Week 4

Ben Derrick

2023-02-21

Note: this worksheet uses the ‘iris’ dataset that exists within R. Only basic statistics and graphics are given in this worksheet to keep the focus on the hypothesis tests. [Advanced week 4 worksheet with alternative example can be used to see well designed summary statistics and graphics]

## Research Question 1. Is the population Sepal Width 3cm?

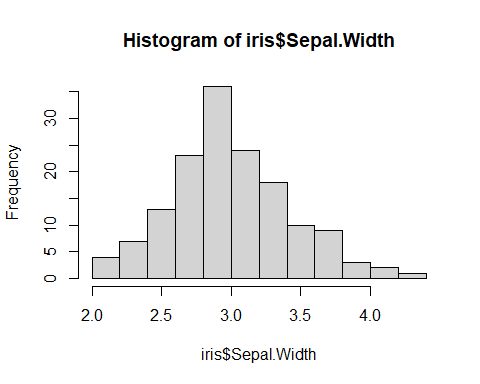
First, we write the null hypothesis formally.

### EDA

I can get a quick summary of this variable (five number summary, mean as well as count of NA values):

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.000 2.800 3.000 3.057 3.300 4.400

One other key piece of EDA will be a histogram.



What does the EDA tell us about the truth (or otherwise) of the null hypothesis?

## A text book hypothesis test

The textbook hypothesis test in this case would be the “one-sample t-test”.

### Assumptions

* no gross outliers
* observations independent
* sample large enough that could be Normally distributed

### The test

t.test(iris$Sepal.Width, mu = 3, alternative = "two.sided")

##   
## One Sample t-test  
##   
## data: iris$Sepal.Width  
## t = 1.611, df = 149, p-value = 0.1093  
## alternative hypothesis: true mean is not equal to 3  
## 95 percent confidence interval:  
## 2.987010 3.127656  
## sample estimates:  
## mean of x   
## 3.057333

### The conclusion

What is the statistical conclusion to this hypothesis test? (Remember this will be a statement about the null hypothesis.)

### The interpretation

1. Are the assumptions met (look at the histogram and the NA count)?
2. If we reject the null hypothesis, is the difference practically important?

## Research Question 2. Do setosa have the same Sepal Widths as versicolor?

Firstly we will filter so that only setos and versicolor species are included.

Filtered\_iris = subset (iris, Species == "setosa" | Species =="versicolor")

The null hypothesis in this case can be written down.

where denotes the Population mean value for setos, and denotes the Population mean value for versicolor.

### EDA

We can obtain group means

summary(Filtered\_iris$Sepal.Width[iris$Species=="versicolor"])

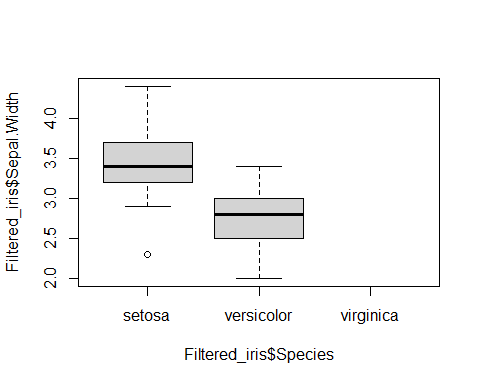
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.000 2.525 2.800 2.770 3.000 3.400

summary(Filtered\_iris$Sepal.Width[iris$Species=="setosa"])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.300 3.200 3.400 3.428 3.675 4.400

And the most important visual will be a boxplot (or a faceted histogram).

boxplot(Filtered\_iris$Sepal.Width ~ Filtered\_iris$Species)



What does the EDA tell us about the truth (or otherwise) of the null hypothesis?

## Another text book hypothesis test

In this case we might use a two-sample t-test (independent samples).

### Assumptions

* no gross outliers,
* observations independent,
* sample large enough that and could be Normally distributed,
* both groups have the same variance

### The test

t.test(Filtered\_iris$Sepal.Width ~ Filtered\_iris$Species, var.equal=TRUE)

##   
## Two Sample t-test  
##   
## data: Filtered\_iris$Sepal.Width by Filtered\_iris$Species  
## t = 9.455, df = 98, p-value = 1.845e-15  
## alternative hypothesis: true difference in means between group setosa and group versicolor is not equal to 0  
## 95 percent confidence interval:  
## 0.519895 0.796105  
## sample estimates:  
## mean in group setosa mean in group versicolor   
## 3.428 2.770

### The conclusion

What is the statistical conclusion to this hypothesis test? (Remember this will be a statement about the null hypothesis.)

### The interpretation

1. Are the assumptions met (look at the histogram and the NA count)?
2. If we reject the null hypothesis, is the difference practically important?

## Introductory exercise, using Dewis

Use the help provided in the Dewis test to:

* perform a paired samples t-test to see if the mean of two groups are equal
* assess the assumption of the paired samples t-test (normality of differences)

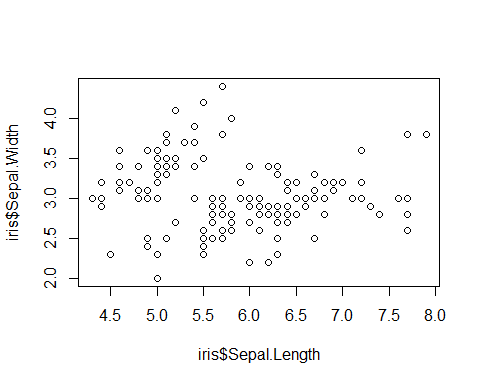
# In here, type in your code for the Dewis practice test on the paired samples t-test

## Research Question 3. Is there a correlation between Sepal Length and Sepal Width?

Note this one carefully, the null hypothesis is that the Population correlation coefficient is zero:

### EDA

plot(iris$Sepal.Length,iris$Sepal.Width)



with(iris, cor(Sepal.Length, Sepal.Width, use="pairwise.complete.obs"))

## [1] -0.1175698

## Hypothesis test

A hypothesis test assumes that Pearson’s product moment correlation coefficient follows a t distribution with degrees of freedom if the samples follow independent normal distributions. What other EDA might we need to check these assumptions?

### The test

with(iris, cor.test(Sepal.Length, Sepal.Width, use="pairwise.complete.obs"))

##   
## Pearson's product-moment correlation  
##   
## data: Sepal.Length and Sepal.Width  
## t = -1.4403, df = 148, p-value = 0.1519  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.27269325 0.04351158  
## sample estimates:  
## cor   
## -0.1175698

### The conclusion

What is the statistical conclusion to this hypothesis test? (Remember this will be a statement about the null hypothesis.)

### The interpretation

1. Are the assumptions met - do you need some more EDA to check this
2. If we reject the null hypothesis, does this tell us that the correlation is practically important?

## Further Dewis practice on correlation

There is a Dewis practice test on testing and interpreting correlation Question 1 is on testing and interpreting correlation Question 2 is extra practice of interpreting statistical tests.

# In here, type in your code for Question 1 of the Dewis practice test on correlation

## Advanced activity using iris

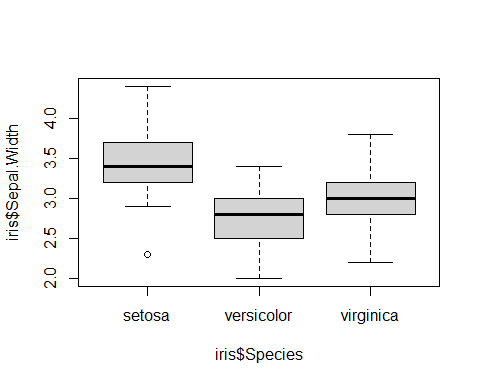
## Research Question 4. Do the three Species differ on Sepal Width

This is an extension of the comparison of means we did for two Species, only now we have more than 2 groups and must use a different test. The null hypothesis:

(where S denotes setosa, V denotes versicolor, Vi denotes virginoca)

### EDA

boxplot(iris$Sepal.Width ~iris$Species)



## Hypothesis test

Because we are comparing more than one population mean, first of all we conduct a so-called “Analysis of variance” (ANOVA) against the null hypothesis that all means are the same.

### Assumptions

1. There are no gross outliers,
2. the observations are independent,
3. the variance is the same in each group and
4. the group means are Normally distributed

### The test

We can perform a classical ANOVA using the aov() command in R. The print() and summary() methods give slightly different information on the ANOVA object you create.

test3 <- aov(Sepal.Width ~ Species, iris)  
print(test3)

## Call:  
## aov(formula = Sepal.Width ~ Species, data = iris)  
##   
## Terms:  
## Species Residuals  
## Sum of Squares 11.34493 16.96200  
## Deg. of Freedom 2 147  
##   
## Residual standard error: 0.3396877  
## Estimated effects may be unbalanced

summary(test3)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Species 2 11.35 5.672 49.16 <2e-16 \*\*\*  
## Residuals 147 16.96 0.115   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1