Aula 7 Ficha de Trabalho 5

Tiago A. Marques

October 28, 2019

Introduction

This is the material for the Ficha de Trabalho 5 of Ecologia Numérica. Please resist the temptation to just copy paste code. You need to understand what the code is doing, and why you do what you do and reach the conclusions that you reach!

Note that the code assumes that the data files are in the same folder as the .Rmd. If they are not the path's might have to be edited.

Exercício 1

Efectue um teste t para avaliar se o comprimento médio dos gastrópodes é igual a 27 mm, partindo do princípio de que são satisfeitos os pressupostos exigidos para realizar esta análise (DataTP5gastropodes.csv).

```
gast <- read.csv("DataTP5gastropodes.csv", sep=";")
```

See if it all went fine

```
head(gast)
```

```
##   Ind Comprimento
## 1   1       26.59
## 2   2       26.19
## 3   3       25.58
## 4   4       24.67
## 5   5       25.50
## 6   6       23.08
```

```
summary(gast)
```

```
##       Ind          Comprimento
##  Min.   :  1.00   Min.   :20.66
##  1st Qu.: 25.75   1st Qu.:23.61
##  Median : 50.50   Median :24.70
##  Mean   : 50.50   Mean   :24.92
##  3rd Qu.: 75.25   3rd Qu.:26.20
##  Max.   :100.00   Max.   :31.02
```

We have 100 observations of 2 variables.

Now, if we change the data, the same line of code as above will return a different output: this is just illustrating the true nature of a dynamic report!

```
gast$v2=rnorm(100)
head(gast)
```

```
##   Ind Comprimento         v2
## 1   1       26.59 -1.7837178
## 2   2       26.19  1.2317063
## 3   3       25.58 -0.5726645
```

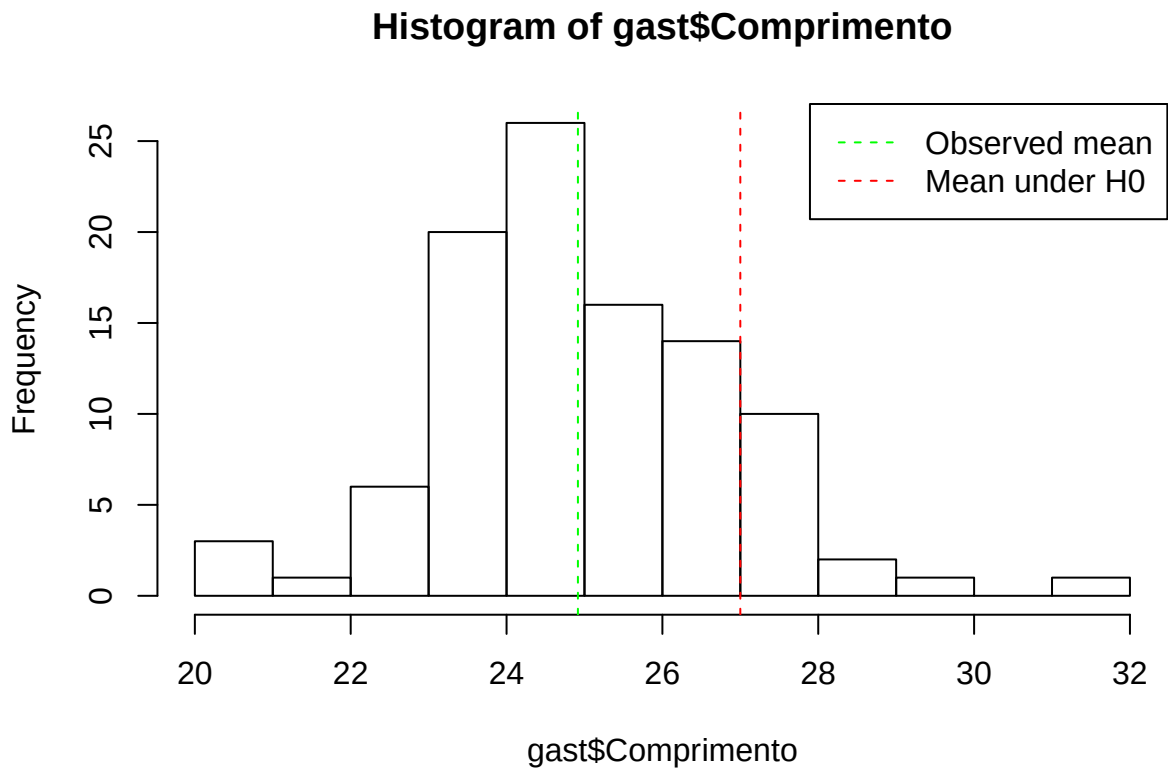
```
## 4   4       24.67  1.1555743
## 5   5       25.50 -0.2511230
## 6   6       23.08  1.0825463
```

We now have the same100 observations, but for 3 variables. Any way, lets clean up what we just did and actually get on with the exercise.

```
gast=gast[,1:2]
```

We start by looking at the data. There are never too many plots when doing statistical analysis.

```
hist(gast$Comprimento)
abline(v=27,col="red",lty=2)
abline(v=mean(gast$Comprimento),col="green",lty=2)
legend("topright",legend=c("Observed mean","Mean under H0"),col=c("green","red"),lty=2)
```



It does seem unlikely that the sample came from a pop with mean 27. We can test that formally with a `t.test`, which has as hypothesis:

- μ=27 mm
- $\mu \neq$ 27 mm

```
testgast=t.test(gast$Comprimento,mu=27)
testgast
```

```
##
##  One Sample t-test
##
## data:  gast$Comprimento
```

```
## t = -11.071, df = 99, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 27
## 95 percent confidence interval:
##  24.54226 25.28934
## sample estimates:
## mean of x
##   24.9158
```

The P-value is 0 therefore I reject H0: μ=27 mm for any of the usual significance levels.

We could have formally tested if the assumption behind the `t.test` holds, namely is the underlying population distributed as a Gaussian. For that we could use say a Kolmogorov-Smirnov test.

```
ks.test(gast$Comprimento,y="pnorm")
```

```
## Warning in ks.test(gast$Comprimento, y = "pnorm"): ties should not be
## present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  gast$Comprimento
## D = 1, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
shapiro.test(gast$Comprimento)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  gast$Comprimento
## W = 0.98671, p-value = 0.4187
```

Interestingly, there would have been reasons to suspect the underlying distribution was not Gaussian, at least given the Kolmogorov-Smirnov test. The Shapiro test is more conservative and does not reject the H0 that the data came from a Gaussian distribution. Nonetheless, the exercise was pretty clear, as it was stated we should use a `t.test`.

Exercicio 2

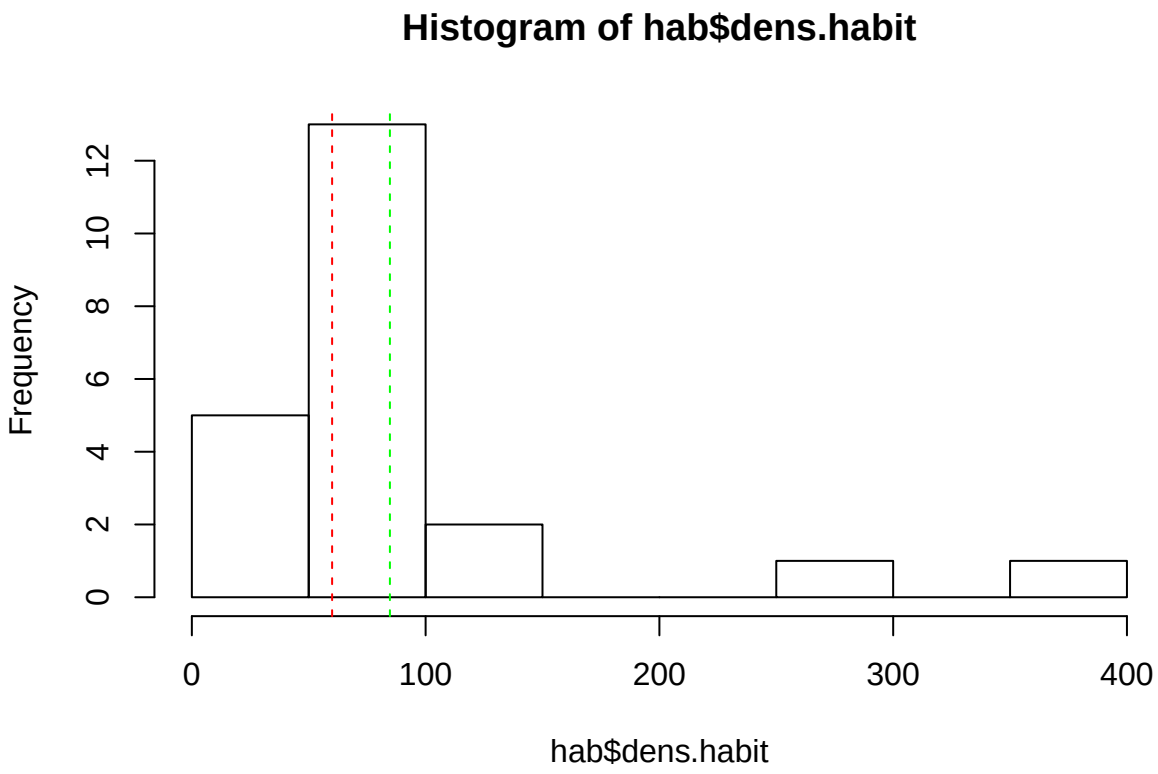
Efectue o teste que lhe pareça adequado para avaliar se a densidade populacional em Concelhos de determinado Distrito é igual a 60 habitantes por km2 (DataTP5habitantes.csv).

As usual, we begin by reading the data in

```
hab<-read.csv("DataTP5habitantes.csv", sep=";")
```

and, as before, look at the data

```
hist(hab$dens.habit)
abline(v=60,col="red",lty=2)
abline(v=mean(hab$dens.habit),col="green",lty=2)
```



We can check what the data looks like, and we see that the mean (84.73) and median (62) are quite different.

```
summary(hab$dens.habit)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   54.00   62.00   84.73   74.75  356.00
```

Given the observed distribution, it is very unlikely that the data is a sample from a Gaussian distribution. We can test it formally

```
ks.test(x=hab$dens.habit,y="pnorm")
```

```
## Warning in ks.test(x = hab$dens.habit, y = "pnorm"): ties should not be
## present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  hab$dens.habit
## D = 1, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
shapiro.test(hab$dens.habit)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  hab$dens.habit
## W = 0.6707, p-value = 8.484e-06
```

We clearly reject H0 for both tests for any of the usual significance levels. Therefore, we need a non-parametric test, which in the case of a test over the localization of a distribution, is the Wilcoxon test.

```
wilcox.test(hab$dens.habit,mu=60)
```

```
## Warning in wilcox.test.default(hab$dens.habit, mu = 60): cannot compute
## exact p-value with ties
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  hab$dens.habit
## V = 148.5, p-value = 0.4845
## alternative hypothesis: true location is not equal to 60
```

We do not reject the H0 that μ=60, for any of the usual significance values.

Note that it would be wrong to state that μ=60. The only thing we can say is that there is not enough information to suggest otherwise. But the best estimate of μ would not be 60! In fact, if we test the H0 that say $\mu = 62$, i.e.the observed median, we get an even larger P-value, hinting to the fact that this would be an even more likely data set under such a H0.

```
wilcox.test(hab$dens.habit,mu=62)
```

```
## Warning in wilcox.test.default(hab$dens.habit, mu = 62): cannot compute
## exact p-value with ties
```

```
## Warning in wilcox.test.default(hab$dens.habit, mu = 62): cannot compute
## exact p-value with zeroes
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  hab$dens.habit
## V = 95, p-value = 0.6951
## alternative hypothesis: true location is not equal to 62
```

Just out of curiosity, we can see what would have happened if we had ignored the failure to respect the assumptions of the t.test and had proceed with the parametric analysis.

```
t.test(hab$dens.habit,mu=60)
```

```
##
##  One Sample t-test
##
## data:  hab$dens.habit
## t = 1.4445, df = 21, p-value = 0.1634
## alternative hypothesis: true mean is not equal to 60
## 95 percent confidence interval:
##   49.1276 120.3269
## sample estimates:
## mean of x
##  84.72727
```

The conclusion would have been the same, but note that the P-value would now be considerably lower than that of the Wilcoxon test.Why? Because the t-test is more likely to find differences, compared to the Wilcoxon test (it is a more powerful test).

Exercício 3

Efectue um teste que lhe permita averiguar se as médias das duas populações a que se referem as amostras são iguais (DataTP5concelhos.csv).

```
conc<-read.csv("DataTP5concelhos.csv", sep=";")
```

We can see what is in the data

```
str(conc)
```

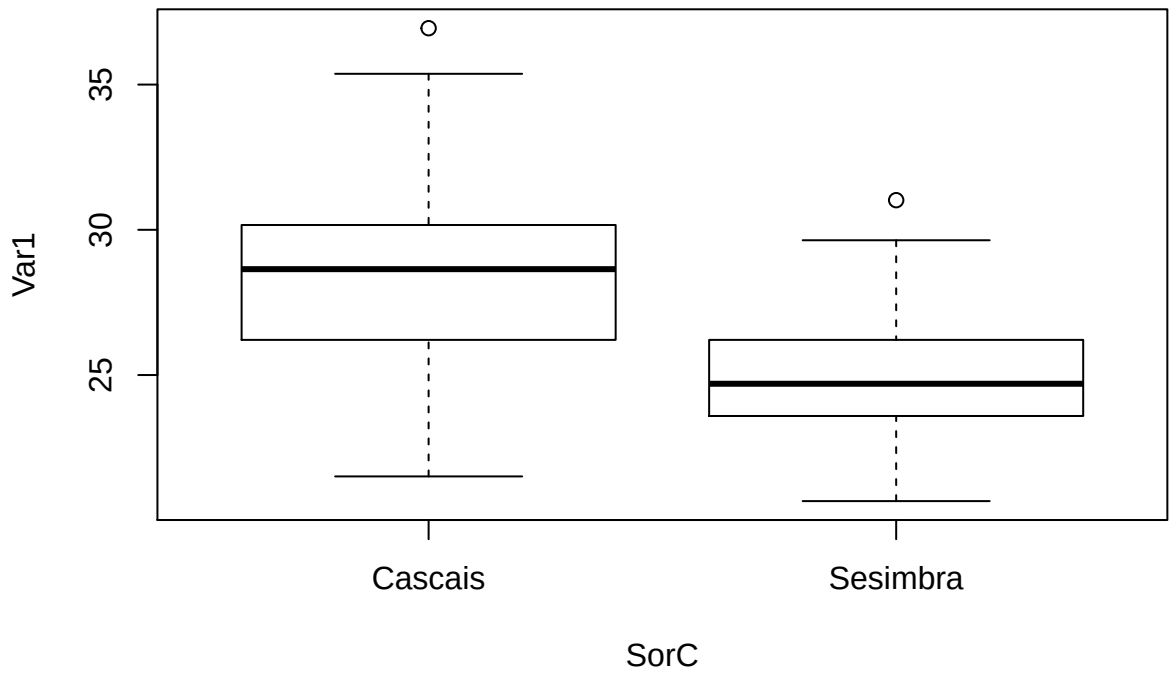
```
## 'data.frame':    100 obs. of  3 variables:
##  $ Concelho: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Sesimbra: num  26.6 26.2 25.6 24.7 25.5 ...
##  $ Cascais : num  31 24.9 27.7 31.2 23.9 ...
```

We have 100 records for each of 3 variables. In fact, this data set could be rearranged in another way, as in fact we simply have 1 variable, a variable for each of 100 locations in 2 "concelhos". Another way of representing this would be with a data.frame with just 2 columns

```
conc2<-data.frame(SorC=rep(c("Sesimbra","Cascais"),each=100),Var1=c(conc[,2],conc[,3]))
```

Now we can compare these one versus the other

```
with(conc2,boxplot(Var1~SorC))
```



These distributions seem symmetric enough to try a parametric test (but you should test it formally to be sure!)

```r
with(conc,t.test(Cascais,Sesimbra))
```

```
##
##  Welch Two Sample t-test
##
## data:  Cascais and Sesimbra
## t = 9.4221, df = 167.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.624583 4.016017
## sample estimates:
## mean of x mean of y
##   28.2361   24.9158
```

The P-value leaves no room for questions. For any of the usual significance levels we would reject the H0 that the means are the same.

OK, let's formally test if the assumptions of the t-test hold

```r
#test for equal variances
#head(conc)
bartlett.test(x =c(conc$Cascais,conc$Sesimbra),g = rep(c("Sesimbra","Cascais"),each=100))
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  c(conc$Cascais, conc$Sesimbra) and rep(c("Sesimbra", "Cascais"), each = 100)
## Bartlett's K-squared = 20.056, df = 1, p-value = 7.52e-06
```

```r
#what about log transform data
#head(conc)
bartlett.test(x =log(c(conc$Cascais,conc$Sesimbra)),g = rep(c("Sesimbra","Cascais"),each=100))
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  log(c(conc$Cascais, conc$Sesimbra)) and rep(c("Sesimbra", "Cascais"), each = 100)
## Bartlett's K-squared = 11.933, df = 1, p-value = 0.0005516
```

```r
#test for Gaussian residuals
```

After all, the t-tes is not adequate, bacuase the variances seem to diiffer. Then I must used a NP alternative. In this case, the WMW test.

Exercício 4

Efectue o teste de Mann-Whitney para avaliar se o comprimento total dos machos é igual ao das fêmeas (DataTP5comprimento.csv).

```r
comp<-read.csv("DataTP5comprimento.csv", sep=";")
```

as before, look at the data structure

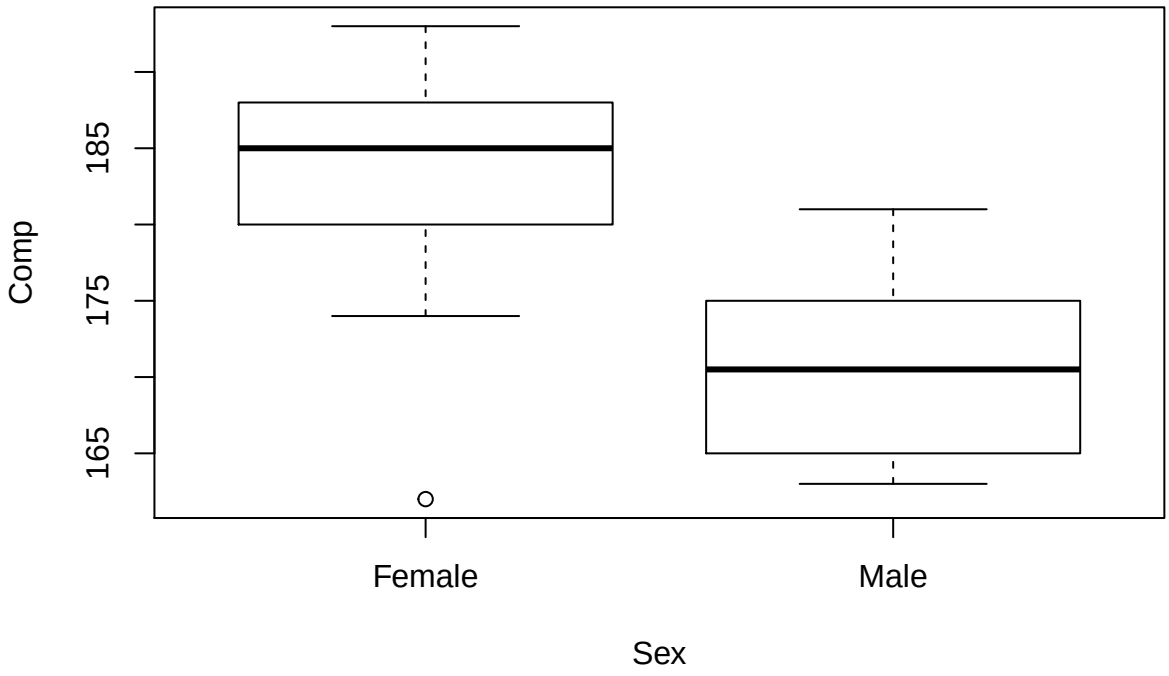
```r
str(comp)
```

```
## 'data.frame':    9 obs. of  3 variables:
##  $ Sample: int  1 2 3 4 5 6 7 8 9
##  $ Male  : int  175 173 168 165 163 181 NA NA NA
```

```
##  $ Female: int  193 188 185 183 180 162 174 192 186
```

we see that we have a different number of males and females. As before, we can look at a boxplot of the data, and to do so we re-code the data first

```
comp2=data.frame(Sex=c(rep("Male",sum(!is.na(comp$Male))),rep("Female",sum(!is.na(comp$Female)))),
Comp=c(comp$Male[1:sum(!is.na(comp$Male))],comp$Female[1:sum(!is.na(comp$Female))]))
with(comp2,boxplot(Comp~Sex))
```



even with a small sample size, females appear bigger than males. With such a small sample size, the precautionary approach should be to use a more conservative (less likely to commit type I errors test), like the Wilkoxon two-sample test.

```
wilcox.test(comp$Male,comp$Female)
```

```
##
##  Wilcoxon rank sum test
##
## data:  comp$Male and comp$Female
## W = 9, p-value = 0.03596
## alternative hypothesis: true location shift is not equal to 0
```

In this case, for the usual significance values of 10% and 5% we would reject H0, but for the significance level of 1% we would not. This would be one of those cases in which the best course of action would be to collect a second, larger sample, to see if the pattern was corroborated.

Note that if by any chance the researchers had a previous indication that females were larger, then a one sided test (instead of a bilateral, as was used in all the previous exercises) would be required. Interestingly, then, because the H0 is then that the mean of the females is equal or smaller than that of the males, i.e. $\mu_F \leq \mu_M$

(that is, the opposite of what we suspect and hence that we would like to prove), vs H1: $\mu_F > \mu_M$

```
wilcox.test(comp$Male,comp$Female,alternative="less")
```

```
##
##  Wilcoxon rank sum test
##
## data:  comp$Male and comp$Female
## W = 9, p-value = 0.01798
## alternative hypothesis: true location shift is less than 0
```

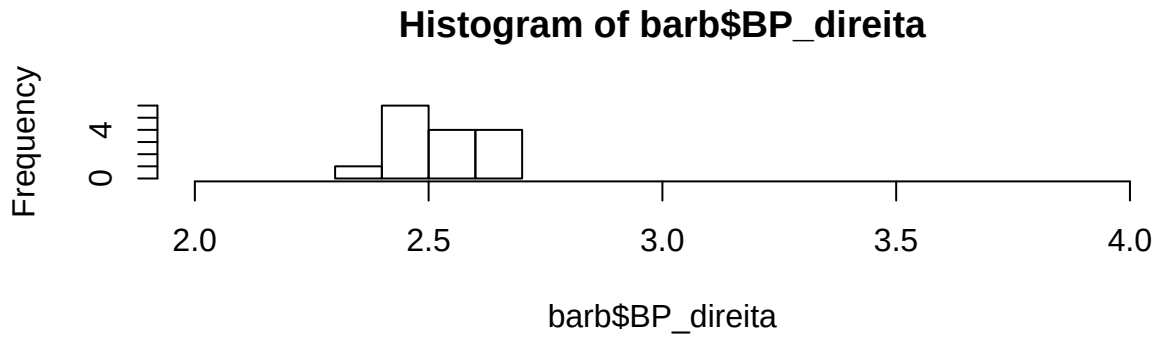
The conclusion is the same, but the test almost becomes significant even at the 1% level (note that the P-value is 50% smaller than before, 1 tail vs. 2 tails!). However, one should resist interpreting almost significant results. These are meaningless. If you decided the significance a priori, then a test is or is not significant given the P-value. There should be no maybe...

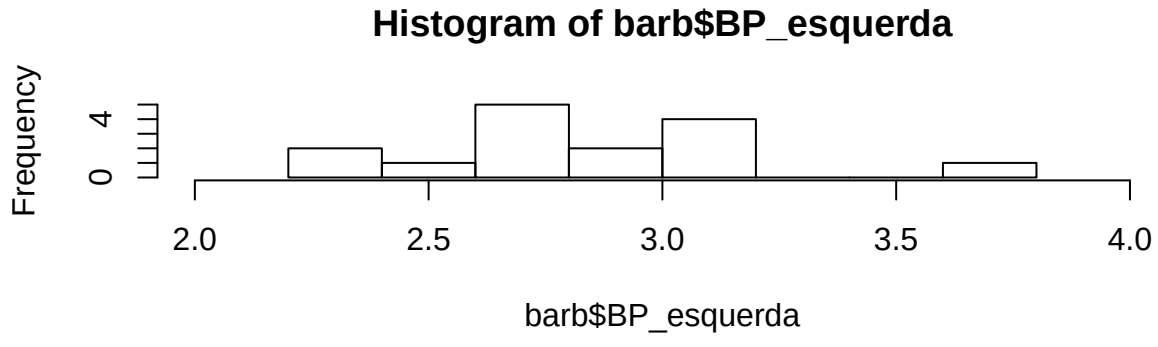
Exercicio 5

Efectue o teste t para amostras emparelhadas para avaliar se os comprimentos da barbatana peitoral direita são iguais ao da esquerda (DataTP5barbatanas.csv).

```
barb <- read.csv("DataTP5barbatanas.csv", sep=";")
```

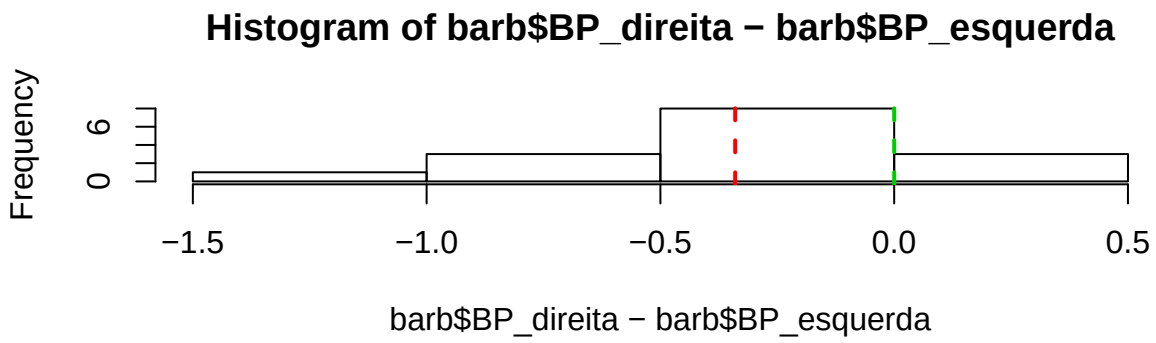
```
par(mfrow=c(2,1))
hist(barb$BP_direita,xlim=c(2,4))
hist(barb$BP_esquerda,xlim=c(2,4))
```





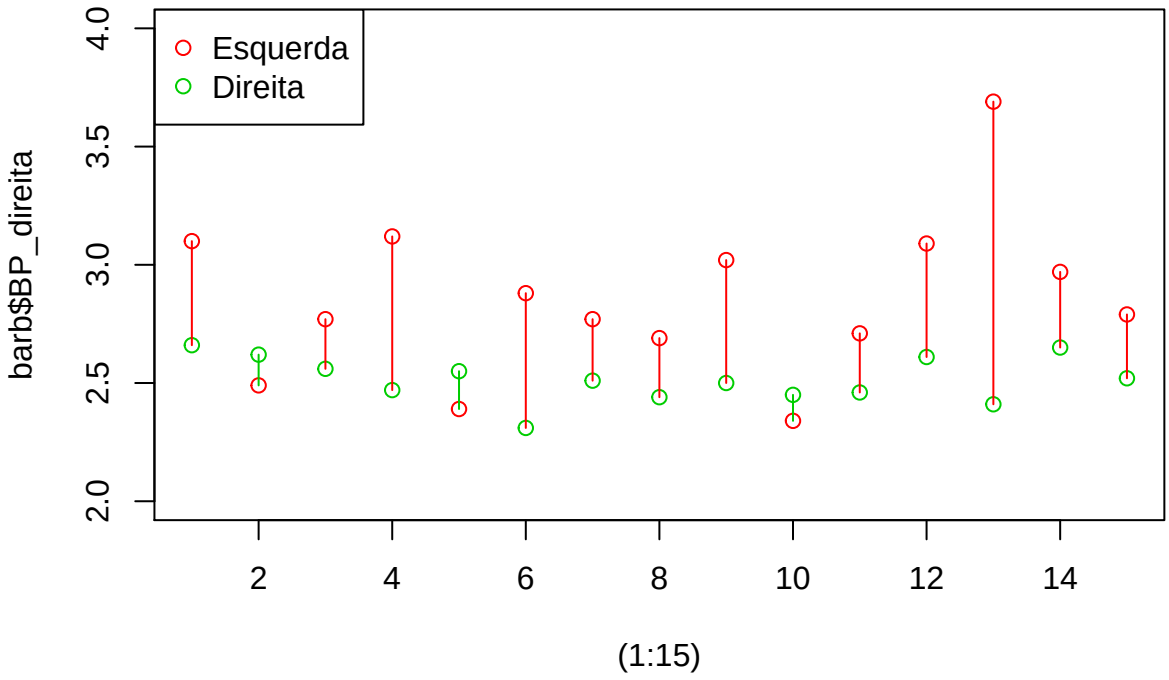
I can now look at the differences

```
par(mfrow=c(2,1))
hist(barb$BP_direita-barb$BP_esquerda)
abline(v=0,lty=2,lwd=2,col=3)
abline(v=mean(barb$BP_direita-barb$BP_esquerda),lty=2,lwd=2,col=2)
```



In fact, a nice way to represent this data is a plot where the difference is shown in a different color depending on which is the larger fin

```
plot((1:15),barb$BP_direita,col=3,ylim=c(2,4))
points((1:15),barb$BP_esquerda,col=2)
segments(x0=1:15,y0=barb$BP_direita,x1=1:15,y1=barb$BP_esquerda,
col=ifelse(barb$BP_direita-barb$BP_esquerda>0,3,2))
legend("topleft",col=c(2,3),legend=c("Esquerda","Direita"),pch=1)
```



We can see that for all but 3 the left fin is larger than the right. That means that quite likely the differences are significant. Let's confirm that formally with a paired `t.test`.

```
t.test(barb$BP_direita,barb$BP_esquerda,paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  barb$BP_direita and barb$BP_esquerda
## t = -3.6594, df = 14, p-value = 0.002576
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5392734 -0.1407266
## sample estimates:
## mean of the differences
##                   -0.34
```

Perhaps not surprisingly, given that we had noticed differences obvious upfront, the differences would be statistically significant for any of the usual significance levels.

Note this would be the same as a simple t test over the differences.

```
t.test(barb$BP_direita-barb$BP_esquerda)
```

```
##
##  One Sample t-test
##
## data:  barb$BP_direita - barb$BP_esquerda
## t = -3.6594, df = 14, p-value = 0.002576
```

```
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.5392734 -0.1407266
## sample estimates:
## mean of x
##     -0.34
```

Exercício 6

Admita que os pressupostos de normalidade e homocedasticidade não são cumpridos no caso anterior. Efectue um procedimento alternativo que lhe permita realizar um teste para avaliar a mesma hipótese.

Under such a case, we need a paired Wilcoxon test

```
wilcox.test(barb$BP_direita,barb$BP_esquerda,paired = TRUE)
```

```
## Warning in wilcox.test.default(barb$BP_direita, barb$BP_esquerda, paired =
## TRUE): cannot compute exact p-value with ties
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  barb$BP_direita and barb$BP_esquerda
## V = 6, p-value = 0.002372
## alternative hypothesis: true location shift is not equal to 0
```

and we reach the exact same conclusions. There's evidence to reject the H0 that the left and right fin are of the same length (H0:$\mu_R = \mu_L$)

Exercício 7

Efectue o teste que julgar adequado aos dados DataTP5densidades.csv para avaliar se as amostras em causa são provenientes da mesma população. Explore os resultados obtidos e apresente-os tal como se pretendesse incluí-los numa apresentação ou publicação científica.

```
dens <- read.csv("DataTP5densidades.csv", sep=";")
```

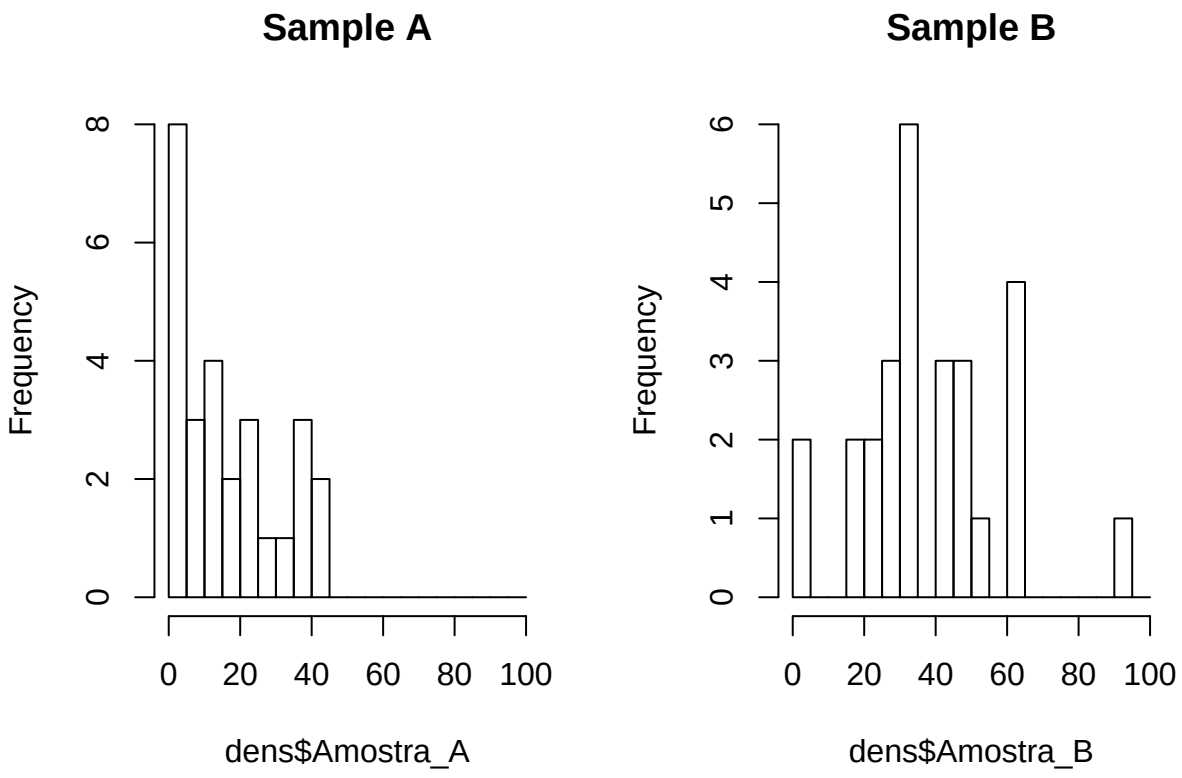
We begin by looking at the data structure

```
str(dens)
```

```
## 'data.frame':    27 obs. of  2 variables:
##  $ Amostra_A: num  4.26 0 22.22 8.33 24.8 ...
##  $ Amostra_B: num  23.1 15.2 46.6 33.6 50.6 ...
```

We can look at the corresponding distributions

```
par(mfrow=c(1,2))
hist(dens$Amostra_A,xlim=c(0,100),breaks=seq(0,100,by=5),main="Sample A")
hist(dens$Amostra_B,xlim=c(0,100),breaks=seq(0,100,by=5),main="Sample B")
```



The sample A values seem clearly lower than the B ones. The data looks nothing like Gaussian, so we use a Wilcoxon test for two samples. There's no indication that these are paired, so we assume independent samples. The H0 is H0:$\mu_A = \mu_B$

```
wilcox.test(dens$Amostra_A,dens$Amostra_B)
```

```
## Warning in wilcox.test.default(dens$Amostra_A, dens$Amostra_B): cannot
## compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  dens$Amostra_A and dens$Amostra_B
## W = 134, p-value = 6.912e-05
## alternative hypothesis: true location shift is not equal to 0
```

Our initial guess is supported by the test. At the usual significance levels we would reject H0. If we would not reject the test, then we could investigate if the variances were the same. If we would not reject that test, then we could conclude there were no support to reject both the equality of means and the equality of variances... But, actually, the request was to test if the two samples were coming from the same underlying distribution.

While this was not presented as such in the Aulas Teóricas, a variant of the Kolmogorov-Smirnov can be used to test if two samples might come from the same underlying distribution. We implement it here

```
ks.test(dens$Amostra_A,dens$Amostra_B)
```

```
## Warning in ks.test(dens$Amostra_A, dens$Amostra_B): cannot compute exact p-
## value with ties
```

```
##
```

```
##  Two-sample Kolmogorov-Smirnov test
##
## data:  dens$Amostra_A and dens$Amostra_B
## D = 0.55556, p-value = 0.0004807
## alternative hypothesis: two-sided
```

Not surprisingly (since even the H0 over the means was rejected) we reject the H0 that the two samples come from the same underlying distribution.

Conclusions

In this Ficha de Trabalho we implemented tests for 1 and 2 samples, both for paired or independent samples, if necessary by first testing the assumptions of the parametric versions of the tests, and if those failed, using the corresponding non-parametric alternatives.

If the P-value of a test is lower than the significance value we defined a priori for the test, we can reject the null hypothesis. If not, we can't reject it, because there is no sufficient information to do so. We should never either accept the null or accept the alternative. Hypothesis can't be accepted (or proven!), they can only be rejected or not. Therefore, we can ONLY reject, or not, the null hypothesis.