutos_algoritmo2
```

```
##   1
## 396
```

```
p2=prop.table(table(algoritmo2))["1"]
p2
```

```
##     1
## 0.792
```

Después de los cálculos preliminares si denotamos las proporciones poblacionales de aciertos de cada algoritmo por p_1 y p_2 respectivamente, el contraste que nos piden es

$$\begin{cases} H_0: & p_1 = p_2 \\ H_1: & p_1 > p_2 \end{cases}$$

estamos ante un diseño de comparación de proporciones con muestras independientes. Con R lo podemos resolver con el `fisher.test` o con el `prop.test`

```
x=matrix(c(aciertos_absolutos_algoritmo1,n1-aciertos_absolutos_algoritmo1,
           aciertos_absolutos_algoritmo2,n2-aciertos_absolutos_algoritmo2),
         ncol=2,byrow=FALSE)
x
```

```
##      [,1] [,2]
## [1,]  396  396
## [2,]  104  104
```

```
fisher.test(x,alternative="greater",conf.level=0.95)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  x
## p-value = 0.531
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  0.7647666       Inf
## sample estimates:
## odds ratio
##          1
```

```
c(aciertos_absolutos_algoritmo1,aciertos_absolutos_algoritmo2)
```

```
##   1   1
## 396 396
```

```
c(n1,n2)
```

```
## [1] 500 500
```

```
prop.test(c(aciertos_absolutos_algoritmo1,aciertos_absolutos_algoritmo2), c(n1,n2),alternative="greater
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(aciertos_absolutos_algoritmo1, aciertos_absolutos_algoritmo2) out of c(n1, n2)
## X-squared = 0.00000000000000000000000000038836, df = 1, p-value = 0.5
## alternative hypothesis: greater
## 95 percent confidence interval:
##  -0.04222324  1.00000000
## sample estimates:
## prop 1 prop 2
##  0.792  0.792
```

Con ambos test obtenemos p valores altos (el más pequeño es el de Fisher y es mayor que 0.4, así que no podemos rechazar que las proporciones de aciertos de los dos algoritmos sean iguales contra que la proporción de aciertos del algoritmo 1 es mejor que la del 2.

El intervalo de confianza asociado a este test es

```
prop.test(c(aciertos_absolutos_algoritmo1,aciertos_absolutos_algoritmo2), c(n1,n2),alternative="greater
```

```
## [1] -0.04222324  1.00000000
## attr(,"conf.level")
## [1] 0.95
```

luego con una probabilidad del 95% la $p_1 - p_2 > -1$ contiene el 0 y no podemos despreciar que sean iguales contra que $p_1 > p_2$.