Exam

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

Part I: Estimating probabilities

Remember to load the `mosaic` package first:

```
library(mosaic)
```

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

```
options(digits = 4)
```

Chile referendum data

In this part we will use the dataset `Chile`. Remember to read the description of the dataset as well as the Wikipedia entry about the background.

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

NB: This dataset has several missing values (`NA`). To remove these when you use `tally` you can add the argument `useNA = "no"`.

- Do a cross tabulation of the variables `vote` and `sex`.

```
TableChile<-tally(~vote+sex, data=Chile, useNA = "no")
```

- Estimate the probability of `vote=N`.

```
suma<-sum(TableChile[2,1:2])/(sum(TableChile[1:4,1:2]))
suma
```

```
## [1] 0.3511
```

```
a<-(sum(TableChile[1:4,1:2]))
a
```

```
## [1] 2532
```

```
VoteN<-(363+526)/2532
VoteN
```

```
## [1] 0.3511
```

- Make a 95% confidence interval for the probability of `vote=N`.

```
'What is pi_hat? is a proportion. the ones voting N among the other ones. the probablity '
```

```
## [1] "What is pi_hat? is a proportion. the ones voting N among the other ones. the probablity "
```

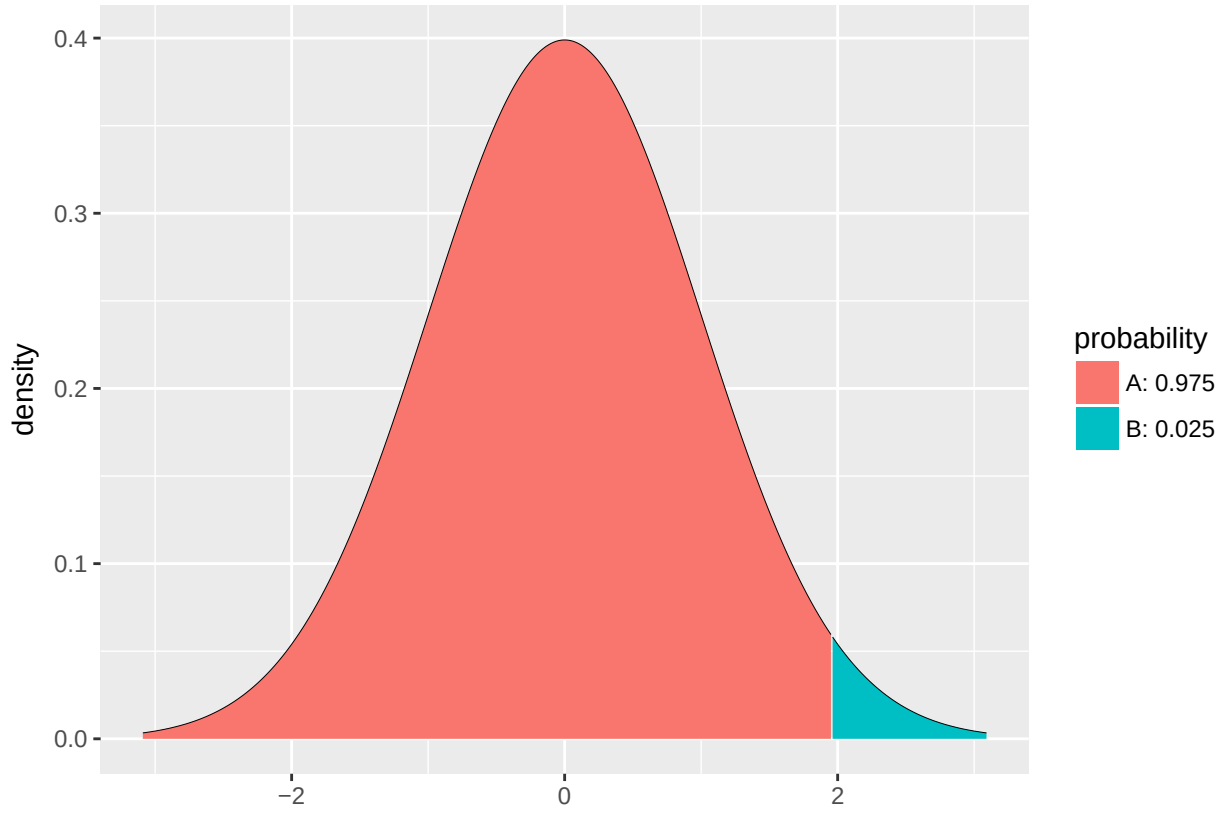
```
n<-2532
pi_hat<-((363+526)/2532)
pi_hat
```

```
## [1] 0.3511
```

```
'p-value: which is 0.05, so you obtain the z value, why divided by 2?? because R only takes the whole a
```

```
## [1] "p-value: which is 0.05, so you obtain the z value, why divided by 2?? because R only takes the
```

```
z1<-qdist("norm", 1-0.05/2)
```




```
z1
```

```
## [1] 1.96
```

```
estimated_standard_error<-sqrt(pi_hat*(1-pi_hat)/(n))
estimated_standard_error
```

```
## [1] 0.009486
```

```
confidence_interval991<-pi_hat+z1*estimated_standard_error
confidence_interval991
```

```
## [1] 0.3697
```

```
confidence_interval992<-pi_hat-z1*estimated_standard_error
confidence_interval992
```

```
## [1] 0.3325
```

```
prop.test(Chile$vote,p=pi_hat, 2532, success="N", correct=FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
```

```
## data:  Chile$vote  [with success = N]
## X-squared = 4.6e-29, df = 1, p-value = 1
## alternative hypothesis: true p is not equal to 0.3511
## 95 percent confidence interval:
##  0.3328 0.3699
## sample estimates:
##      p
## 0.3511
```

```
'we use prop test for a proportion. in this case it calculates all the parameters for N, saying that th
1-pvalue>0.05 so we can say that the zero hyp is true. 2-we calculate the interval confidence which is

So you are sure, if you do another experiment with different samples that your proportion(pihat)will be
```

```
## [1] "we use prop test for a proportion. in this case it calculates all the parameters for N, saying
```

- Estimate the probability of `vote=N`, given that `sex=F`.

```
'Conditional probability (probablity of vote N having a condiotion, in this case those which are N have
```

```
## [1] "Conditional probability (probablity of vote N having a condiotion, in this case those which are
```

```
ProbabilityNFemale<-363/(sum(TableChile[1:4,1:1]))
ProbabilityNFemale
```

```
## [1] 0.2773
```

```
#or you can do it this way:
sumaF<-sum(TableChile[2,1:1])/(104+363+362+480)
sumaF
```

```
## [1] 0.2773
```

- What would these probabilities satisfy if `vote` and `sex` were statistically independent?

```
'Now both probabilities are independent, so the answer is the probability to be a woman and have voted
```

```
## [1] "Now both probabilities are independent, so the answer is the probability to be a woman and have
```

```
P_independent<-((363+526)/2532)*(1309/2532)
P_independent
```

```
## [1] 0.1815
```

Part II: Sampling distributions and the central limit theorem

This is a purely theoretical exercise where we investigate the random distribution of samples from a known population.

House prices in Denmark

The Danish real estate agency HOME has a database containing approximately 80.000 house prices for one-family houses under DKK 10 million for the period 2004-2016. The house prices (without all the additional information such as house size, address etc.) are available as a R data file `Home.RData` on the course webpage. If you download it you can load it using `load("Home.RData")` assuming you have saved it in the same directory as this Rmarkdown document. This will add the vector `price` to your work space. Alternatively,

you can add it directly from the course website (this will download it every time you run the Rmarkdown document, so make sure you have a decent internet connection):

```
load(url("http://asta.math.aau.dk/dan/static/datasets?file=Home.RData"))
```

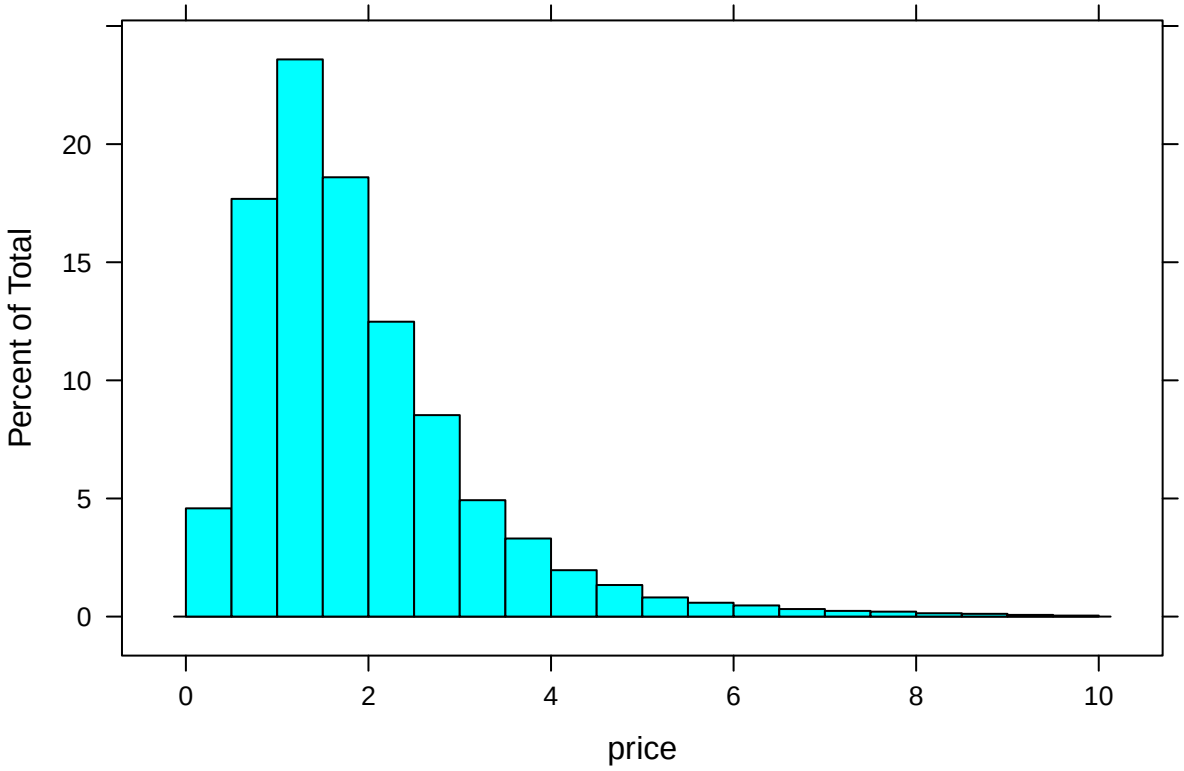
Make a histogram of all the house prices using a command like

`histogram(price, breaks = 30)` inserted in a new code chunk (try to do experiments with the number of breaks):

```
'this command makes an histogram with 30 columns, if you change the number30 to 100 it will make 100 co
```

```
## [1] "this command makes an histogram with 30 columns, if you change the number30 to 100 it will make
```

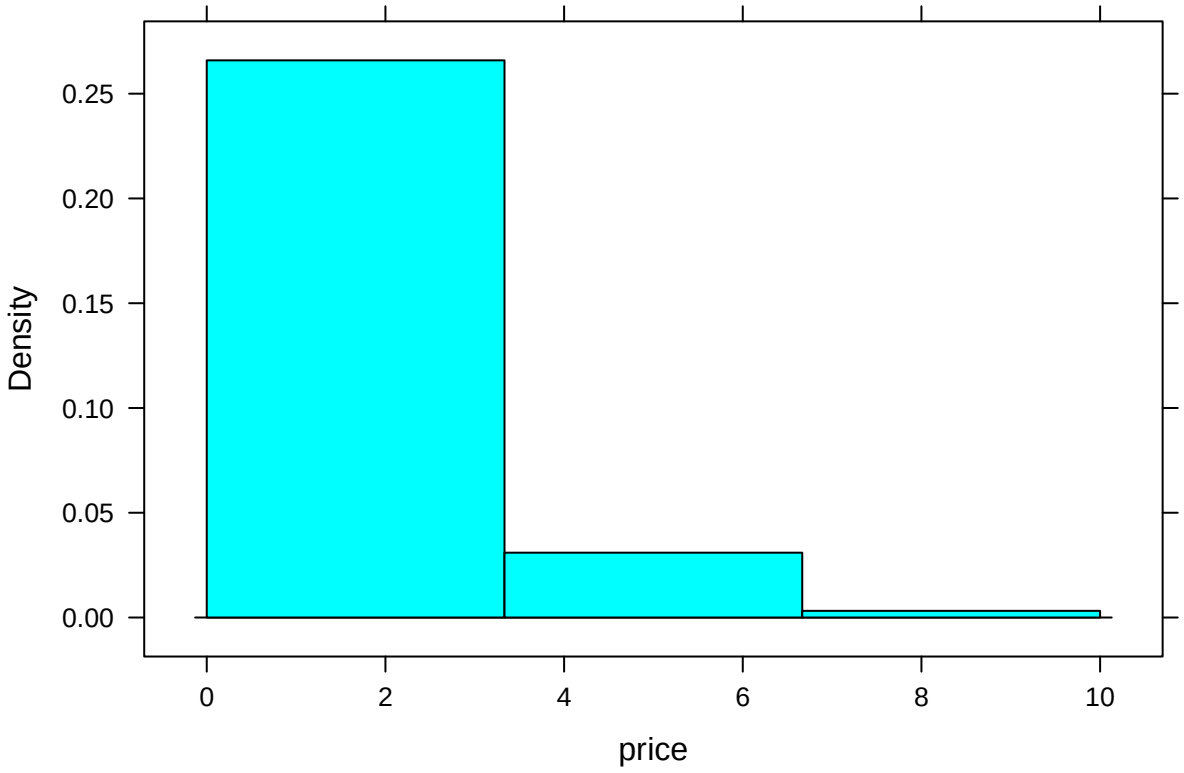
```
histogram(price, breaks =30)
```




```
'or you can use the next command to put the columnns, for example for 3 columns'
```

```
## [1] "or you can use the next command to put the columnns, for example for 3 columns"
```

```
histogram(price, breaks=c(0,3.333,6.666,10))
```



- Explain how a histogram is constructed.

```
'A histogram is an accurate graphical representation of the distribution of numerical data. It is an es

If the bins are of equal size, a rectangle is erected over the bin with height proportional to the frequ
```

```
## [1] "A histogram is an accurate graphical representation of the distribution of numerical data. It is
```

- Does this histogram look like a normal distribution?

```
'no, because it does not have the characteristic bell shaped curve'
```

```
## [1] "no, because it does not have the characteristic bell shaped curve"
```

In this database (our population) the mean price is 1.929 and the standard deviation is 1.2744.

```
mean_value<-mean(price)
mean_value
```

```
## [1] 1.929
```

```
standard_deviation<- sd(price)
standard_deviation
```

```
## [1] 1.274
```

In many cases access to such databases is restrictive and in the following we imagine that we are only allowed access to a random sample of 40 prices and the mean of this sample will be denoted `y_bar`.

Before obtaining this sample we will use the Central Limit Theorem (CLT) to predict the distribution of `y_bar`:

```
'(page 97) For random sampling with a large sample size n, the sampling distribution of the sample
mean y is approximately a normal distribution.

http://onlinestatbook.com/stat_sim/sampling_dist/

The approximate normality of the sampling distribution applies no matter what
the shape of the population distribution. This is quite remarkable. For large
random samples, the sampling distribution of y is approximately normal even
if the population distribution is highly skewed, U shaped.'
```

```
## [1] "(page 97) For random sampling with a large sample size n, the sampling distribution of the samp
```

- What is the expected value of `y_bar`?

```
y_bar<-mean(price)
y_bar
```

```
## [1] 1.929
```

- What is the standard deviation of `y_bar` (also called the standard error)?

```
'see example page 97'
```

```
## [1] "see example page 97"
```

```
se<-sd(price)/sqrt(40)
se
```

```
## [1] 0.2015
```

- What is the approximate distribution of `y_bar`?

```
'it is a normal distribution with a mean of 1.929 and a standard deviation of 0.2015'
```

```
## [1] "it is a normal distribution with a mean of 1.929 and a standard deviation of 0.2015"
```

the aproximate distribution of y_bar is a normal distribution with the mean od the sample and the standard deviation eual to the standard deviation of the samples over the square root of the number of sampas

Now make a random sample of 40 house prices and calculate the sample mean:

```
'First of all we make a random sample (40 elements) of the price vector, with the next command:'
```

```
## [1] "First of all we make a random sample (40 elements) of the price vector, with the next command:"
```

```
y <- sample(price, 40)
y
```

```
##  [1] 3.200 2.195 1.660 0.815 1.625 2.645 1.995 1.198 0.315 2.195 1.700
## [12] 0.690 1.950 3.400 2.895 1.395 3.995 1.575 1.885 4.200 4.700 0.860
## [23] 2.450 1.115 1.565 0.910 0.950 1.575 2.700 1.125 0.785 1.225 1.200
## [34] 2.800 1.245 2.150 1.998 2.000 1.837 0.675
```

```
'and we calculate the mean for this new sample'
```

```
## [1] "and we calculate the mean for this new sample"
```

```
mean_valuey<-mean(y)
mean_valuey
```

```
## [1] 1.885
```

```
'this is the actual mean of the other sample(80000 elements) calculated previously:'
```

```
## [1] "this is the actual mean of the other sample(80000 elements) calculated previously:"
```

```
mean_value<-mean(price)
mean_value
```

```
## [1] 1.929
```

Repeat this command a few times. Is each mean price close to what you expected?

Use `replicate` to repeat the sampling 200 times and save each mean value in the vector `y_bar`:

```
'so now we take 200 means of a vector of different 40 samples.'
```

```
## [1] "so now we take 200 means of a vector of different 40 samples."
```

```
y_bar2 <- replicate(200, mean(sample(price, 40)))
y_bar2
```

```
##   [1] 2.059 1.915 1.849 1.957 1.875 1.772 1.816 2.121 2.428 1.998 2.118
##  [12] 2.142 1.702 1.959 1.953 2.011 2.021 1.841 1.849 1.940 1.764 1.607
##  [23] 1.831 1.936 2.164 1.972 1.978 2.142 2.093 2.058 1.929 1.923 1.887
##  [34] 1.777 1.715 2.133 1.745 1.599 1.600 1.924 2.040 1.633 2.219 2.075
##  [45] 1.909 2.009 2.299 1.956 1.882 1.858 1.657 1.826 2.127 1.754 2.200
##  [56] 2.235 2.294 1.672 2.059 2.231 1.829 2.018 2.312 2.143 1.973 2.123
##  [67] 1.958 1.919 2.031 1.906 2.034 2.290 1.724 1.482 1.756 2.015 1.911
##  [78] 2.111 1.671 1.639 1.985 2.029 1.892 1.897 1.579 1.705 1.554 1.797
##  [89] 1.825 1.977 2.040 1.974 2.284 2.045 1.987 1.905 1.958 1.981 2.199
## [100] 1.678 1.891 1.584 1.724 1.970 1.864 2.112 1.586 2.067 1.656 1.700
## [111] 2.169 1.511 1.532 1.724 2.301 2.079 1.945 2.291 2.350 2.066 1.831
## [122] 1.950 1.791 2.075 1.836 2.072 2.163 1.690 1.890 1.850 1.994 1.689
## [133] 1.700 1.863 1.442 1.827 1.948 2.415 1.961 1.756 1.769 1.945 1.744
## [144] 1.943 1.940 1.814 1.782 2.226 2.043 1.723 1.670 1.766 1.631 2.164
## [155] 1.701 1.957 2.339 1.906 1.818 1.862 2.039 1.843 1.708 2.270 1.596
## [166] 1.770 1.568 2.327 2.055 1.585 2.065 1.615 1.874 2.081 1.804 2.215
## [177] 2.090 1.828 1.843 2.160 1.961 2.115 1.874 1.708 1.846 2.065 1.803
## [188] 1.688 1.811 1.976 1.929 2.013 2.782 1.872 1.987 1.795 1.945 2.316
## [199] 1.906 1.645
```

Calculate the mean and standard deviation of the values in `y_bar`.

```
Final_mean<-mean(y_bar2)
Final_mean
```

```
## [1] 1.926
```

```
Final_standard_deviation<-sd(y_bar2)
Final_standard_deviation
```

```
## [1] 0.2113
```

```
'Now we compare these values to the first 80000 elements mean and standard deviation and we expect that
page100 example interestinf'
```

```
## [1] "Now we compare these values to the first 80000 elements mean and standard deviation and we expe
```

```
y_bar
```

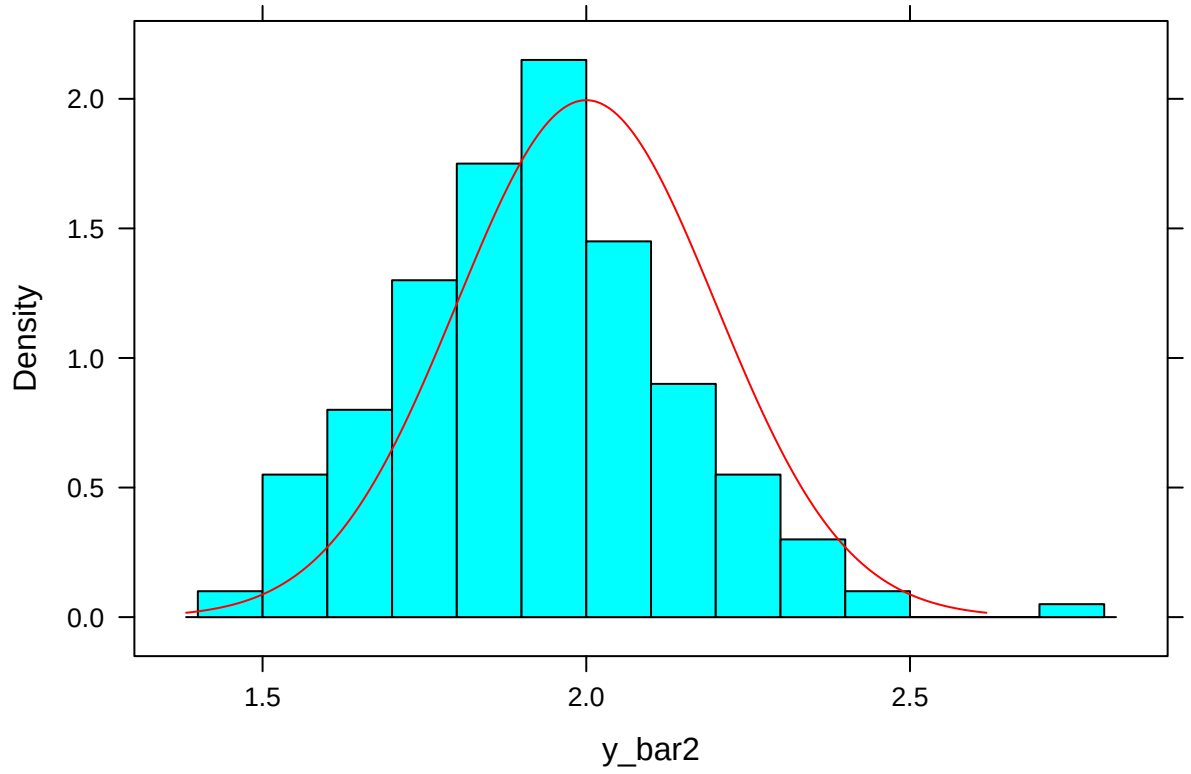
```
## [1] 1.929
```

```
se
```

```
## [1] 0.2015
```

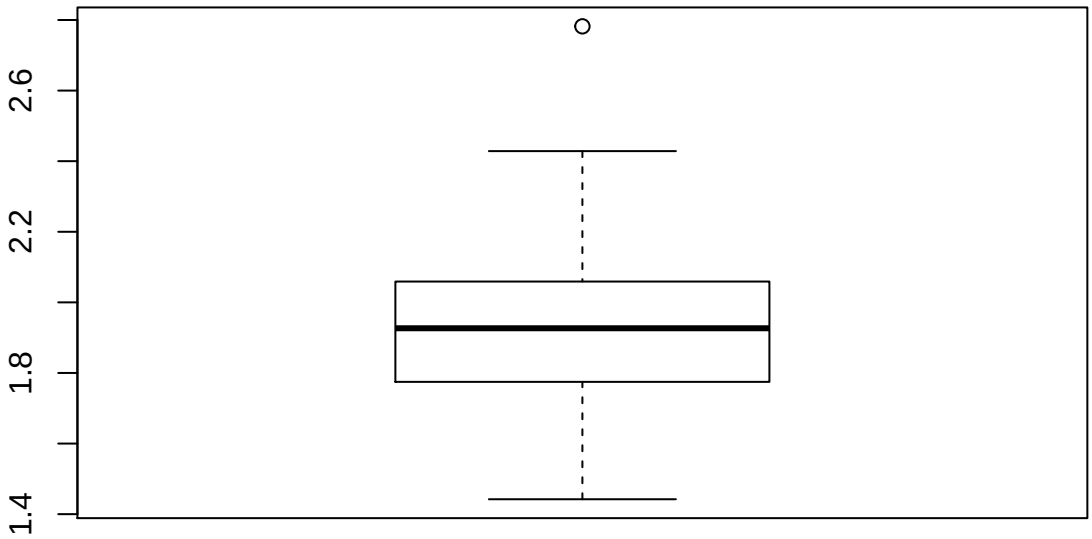
- How do they match with what you expected?
- Make a histogram of the values in `y_bar` and add the density curve for the approximate distribution you predicted previously using `plotDist` with the argument `add = TRUE`. For example if you predicted a normal distribution with mean 2 and standard deviation 0.2:

```
histogram(y_bar2, breaks = 15, type = "density")
plotDist("norm", mean = 2, sd = 0.2, add = TRUE, col = "red")
```



- Make a boxplot of `y_bar` and explain how a boxplot is constructed.

```
boxplot(y_bar2)
```



Part III: Theoretical boxplot for a normal distribution

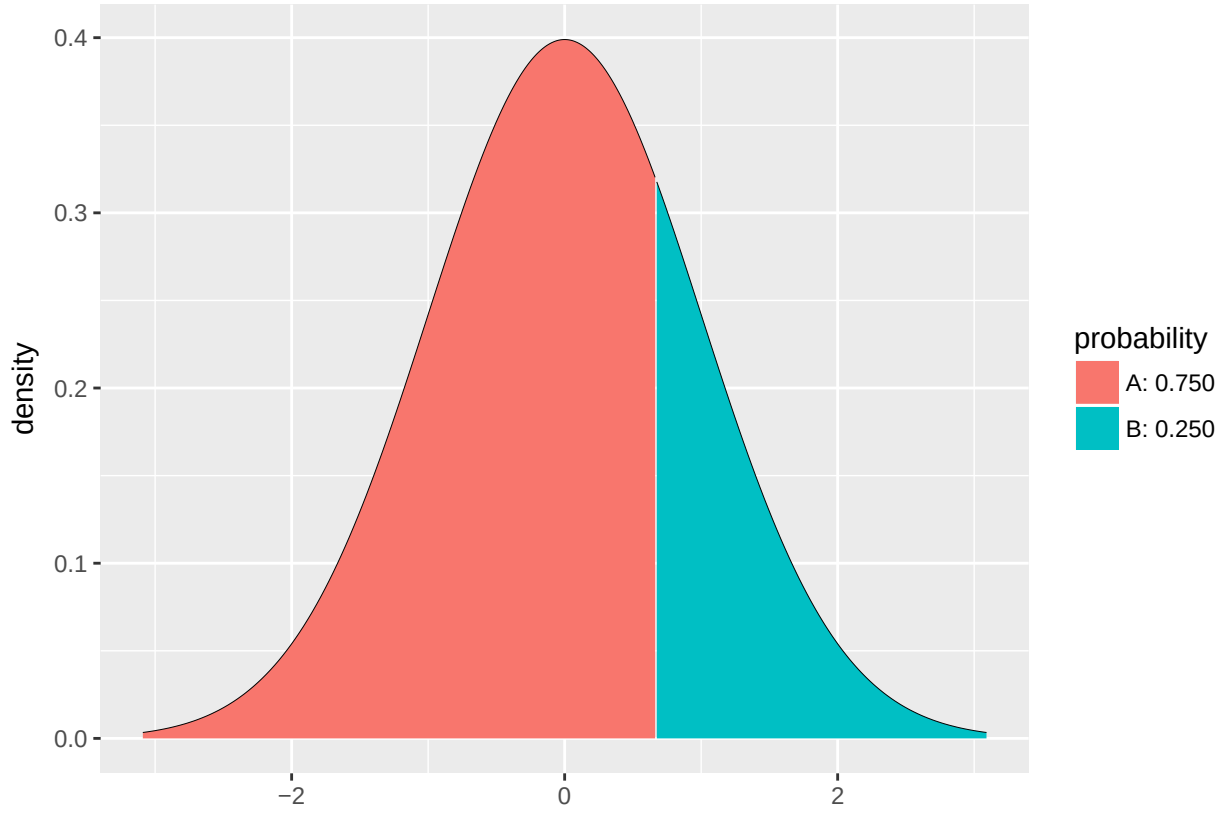
Finally, consider the theoretical boxplot of a general normal distribution with mean μ and standard deviation σ, and find the probability of being an outlier according to the 1.5·IQR criterion:

- First find the z-score of the lower/upper quartile. I.e. the value of z such that $\mu \pm z\sigma$ is the lower/upper quartile.

```
'lowe quartile is the first quartile Q1 (25%) and the upper quartile is the thir onde Q3(%75)'
```

```
## [1] "lowe quartile is the first quartile Q1 (25%) and the upper quartile is the thir onde Q3(%75)"
```

```
z1<-qdist("norm", 0.75)
```




```
'Z will be the same for a normal distribution (mu=0, se=1) for the upper Q3 and lower Q1 quartil'
```

```
## [1] "Z will be the same for a normal distribution (mu=0, se=1) for the upper Q3 and lower Q1 quartil
```

- Use this to find the IQR (expressed in terms of σ).

```
'inter quartil range=coeficiente de los quartiles: En estadística descriptiva, se le llama rango interc
```

```
## [1] "inter quartil range=coeficiente de los quartiles: En estadística descriptiva, se le llama rango
```

```
mu=0
sigma=1
IQR=(mu+sigma*z1)-(mu-sigma*z1)
IQR
```

```
## [1] 1.349
```

```
'si estas 1.5 veces el IQR significa q estas fuera de tu rango '
```

```
## [1] "si estas 1.5 veces el IQR significa q estas fuera de tu rango "
```

- Now find the z-score of the maximal extent of the whisker. I.e. the value of z such that $\mu \pm z\sigma$ is the endpoint of lower/upper whisker.

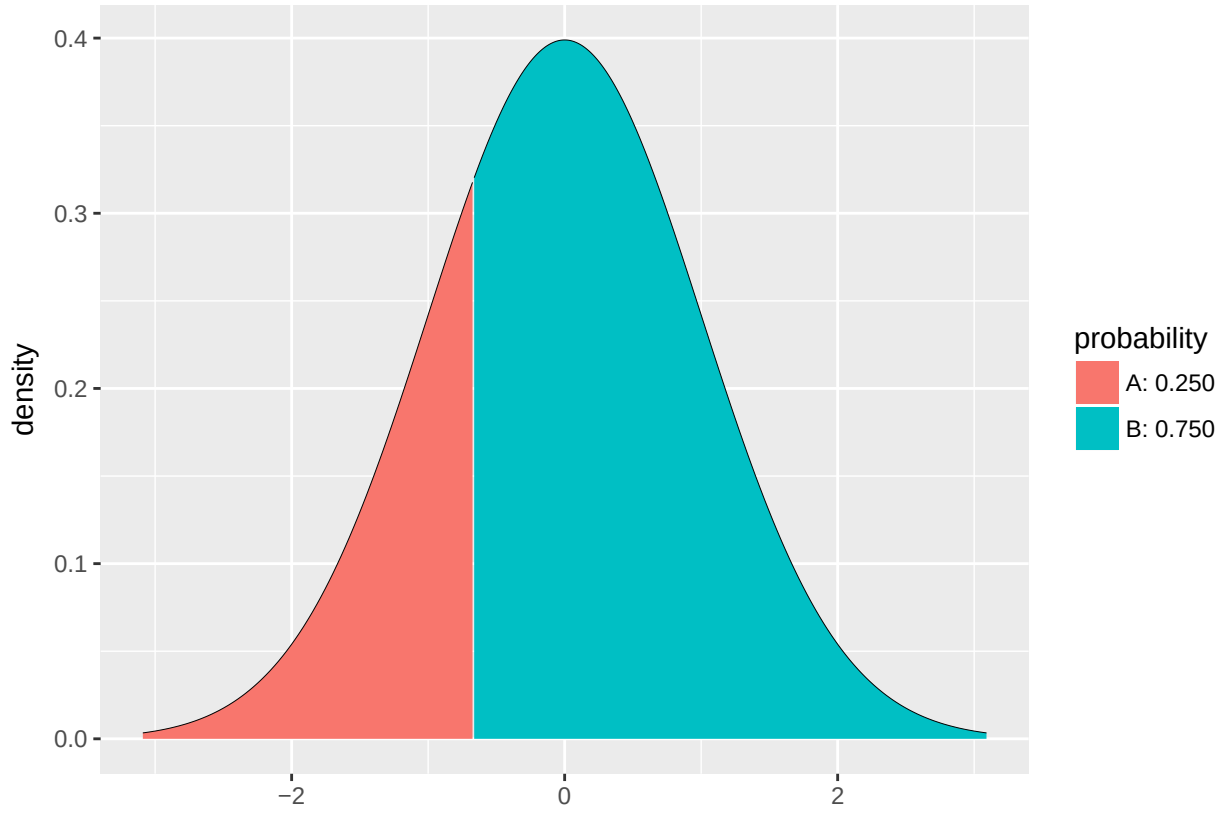
```
mu=0
sigma=1
IQR=(mu+sigma*z1)-(mu-sigma*z1)
```

- Find the probability of being an outlier.

```
'la poabilidad de estar en los limites. Asiq vas a tneer una probabilidad q es el area (%2.2 mas o menos
```

```
## [1] "la poabilidad de estar en los limites. Asiq vas a tneer una probabilidad q es el area (%2.2 mas
```

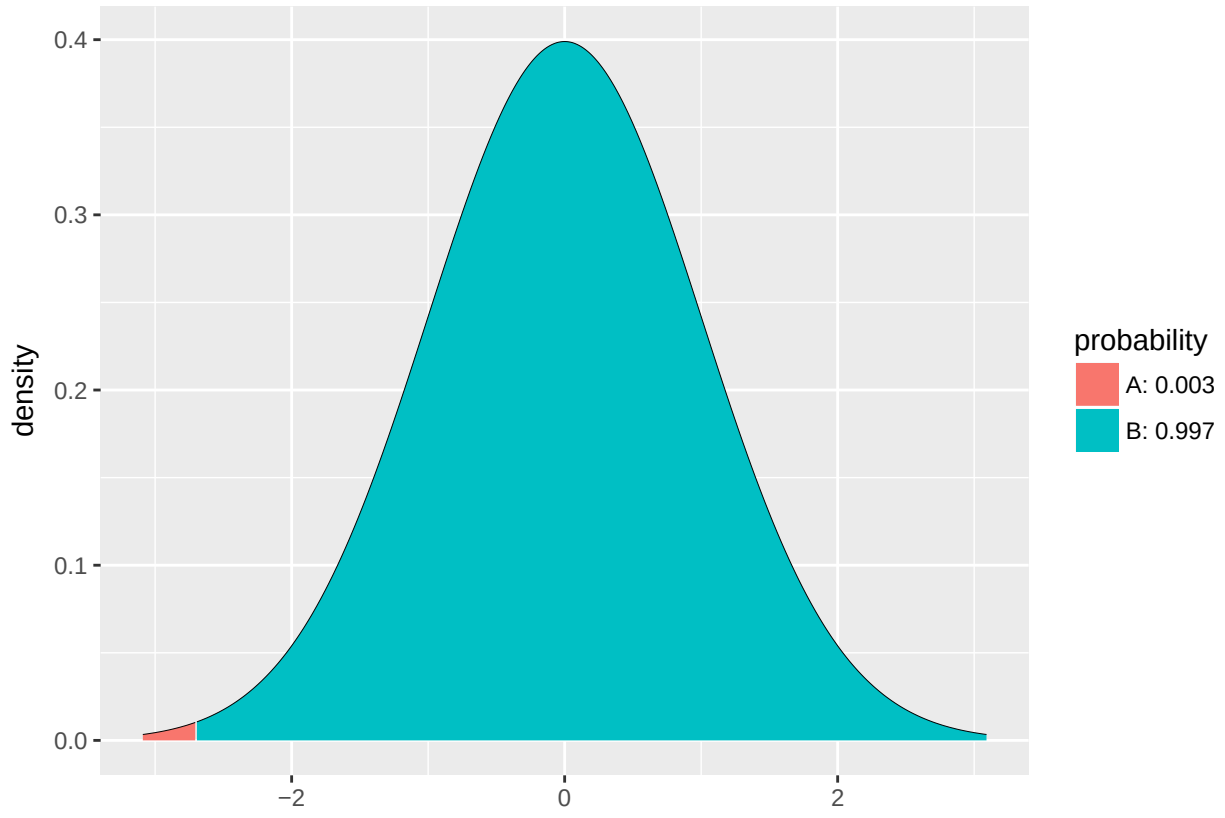
```
z1<-qdist("norm", 0.25)
```




```
z1
```

```
## [1] -0.6745
```

```
outlier<-2*pdist("norm", z1-1.5*IQR)
```




```
outlier
```

```
## [1] 0.006977
```