Exam exercise: Probalilities, CLT and boxplots

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)
options(digits = 4)
```

EU climate data

In a recent survey from Eurobarometer you can extract data for response to the following question:

Do you consider climate change to be the single most serious problem facing the world as a whole?

The data are divided according to whether the respondent comes from Denmark or not.

```
climate <- as.table(matrix(c(309,4918,1010,25830),2))
dimnames(climate) <- list(origin=c("Denmark","Rest of EU"),answer=c("Yes","No"))
climate
```

```
##             answer
## origin         Yes    No
##   Denmark      309  1010
##   Rest of EU  4918 25830
```

- Estimate the probability of answering "Yes" to the question.
- Make a 95% confidence interval for the probability of answering "Yes".
- Estimate the probability of answering "Yes" given that you come from Denmark.
- What would the true population probabilities satisfy if `origin` and `answer` were statistically independent? Based on your results do you think they are independent?

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("https://asta.math.aau.dk/datasets?file=Home.RData"))
```

Make a histogram of all the house prices using a command like `gf_histogram(~price, bins = 30)` inserted in a new code chunk (try to do experiments with the number of bins):

- Explain how a histogram is constructed.

- Does this histogram look like a normal distribution?

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

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`:

- What is the expected value of `y_bar`?
- What is the standard deviation of `y_bar` (also called the standard error)?
- What is the approximate distribution of `y_bar`?

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

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

```
## [1] 1.746
```

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

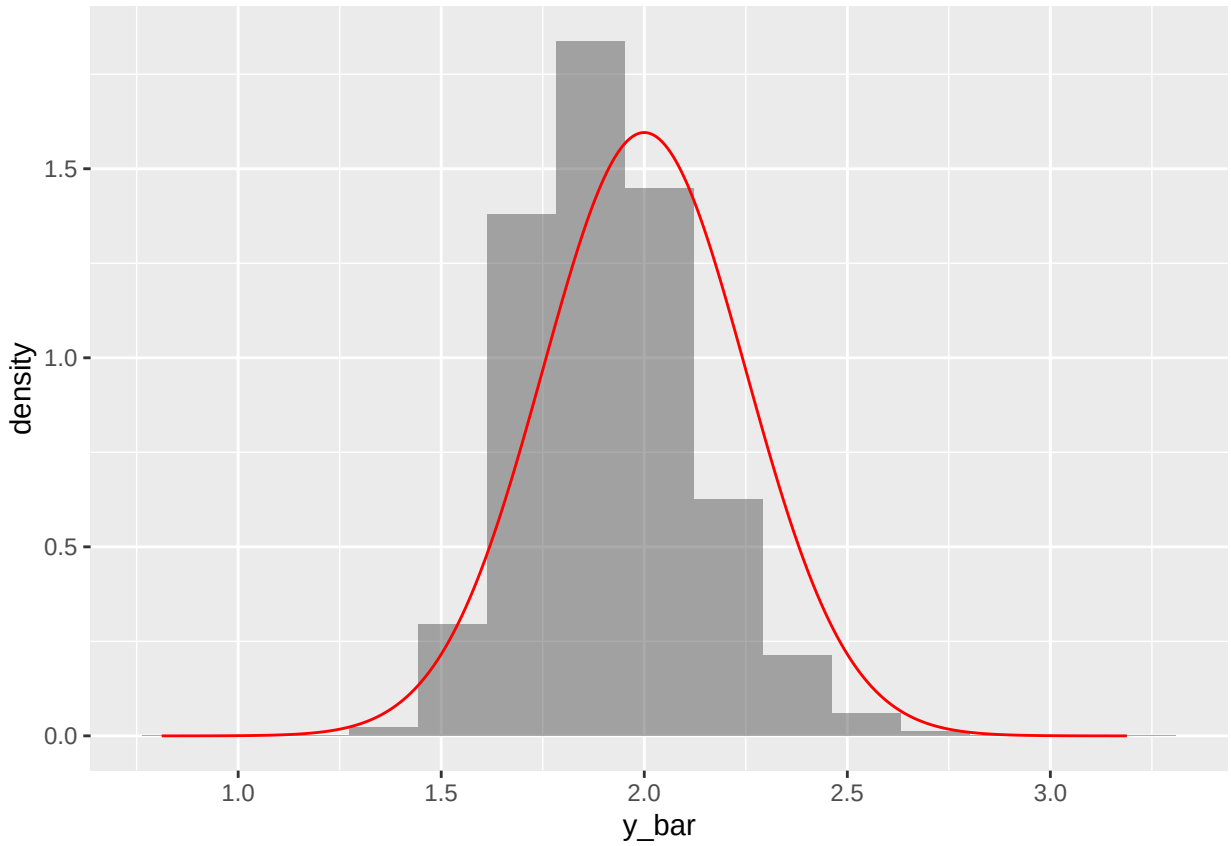
Use `replicate` to repeat the sampling 500 times and save each mean value in the vector `y_bar`:

```
y_bar <- replicate(500, mean(sample(price, 40)))
```

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

- 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 `gf_dist`. For example if you predicted a normal distribution with mean 2 and standard deviation 0.25:

```
gf_dhistogram( ~ y_bar, bins = 15) %>%
gf_dist("norm", mean = 2, sd = 0.25, col = "red")
```



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

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.
- Use this to find the IQR (expressed in terms of σ).
- 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.
- Find the probability of being an outlier.