Q2.1: Analysis of 3 normal populations with Anova

Yaëlle Pihan, Marco Carega

13/04/2022

Introduction

This analysis consists in identifying if the 3 populations of normal vectors/3 groups, are different, or not, considering their means.
In order to do this, we use Anova statistical test which test if the means of each groups are equal.

Table of contents

1. Generation of 3 populations following a normal distribution

1. Importing libraries

```
library(car)
library(carData)
library(sandwich)
library(lmtest)
library(RcmdrMisc)
library(agricolae)
library(dplyr)
```

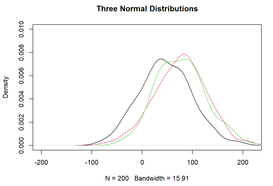
2. Generation of 3 vectors following a normal distribution and sharing the same variance (sd²) and storing them into a csv file

```
#v1=rnorm(200, mean=50, sd=50)
#v2=rnorm(200, mean=80, sd=50)
#v3=rnorm(200, mean=80, sd=50)

#df_create <- data.frame(v1,v2,v3)
#write.csv(df,"vectors.csv", row.names = FALSE)
```

3. Reading of csv file to get the vectors and creation of a dataframe with 2 columns x1 and x2:
   - x1 stores the values of each vectors
   - x2 stores the categorical value / factor that identifies each vector as a group: v1, v2 or v3

For this study we choose to generate 2 vectors with the same mean equal to 80, and an the third one with the mean equal to 50.

```
df_read <- read.csv("vectors.csv")

v1 <- pull(df_read,v1)
v2 <- pull(df_read,v2)
v3 <- pull(df_read,v3)

createData=function(x,y,z)
{

  plot(density(v1),xlim=c(-200,220),ylim=c(0,0.01),main="Three Normal Distributions")
  lines(density(v2),col=2)
  lines(density(v3),col=3)

  v1n=data.frame(x1=x, x2="v1")
  v2n=data.frame(x1=y, x2="v2")
  v3n=data.frame(x1=z, x2="v3")

  data=mergeRows(v1n, v2n, common.only=FALSE)
  data=mergeRows(as.data.frame(data), v3n, common.only=FALSE)
  print("Data summary\n\n")
  print(summary(data))
  print("Head data")
  print(head(data))
  return(data)
}

data=createData(v1,v2,v3)
```




```
## [1] "Data summary\n\n"
##        x1              x2
##  Min.   :-83.46   Length:600
##  1st Qu.: 29.07   Class :character
##  Median : 67.18   Mode  :character
##  Mean   : 65.80
##  3rd Qu.:100.87
##  Max.   :221.18
## [1] "Head data"
##         x1 x2
## 1 80.262254 v1
## 2 89.990510 v1
## 3 13.865358 v1
## 4 28.674051 v1
## 5  2.265043 v1
## 6 27.097767 v1
```

2. Expectations

We expect the 3 assumption tests to be verified, allowing us to apply Anova test on the data.

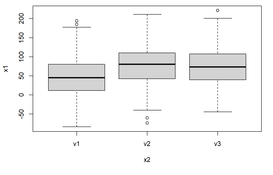
Indeed, the 3 groups are supposed to follow a normal distribution, to be independent from each other and to show homoscedasticity (homogeneity of variance).

Furthermore, as 2 means upon 3 are equal, we expect the Anova test to reject the equality of means and the Posthoc Tukey test to identify from which group the difference comes from.

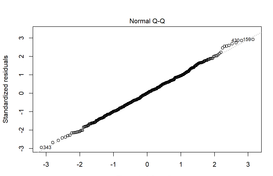
3. Tests

```
tests=function(x1,x2,data)
{
  print("----------------------Anova test-----------------------")
  formula=x1~x2
  model=aov(formula,data=data) # anova
  print(summary(model))
  Boxplot(x1~x2,data=data,id=FALSE)
  print("----------------------Assumption tests-----------------")
  print(shapiro.test(residuals(model))) #normality
  plot(model,2) #QQ plot to test normality
  print(dwtest(model,alternative = "two.sided")) #independence
  print(bptest(model)) #homoscedasticity
  plot(model,1) #Residuals plot to test homoscedasticity
  print("----------------------Posthoc test-----------------------")
  TukeyHSD(model)
}

tests(data$x1,data$x2,data)
```

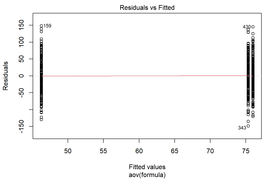
```
## [1] "----------------------Anova test-----------------------"
##              Df  Sum Sq Mean Sq F value   Pr(>F)
## x2            2  115349   57674  22.33 4.46e-10 ***
## Residuals   597 1542154    2583
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




```
## [1] "----------------------Assumption tests-----------------"
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model)
## W = 0.99845, p-value = 0.8808
```




```
##
##  Durbin-Watson test
##
## data:  model
## DW = 1.952, p-value = 0.5026
## alternative hypothesis: true autocorrelation is not 0
##
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 0.55505, df = 2, p-value = 0.7577
```




```
## [1] "----------------------Posthoc test-----------------------"

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = formula, data = data)
##
## $x2
##             diff       lwr       upr     p adj
## v2-v1 29.0753426  17.13367 41.017312 0.0000001
## v3-v1 29.7351213  17.79745 41.673093 0.0000000
## v3-v2  0.6637787 -11.27790 12.605545 0.9906408
```

4. Analysis

1. Assumption analysis

- Normality
  - Density plot shows 3 lines looking normal
  - Shapiro Wilk test gives a p_value = 0.8808 > 0.05
  - QQ plot shows that the residuals follow a linear regression
    Normality is verified.
- Independence
  - Durbin-Watson test test has a p_value = 0.5026 > 0.05 and 1.5<DW<2.5
    The residuals are not linearly auto-correlated, independence is verified.
- Homoscedasticity
  - studentized Breusch-Pagan test has a p_value = 0.7577 > 0.05
  - Thanks to Residuals vs Fitted plot we can see that there is no evident relationship between residuals and fitted values(means of each groups)
    Homoscedasticity is verified.

2. Anova analysis
   Anova test gives a p_value < 0.05 and indicates through 3 stars (***) that the difference of means is significant. However we cannot identify from which group the difference of mean comes from without Tukey test analysis.

3. Posthoc analysis
   Tukey test shows that the only p_value > 0.05 results from the comparison of v3 and v2. On the contrary, the comparison of v2 with v1 and v3 with v1 returns p_values = 0 < 0.05.

Therefore we can do the hypothesis that the difference of means comes form the group/population v1. This difference is also visible on the boxplot.

Conclusion

To conclude, we find the results expected: anova assumptions are verified and the mean difference comes logically from v1: the normal vector with a mean different from the others (mean = 50 instead of 80).