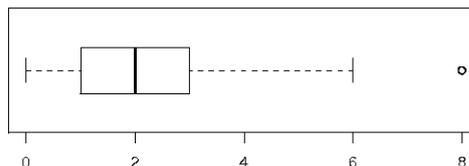


Instituto Superior de Agronomia
Modelos Matemáticos e Aplicações – 2018/2019
Módulo I – Exercícios

1. Registou-se o número de cogumelos observados em 200 passeios. Fez-se um estudo descritivo dessa amostra utilizando o *software* . Os resultados encontram-se abaixo.

```
> cogumelo<-c( ) # neste vector foram colocados os valores observados
> ni<-table(cogumelo); ni
cogumelo
 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)
> # apresenta-se agora a tabela de frequências
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(cogumelo)
[1] D
> quantile(cogumelo,type=2)
 0%  25%  50%  75% 100%
 0   1   E   3   8
> Fi[3]
  F
```

- Complete a tabela de frequências, obtendo os valores de A, B e C.
- Complete os restantes valores em falta no *output*, i.e., D, E e F.
- Poderá o *boxplot* apresentado abaixo corresponder aos valores observados? Justifique.



2. O vector *precip* in  contém a precipitação média (em polegadas) para cada uma das 70 cidades nos Estados Unidos e em Porto Rico
- Dê uma olhadela nos dados. Veja sua estrutura. Qual é a variável em estudo? De que tipo é?
 - Veja os seguintes comandos e explique o que cada um executa

```

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="")

#para escrever uma tabela mais leve, arredondemos fi e Fi para 3 casas decimais
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=","

```

Veja que uma tabela de frequências absolutas e relativas foi construída e escrita num ficheiro, que pode ser aberta pelo excel. Leia essa tabela.

- c) Desenhe dois histogramas considerando diferentes classes de intervalos. Comente a forma e tente localizar as cidades que parecem mostrar valores diferentes relativamente aos outros.
 - d) Desenhe o *boxplot*, comente brevemente a simetria dos valores observados e escreva alguns comentários que considere relevantes.
3. Considere novamente o mesmo conjunto de dados.
- a) Faça uma primeira análise exploratória, calculando todos os indicadores que conhece.
 - b) Faça uma análise exploratória da normalidade dos dados.
 - c) Agora escolha um teste adequado para testar a normalidade.
4. As alíneas que se seguem referem-se todas às variáveis contidas no conjunto de dados *iris* disponível no R. Consulte o **output** e escolha as respostas adequadas a cada uma das perguntas seguintes:
- a) Indique valores dos principais indicadores descritivos.
 - b) O que resulta da aplicação do comando identificado por I??
 - c) Alguns gráficos apresentados podem ser usados para fazer uma primeira análise da normalidade da variável estudo. Quais? Qual a interpretação que faz?
 - d) Apresente os comandos necessários para obter as caixas bigodes de *Sepal.Length* para cada uma das espécies?

```

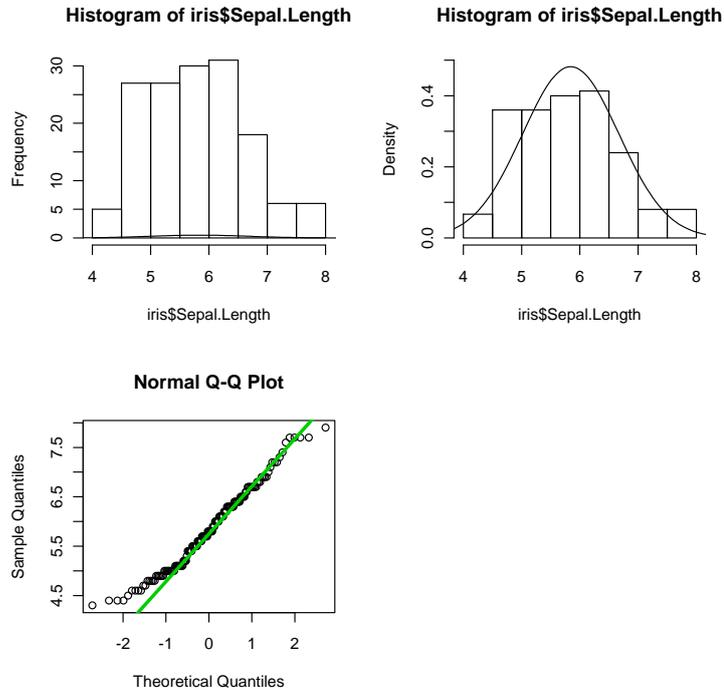
>iris
> dim(iris)
[1] 150  5
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1          3.5          1.4          0.2  setosa
2          4.9          3.0          1.4          0.2  setosa

```

```

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)

```



5. Considere uma variável aleatória contínua X , da qual se conhece a lei, mas que depende de um parâmetro desconhecido θ , i.e.

$$f(x|\theta) = (\theta + 1)x^\theta, \quad \text{se } 0 < x < 1, \quad \text{e nula nos restantes valores de } x$$

Considere que se tem uma amostra aleatória de dimensão n associada a esta variável, i.e, temos (X_1, X_2, \dots, X_n)

- Calcule o estimador de θ pelo método dos momentos.
Nota: Recorde que $E[X] = \int_0^1 x(\theta + 1)x^\theta dx = \frac{\theta + 1}{\theta + 2}$.
- Calcule o estimador de máxima verosimilhança para θ .
- Suponha que observou a seguinte amostra de 15 valores da variável 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

Escreva os comandos de R necessários à obtenção das estimativas de θ usando os dois estimadores construídos acima. Corra o R e indique os resultados que obteve.

6. Considere os seguintes comandos e os gráficos apresentados abaixo, relativos ao conjunto de dados `faithful` contidos no R:

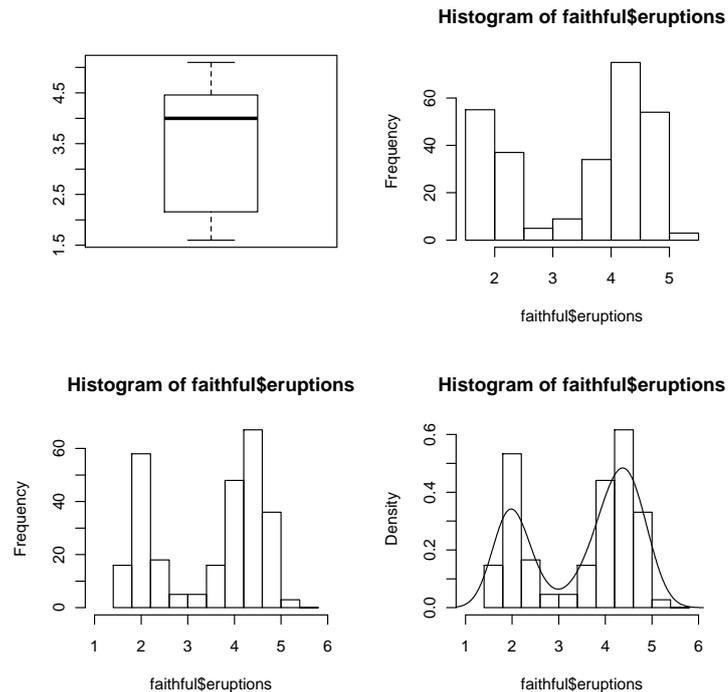
```
data(faithful)
faithful
head(faithful)
```

```

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) Interprete cada um dos comandos e verifique e explique as diferenças de cada um dos gráficos.



b) Consulte os slides 64-70 e interprete agora cada um dos comandos. Tente executá-los no R e interprete os resultados obtidos.

```

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

```

7. Suponha que tem um terreno onde estão plantadas árvores em 10 filas. Porém as filas têm números diferentes de árvores, i.e. têm 5, 12, 3, 6, 8, 10, 7, 12, 6 e 8 árvores respectivamente.

Elabore um procedimento em R para obter uma amostra simples sem reposição de 20 árvores.

8. Introduza no R e interprete todos os exercicios apresentados nos slides da aula4.
9. Uma empresa agrícola pretende adquirir um gerador. Face a dois tipos de geradores (que designamos por ger1 e ger2), de preços semelhantes, quer decidir comparando a produção de energia eléctrica, em kWh, de cada um dos geradores. Recolheu uma amostra de 25 observações da quantidade de energia eléctrica produzida por cada gerador, em diferentes situações. Os dados foram introduzidos no *software* R. Considere os seguintes resultados referentes ao gerador tipo 1:

```
>var(ger1); mean(ger1)           >shapiro.test(ger1)
[1] 18.90528                     Shapiro-Wilk normality test
[1] 9.6056                       data: ger1
                                W = 0.9514, p-value = 0.2698
```

- a) Indique uma estimativa da variância da produção de energia eléctrica do gerador 1. Explícite o estimador que lhe está associado.
- b) Para obter um intervalo de confiança a 95% para a diferença entre as produções médias de energia eléctrica dos geradores 1 e 2, usou-se o comando do R que se apresenta de seguida e respectivo resultado (no qual três valores foram substituídos por ????):

```
>t.test(ger1,ger2,paired=????,var.equal=T)
```

Two Sample t-test

```
data: ger1 and ger2
t = -0.6632, df = ????, p-value = 0.5104
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.009280  ????
sample estimates:
mean of x mean of y
 9.6056   10.3520
```

- i) Qual a diferença entre as médias observadas para a produção de energia eléctrica dos dois geradores? Qual o erro máximo que se cometeria, a 95% de confiança, ao identificar esta diferença com a verdadeira diferença entre as produções médias de energia eléctrica dos geradores?
- ii) Complete o *output* indicando a informação em falta nos três locais assinalados por ????.
- iii) Que pressupostos foi necessário verificar para construir o intervalo de confiança dado no *output*?
- iv) O que pode afirmar, a 95% de confiança, sobre a diferença entre as produções médias de energia eléctrica dos geradores?
- v) A conclusão da alínea anterior seria alterada se considerasse uma confiança de 99%? Justifique.
10. Seja X uma variável aleatória com valor esperado μ e variância σ^2 desconhecidos. Recolheram-se aleatoriamente 15 dados da população X , que se introduziram no *software* R no objecto x . Responda às seguintes questões utilizando, sempre que possível, o *output* apresentado abaixo.
- a) Com base na amostra recolhida indique estimativas do valor médio de X e da variância de \bar{X} .
- b) Recolheu-se uma amostra de dimensão 15 de outra população Y . Os dados foram introduzidos no objecto y do R.
- i) Indique um intervalo de confiança a 95% para a diferença das médias das populações X e Y . Explícite e valide os pressupostos necessários.
- ii) Os dados recolhidos evidenciam que as médias das duas populações são iguais? Justifique convenientemente.

```

> mean(x)           > mean(y)
[1] 5.686667        [1] 4.646667

> var(x)            > var(y)
[1] 1.084095        [1] 1.855524

> shapiro.test(x)           > shapiro.test(y)
      Shapiro-Wilk normality test           Shapiro-Wilk normality test
data:  x                                   data:  y
W = 0.9747, p-value = 0.6733             W = 0.9577, p-value = 0.2704

> shapiro.test(x-y)
      Shapiro-Wilk normality test
data:  x - y
W = 0.924, p-value = 0.2214

> var.test(x,y)

      F test to compare two variances

data:  x and y
F = 0.5843, num df = 14, denom df = 14, p-value = 0.3261
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.196151 1.740249
sample estimates:
ratio of variances
 0.5842529

> t.test(x,y,paired=T)
      Paired t-test
data:  x and y
t = 2.1864, df = 14, p-value = 0.04627
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.01977868 2.06022132
sample estimates:
mean of the differences
          1.04

> t.test(x,y,paired=F,var.equal=T)
      Two Sample t-test
data:  x and y
t = 2.3493, df = 28, p-value = 0.0261
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1331903 1.9468097
sample estimates:
mean of x mean of y
 5.686667  4.646667

```