

Instituto Superior de Agronomia
Modelos Matemáticos e Aplicações– 2023/2024
Module I – Exercises

1. The vector `precip` in `R` contains the average amount of rainfall (in inches) for each of 70 cities in the United States and Puerto Rico.

- a) Have a look at the data. See its structure. What is the variable under study. What type is it?
- b) See the following commands and explain what each one performs

```
precip
hist(precip)
length(precip)
histograma<-hist(precip,plot=FALSE)
str(histograma)
histograma$breaks
histograma$counts
n<-length(precip)
ni<-histograma$counts
nclass<-length(ni)
fi<-ni/n;fi
Ni<-cumsum(ni);Ni
Fi<-Ni/sum(ni);Fi
ci<- paste ("",histograma$breaks[1:nclass],",",
           histograma$breaks[2:(nclass+1)] , """,sep="")


#for writting a lighter table let us round fi and Fi to 3 decimal places
fi<-round(fi,3);fi
Fi<-round(Fi,3);Fi

tabela<-data.frame(ci,ni,fi,Fi,Ni)
write.table(tabela,"tabela.csv",sep=";",row.names=FALSE)

#Se o separador do excel é ; esta tabela abre com o excel
# se não for deverão escrever sep=","
```

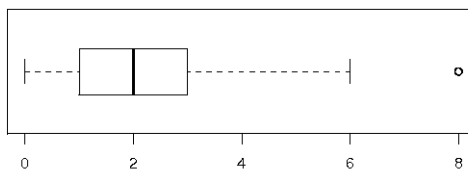
See that a table of absolute and relative frequencies has been constructed and written in a file, that can be opened by excel. Read that table.


- c) Draw two histograms considering different classes of intervals. Comment the shape and try to locate the cities that seem to show different values relatively to the others.
 - d) Draw a boxplot, comment briefly the symmetry of the observed values and write some comments you consider relevant.
2. Consider again the same dataset.
- a) Do a first exploratory analysis calculating all the indicators you know.
 - b) Do an exploratory analysis of the normality of the data.
 - c) Now choose an adequate test for testing the normality.
3. Suppose, a group of 25 people are surveyed as to their beer-drinking preference. The categories were (1) Domestic can, (2) Domestic bottle, (3) Microbrew and (4) import. The raw data is 3 4 1 1 3 4 3 3 1 3 2 1 2 1 2 3 2 3 1 1 1 1 4 3 1

- a) Write the commands that are necessary to draw a plot for both frequencies and proportions.
 - b) Perform a small exploratory study.
4. Consider the number of wild plants of a given species observed in 200 walks. Using the  package a descriptive study was done. Some results are shown below:

```
> plant<-c( ) # here we have the number of plants counted
> ni<-table(plant); ni
plant
 0  1  2  3  4  6  8
15 44 60 50 18 12  1
> Ni<-cumsum(ni); fi<-ni/sum(ni); Fi<-round(Ni/sum(ni),3)
> # here we have a table of frequencies
xi  ni   Ni    fi     Fi
0   15   15   0.075  0.075
1   44   A    0.220  0.295
2   60  119   0.300  0.595
3   50  169   0.250    C
4   18  187   0.090  0.935
6   12  199     B  0.995
8    1  200   0.005  1.000
>mean(plant)
[1] D
> quantile(plant,type=2)
 0%  25%  50%  75% 100%
 0   1   E   3   8
> Fi[3]
  F
```

- a) Fill in the table by calculating the values of A, B and C.
- b) Calculate the other values not given in the *output*, i.e., D, E e F.
- c) Do you think that the *boxplot* shown below can correspond to the observed values? Please justify.

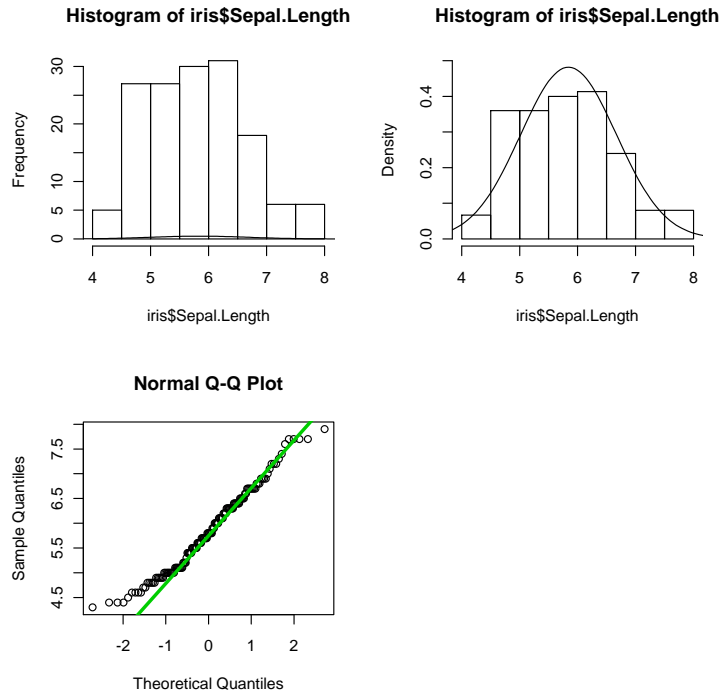


5. The following questions refer to the variables contained in the dataset `iris` available in . Use the **output** for choosing the adequate answers to each one of the questions listed below:
- a) Please indicate values for the main descriptive indicators.
 - b) When the command denoted by I is executed, what do you expect as the result?
 - c) Some graphical displays can be used as a first analysis of the normality of the characteristic under study. Which of them? Interpret them.
 - d) Write the necessary commands to obtain the boxplots for `Sepal.Length` associated to each one of the species?

```

>iris
> dim(iris)
[1] 150 5
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1          3.5          1.4          0.2  setosa
2          4.9          3.0          1.4          0.2  setosa
3          4.7          3.2          1.3          0.2  setosa
4          4.6          3.1          1.5          0.2  setosa
5          5.0          3.6          1.4          0.2  setosa
6          5.4          3.9          1.7          0.4  setosa
> iris[,1]
 [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9 5.4 4.8 4.8 4.3
[15] 5.8 5.7 5.4 5.1 5.7 5.1 5.4 5.1 4.6 5.1 4.8 5.0 5.0 5.2
[29] 5.2 4.7 4.8 5.4 5.2 5.5 4.9 5.0 5.5 4.9 4.4 5.1 5.0 4.5
[43] 4.4 5.0 5.1 4.8 5.1 4.6 5.3 5.0 7.0 6.4 6.9 5.5 6.5 5.7
[57] 6.3 4.9 6.6 5.2 5.0 5.9 6.0 6.1 5.6 6.7 5.6 5.8 6.2 5.6
[71] 5.9 6.1 6.3 6.1 6.4 6.6 6.8 6.7 6.0 5.7 5.5 5.5 5.8 6.0
[85] 5.4 6.0 6.7 6.3 5.6 5.5 5.5 6.1 5.8 5.0 5.6 5.7 5.7 6.2
[99] 5.1 5.7 6.3 5.8 7.1 6.3 6.5 7.6 4.9 7.3 6.7 7.2 6.5 6.4
[113] 6.8 5.7 5.8 6.4 6.5 7.7 7.7 6.0 6.9 5.6 7.7 6.3 6.7 7.2
[127] 6.2 6.1 6.4 7.2 7.4 7.9 6.4 6.3 6.1 7.7 6.3 6.4 6.0 6.9
[141] 6.7 6.9 5.8 6.8 6.7 6.7 6.3 6.5 6.2 5.9
> mean(iris[,1])
[1] 5.843333
> summary(iris$Sepal.Length)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 4.300  5.100  5.800  5.843  6.400  7.900
>
> by(iris$Sepal.Length,iris$Species,summary)
iris$Species: setosa
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 4.300  4.800  5.000  5.006  5.200  5.800
-----
iris$Species: versicolor
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 4.900  5.600  5.900  5.936  6.300  7.000
-----
iris$Species: virginica
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 4.900  6.225  6.500  6.588  6.900  7.900
>
> by(iris$Sepal.Length,iris$Species,var) # I
> hist(iris$Sepal.Length)
> val<-seq(2,10,by=0.05)
> lines(val,dnorm(val,mean(iris$Sepal.Length),sd(iris$Sepal.Length)))
> hist(iris$Sepal.Length,freq=F,ylim=c(0,0.5))
> val<-seq(2,10,by=0.05)
> lines(val,dnorm(val,mean(iris$Sepal.Length),sd(iris$Sepal.Length)))
> qqnorm(iris$Sepal.Length)
> qqline(iris$Sepal.Length,col=3,lwd=3)

```



6. Consider a continuous random variable X , which law is assumed as being known, but depending on an unknown parameter, $\theta > 0$, i.e.

$$f(x|\theta) = (\theta + 1)x^\theta, \quad \text{if } 0 < x < 1, \quad \text{and } f(x) = 0 \text{ for other values of } x$$

Let (X_1, X_2, \dots, X_n) be a random sample of size n , associated to that variable.

- a) Please calculate the Moments Estimator of θ .


Hint: Remember that $E[X] = \int_0^1 x(\theta + 1)x^\theta dx = \frac{\theta + 1}{\theta + 2}$.

- b) Please obtain the Maximum Likelihood Estimator of θ .

- c) Suppose you have observed the following sample, with size $n = 15$, from X ,

0.34 0.52 0.99 0.90 0.89 0.75 0.76 0.66 0.20 0.24 0.78 0.93 0.89 0.88 0.87

Write the necessary commands for obtaining the estimates for θ , as values of the estimators derived above.

7. Please consider commands and figures displayed below, related with the  dataset `faithful`:

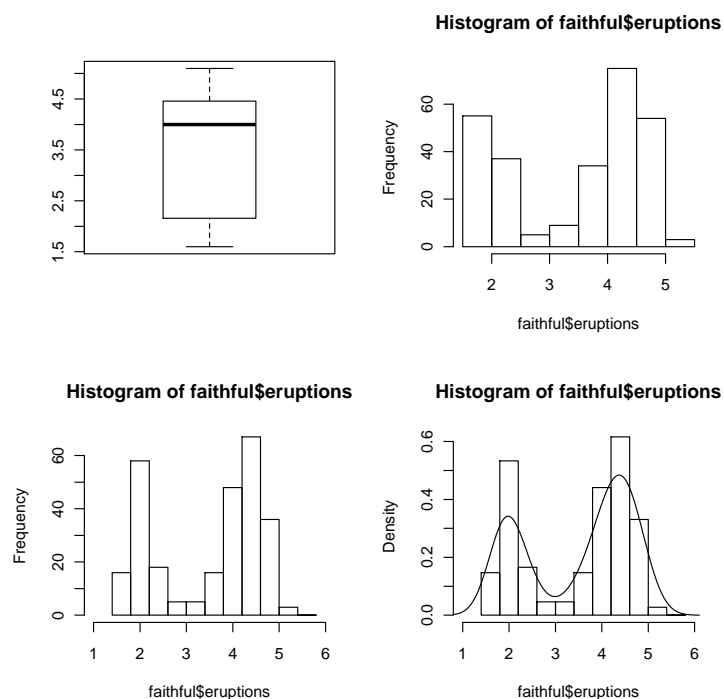
```
data(faithful)
faithful
head(faithful)
help(faithful)
summary(faithful$eruptions)
```

```

par(mfrow=c(2,2))
boxplot(faithful$eruptions)
hist(faithful$eruptions)
hist(faithful$eruptions,breaks=seq(1.4,6,0.4),xlim=c(1,6))
hist(faithful$eruptions,breaks=seq(1.4,6,0.4),xlim=c(1,6),freq=F)
lines(density(faithful$eruptions))

```

- a) Interpret each one of the commands, verify and explain the differences among the graphics displayed.



- b) What do you think that the following commands perform? Try to execute them.

```

cor(faithful$eruptions,faithful$waiting)
lm(faithful$eruptions~faithful$waiting)
plot(faithful$waiting,faithful$eruptions)

```

8. The normal plot is a fancy way of checking if the distribution looks normal. As we referred to in our classes a more primitive one is to check the rule of thumb that 68% of the data is 1 standard deviation from the mean, 95% within 2 standard deviations and 99.8% within 3 standard deviations. Create 100 random numbers from a standard normal with mean 0 and standard deviation 1. What percent are within 1 standard deviation of the the mean? Two standard deviations, 3 standard deviations? Is your data consistent with the normal?

Hint: The data is supposed to have mean 0 and variance 1. To check for 1 standard deviation we can do

```

> x<-c() #vector with the simulated values
> x<-rnorm(100);x<-1;sigma<-1
> n <-length(x)
> x<-rnorm(100);x
> int2<-sum( -k*sigma <x & x< k*sigma)/n # here we are assuming mu=0; sigma=1

```

what happen if we estimate?

```

> mu<-mean(x);sigma <-sd(x);mu;sigma
> int2<-sum( mean(x)-k*sigma <x & x< mean(x)+k*sigma)/n
> int2
> hist(x)

```

9. In a study of the effectiveness of certain exercises in weight reduction, a group of 16 persons did these exercises for one month and showed the following results:

Weight before	Weight after	Weight before	Weight after
211	198	172	166
180	173	155	154
171	172	185	181
214	209	167	164
182	179	203	201
194	192	181	175
160	161	245	233
182	182	146	142

Do we have reasons to think that, actually, the exercises are effective in weight reduction? Give a complete answer.