Exam 2

You can download the combined lecture notes for this module at: http://asta.math.aau.dk/dan/2017f/asta/?file=handouts/module-B.pdf

It is highly recommended that you answer the exam using Rmarkdown (you can simply use the exam Rmarkdown file as a starting point).

Remember to load the `mosaic` package first:

```
library(mosaic)
```

```
## Warning: package 'mosaic' was built under R version 3.4.2
```

Part I: Directed reading activities

An educator conducted an experiment to test whether new directed reading activities in the classroom will help elementary school pupils improve some aspects of their reading ability.

She arranged for a third grade class of 21 students to follow these activities for an 8-week period. A control classroom of 23 third graders followed the same curriculum without the activities. At the end of the 8 weeks, all students took a Degree of Reading Power (DRP) test, which measures the aspects of reading ability that the treatment is designed to improve.

Read in the data:

```
reading <- read.table("http://asta.math.aau.dk/dan/static/datasets?file=reading.dat", header=TRUE)
head(reading)
```

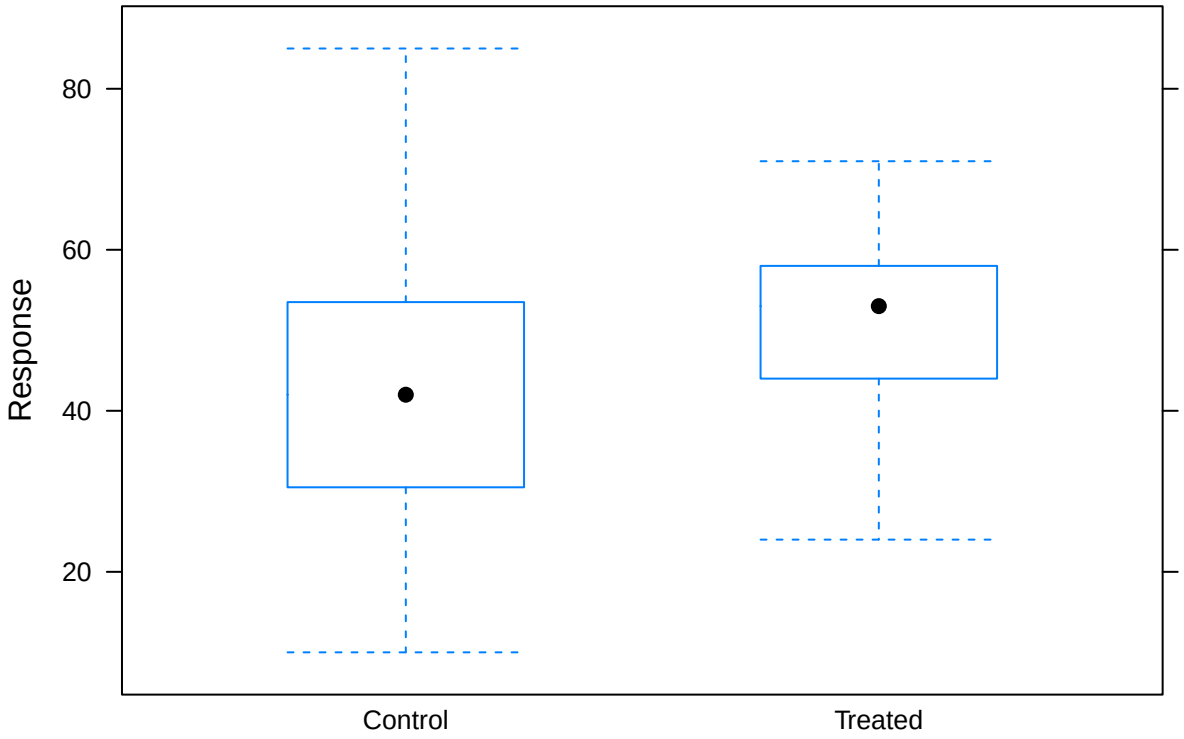
```
##   Treatment Response
## 1   Treated       24
## 2   Treated       43
## 3   Treated       58
## 4   Treated       71
## 5   Treated       43
## 6   Treated       49
```

Use a boxplot to compare the of measurements of `DRP` for `Treated`(direct reading activities) and `Control` visually.

```
'we want to see the score of each group, DRP=the experiment they have done'
```

```
## [1] "we want to see the score of each group, DRP=the experiment they have done"
```

```
bwplot(Response~Treatment,data=reading)
```




```
'WHAT CAN BE SEEN IN THIS BOXPLOT?
1-treated group all the quartos are higehr than in the control
some childresn score higher because the y axes is higher, but the treated perform better, this is not s
```

```
## [1] "WHAT CAN BE SEEN IN THIS BOXPLOT? \n1-treated group all the quartos are higehr than in the cont
```

Use `favstats` to make a numerical summary of the measurements for `Treated` and `Control`.

```
a<-favstats(Response ~ Treatment, data = reading)
a
```

```
##   Treatment min   Q1 median   Q3 max     mean       sd  n missing
## 1   Control  10 30.5     42 53.5  85 41.52174 17.14873 23       0
## 2   Treated  24 44.0     53 58.0  71 51.47619 11.00736 21       0
```

```
'now we see the response in relation to treatmeant, so the mean is 10 points higher!, the min, q1 compa
favstats(Respose~Treatment,)'
```

```
## [1] "now we see the response in relation to treatmeant, so the mean is 10 points higher!, the min, q
```

```
Control <- subset(reading, Treatment == "Control")
Treated <- subset(reading, Treatment == "Treated")
y1 = mean(~ Response, data = Control)
y2 = mean(~ Response, data = Treated)
y1
```

```
## [1] 41.52174
```

```
y2
```

```
## [1] 51.47619
```

```
s1 = sd(~ Response, data = Control)
s2 = sd(~ Response, data = Treated)
s2
```

```
## [1] 11.00736
```

```
s1
```

```
## [1] 17.14873
```

```
'S es la variacion y s^2 es la S=Sd desviacion tipica'
```

```
## [1] "S es la variacion y s^2 es la S=Sd desviacion tipica"
```

- Write down a point estimate of the mean of DRP for students following the new *directed reading activities* and explain how this is calculated.

$$\overline{y}_{Treated} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i; \text{with } y_1 \text{in the treated group}$$

```
'we just need to multiply each observation and divide it by the size of our variable.'
```

```
## [1] "we just need to multiply each observation and divide it by the size of our variable."
```

$$se = \sqrt{\frac{\pi(1-\pi)}{n}}$$

```
the sum of all the obs, divided by the obs of the group, explain the formula
'how the mean is calc'
```

- Write down a point estimate of the standard deviation of DRP for this group and explain how this is calculated.

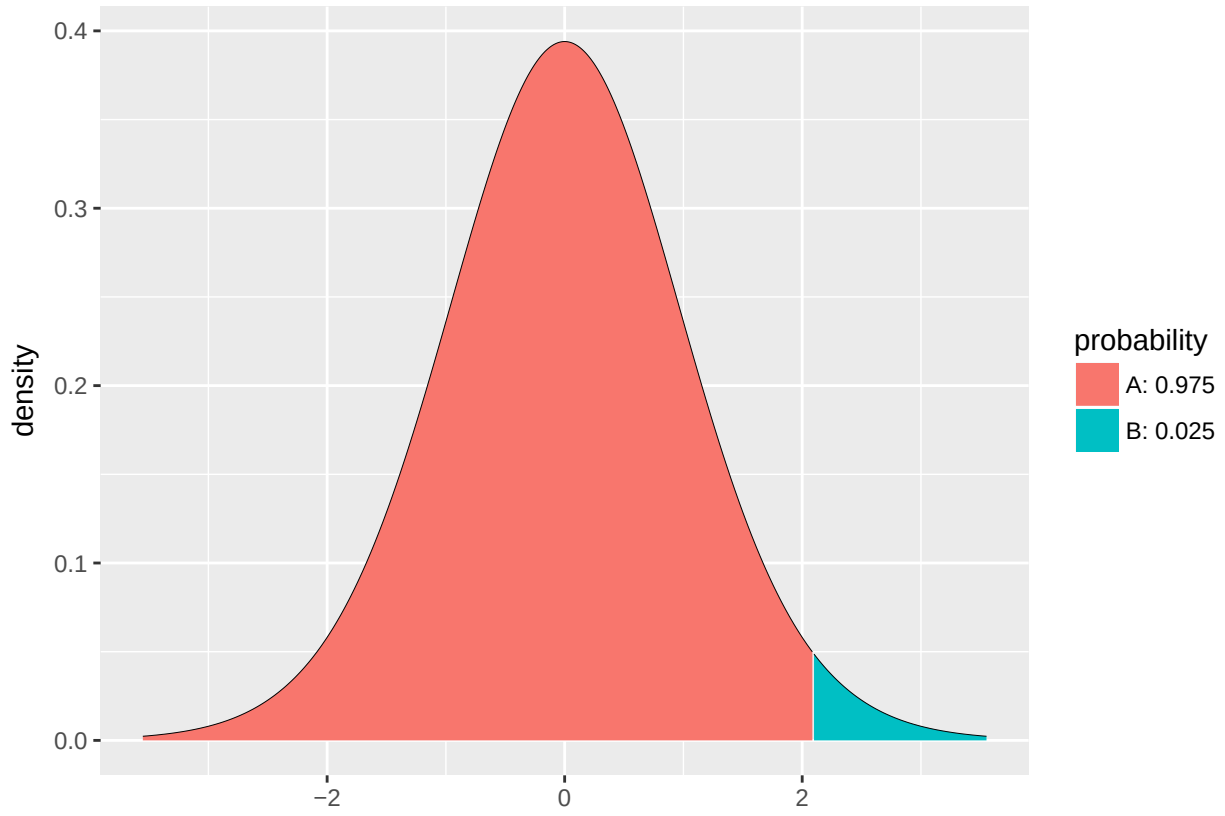
$$S_{treated} = \sqrt{\frac{\sum(yi-\overline{y})^2}{n-1}}$$

- Write down a 95% confidence interval for the mean of DRP for this group and explain how this is calculated.

```
'we use t score because the number of samples is less than 30'
```

```
## [1] "we use t score because the number of samples is less than 30"
```

```
yvar=51.447619
s=11.00736
n=21
t<-qdist("t", 1-0.025, df=n-1)
```




```
t
```

```
## [1] 2.085963
```

```
'remember that s/sqrt(n) is equal to se'
```

```
## [1] "remember that s/sqrt(n) is equal to se"
```

```
interval951<-yvar-t*s/sqrt(n)

interval952<-yvar+t*s/sqrt(n)
interval951
```

```
## [1] 46.43713
```

```
interval952
```

```
## [1] 56.45811
```

```
' So you are 95% confident that the real mean will be between those interval=56.46 and 46.44'
```

```
## [1] " So you are 95% confident that the real mean will be between those interval=56.46 and 46.44"
```

All those results can be obtained by the comand `t.test`

```
'So we use t.test to comare it with the actual results calculated previously and we can say that in int
```

```
## [1] "So we use t.test to comare it with the actual results calculated previously and we can say that
```

```
t.test( ~Response, data = Treated, conf.level = 0.95)
```

```
## ~Response
```

```
##
##  One Sample t-test
##
## data:  Response
## t = 21.431, df = 20, p-value = 2.877e-15
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  46.46570 56.48668
## sample estimates:
## mean of x
##  51.47619
```

Use the command `t.test` to compare the mean DRP of the two groups.

```
t.test(Response ~ Treatment , data = reading)
```

```
## Response ~ Treatment
```

```
##
##  Welch Two Sample t-test
##
## data:  Response by Treatment
## t = -2.3109, df = 37.855, p-value = 0.02638
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.67588  -1.23302
## sample estimates:
## mean in group Control mean in group Treated
##              41.52174              51.47619
```

```
'If Pvalue is less than 0.05 we can say that we reject the zero hypothesis as it is unprobable to occur

Dentro del resultado t test el intervalo de confianza del 95% es para la hypothesis, If 0 is inside the
```

```
## [1] "If Pvalue is less than 0.05 we can say that we reject the zero hypothesis as it is unprobable t
```

```
'What does the -18 and -1 confidence interval mean?? that if you make another experiment with new sampl
```

```
## [1] "What does the -18 and -1 confidence interval mean?? that if you make another experiment with ne
```

Go through the details of the output from t.test. Your analysis must include an account of

- What the relevant null hypothesis and the corresponding alternative hypothesis is.

```
'
dependent sample: if we take the wiegh of our class one day, then we ate some chocholoate and see the w
independent sample: if we take the weights of different classes.

WE ASSUME THAT IT IS INDEPENDENT for the calculations.the first 3 questions are related making a hypoth
```

```
## [1] " \ndependent sample: if we take the wiegh of our class one day, then we ate some chocholoate an
```

The null hypothesis is

$$H_0 : \hat{\mu}_1 = \hat{\mu}_2$$

and the alternative hypothesis is

$$H_a : \hat{\mu}_1 \neq \hat{\mu}_2$$

- Choice and calculation of test statistic.

```
'For obtaining the test statistic, it is neccesary to calculate the standard error of the difference, as
'
```

```
## [1] "For obtaining the test statistic, it is neccesary to calculate the standard error of the differe
```

$$se_d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

```
##   Treatment min   Q1 median   Q3 max     mean       sd  n missing
## 1   Control  10 30.5     42 53.5  85 41.52174 17.14873 23       0
## 2   Treated  24 44.0     53 58.0  71 51.47619 11.00736 21       0
```

```
## [1] 11.00736
```

```
## [1] 17.14873
```

```
## [1] 4.307628
```

```
## [1] "es lo q se desvia de la diferencia por arriba y por abajo. en este caso la d= es 10 mas o menos
```

The value of t for the t.test is calculated as follows:

$$t = \frac{\overline{y}_2 - \overline{y}_1}{se_d} \; \mu_1 and \mu_2 \text{the difference is equal to } 0$$

```
## [1] -2.310889
```

```
## [1] "tenemos una z , teniendo la t=z (eje X) miramos la probabilidad de estar a la derecha de esa obs
```

- Calculation of p-value and its interpretation in connection to a conclusion of the analysis.

```
'qdits("norm") we use it for the z score
qdits("t") we use it for the z score
degree od freedom because the t score varies with the degree of freedom.
n=number of observations bigger than 30  you use a z score
n=number of observations less than 30 you use a t score'
```

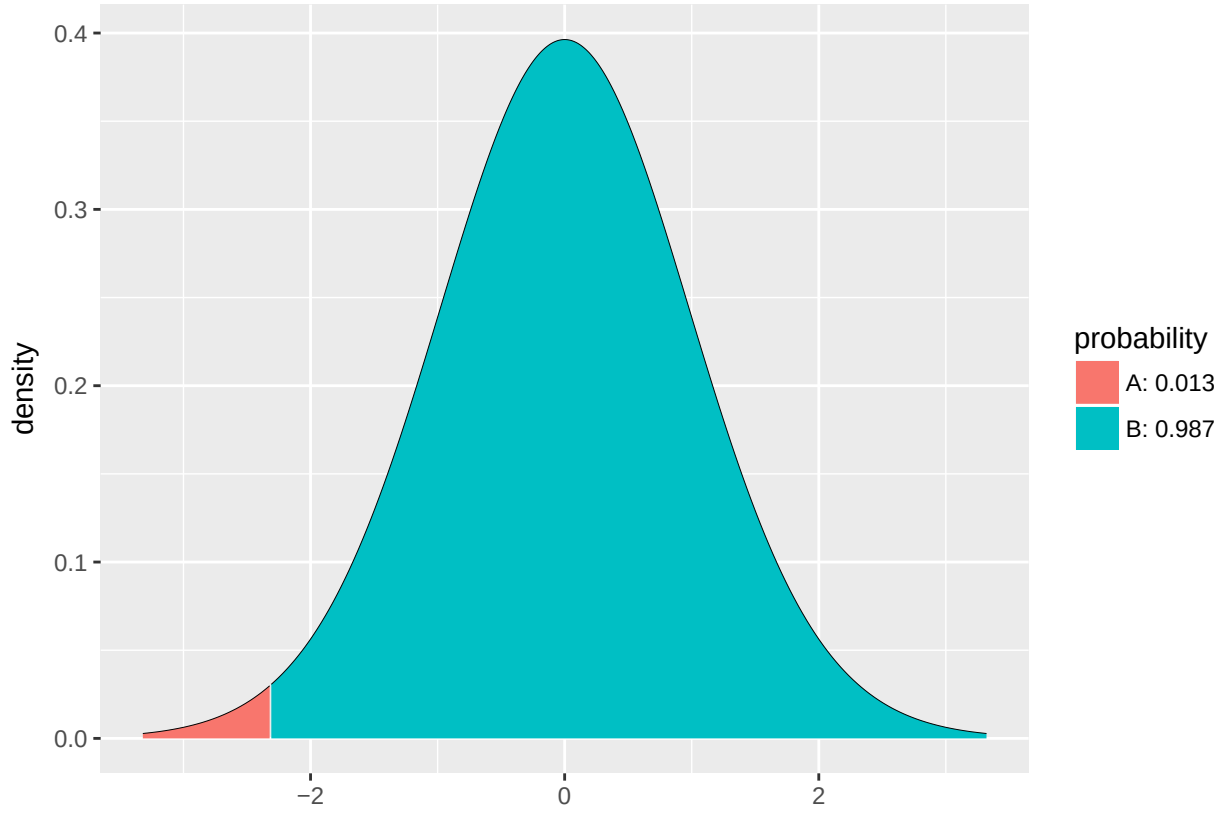
```
## [1] "qdits(\"norm\") we use it for the z score\nqdits(\"t\") we use it for the z score\ndegree od fr
```

Once obtained the t value, the probability of occurrence of this value is obtained by the t curve. The p-value is defined as the left probability of this value times two. When the p-value is small, we can consider rejecting the H_0 hypothesis.

```
## [1] "you obtain the degrees of freedom df from the t.test previous exercises, which is 38, t.test(Res
```

```
## Response ~ Treatment
```

```
##
##  Welch Two Sample t-test
##
## data:  Response by Treatment
## t = -2.3109, df = 37.855, p-value = 0.02638
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.67588  -1.23302
## sample estimates:
## mean in group Control mean in group Treated
##              41.52174              51.47619
```




```
## [1] 0.02638247
```

and if we put t for the t score - Calculation and interpretation of a relevant confidence interval.

```
'independent they are not all the same students'
```

```
## [1] "independent they are not all the same students"
```

Part II: Determining sample size

In this part there is no dataset to load into R and analyze. You should just use R as a calculator when you apply the relevant formulas (which are towards the end of the lecture notes for module 1).

To estimate the proportion of danish companies with less than 10 employees determine the necessary sample size for the estimate to be accurate to within 0.06 with probability 0.90. Based on results from a previous study in 2013, we expect the proportion to be about 0.70.

```
#how many companies you have to take to be ass aquerrate as 0.7 with a prob of 0.9'
#90% means you can be on the right 5% and to the left %5'
```

The margin of error when estimating a proportion is

$$M = z\sqrt{\frac{\pi(1-\pi)}{n}}$$

Solving the equation for obtaining n prop.test(Chile$vote, 2532, success="N", correct=FALSE)

$$n = \pi(1-\pi)(\frac{z}{M})^2$$

With the given values, the sample size can be calculated

```
M = 0.06
pi = 0.7
z<-qdist("norm", p= 0.05, mean = 0, sd = 1, plot = FALSE)
'with q dist you calculate the z value for a probability of 10 %. 5% to the right and 5% to the left'
```

```
## [1] "with q dist you calculate the z value for a probability of 10 %. 5% to the right and 5% to the l
```

```
n = pi*(1-pi)*(z/M)^2
n
```

```
## [1] 157.8234
```

If we want to be more conservative, we can use the value of $\pi = 0.5$

$$n = (\frac{z}{2M})^2$$

```
## [1] "now if we decrease the pi we have less information about the proportion, so if we want a 90 perc
```

```
## [1] 187.885
```