MS-C1620 Statistical Inference

Exercise 1

Class exercise

Note: the internet is full of good tutorials for learning R:

- *In case you want to brush up on your R skills outside this course, good starting points can be found e.g. on RStudio's online learning page.*
- *Problems related to the use of particular functions are often most efficiently solved by searching the R section of Stack overflow where someone else has with high probability encountered the same issue before.*
- *The internal documentation of R can also sometimes be helpful. To reach it, use either the help tab on the lower right corner of RStudio or directly ask for the help page of a particular function using e.g.* `?rnorm` *or* `help(rnorm)`.

---

1. a. Visit the website https://data.oecd.org/ and pick three data sets that you find interesting. What kind of tools are used to summarize/plot the data? Are the summaries/plots clear and easy to read?
   b. Search online for statistics about the income distribution in different countries. How are the data typically summarized/plotted? Are the general trends and patterns easy to spot from the used summaries/plots?

---

2. a. Generate a sample of 100 observations from the standard normal distribution and save it as the vector `x`
   b. Calculate the sample mean and sample standard deviation of `x` using the functions `mean` and `sd`.
   c. Find (or code!) a function that will compute the sample variance of a vector of values.
   d. Generate three samples from a normal distribution with expected value 1 and standard deviation 3, one with 10 observations, one with 100 and one with 1000.
   e. Compute the sample means and sample standard deviations of the three samples in part *d*. How do the statistics behave when the sample size is increased? What causes this?

```
# a.
x <- rnorm(100) # Or, more explicitly, x <- rnorm(100, 0, 1)

# b.
mean(x); sd(x)
```

```
## [1] -0.2536933
```

```
## [1] 1.121438
```

```
# c.
var(x)
```

```
## [1] 1.257622
```

```
# d.
x_10 <- rnorm(10, 1, 3) # Note that rnorm takes standard deviation, not variance, as its argument
x_100 <- rnorm(100, 1, 3)
x_1000 <- rnorm(1000, 1, 3)

# e.
mean(x_10); sd(x_10)
```

```
## [1] 1.45416
```

```
## [1] 3.299392
```

```
mean(x_100); sd(x_100)
```

```
## [1] 1.482397
```

```
## [1] 2.79772
```

```
mean(x_1000); sd(x_1000)
```

```
## [1] 0.8295466
```

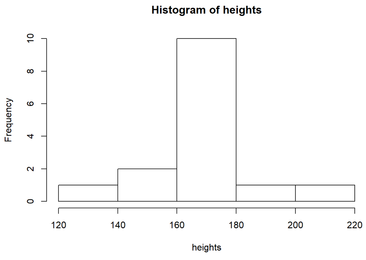
```
## [1] 2.998261
```

```
# With increasing sample size, the statistics converge to their population counterparts by the law of large numbers.
```

---

3. a. Collect together a random sample, 10-15 observations, of the *heights* of the students in the class and save the data as a vector.
   b. Explore how the heights of the students are distributed by finding out how the function `hist` works and drawing a histogram of the heights.
   c. How do you think the histogram would change if we had sampled the whole classroom? How about if we had sampled 2000 random students from the Aalto university?
   d. Find out how to change the number of bins in the histogram and experiment with it to see how the plot changes.

```
# a.
heights <- rnorm(15, 170, 15) # Fill in here.

# b.
hist(heights)

# c.
# The histogram would most likely approach the true distribution of all students in Aalto university. Human height is often
well approximated by the normal distribution, and assuming that this holds for the student population of Aalto University, a
t least in the 2000 student sample the resulting histogram would most likely look like something resembling a normal distrib
ution. However, even if this holds, the histograms of the classroom samples can still look quite non-normal as the sample si
zes are small (and the classroom sample might not be representative of the whole Aalto student population).

# d.
# The number of bins (breakpoints) is controlled by the argument "breaks"
hist(heights, breaks = 3)
```

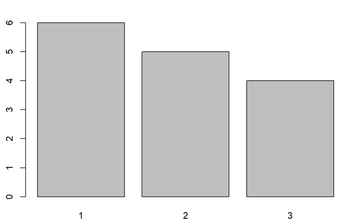



```
# For small sample sizes, the histograms can change quite a lot when the number of bins is varied.
```

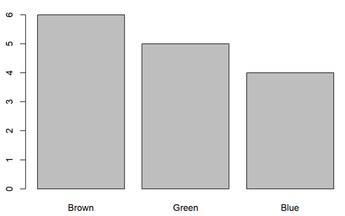
---

4. a. Collect together a random sample, 10-15 observations, of the *eye colors* of the students in the class and save the data as a vector. Code the different colors either using different numbers (for example, blue is `1` etc.) or, if you know how, using the `factor` class in R.
   b. Find out which function draws a bar chart and use it to plot the data. The function `table` might also prove useful.
   c. How do you think the bar chart would change if we had sampled the whole classroom? How about if we had sampled 2000 random students from the Aalto university?

```
# a.
eyecolors <- sample(1:3, 15, replace = TRUE) # Fill in here.
eyecolors_2 <- factor(eyecolors, labels = c("Brown", "Green", "Blue"))

# b.
barplot(table(eyecolors))
```




```
barplot(table(eyecolors_2))
```




```
# c.
# See answer c. in the previous problem. Here the true distribution is a discrete distribution of all possible eye colors en
countered amongst Aalto students. Also, sampling a larger population could reveal some rare eye colors that are not found in
the sample, thus adding more bars to the chart.
```

---

5. **(Optional)** Install both R and RStudio on your personal computer and use them to experiment with the lecture and exercise codes throughout the course. If you have troubles in the installation, you can bring your laptop with you to the next exercise session and ask the course assistant to help you.

---

6. **(Optional)** The RStudio website is a home to numerous useful cheat sheets which list the key commands of various packages and tasks (plotting, data import etc.) Check them out, paying attention especially to the "RStudio IDE Cheat Sheet".

---

7. **(Optional)** If you prefer learning R hands-on, check out the R-package `swirl`, a real-time tutorial of R inside R.

```
install.packages("swirl")
library(swirl)
swirl()
```

---

8. **(Optional)** This exercise sheet was created using R Markdown. Try it out yourself by choosing "File -> New File -> R Markdown..." in RStudio. Try "knitting" the document into a .html of .pdf file by pressing the Knit button in the toolbar. Use the R Markdown Cheat Sheet to experiment with different formatting in your R Markdown document.

MS-C1620 Statistical Inference

Exercise 2

Homework exercise

To be solved at home before the exercise session.

1. Visit the website https://dataviz-project.com/ and pick one data visualization/plot that interests you. Find out how it is drawn and what aspects of the data the different components represent. Be prepared to explain how your visualization of choice works in the class.
2. Type the command `data()` in **R** to show all data sets currently available in your installed packages. Go through the data sets and pick one that interests you. Check the help file of the data set using the command `?packagename` for more detailed information. Be prepared to describe your answers to the following questions in the class:
    - What is the purpose of the data? What kind of phenomenon does it describe?
    - What kind of study is behind the data (observational, controlled, simulation, survey or something else)?
    - How is the data represented in R (univariate, multivariate, time series...)?
    - What kind of plots would you use to best summarize the data?
    - What kind of numerical statistics would you use to best summarize the data?

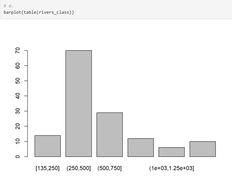
Class exercise

To be solved at the exercise session.

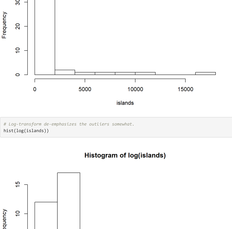
Note: all the needed data sets are available in base R

1. The data set `rivers` contains the lengths of 141 major rivers in North America.
    a. Find a suitable way to visualize the data and plot it.
    b. How are the lengths distributed based on your plot?
    c. Discretize the lengths into six classes: [min, 250), [250, 500), (500, 750], (750, 1000], (1000, 1250], (1250, max]. The function `cut` may prove helpful.
    d. Find a suitable way to visualize the discretized data and plot it.
    e. Which of the two visualizations is more informative?

```
# a.
boxplot(rivers)
```



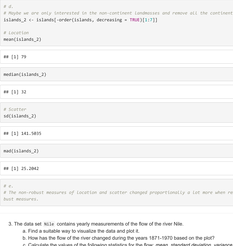

```
# b.
# Distribution seems strongly skewed to the right.

# c.
rivers_class <- cut(rivers, breaks = c(min(rivers), 250, 500, 750, 1000, 1250, max(rivers)), include.lowest = TRUE)

# d.
barplot(table(rivers_class))
```



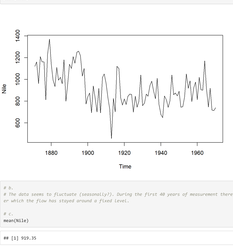

```
# e.
# Depends on the purpose but the barplot at least shows the main bulk of the data more clearly.
```

2. The data set `islands` contains the areas of all landmasses in the world which exceed a certain threshold.
    a. Find a suitable way to visualize the data and plot it.
    b. How are the landmasses distributed based on your plot?
    c. Compute both robust and non-robust measures of location and scatter for the data.
    d. Remove some of the outliers (and think of a possible reason for justifying that) from the data and compute the same measures as in part c.
    e. Compare the results of part c and part d.

```
# a.
# Not very informative due to the outliers.
# Maybe some better options exist...
hist(islands)
```




```
# Log-transform de-emphasizes the outliers somewhat.
hist(log(islands))
```




```
# b.
# The distribution is strongly skewed to the right.

# c.
# Location
mean(islands)
```

```
## [1] 1252.729
```

```
median(islands)
```

```
## [1] 41
```

```
# Scatter
sd(islands)
```

```
## [1] 3371.146
```

```
mad(islands)
```

```
## [1] 39.2889
```

```
# d.
# Maybe we are only interested in the non-continent landmasses and remove all the continents
islands_2 <- islands[-order(islands, decreasing = TRUE)[1:7]]

# Location
mean(islands_2)
```

```
## [1] 79
```

```
median(islands_2)
```

```
## [1] 32
```

```
# Scatter
sd(islands_2)
```

```
## [1] 141.5835
```

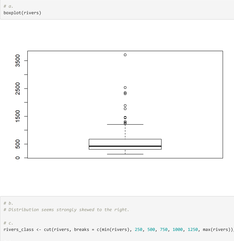
```
mad(islands_2)
```

```
## [1] 25.2042
```

```
# e.
# The non-robust measures of location and scatter changed proportionally a lot more when removing the "outliers" than the ro
bust measures.
```

3. The data set `Nile` contains yearly measurements of the flow of the river Nile.
    a. Find a suitable way to visualize the data and plot it.
    b. How has the flow of the river changed during the years 1871-1970 based on the plot?
    c. Calculate the values of the following statistics for the flow: *mean, standard deviation, variance, minimum, maximum, median, median absolute deviation, mode, skewness and kurtosis.*
    d. How are each of the statistics in part c visible in the plot of part a?

```
# a.
# The data are time series and thus time should be a part of the plot
# "Nile" is already saved as a time series object in R and the plain "plot" command produces a plot of flow vs. time.
plot(Nile)
```




```
# b.
# The data seems to fluctuate (seasonally?). During the first 40 years of measurement there was a general downward trend aft
er which the flow has stayed around a fixed level.

# c.
mean(Nile)
```

```
## [1] 919.35
```

```
sd(Nile)
```

```
## [1] 169.2275
```

```
var(Nile)
```

```
## [1] 28637.95
```

```
min(Nile)
```

```
## [1] 456
```

```
max(Nile)
```

```
## [1] 1370
```

```
median(Nile)
```

```
## [1] 893.5
```

```
mad(Nile)
```

```
## [1] 179.3946
```

```
# All four first values below are modes
sort(table(Nile), decreasing = TRUE)
```

```
## Nile
##  845 1020 1100 1140  744  874 1040 1050 1120 1140 1210  456  649  676  692
##    3    3    3    3    2    2    2    2    2    2    2    1    1    1    1
##  694  698  701  702  714  718  726  740  742  746  749  759  764  768  771
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
##  774  781  796  797  799  801  812  813  815  821  822  824  831  832  833
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
##  838  840  846  848  860  862  864  865  890  897  901  906  912  916  918
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
##  919  923  935  940  944  958  960  963  969  975  984  986  994  995 1010
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
## 1050 1120 1150 1170 1180 1220 1230 1250 1260 1370
##    3    1    1    1    1    1    1    1    1    1
```

```
library(moments)
skewness(Nile)
```

```
## [1] 0.3213697
```

```
kurtosis(Nile)
```

```
## [1] 2.695093
```

```
# d.
# Mean = the average level around which the data fluctuates
# SD & Var = the average size of these fluctuations
# Min = the lowest point of the curve
# Max = the lowest point of the curve
# Median = the average level around which the data fluctuates (robust)
# MAD = the average size of these fluctuations (robust)
# (Mode = difficult to see...)
# (Skewness = difficult to see...)
# (Kurtosis = difficult to see...)
```

4. **(Optional)** Try out the 3d-visualization tools in the package `rgl`. The following code plots an interactive 3d-scatter plot of the first three variables in the `iris` data. Find out how you can colour the points in the plot according to the variable `Species`.

```
install.packages("rgl")
library(rgl)

# Opens in a new window
open3d()
plot3d(iris[, 1:3])
```

5. **(Optional)** Pretty plots are often cumbersome to produce with base **R** and numerous packages offer various more attractive approaches. Try out the package `ggplot2` by running the following code.

```
install.packages("ggplot2")
library(ggplot2)

ggplot(data = mpg, aes(x = hwy, y = cty)) +
  geom_point() +
  labs(x = "Highway miles per gallon", y = "City miles per gallon")
```

What does the plot represent? Experiment with the code and find out how you can color the points according to the `class` of the car.

Numerous tutorials about `ggplot2` can be found online. Check out at least https://r4ds.had.co.nz/data-visualisation.html and http://r-statistics.co/ggplot2-Tutorial-With-R.html.

MS-C1620 Statistical Inference

Exercise 3

Homework exercise

To be solved at home before the exercise session.

1. Consider the confidence interval for the expected value of the normal distribution on page 2.9 of the lecture notes. Describe what will (most likely) happen to the width of the confidence interval (does it get smaller, larger or stay the same?) if we.
   a. Increase the sample size n.
   b. Decrease the confidence level $100(1-\alpha)$.
   c. Increase the variance σ^2.
   d. Decrease the expected value μ.

```
# a.
# The width of the interval is proportional to S1/sqrt(n)S, and thus increasing $n$ will make the interval shorter.
# Heuristic explanation: larger sample = more information = larger precision = shorter interval.

# b.
# Decreasing $100(1 -alpha)$ (increasing alpha) will decrease the corresponding quantile of the t-distribution (we move clo
ser to the center from the tail). As the width of the interval is proportional to the quantile, decreasing the confidence le
vel will make the interval shorter.
# Heuristic explanation: Lower confidence level = we are satisfied with smaller probability to capture the true value = shor
ter interval.

# c.
# Increasing the population variance will most likely also increase the sample variance $s^2$, and consequently the sample s
tandard deviation. As the width of the interval is proportional to $s$, increasing the variance will make the interval wide
r.
# Heuristic explanation: larger variance = data is less accurate = capturing the true value is more difficult = need larger
interval

# d. Increasing the expected value will only affect the location of the interval in the x-axis. The width will stay the sam
e.
```

2. Consider the following four hypothesis testing scenarios. For each scenario, describe what the Type I error and Type II error mean in that particular context. Comment also on the possible consequences of the two errors in each case (which one of the errors is more "dangerous"?). For part d. come up with a typical hypothesis testing scenario from your own field of science.
   a. A suspect is brought before a judge.
      - H0: The suspect is innocent.
      - H1: The suspect is guilty.
   b. A new experimental cancer treatment is compared to placebo.
      - H0: The new treatment is no better than placebo.
      - H1: The new treatment is better than placebo.
   c. An automated security screening scans passengers at the airport.
      - H0: The passenger is not carrying dangerous items.
      - H1: The passenger is carrying dangerous items.
   d. Your own scenario here!

```
# a.
# Type I: Innocent person goes to jail. <- typically thought as more "dangerous"
# Type II: Guilty person walks free.

# b.
# Type I: New treatment does not work but we still start giving it to patients (as we are under the impression that it work
s)
# Type II: The new treatment works but we do not realize it, and consequently it never reaches the market.
# It's difficult to say which one is more dangerous, both can possibly lead to fatalities.

# c.
# Type I: A passenger not carrying dangerous items is taken to further inspection.
# Type II: A passenger carrying dangerous items goes through unnoticed. <- More dangerous.
```

Class exercise

To be solved at the exercise session.

Note: all the needed data sets are either given below or available in base R.

1. The data set `iris` contains measurements of the sepal length and width and petal length and width for 50 flowers from each of 3 species of iris. We want to study the distribution of the ratio between sepal length and sepal width of an iris of the species `setosa`.
   a. Create a new 1-dimensional data set which contains only the ratios `Sepal.Length/Sepal.Width` for the irises of the species `setosa`.
   b. Find a suitable way to visualize the ratio.
   c. Use bootstrapping to construct a 95% confidence interval for the expected value of the ratio.
   d. Add the confidence interval end points to the plot of part b.
   e. What does the confidence interval tell us about the distribution of the ratio?
   f. What assumptions did the confidence interval in part c make?

```
# a.
iris_temp <- iris[iris$Species == "setosa", ]
iris_ratio <- iris_temp[, 1]/iris_temp[, 2]

# b.
hist(iris_ratio)

# c.
boots <- 10000

boot_res <- replicate(boots, mean(sample(iris_ratio, length(iris_ratio), replace = TRUE)))
boot_ci <- quantile(boot_res, probs = c(0.025, 0.975))

# d.
hist(iris_ratio)
abline(v = boot_ci, col = 2, lwd = 3)
```


```
# e.
# The expected value of the ratio is (in a certain sense) likely to be in the interval given by boot_ci.

# f.
# That the 50 setosas constitute an i.i.d. sample from some population of irises of species setosa. No assumption regarding
any distributions were made.
```

2. The data set below contains the annual salaries (in dollars) of 8 American women and 8 American men. The observations are paired such that each woman is matched with a man having similar background (age, occupation, level of education, etc). We are interested in studying whether the expected values of the salaries of women and men differ.
   a. Find a suitable way to visualize the data.
   b. Which test is appropriate in studying our question of interest?
   c. State the hypotheses of your chosen test and conduct it on the significance level 10%.
   d. What is the conclusion of the test?
   e. What assumptions did the test in part c make? Are they justifiable?

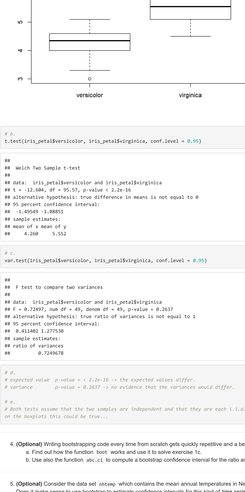
```
salary <- data.frame(women = c(42600, 43600, 49300, 42300, 46200, 45900, 47500, 41300),
                     men = c(46200, 44700, 48400, 41700, 48600, 49300, 48300, 44300))
```

```
# a.
salary <- data.frame(women = c(42600, 43600, 49300, 42300, 46200, 45900, 47500, 41300),
                     men = c(46200, 44700, 48400, 41700, 48600, 49300, 48300, 44300))

boxplot(salary)
```


```
# b.
# The data is paired (and the pairs are not independent), making paired t-test a good choice.
# Note that using the two-sample t-test is not justified as it assumes the independence of the two samples.

# c.
# H0: \mu_women <= \mu_men (or \mu_women - \mu_men <= 0)
# H1: \mu_women > \mu_men (or \mu_women - \mu_men > 0)
t.test(salary$women - salary$men, mu = 0, conf.level = 0.9)
```

```
##
##  One Sample t-test
##
## data:  salary$women - salary$men
## t = -2.5632, df = 7, p-value = 0.03738
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  -2782.6221  -417.3779
## sample estimates:
## mean of x
##     -1600
```

```
# d.
# The p-value equals 0.037 -> we reject the null hypothesis, there is a difference in the expected values of the salaries.

# e.
# We assumed the normality of the salary difference. The distribution of the data looks to be multimodal, so the assumption m
ight not be justified. However, the sample size is small and the pattern could be caused by randomness. In any case, better
alternatives are given by the non-parametric methods studied next week.
plot(salary$women - salary$men, rep(1, 8))
```



3. Consider again the `iris` data set from exercise 1. Study whether the expected values and variances of `Petal.Length` differ between irises of the species `versicolor` and `virginica`
   a. Find a suitable way to visualize the data.
   b. Test whether the expected values differ using the two-sample t-test on a significance level 5%.
   c. Test whether the variances differ using the variance comparison test on a significance level 5%.
   d. What are the conclusions of the tests?
   e. What assumptions did the tests in parts b and c make? Are they justifiable?

```
# a.
iris_temp <- iris[iris$Species != "setosa", c(3, 5)]
iris_petal <- data.frame(versicolor = iris_temp[iris_temp[, 2] == "versicolor", 1],
                         virginica = iris_temp[iris_temp[, 2] == "virginica", 1])

boxplot(iris_petal)
```



```
# b.
t.test(iris_petal$versicolor, iris_petal$virginica, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  iris_petal$versicolor and iris_petal$virginica
## t = -12.604, df = 95.57, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.49549 -1.08851
## sample estimates:
## mean of x mean of y
##     4.260     5.552
```

```
# c.
var.test(iris_petal$versicolor, iris_petal$virginica, conf.level = 0.95)
```

```
##
##  F test to compare two variances
##
## data:  iris_petal$versicolor and iris_petal$virginica
## F = 0.72497, num df = 49, denom df = 49, p-value = 0.2637
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.411482 1.277530
## sample estimates:
## ratio of variances
##          0.7249678
```

```
# d.
# expected value  p-value = < 2.2e-16 -> the expected values differ.
# variance        p-value = 0.2637 -> no evidence that the variances would differ.

# e.
# Both tests assume that the two samples are independent and that they are each i.i.d. from some normal distribution. Based
on the boxplots this could be true...
```

4. **(Optional)** Writing bootstrapping code every time from scratch gets quickly repetitive and a better idea is to use the package `boot`.
   a. Find out how the function `boot` works and use it to solve exercise 1c.
   b. Use also the function `abc.ci` to compute a bootstrap confidence interval for the ratio and compare the results.

5. **(Optional)** Consider the data set `nhtemp` which contains the mean annual temperatures in New Haven, Connecticut, from 1912 to 1971. Does it make sense to use bootstrap to estimate confidence intervals for this kind of time series data? (Hint: are the bootstrap samples similar to the original sample in a meaningful way?)

6. **(Optional)** Let $x_1, \ldots, x_{100}$ be a random sample from the exponential distribution with the unknown rate parameter λ. We test the hypotheses,
   - H0: $\lambda = 1$,
   - H1: $\lambda \neq 1$,

   using $t = \left(\hat{\mu}\right)^{-1}$ as a test statistic. Using simulations,
   a. find an approximate 95% critical region for the test,
   b. find the approximate Type II error probability when the true value of the parameter is $\lambda = 2$ and we use the critical region from part a.

MS-C1620 Statistical Inference

Exercise 4

Homework exercise

To be solved at home before the exercise session.

1. a. Let $x_1, \ldots, x_n$ be a random sample (iid) from some distribution F_θ with the unknown parameter θ. Which of the three one-sample tests (*t*-test, sign test or signed rank test) would you use (and why!) to test whether the location (expected value/median) of the data is equal to 1 if we know for certain that the distribution F_θ is
   i. an exponential distribution with unknown rate parameter θ.
   ii. a normal distribution with variance 2 and unknown expected value θ.
   iii. a Laplace distribution with known scale parameter 5 and unknown location parameter θ.
   iv. a Poisson distribution with unknown parameter θ?

```
# In each case we want to use a test which makes the strictest assumptions (such that they are still satisfied). This gives us maximal power (lowest Type 2 error rate), as we "use more information" about the data. See the lecture examples of week 3.
# The assumptions the three tests make besides iid data are:
# t-test: normality                          <- strictest
# signed rank test: symmetric continuous distribution    <- less strict
# sign test: continuous distribution                     <- even less strict

# a.
# Exponential distributions are not symmetric -> sign test.

# b.
# t-test

# c.
# Laplace distributions are not normal but they are symmetric -> signed rank test.

# d.
# Poisson distribution is neither continuous nor symmetric so, being strict, none of the tests apply. However, sign test is regularly applied to discrete data as well (e.g. using the conventions of slide 3.7).
```

b. The data set `airmiles` lists the passenger miles flown by commercial airlines in the United States for each year from 19 37 to 1960. To inspect whether the yearly passenger miles equal 10000 on average, a researcher performed a sign test to test the null hypothesis $\text{Med}_\theta = 10000$ on significance level 5% with the results shown below and concluded that there is no evi dence against the null hypothesis. Do you agree with the researcher's conclusion?

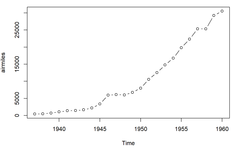
```
airmiles
```

```
## Time Series:
## Start = 1937
## End = 1960
## Frequency = 1
##  [1]  412  480  683 1052 1385 1418 1634 2178 3362 5948 6389
## [12]  5981 6753 8003 10566 12528 14760 16769 19819 22362 25340 25343
## [23] 29269 30514
```

```
# Sign test
binom.test(sum(airmiles > 10000), length(airmiles))
```

```
##
##  Exact binomial test
##
## data:  sum(airmiles > 10000) and length(airmiles)
## number of successes = 10, number of trials = 24, p-value = 0.5413
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.2218068 0.6335694
## sample estimates:
## probability of success
##              0.4166667
```

```
# The sign test assumes that the observations x1, x2, ..., xn form an iid. sample from some particular continuous distribution, while it could be plausible to view the yearly passenger miles as realisations of identically distributed random variab. as from a continuous distribution, they are certainly not independent. If the passenger miles go up one year, it is likely t hat they continue going up in the coming years as well (the technology develops etc.). This can be seen in the time series p lot of the data:
plot(airmiles, type = "b")
```




```
# In this kind of situation methods of time series analysis are needed (not covered in this course).
```

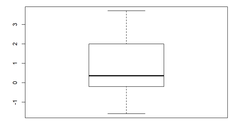
Class exercise

To be solved at the exercise session.

Note: all the needed data sets are either given below or available in base R.

1. The data set `sleep` shows the effect of two soporific drugs (increase in hours of sleep compared to control) on 10 patients. We are interested in studying whether drug 1 helps in increasing the number of hours slept compared to placebo.
   a. Extract the increases in hours of sleep of the patients who received drug 1 (`group == 1`).
   b. Visualize the data.
   c. Conduct an appropriate test to evaluate whether the location (expected value/median) of the increase in hours of sleep differs from 0 on significance level 5%.
   d. Draw conclusions.

```
# a.
sleep_1 <- sleep[sleep$group == 1, 1]

# b.
boxplot(sleep_1)
```




```
# c.
# With so few observations it is difficult to say whether the data comes from a normal, or even a symmetric, distribution. S ign test seems like the safest choice.
binom.test(sum(sleep_1 > 0), length(sleep_1))
```

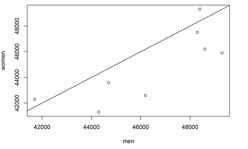
```
##
##  Exact binomial test
##
## data:  sum(sleep_1 > 0) and length(sleep_1)
## number of successes = 7, number of trials = 10, p-value = 1
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.1878865 0.8129314
## sample estimates:
## probability of success
##                    0.5
```

```
# p-value = 1

# The highest possible p-value -> no evidence against H0 -> the drug 1 is no better than placebo.
```

2. The data set below contains the annual salaries (in dollars) of 8 American women and 8 American men (recall exercise 3.2). The observations are paired such that each woman is matched with a man having similar background (age, occupation, level of education, etc). We are interested in studying whether the locations of the salaries of women and men differ (recall that last time paired *t*-test concluded that the salaries differ).
   a. Begin again by visualizing the data.
   b. Which two non-parametric tests are appropriate in studying our question of interest?
   c. State the hypotheses of the tests and conduct them on the significance level 10%.
   d. What are the conclusions of the tests?
   e. What assumptions did the test in part c make? Are they justifiable?

```
salary <- data.frame(women = c(42600, 43600, 45300, 42300, 46200, 45900, 47500, 41300),
                     men = c(46200, 44700, 48400, 41700, 48600, 49300, 48300, 44300))
```

```
# a.
salary <- data.frame(women = c(42600, 43600, 45300, 42300, 46200, 45900, 47500, 41300),
                     men = c(46200, 44700, 48400, 41700, 48600, 49300, 48300, 44300))

# Alternative visualization to the last time
plot(women ~ men, data = salary)
abline(a = 0, b = 1)
```



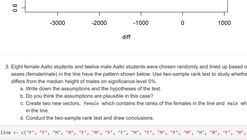

```
# Most points are below the y=x -line, meaning that the salary of the man in a pair is more often larger than that of the wo man

# b.
# The data is paired (and the pairs are not independent), making paired sign test and paired signed rank test appropriate ch oices.
# Note that using a two-sample rank test is not justified as it assumes the independence of the two samples.

# c.
# Both tests have the same hypotheses
# H0: med_women - med_men = 0
# H1: med_women - med_men != 0
diff <- salary$women - salary$men

# Paired sign test
binom.test(sum(diff > 0), length(diff))
```

```
##
##  Exact binomial test
##
## data:  sum(diff > 0) and length(diff)
## number of successes = 2, number of trials = 8, p-value = 0.2891
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.03185403 0.65085179
## sample estimates:
## probability of success
##                   0.25
```

```
# Paired signed rank test
wilcox.test(diff, mu = 0)
```

```
##
##  Wilcoxon signed rank test
##
## data:  diff
## V = 4, p-value = 0.05469
## alternative hypothesis: true location is not equal to 0
```

```
# d.
# Paired sign test does not reject the null but the paired signed rank test does.

# e.
# Both tests assume that the differences d1, d2, ..., d8 form an iid. sample from some particular continuous distribution. P aired signed rank test furthermore assumes that this distribution is symmetric. It is difficult to say whether the underlyin g distribution is symmetric based on a so small sample so the paired sign test might be a better choice -> no difference in value of p.
plot(diff, rep(1, 8))
```



3. Eight female Aalto students and twelve male Aalto students were chosen randomly and lined up based on their heights (shortest first). The sexes (female/male) in the line have the pattern shown below. Use two-sample rank test to study whether the median height of females differs from the median height of males on significance level 5%.
   a. Write down the assumptions and the hypotheses of the test.
   b. Do you think the assumptions are plausible in this case?
   c. Create two new vectors, `female` which contains the ranks of the females in the line and `male` which contains the ranks of the males in the line.
   d. Conduct the two-sample rank test and draw conclusions.

```
line <- c("F", "F", "M", "M", "F", "M", "F", "F", "M", "F", "M", "M", "M", "F", "M", "M", "M")
```

```
# a.
# The test assumes that the female and male samples are mutually independent iid samples from the continuous distributions F x and Fy, respectively. Moreover, the distributions Fx and Fy are assumed to be equal up to location shift ("same-shaped hi st.").

# The null hypothesis is that the medians of Fx and Fy (and consequently the distributions itself) are equal (and the altern ative hypothesis is the opposite of that).

# b.
# As the samples where chosen randomly, it is plausible that the samples are independent and iid. Also, googling "female vs. male height distribution" shows that the male distribution of heights is slightly wider than for females. We assume that thi s difference in scales is small enough that the test can still be used.

# c.
line <- c("F", "F", "M", "M", "F", "M", "F", "F", "M", "F", "M", "M", "M", "F", "M", "M", "M")

female <- (1:20)[line == "F"]
male <- (1:20)[line == "M"]

# d.
wilcox.test(female, male)
```

```
##
##  Wilcoxon rank sum test
##
## data:  female and male
## W = 25, p-value = 0.0025
## alternative hypothesis: true location shift is not equal to 0
```

```
# p-value = 0.0025 -> not enough evidence to reject H0 on significance level 5% -> no difference in medians. As we "know" th at there should be a difference, then either the sample size was too small, the result was caused by randomness or the assum ptions weren't justified.
```

4. **(Optional)** Data manipulation using just functions in base R does not always produce the most readable code. The task in 1a. can be achieved more transparently using the package `dplyr` as follows.

```
# install.packages("dplyr")
library(dplyr)

sleep_1 <- sleep %>%
  filter(group == 1) %>%
  select(extra)
```

Find out how the package and the piping operator `%>%` work by going through an online tutorial.

MS-C1620 Statistical Inference

Exercise 5

Homework exercise

To be solved at home before the exercise session.

1. a. A simple sample size calculation can be performed for binary proportion confidence intervals as follows. We bound the standard deviation estimate from above as $\sqrt{\hat{p}(1-\hat{p})} \leq 0.5$ to obtain the conservative confidence interval,

$$\left(\hat{p} - z_{\alpha/2}\frac{0.5}{\sqrt{n}}, \hat{p} + z_{\alpha/2}\frac{0.5}{\sqrt{n}}\right).$$

   The half-width of a confidence interval is known as its *margin of error* and for the conservative confidence interval the margin of error does not depend on the proportion of "successes". Thus we can compute a universal sample size for which a certain desired margin of error is reached.

      i. Compute the required sample sizes to obtain the margins of error of 0.01, 0.02 and 0.03 for a 95% conservative confidence interval.

      ii. Study how much the calculations in part i over-estimate the required sample sizes when the proportion of successes is small $\hat{p} = 0.05$. That is, redo part i using the regular binary confidence interval in slide 4.6.

```
# i. The half-width of the conservative 95% interval is 1.96*0.5/sqrt(n). This equals 0.01*a if
# 1.96*0.5/sqrt(n) = 0.01*a <=> 1.96*0.5/(0.01*a) = sqrt(n) <=> n = 9604/a^2.
# That is, the required sample sizes are n = 9604, 2401, 1068.
#
# ii. The half-width of the standard 95% interval for that p = 0.05 is 1.96*sqrt(0.05*0.95)/sqrt(n). As in part i, we obtain
# n = 1824.76/a^2 and the true required sample sizes are 81% (= 1 - 1824.76/9604) smaller than those approximated in part i.
```

```
b. A manufacturer claims that only 6% of their products are faulty. To investigate this, a customer picks a random sample of
size $n$ of products and observes the proportion of faulty ones to be $\hat{p} = 0.09$. He tests the manufacturer's claim us
ing the asymptotic one-sample proportion test in slide 4.9. Is the p-value of the test smaller for sample size $n = 100$ or
$n = 200$?
```

```
# The 2-value of the test is proportional to the square root of the sample size n. Thus increasing the sample
# size increases the Z-value and consequently pushes it towards the tail of the distribution, decreasing the
# p-value. Thus the p-value for n = 200 is smaller
#
# An intuitive reasoning for the result is that the difference between 0.06 and 0.09 is "proportionally"
# larger for n = 200 than for n = 100 (as larger n implies increased accuracy) and as such also more
# deviating.
```

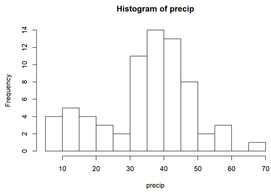
Class exercise

To be solved at the exercise session.

Note: all the needed data sets are either given below or available in base R.

1. The data set `precip` describes the average annual amounts of precipitation (rainfall) in inches for 70 United States (and Puerto Rico) cities. A city is said to be *dry* if its average annual rainfall is less than 20 inches. Treat the data as a random sample amongst all US cities and estimate a confidence interval for the proportion of dry cities in the US.
   a. Visualize the data.
   b. Create a new variable which takes the value `1` if the city is dry and `0` otherwise.
   c. Compute an approximate 95% confidence interval for the proportion of dry cities.
   d. What is the interpretation of the confidence interval in part c?

```
# a.
# Distribution seems to be bi-modal.
hist(precip, breaks = 10)
```




```
# b.
precip_dry <- 1*(precip < 20)
precip_dry
```

```
##             Mobile             Juneau            Phoenix
##                  0                  0                  1
##        Little Rock        Los Angeles         Sacramento
##                  0                  1                  1
##      San Francisco             Denver          Hartford
##                  0                  1                  0
##         Wilmington         Washington       Jacksonville
##                  0                  0                  0
##              Miami            Atlanta            Honolulu
##                  0                  0                  0
##              Boise            Chicago             Peoria
##                  1                  0                  0
##       Indianapolis          Des Moines            Wichita
##                  0                  0                  0
##          Louisville        New Orleans           Portland
##                  0                  0                  0
##          Baltimore             Boston            Detroit
##                  0                  0                  0
##  Sault Ste. Marie           Duluth Minneapolis/St Paul
##                  0                  0                  0
##            Jackson        Kansas City          St Louis
##                  0                  0                  0
##       Great Falls              Omaha               Reno
##                  1                  0                  1
##          Concord       Atlantic City      Albuquerque
##                  0                  0                  1
##            Albany            Buffalo           New York
##                  0                  0                  0
##         Charlotte            Raleigh            Bismark
##                  0                  0                  1
##        Cincinnati          Cleveland           Columbus
##                  0                  0                  0
##      Oklahoma City           Portland       Philadelphia
##                  0                  0                  0
##        Pittsburgh         Providence           Columbia
##                  0                  0                  0
##       Sioux Falls            Memphis          Nashville
##                  0                  0                  0
##            Dallas             El Paso            Houston
##                  0                  1                  0
##      Salt Lake City          Burlington            Norfolk
##                  1                  0                  0
##            Richmond       Seattle Tacoma            Spokane
##                  0                  0                  1
##          Charleston          Milwaukee          Cheyenne
##                  0                  0                  1
##           San Juan
##                  0
```

```
# c.
# Mean of the dummy variable gives the proportion
phat <- mean(precip_dry)
n <- length(precip_dry)

# The approximate 95% CI
ci <- c(phat - 1.96*sqrt(phat*(1 - phat))/sqrt(n), phat + 1.96*sqrt(phat*(1 - phat))/sqrt(n))

# d.
# For every 100 random samples of US cities of size 70, in roughly 95 of them the confidence interval
# computed as in part c contains the true proportion of dry cities in the US. We hope that the single
# interval we have is one of these 95.
```

2. In 2018, a proportion $p_0 = 0.098$ of people living in Finland had their last name beginning with a vowel. Treat the previous fact as a hypothesis and test it using the participants of the exercise session as a sample.
   a. Observe the sample size n and the observed proportion $\hat{p}$ of participants having last names beginning with a vowel.
   b. Write down the assumptions and hypotheses of the one-sample proportion test.
   c. Conduct the test, using the exact version of the test if the requirements of the approximative test on slide 4.9 are not fulfilled.
   d. What is the conclusion of the test? Can this conclusion be taken as evidence against/for the "hypothesis"?

```
# a.
# Substitute the real values in place of the defaults:
n <- 35
x <- 5
phat <- x/n

# b.
# Assumptions:
# The sample is (id from Bernoulli with parameter value p where p is the proportion of people living in
# Finland with last name starting with a vowel.
# (That is, everyone has their last name beginning with a vowel with equal proability and independently
# of each other)
#
# Hypotheses:
# H0: p == 0.098
# H1: p != 0.098

# c.
# Exact test if n*phat <= 10 or n*(1 - phat) <= 10
binom.test(x, n, p = 0.098)
```

```
##
##  Exact binomial test
##
## data:  x and n
## number of successes = 5, number of trials = 35, p-value = 0.3855
## alternative hypothesis: true probability of success is not equal to 0.098
## 95 percent confidence interval:
##  0.04800478 0.30537135
## sample estimates:
## probability of success
##              0.1428571
```

```
# Asymptotic test else
prop.test(x, n, p = 0.098, correct = FALSE)
```

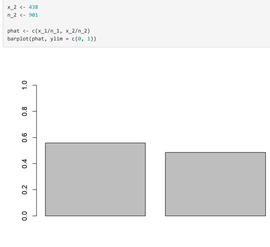
```
## Warning in prop.test(x, n, p = 0.098, correct = FALSE): Chi-squared
## approximation may be incorrect
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  x out of n, null probability 0.098
## X-squared = 0.77671, df = 1, p-value = 0.3721
## alternative hypothesis: true p is not equal to 0.098
## 95 percent confidence interval:
##  0.06260231 0.29375554
## sample estimates:
##         p
## 0.1428571
```

```
# d.
# Substitute conclusions here. The conclusions can most likely not be used to draw inference on the
# proportion of people in the whole Finland as the session participants make a poor "random" sample of this population.
# At best, the participants could be considered a random sample of all Aalto students in
# particular programmes.
```

3. In the beginning of the year a total of $n_1 = 963$ people were polled and $x_1 = 537$ out of them expressed their support for a certain presidential candidate. In a poll organized one month later $x_2 = 438$ people out of $n_2 = 901$ people claimed to support the candidate. Based on the data, has the support for the candidate decreased?
   a. Visualize the data.
   b. Write down the hypotheses for a two-sample proportion test and conduct it on a significance level 5%.
   c. What are the conclusions of the test?
   d. What assumptions were required by the test in part b? How can a poll-organizer ensure that they are satisfied?

```
# a.
x_1 <- 537
n_1 <- 963

x_2 <- 438
n_2 <- 901

phat <- c(x_1/n_1, x_2/n_2)
barplot(phat, ylim = c(0, 1))
```




```
# b.
# H0: p_1 = p_2
# H1: p_1 > p_2
# where p_1, p_2 are the success probabilities (probabilities to support the candidate) in the two samples
# which are assumed to be independent of each other and (id from two Bernoulli distributions.

prop.test(c(x_1, x_2), c(n_1, n_2), alternative = "greater", correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(x_1, x_2) out of c(n_1, n_2)
## X-squared = 9.5406, df = 1, p-value = 0.001005
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.0315168 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.5576324 0.4861265
```

```
# c.
# The one-sided p-value ~= 0.001 < 0.05 -> we reject the null hypothesis in favor of the alternative.
# That is, the support has decreased.

# d.
# (The assumptions are stated above in the answer to b.) To ensure that the samples are independent and (id
# and representative of the nationwide support level, the pollmaker should draw the samples perfectly randomly
# from amongst all eligible voters.
```

4. **(Optional)** Find out how the *Wilson score confidence interval* for a binary proportion is computed and locate an R package which computes it (there are several). The Wilson interval gives a better coverage probability than the standard CI given in slide 4.6 for small sample sizes. Find out how large this improvement is by conducting a simulation study. For example, simulate `m = 10000` samples from the binomial distribution with $n = 15$ and $p = 0.3$ and compute for both intervals the proportion of the samples in which the interval contains the true parameter value.

MS-C1620 Statistical Inference

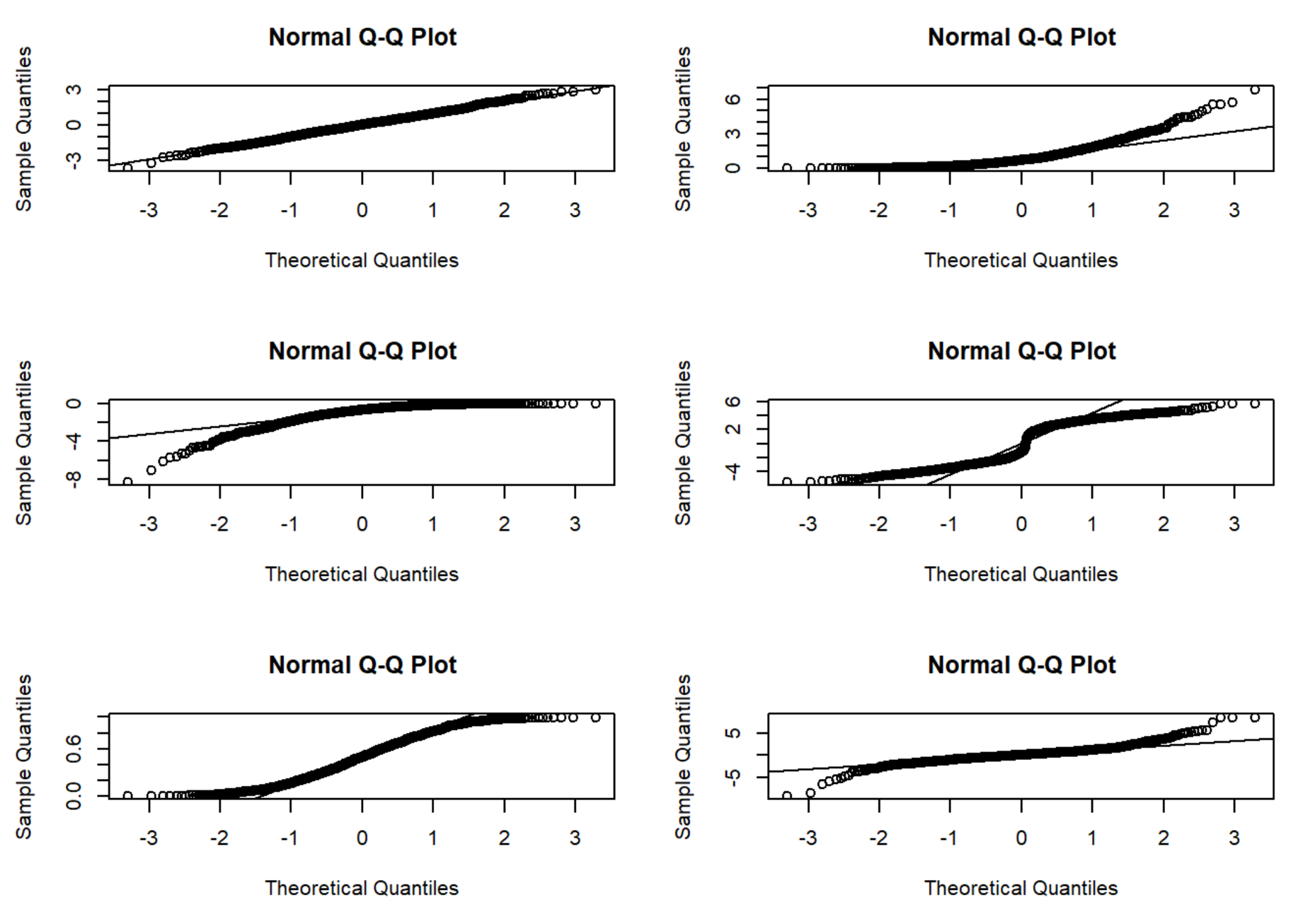
Exercise 6

Homework exercise

To be solved at home before the exercise session.

1. a. Assume that we have an iid. random sample $x_1, \ldots, x_{1000}$ and we'd like to use the normal Q-Q plot to assess whether the sample came from a normal distibution. How do you expect the normal Q-Q plot to roughly look like (i.e. what general features do you expect it to have and *why*), if the true distribution of the data is
    i. a normal distribution,
    ii. a right-skew distribution,
    iii. a left-skew distribution,
    iv. a bimodal distribution,
    v. a distribution with light tails,
    vi. a distribution with heavy tails?

```
# i. Straight line: sample quantiles depend approximately linearly on normal quantiles

# ii. U-shaped curve: the high values are too high (long right tail) and the low values
# are too high (short left tail).

# iii. Inverse U-shaped curve: for the opposite reasons as in ii.

# iv. S-shape near the middle of the plot: the points left of median are too small and the points right of median are too la
rge (too little mass near median)

# v. S-shape in the tails: the low values are too high (the left tail is too short) and the high values
# are too low (the right tail is too short).

# vi. Something resembling a cubic function: the low values are too low (the left tail is too long) and
# the high values are too high (the right tail is too long).

# Sample plots:
par(mfrow = c(3, 2))

x <- rnorm(1000)
qqnorm(x)
qqline(x)

x <- rexp(1000, 1)
qqnorm(x)
qqline(x)

x <- -1*rexp(1000, 1)
qqnorm(x)
qqline(x)

b <- rbinom(1000, 1, 1/2) == 1
x <- c(rnorm(1000, 3)[b], rnorm(1000, -3)[!b])
qqnorm(x)
qqline(x)

x <- runif(1000)
qqnorm(x)
qqline(x)

x <- rt(1000, 3)
qqnorm(x)
qqline(x)
```



```
par(mfrow = c(1, 1))
```

```
b. Recall the differences between the interpretations of the $\chi^2$ homogeneity test and $\chi^2$ test for independence. C
ome up with a practical situation where the collected data can be expressed as a 2-by-2 table and a related research questio
n for which the correct interpretation is through
    i. the $\chi^2$ homogeneity test,
    ii. the $\chi^2$ test for independence.
```

```
# The key difference between the two tests is in how the data is sampled, i.e., are the margins fixed or not.

# E.g. assume we're interested in studying whether sex (female/male) has an effect on the voting preference
# (democrat/republican) in the US and for this we interview n people in the street.
# These data can be collected into a two-by-two table such that the row variable is sex and the column
# variable is voting preference.

# i. If we choose beforehands that we will interview n1 females and n2 males, then studying the independence of the two vari
ables will be questionable (since sex is not fully random anymore with its marginal frequencies fixed). The correct interpre
tation is through the homogeneity test which compares two populations, in this case female and male, in their voting prefere
nces.

# ii. If we do not choose beforehands the marginal numbers of females and males, sex is a random variable and we can measure
its independence with the voting behavior. The correct interpretation is now through the test for independence.
```

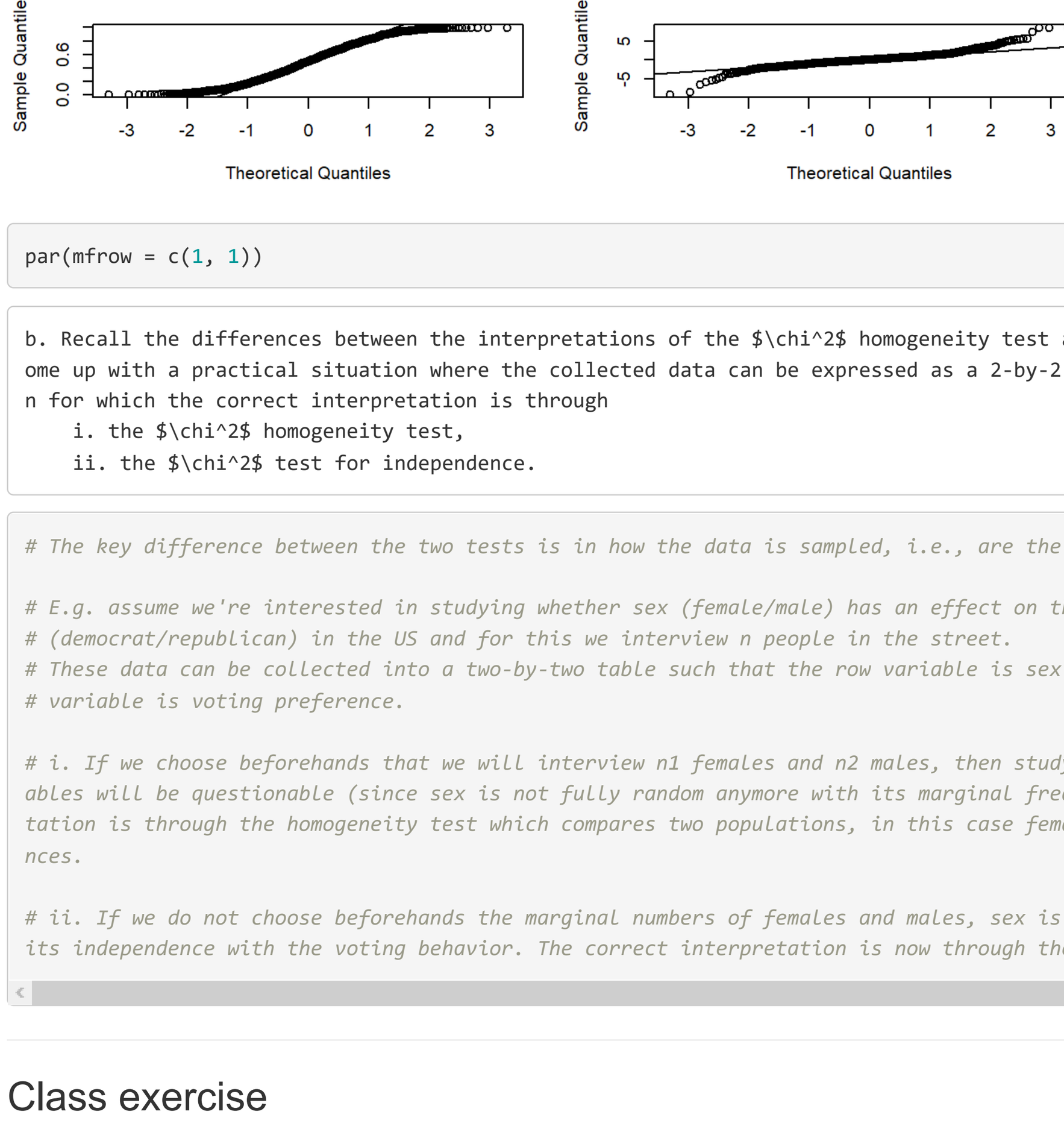
Class exercise

To be solved at the exercise session.

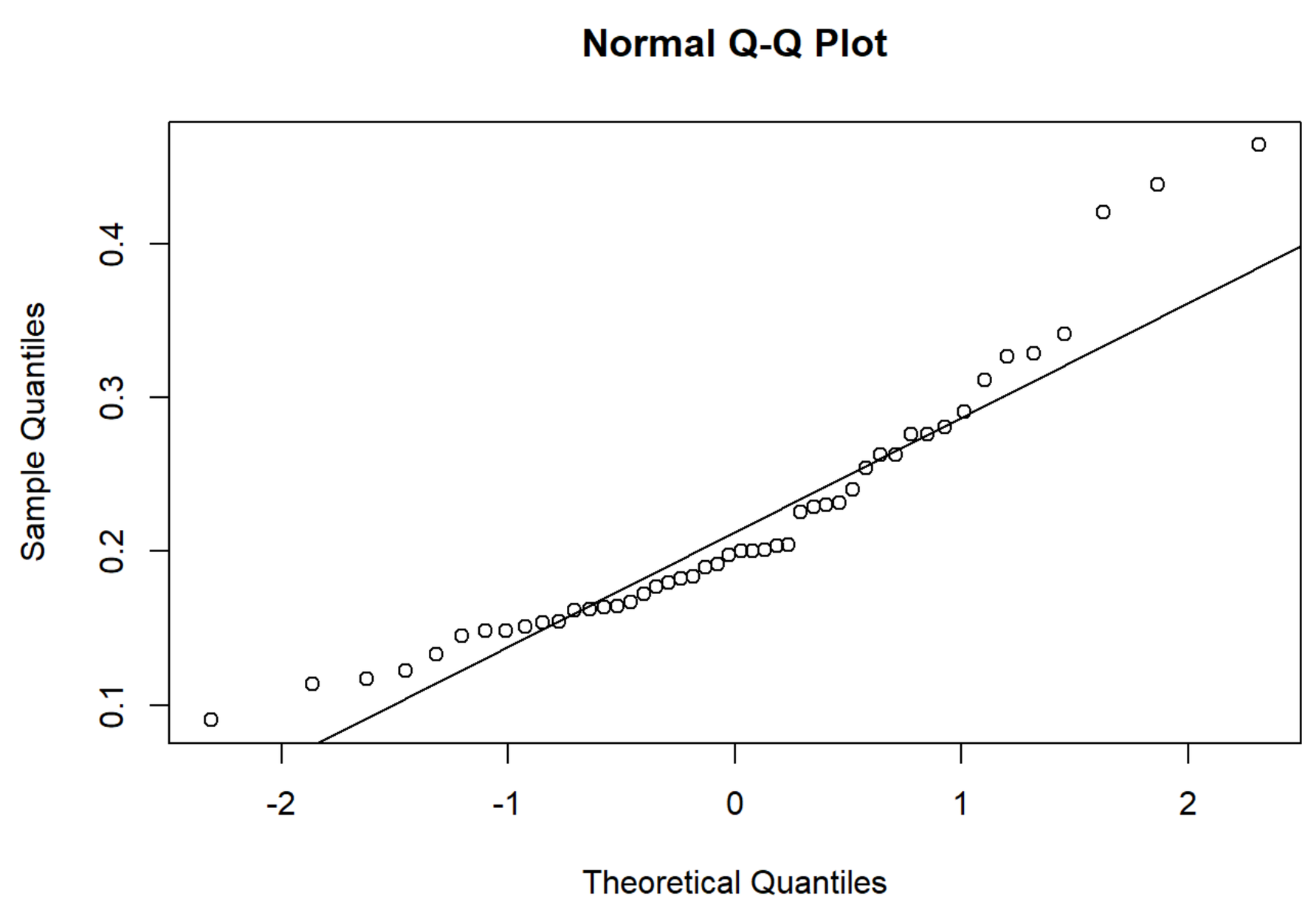
*Note: all the needed data sets are either given below or available in base **R**.*

1. The data set `rock` contains measurements on 48 rock samples from a petroleum reservoir. Treat the data as an iid. random sample from some distribution and test whether the distribution of `shape` is normal.
    a. Visualize the data to obtain a preliminary idea of the possible normality of the data.
    b. Use the normal Q-Q plot to gain more evidence on the normality/non-normality of the data.
    c. Conduct the Bowman-Shenton (Jarque-Bera) and the Shapiro-Wilk tests of normality on significance level 0.05.
    d. After all the previous, would you conclude the data to be normal (or normal enough for methods with normality assumptions)?
    e. Why is the data not really iid.?

```
# a.
x <- rock[, 3]
# The histogram seems a bit skew but otherwise not too far from normal?
hist(x)
```



```
# b.
# The normal Q-Q plot gives more evidence of positive skewness (it is reminiscent of part ii in Homework 1a.)
qqnorm(x)
qqline(x)

# c.
# Bowman-Shenton/Jarque-Bera
library(tseries)
```



```
jarque.bera.test(x)
```

```
##
##  Jarque Bera Test
##
## data:  x
## X-squared = 13.402, df = 2, p-value = 0.00123
```

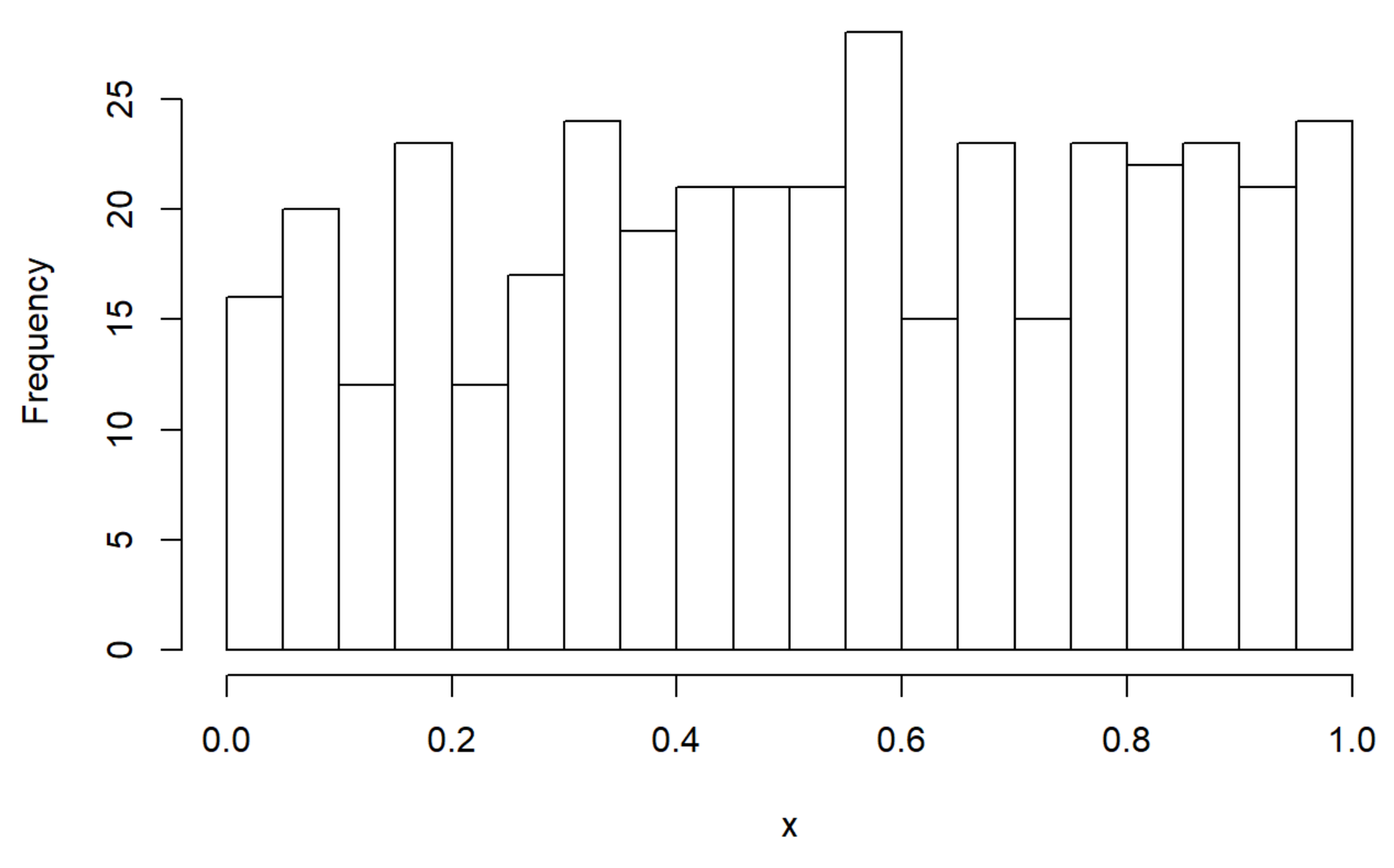
```
# Shapiro-Wilk
shapiro.test(x)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.90407, p-value = 0.0008531
```

```
# d.
# Both tests in c. reject their null hypotheses of normality. Based on all the previous evidence, the data can not be deemed
normal enough to rely on normality assumptions in any further analyses.
# (Note that it is a different matter whether the next analysis steps involve methods that allow the normality assumption to
be "covered" by large enough sample size (by the central limit theorem).)

# e.
# See the help file of the dataset: The sample is not iid. as, to obtain the 48 measurements, first 12 "core samples" were o
btained (randomly?) and then from each of these 4 observations were taken to yield the final 48 observations. Thus, the sets
of 4 observations come from a same core sample and as such are not independent, even if the different core samples were.
```

2. The data set `randu` contains 400 triples of successive random numbers from the random number generator RANDU. Use the χ^2 goodness-of-fit test to assess whether the first elements in the triplets really obey the uniform distribution on $[0, 1]$.
    a. Extract the first elements in the triplets and visualize their sample distribution.
    b. Discretize the values into a suitable number of categories and calculate the observed category frequencies.
    c. Compute the corresponding expected category probabilites under the uniform distribution on $[0, 1]$.
    d. Recall the hypotheses of the test and conduct it on significance level 0.05.
    e. What are the conclusions of the test? Compare your results with someone who used a different choice of categories for the discretization.

```
# a.
x <- randu[, 1]

# For data coming from a uniform distribution, we expect the histogram bars to be approximately equal in height.
# Seems like this could be the case with the current data.
hist(x, breaks = 20)
```



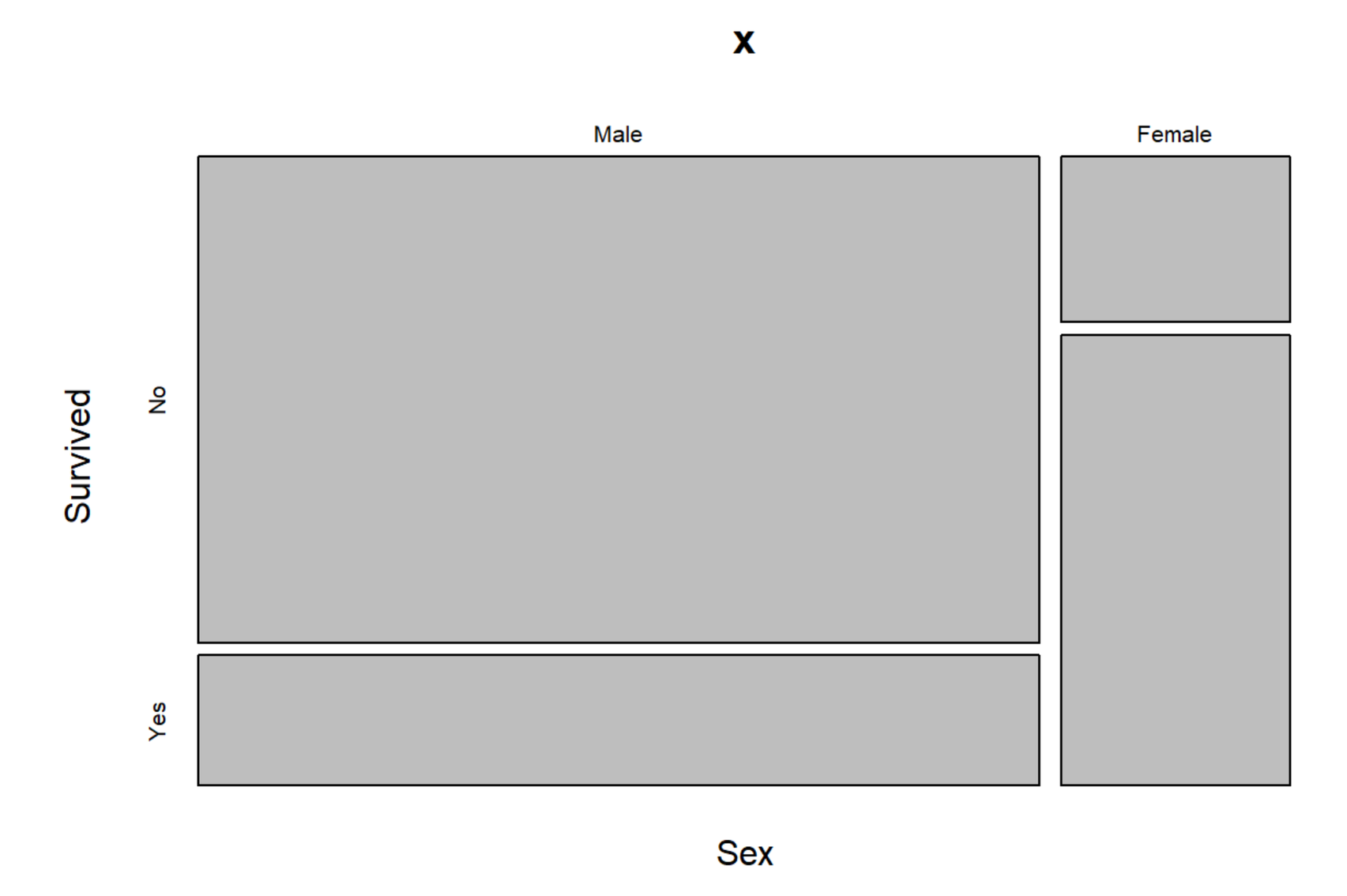
```
# b.
# The categories: [0, 0.1], (0.1, 0.2],...
obs_x <- diff(sapply(seq(0, 1, by = 0.1), function(i) sum(x <= i)))

# c.
# Each category has the same width 0.1 and the postulated distribution is uniform on [0, 1]
# -> the expected category probabilities are each 0.10 (40 obs. per category)
p_x <- rep(0.1, 10)

# d.
# The null hypothesis is that the data was generated by the uniform distribution on [0, 1]
# and the alternative that it was not.
chisq.test(x = obs_x, p = p_x)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  obs_x
## X-squared = 7.85, df = 9, p-value = 0.5493
```

```
# e.
# The test p-value is 0.5493
# -> no evidence against H0, it is still plausible that the data is from Uniform[0, 1].
# By choosing the categories suitably, it is most likely possible to get the opposite result
# (recall the Type I and II errors). However, this should not be taken advantage of in practice...
```

3. The data set `Titanic` contains information on the fate of passengers on the fatal maiden voyage of the ocean liner "Titanic". We use the data to study whether there is a connection between the sex (Male/Female) of a passenger and surviving from the ship (No/Yes).
    a. Extract a marginal table containing only the cross-tabulation of the variables `Sex` and `Survived`.
    b. Find a suitable way to visualize the data.
    c. Which test is appropriate for these data (and why?), χ^2 homogeneity test or the χ^2 test for independence?
    d. Conduct your chosen test on significance level 0.05 and state your conclusions.

```
# a.
x <- margin.table(Titanic, c(2, 4))

# b.
# Mosaic plot reveals the proportions of survived passengers for both sexes.
# -> proportionally fewer male passengers survived than female. To see whether this effect is statistically
# significant (and not caused just by randomness), we conduct an appropriate test.
mosaicplot(x)
```



```
# c.
# It seems plausible that there were no quotas on Female/Male passengers on the ship. As such, both factors had
# their margins non-fixed and the correct test is the test for independence.

# d.
chisq.test(x)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  x
## X-squared = 454.5, df = 1, p-value < 2.2e-16
```

```
# Very low p-value
# -> sex and survival status are not independent -> females had stat. significant higher chance of surviving.
```

4. **(Optional)** Choose your favorite non-normal distribution and use simulations to study the Type II error probabilities of the Bowman-Shenton (Jarque-Bera) and Shapiro-Wilk tests of normality for that distribution on different sample sizes (e.g. $n = 10, 100, 1000, 10000$). That is, find out the probabilty of falsely concluding that the data comes from a normal distribution when it does not.

MS-C1620 Statistical Inference

Exercise 7

Homework exercise

To be solved at home before the exercise session.

1. a. Go to the website which lists pairs of variables that have no causal relationship but still exhibit a large correlation. Pick one of the datasets and figure out how the data is presented, i.e., how are the plots constructed from the (x_i, y_i)-data (the plots are *not* scatter plots of the two variables in question), how are individual pairs (x_i, y_i) represented in the plots and what are the lines going through the points?

```
# a.
# In the plots:
# * x-axis is time
# * each time point corresponds to a single pair (x_i, y_i)
# * the x_i-value (y_i-value) of a pair is plotted on the corresponding time point in black (red)
# * the "Correlation" is calculated between the x_i-values and y_i-values
# * the lines are simply smoothed curves running through the x_i-values and y_i-values (they try to visualize the marginal t
rends).
# * Note also that the best fitting line ("y_i = a x_i + b") could not be drawn in the plot in the usual way.
```

```
b. Let $x, y, \varepsilon$ be random variables such that, $$y = x + \varepsilon,$$ where $\mathrm{Var}(x) = 1$, $\mathrm{Va
r}(\varepsilon) = \sigma^2 > 0$ and $x$ and $\varepsilon$ are independent (interpretation: $x$ and $y$ have a perfect linear
relationship but the observed value of $y$ is contaminated with the noise/measurement error $\varepsilon$ having variance
$\sigma^2$). Compute the Pearson correlation $\rho$ between $x$ and $y$ and investigate how it behaves when $\sigma^2$ is in
creased. Interpret this behavior.
```

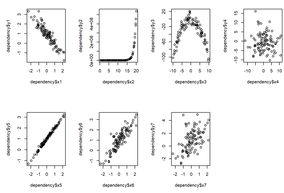
```
# rho(x, y) = cov(x, y)/(sd(x)*sd(y))
#
# Now, cov(x, y) = cov(x, x + e) = cov(x, x) + cov(x, e) = var(x) + 0 = 1,
# where the second equality uses the linearity of covariance and the third equality uses the fact that
# x and e are independent
#
# Also, sd(x) = sqrt(var(x)) = 1 and sd(y) = sqrt(var(x + e)) = sqrt(var(x) + var(e)) = sqrt(1 + sigma^2),
# again by the independence of x and e.
#
# Thus rho(x, y) = 1/sqrt(1 + sigma^2), which decreases towards zero as sigma^2 increases.
# The interpretation for this is that increasing the noise strength (variance) masks the true perfect relationship and the c
orrelation gets weaker (the point-pairs deviate more and more from the straight line). See class exercise 1 for visual versi
on of the same phenomenon.
```

Class exercise

To be solved at the exercise session.

1. The file `data_dependency.txt` contains seven bivariate data sets (the columns `xi` and `yi`, where $i = 1, 2, \ldots, 7$, always form a pair).
   a. Read the file into R using the command `read.table`.
   b. Draw a scatter plot for each pair of variables.
   c. Calculate the Pearson and Spearman correlations of the pairs and compare them to the scatter plots.
   d. The underlying distributions of the samples 5-7 are the same up to the variance of `yi` (the variance is highest in sample 7). What happens to the correlation coefficients as the variance increases and why?

```
# a.
# Replace "params$your_path_here_1" with your path to the .txt file in the code below (and remember that "/" is used to navi
gate sub-folders in R)
dependency <- read.table(params$your_path_here_1, sep = "\t", header = TRUE)


# b.
par(mfrow = c(2, 4))
plot(dependency$x1, dependency$y1)
plot(dependency$x2, dependency$y2)
plot(dependency$x3, dependency$y3)
plot(dependency$x4, dependency$y4)
plot(dependency$x5, dependency$y5)
plot(dependency$x6, dependency$y6)
plot(dependency$x7, dependency$y7)
par(mfrow = c(1, 1))
```



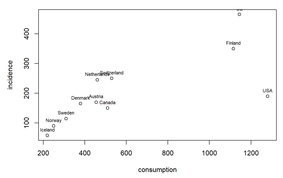

```
# c.
cors <- function(i){
  cor_p <- cor(dependency[, (2*i - 1)], dependency[, 2*i])
  cor_s <- cor(dependency[, (2*i - 1)], dependency[, 2*i], method = "spearman")
  print(paste0("Data set ", i, ": Pearson: ", round(cor_p, 3), ", Spearman: ", round(cor_s, 3)))
}

for(i in 1:7){
  cors(i)
}
```

```
## [1] "Data set 1, Pearson: -0.924, Spearman: -0.922"
## [1] "Data set 2, Pearson: 0.51, Spearman: 1"
## [1] "Data set 3, Pearson: -0.035, Spearman: -0.028"
## [1] "Data set 4, Pearson: -0.05, Spearman: 0.004"
## [1] "Data set 5, Pearson: 0.994, Spearman: 0.993"
## [1] "Data set 6, Pearson: 0.879, Spearman: 0.855"
## [1] "Data set 7, Pearson: 0.582, Spearman: 0.587"
```

```
# Interpretation:
# 1: almost perfect decreasing (linear/monotone relationship
# 2: no clear linear relationship but perfect increasing monotone relationship
# 3: symmetric increasing-decreasing relationship -> both correlations zero ("increase masks decrease")
# 4: no discernible relationship -> both correlations zero
# 5-7: increasing linear/monotone relationship which gets more and more difficult to see because of the increasing y-varian
ce. That is, increasing the y-variance hides the linear relationship under the added "noise", decreasing the correlations. T
his is the same phenomenon as in homework problem 1b.


# d.
# See above.
```

2. The file `data_tobacco.txt` contains data on cigarette consumption and lung cancer incidences from 11 different countries. The variable `consumption` describes the yearly consumption of cigarettes per capita in 1930 and the variable `incidence` tells the lung cancer incidence rates per 100 000 people in 1950. We use correlation to study the connection between these two.
   a. Read the file into R using the command `read.table`.
   b. Draw a scatter plot of `consumption` and `incidence` which also shows the country names.
   c. Using the scatter plot, make an educated guess on the signs and magnitudes of the Pearson and Spearman correlations of the two variables.
   d. Calculate the Pearson and Spearman correlations.
   e. Use permutation test to test whether the two correlations differ significantly from zero, using the significance level 5%.
   f. Drop USA from the data, redo the previous analysis and compare the results to those obtained with the full data. What happened?

```
# a.
# Replace "params$your_path_here_2" with your path to the .txt file in the code below (and remember that "/" is used to navi
gate sub-folders in R)
tobacco <- read.table(params$your_path_here_2, sep = "\t", header = TRUE)


# b.
plot(incidence ~ consumption, data = tobacco)
text(tobacco$consumption, tobacco$incidence, labels = tobacco$country, cex= 0.7, pos=3)
```



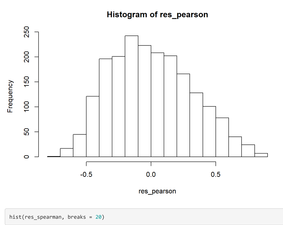

```
# c.
# Apart from USA, the data lies very accurately on a line.
# The outlier most likely decreases the value of Pearson correlation (which is sensitive to outliers) somewhat, and it could
 then equal -0.80. The Spearman correlation is more tolerant to outliers (because it ignores the magnitudes of the outliers)
 and could then equal something bit larger, e.g. -0.90.


# d.
cor(tobacco$consumption, tobacco$incidence)
```

```
## [1] 0.7409723
```

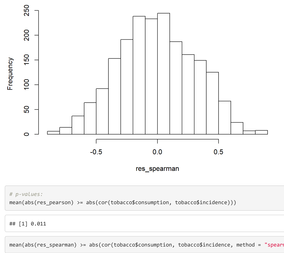
```
cor(tobacco$consumption, tobacco$incidence, method = "spearman")
```

```
## [1] 0.8454545
```

```
# Quite close...


# e.
# Permutation tests
n <- nrow(tobacco)

B <- 2000
res_pearson <- rep(0, B)
res_spearman <- rep(0, B)

for(b in 1:B){
  res_pearson[b] <- cor(tobacco$consumption, sample(tobacco$incidence, n, replace = FALSE))
  res_spearman[b] <- cor(tobacco$consumption, sample(tobacco$incidence, n, replace = FALSE), method = "spearman")
}

# Distributions of the permutation test replicates (distribution of the test statistic under H0)
hist(res_pearson, breaks = 20)
```




```
hist(res_spearman, breaks = 20)
```




```
# p-values:
mean(abs(res_pearson) >= abs(cor(tobacco$consumption, tobacco$incidence)))
```

```
## [1] 0.011
```

```
mean(abs(res_spearman) >= abs(cor(tobacco$consumption, tobacco$incidence, method = "spearman")))
```

```
## [1] 0.002
```

```
# Both smaller than 0.05 -> both correlations differ significantly from 0


# f.
tobacco <- tobacco[-7, ]
# Running the previous code to remove USA and then redoing the steps yields
# Pearson correlation: 0.941, p-value: 0
# Spearman correlation: 0.927, p-value: 0.001
# The correlations are higher and more significant without the "outlier" which masked the "true" relationship.

# Note: In practice, the labeling of USA as an outlier and the consecutive removal of it should be somehow
# justified.
```

3. **(Optional)** Use also the tests given on slides 6.16 and 6.20 to test the null hypothesis $H_0 : \rho = 0$ for Pearson correlation in problem 2e. How do the results compare to the permutation test?

4. **(Optional)** Simulate the distribution of the sample Pearson correlation $\hat{\rho}$ under normality by generating multiple datasets of size n from a bivariate normal distribution of your choice. Then transform the sample Pearson correlations as $\hat{\rho} \mapsto \operatorname{arctanh}(\hat{\rho})$ and inspect the distribution of the transformation. Does it look normal? (it should for large n, as per slide 6.13)

MS-C1620 Statistical Inference

Exercise 8

Homework exercise

To be solved at home before the exercise session.

1. a. Show that if in simple linear regression both the explanatory variable x and the response y have been marginally standardized such that $\bar{x} = 0, s_x = 1$ and $\bar{y} = 0, s_y = 1$, then the estimated least squares regression model is simply,

$$\hat{y}_i = \hat{\rho}(x, y) x_i.$$

That is, the regression coefficient of x equals the sample correlation between x and y.

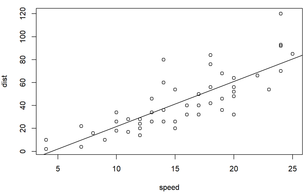
```
# Plug in the means and sample standard deviations to the formula in slide 7.15 to obtain the result.
```

```
b. The `cars` data give the speeds of cars (`speed`, in mph) and the corresponding distances taken to stop (`dist`, in fee
t). The below shows the model summary of a simple linear regression model fit using `speed` as an explanatory variable and `
dist` as a response. Interpret the model results.
```

```
cars_lm <- lm(dist ~ speed, data = cars)
summary(cars_lm)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

```
plot(dist ~ speed, data = cars)
abline(cars_lm)
```



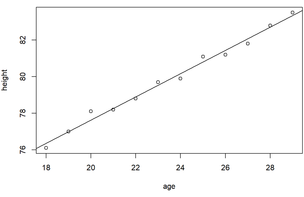

```
# Regression coefficient ~3.9: an increase of 1 mph in speed increases the expected stopping distance by 3.9 feet (and based
on the low p-value, this relationship is not caused by randomness (assuming the model assumptions hold)).
# R-squared 0.6511: the model manages to explain around two thirds of the variation in the response variable, indicating a g
ood fit.
# Based on the scatter plot, the relationship indeed looks to be linear.
```

Class exercise

To be solved at the exercise session.

1. The file `data_children.txt` contains data on children's ages (`age`, in months) and heights (`height`, in centimeters). Investigate whether there is a linear relationship between the two variables.
   a. Read the file into R using the command `read.table`.
   b. Draw a scatter plot of `age` and `height`.
   c. Fit a linear model to the data using `height` as a response variable.
   d. Add the fitted regression line to the scatter plot. Does the fit appear good?
   e. Interpret the estimated regression coefficient of `age` and the R^2-value of the model.

```
# a.
# Replace "params$your_path_here_1" with your path to the .txt file in the code below (and remember that "/" is used to navi
gate sub-folders in R)
children <- read.table(params$your_path_here_1, sep = "\t", header = TRUE)


# b.
plot(children)


# c.
children_lm <- lm(height ~ age, data = children)


# d.
# The line appears to fit the data well.
abline(children_lm)
```



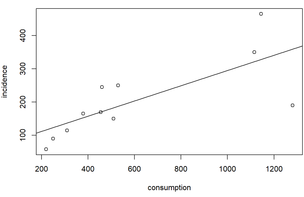

```
# e.
summary(children_lm)
```

```
##
## Call:
## lm(formula = height ~ age, data = children)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27238 -0.24248 -0.02762  0.16014  0.47238
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  64.9283     0.5084  127.71  < 2e-16 ***
## age           0.6350     0.0214   29.66 4.43e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.256 on 10 degrees of freedom
## Multiple R-squared:  0.9888, Adjusted R-squared:  0.9876
## F-statistic:   880 on 1 and 10 DF,  p-value: 4.428e-11
```

```
# The regression coefficient is about 0.64, meaning that for an increase of a single month in a child's age, the expected va
lue of her/his height goes up by 0.64cm.

# The coefficient of determination is ~0.99, indicating an excellent fit.
```

2. The file `data_tobacco.txt` contains data on cigarette consumption and lung cancer incidences from 11 different countries. The variable `consumption` describes the yearly consumption of cigarettes per capita in 1930 and the variable `incidence` tells the lung cancer incidence rates per 100 000 people in 1950. *(Recall exercise 7.2)*
   a. Read the file into R using the command `read.table`.
   b. Draw a scatter plot of `consumption` and `incidence`.
   c. Fit a linear model to the data using `incidence` as a response variable.
   d. Add the fitted regression line to the scatter plot. Does the fit appear good?
   e. Interpret the estimated regression coefficient and p-value of `consumption`.
   f. Interpret the R^2-value of the model.
   g. Drop USA from the data, redo the previous analysis and compare the results to those obtained with the full data. What happened?

```
# a.
# Replace "params$your_path_here_2" with your path to the .txt file in the code below (and remember that "/" is used to navi
gate sub-folders in R)
tobacco <- read.table(params$your_path_here_2, sep = "\t", header = TRUE)


# b.
plot(incidence ~ consumption, data = tobacco)
# text(tobacco$consumption, tobacco$incidence, labels = tobacco$country, cex= 0.7, pos=3)


# c.
tobacco_lm <- lm(incidence ~ consumption, data = tobacco)


# d.
# The line appears to miss most of the points in favour of trying to reach closer to the outlier in the lower right corner.
abline(tobacco_lm)
```




```
# e.
summary(tobacco_lm)
```

```
##
## Call:
## lm(formula = incidence ~ consumption, data = tobacco)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -169.016  -32.813    0.004   45.804  136.914
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 65.74886   48.95871   1.343  0.21217
## consumption  0.22912    0.06921   3.310  0.00908 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 84.13 on 9 degrees of freedom
## Multiple R-squared:  0.549,  Adjusted R-squared:  0.4989
## F-statistic: 10.96 on 1 and 9 DF,  p-value: 0.009081
```

```
# The regression coefficient is about 0.23, meaning that for an increase of a single smoked cigarette in year per capita, th
e expected value of the incidence of lung cancer (20 years afterwards) goes up by 0.23 units.
# The p-value related to the coefficient is small (below the standard 0.05), meaning that this effect is most likely real, a
nd not just caused by randomness (assuming the model assumptions hold).

# f.
# The R-squared is somewhat high (~0.55), meaning that the model succeeds in explaining a bit over half of the variation of
the response variable (however, compare this to the R-squared without the outlier in part g.)


# g.
tobacco <- tobacco[-7, ]
# Running the previous code to remove USA and then redoing the steps yields:
# Examination of the scatter plot shows that the line fits the data much better now
# Coefficient of "consumption" ~ 0.36 (the outlier had "contaminated" the value for the full data)
# p-value is much smaller (removing the outlier helped the model see clearer that the perceived effect is not just randomnes
s)
# R-squared increased to ~0.89, a much better fit.

# NOTE: Again, in practice we should have some clear reason for removing U.S. from the data. We'll investigate this closer n
ext time.
```

3. **(Optional)** Investigate how much a single outlier can affect the results of a linear model: Create a small data set that has a perfect linear relationship between its two variables (such a model has the explanatory variable p-value equal to 0 and the coefficient of determination equal to 1). Then, add a single outlying data point and see how much you can change the p-value and the coefficient of determination by varying the outlier's value.

MS-C1620 Statistical Inference

Exercise 10

Homework exercise

To be solved at home before the exercise session.

1. a. Consider the following linear model,

$$\mathbb{E}(y_i \mid \mathbf{x}_i) = \beta_0 + \beta_1 wrx_i + \beta_2 agr_i + \beta_3 (wrx_i \times agr_i),$$

where wrx_i is a binary variable ($1 = \text{male}$, $1 = \text{female}$) and agr_i is a continuous variable. Write down the model separately for males and females and using the two models give interpretations for the four parameters.

```
# For males, sex = 0 and we have
# E(y|x) = b0 + 0 + b2*age + b3*(0 * age) = b0 + b2*age

# For females, sex = 1 and we have
# E(y|x) = b0 + b1*1 + b2*age + b3(1 * age) = (b0 + b1) + (b2 + b3)*age

# That is,
# b0 = the intercept of males
# b1 = the difference between the intercepts of females and males
# b2 = the slope of age for males
# b3 = the difference between the age-slopes of females and males (e.g., if b3 = 0, the effect of age on the response is the same for both sexes).
```

b. The data set `galaxy` from the package `tlimtatcaaro` contains measurements on the position and radial velocity of the galaxy NGC7531. Fitting a model with the latter as a response, we get the following model summary and residual plot. Does the model fit well? If not, what could be tried next?

```
library(tlimtatcaaro)
library(car)
```

```
## Loading required package: carData
```

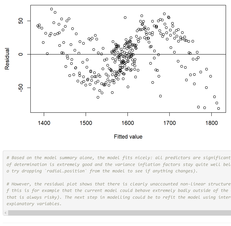
```
lm_galaxy <- lm(velocity ~ ., data = galaxy)
summary(lm_galaxy)
```

```
##
## Call:
## lm(formula = velocity ~ ., data = galaxy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -80.988 -21.673   0.442  22.770  67.527
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1589.42295   3.91939 405.496  < 2e-16 ***
## east.west        0.77458   0.11281   2.481  0.01363 *
## north.south     -1.33179   0.09607 -32.554  < 2e-16 ***
## angle            0.12454   0.04356   2.859  0.00451 **
## radial.position  0.90118   0.10842   5.618 4.23e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.13 on 318 degrees of freedom
## Multiple R-squared:  0.8993,  Adjusted R-squared:  0.8979
## F-statistic: 708.6 on 4 and 318 DF,  p-value: < 2.2e-16
```

```
vif(lm_galaxy)
```

```
##      east.west    north.south          angle radial.position
##       4.996114       1.747546       1.002817       6.118775
```

```
plot(fitted(lm_galaxy), resid(lm_galaxy), xlab = "Fitted value", ylab = "Residual")
abline(h = 0)
```




```
# Based on the model summary alone, the model fits nicely: all predictors are significant on the 0.05 level, the coefficient of determination is extremely good and the variance inflation factors stay quite well below 10 (although it would not hure t o try dropping `radial.position` from the model to see if anything changes).

# However, the residual plot shows that there is clearly unaccounted non-linear structure in the response (one consequence o f this is for example that the current model could behave extremely badly outside of the current range of x-values, although that is always risky). The next step in modelling could be to refit the model using interactions and transformations of the explanatory variables.
```

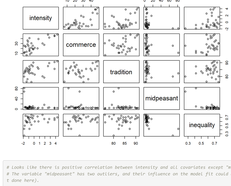
Class exercise

To be solved at the exercise session.

1. The data set `Chirot` from the package `carData` contains statistics on the 1907 Romanian peasant rebellion. Each row of the data is a county for which the `intensity` of the rebellion has been measured, along with various socio-economic variables. Investigate using linear regression whether there is dependency between `intensity` and the explanatory variables.
   a. Visualize the data.
   b. Fit a linear regression model to the data.
   c. Assess the adequacy of the model and its assumptions through the model summary, VIFs and model diagnostics.
   d. Make changes to the model, if needed.
   e. Interpret the results.

```
library(carData)
```

```
# a.
pairs(Chirot)
```



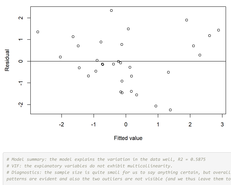

```
# Looks like there is positive correlation between intensity and all covariates except "midpeasant".
# The variable "midpeasant" has two outliers, and their influence on the model fit could be interesting to explore later (no t done here).


# b.
lm_chirot <- lm(intensity ~ ., data = Chirot)

# c.
summary(lm_chirot)
```

```
##
## Call:
## lm(formula = intensity ~ ., data = Chirot)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2468 -0.6781 -0.1013  0.8025  2.3378
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12.919818   5.507499  -2.346 0.026587 *
## commerce      0.091240   0.026268   4.457 0.000114 ***
## tradition    -0.116787   0.060868  -1.919 0.064996 .
## midpeasant   -0.007342   0.017695  -0.189 0.851625
## inequality    1.117978   2.050304   0.399 0.693815
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.227 on 27 degrees of freedom
## Multiple R-squared:  0.7836,  Adjusted R-squared:  0.7519
## F-statistic: 9.462 on 4 and 27 DF,  p-value: 6.476e-05
```

```
vif(lm_chirot)
```

```
##  commerce  tradition midpeasant inequality
##  1.281808  1.133518  1.948342  2.011226
```

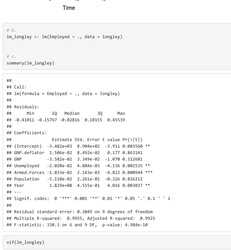
```
plot(fitted(lm_chirot), resid(lm_chirot), xlab = "Fitted value", ylab = "Residual")
abline(h = 0)
```




```
# Model summary: the model explains the variation in the data well, R2 = 0.7875
# VIF: the explanatory variables do not exhibit multicollinearity.
# Diagnostics: the sample size is quite small for us to say anything certain, but overall the residuals look fine. No weird patterns are evident and also the two outliers are not visible (and we thus leave them to the model).


# d.
# No changes.


# e.
# The significant predictor "commerce" has a coefficient = 0.09:
# - a single unit increase in the commercialization of agriculture increases the intensity of the rebellion by 0.09 units

# See "Chirot, D. and C. Ragin (1975) The market, tradition and peasant rebellion: The case of Romania. American Sociologica l Review 40, 428-444 [Table 1]" for more information on the data.
```

2. The data set `longley` contains measurements of economic variables from the years 1947-1962. We are interested in predicting the number of people employed (`Employed`, in thousands) using the other variables.
   a. Visualize the data.
   b. Fit a linear regression model to the data.
   c. Assess the adequacy of the model through the model summary and VIFs.
   d. Make changes to the model, if needed.

```
# a.
# The data is a time series so a possibly useful visualization could be:
plot.ts(longley)
```




```
# b.
lm_longley <- lm(Employed ~ ., data = longley)

# c.
summary(lm_longley)
```

```
##
## Call:
## lm(formula = Employed ~ ., data = longley)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.41011 -0.15767 -0.02816  0.10155  0.45539
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.482e+03  8.904e+02  -3.911 0.003560 **
## GNP.deflator  1.506e-02  8.492e-02   0.177 0.863141
## GNP          -3.582e-02  3.349e-02  -1.070 0.312681
## Unemployed   -2.020e-02  4.884e-03  -4.136 0.002575 **
## Armed.Forces -1.033e-02  2.143e-03  -4.822 0.000944 ***
## Population   -5.110e-02  2.261e-01  -0.226 0.826212
## Year          1.829e+00  4.555e-01   4.016 0.003037 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3049 on 9 degrees of freedom
## Multiple R-squared:  0.9955,  Adjusted R-squared:  0.9925
## F-statistic: 330.3 on 6 and 9 DF,  p-value: 4.984e-10
```

```
vif(lm_longley)
```

```
## GNP.deflator          GNP    Unemployed Armed.Forces   Population
##    135.53244   1788.51348     33.61889      3.58893    399.15102
##         Year
##    758.98060
```

```
# The model predicts the response extremely well (R2 ~ 1) but there is some significant multicollinearity between the predic tors.


# d.
# We drop one-by-one the variables with the highest VIF, until no VIF is > 10, resulting into the model
lm_longley_2 <- lm(Employed ~ Unemployed + Armed.Forces + Population, data = longley)
summary(lm_longley_2)
```

```
##
## Call:
## lm(formula = Employed ~ Unemployed + Armed.Forces + Population,
##     data = longley)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3835 -0.2868 -0.1353  0.1596  1.1382
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.523091   4.211566  -0.314  0.75800
## Unemployed  -0.012292   0.003154  -3.895  0.00214 **
## Armed.Forces -0.001893   0.000516  -3.538  0.00473 **
## Population   0.605146   0.047617  12.709 2.72e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6843 on 12 degrees of freedom
## Multiple R-squared:  0.9696,  Adjusted R-squared:  0.962
## F-statistic: 127.7 on 3 and 12 DF,  p-value: 2.275e-09
```

```
# The resulting model has still very high R2 = 0.9696.

# Diagnostics could now be checked and the model coefficients could be interpreted, although the latter is difficult due to the help file not properly defining the variables and their units...
```

3. **(Optional)** While general non-linear regression is beyond this course, fitting such models with R is quite straightforward. Try out the following code where a non-linear Generalized Additive Model (GAM) is fitted between temperature and ozone content in the `airquality` data.

```
# install.packages("mgcv")
library(mgcv)

x <- data.frame(ozone = airquality[, 1], temp = airquality[, 4])
gam_1 <- gam(temp ~ s(ozone), data = x)

plot(x, main = "Ozone", ylab = "Temperature")
ozone_grid <- data.frame(ozone = seq(min(x$ozone, na.rm = TRUE), max(x$ozone, na.rm = TRUE),length.out = 1000))
points(ozone_grid[, 1], predict(gam_1, ozone_grid), type = 'l')
```

Investigate especially what the final three lines do and what is the meaning of `ozone_grid`. Try also to fit a non-linear model to some other data set using the above as a template.

MS-C1620 Statistical Inference

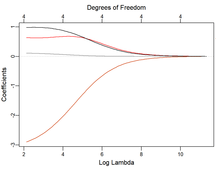
Exercise 11

Homework exercise

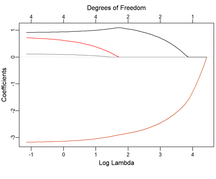
To be solved at home before the exercise session.

1. a. The data set `galaxy` from the package `ElemStatLearn` contains measurements on the position and radial velocity (the response) of the galaxy NGC7531. The following two plots show the ridge and LASSO coefficient profiles of the four explanatory variables. Compare the two plots and interpret them (for example, deduce which of the explanatory variables are the most ``important''?)

```
library(ElemStatLearn)
library(glmnet)
library(plotmo)

ridge_galaxy <- glmnet(as.matrix(galaxy[, 1:4]), as.matrix(galaxy[, 5]), alpha = 0)
plot_glmnet(ridge_galaxy, xvar = "lambda", label = TRUE)
```




```
lasso_galaxy <- glmnet(as.matrix(galaxy[, 1:4]), as.matrix(galaxy[, 5]), alpha = 1)
plot_glmnet(lasso_galaxy, xvar = "lambda", label = TRUE)
```




```
# We begin with LASSO as it is easier to interpret. Going from right to left, the variables enter the plot in the order of t
heir importance to the model fit. The most important is the north-south coordinate (`nrth.sth`) and second most important th
e radial position (`rdl.pstn`). The finnl two variables are almost equally important but the east-west coordinate (`east.wst
`) enters the model slightly before the angle.

# The interpretation of the ridge profiles is more difficult. For example, it looks like angle is the least important variab
le but as the coefficients never hit zero its small value might simply be an indication of it having a different scale than
the other variables. Additionally, the profiles of `rdl.pstn` and `east.wst` look mostly similar but the LASSO plot shows th
at the former is much more important. The Lesson: do the variable selection and interpretation using LASSO.
```

b. Consider a regression model where the response `Y` is explained using the covariates `X1`, `X2` and `X3`. The p-values corresponding to the models of all possible combinations of the covariates are listed below. Use them to perform variable selection with both backward and forward selection with the p-value cutoff $\alpha_0 = 0.05$.

```
##    X1
## 0.0071
```

```
##    X2
## 0.4221
```

```
##    X3
## 0.0014
```

```
##    X1    X2
## 0.0055 0.2809
```

```
##    X1    X3
## 0.0021 0.0004
```

```
##    X2    X3
## 0.1267 0.0006
```

```
##    X1    X2    X3
## 0.0010 0.0516 0.0001
```

```
# Backward selection: begin with the full model and drop always the least significant variable until all have p-value < 0.0
5.

# X1 + X2 + X3 -> X1 + X3

# Forward selection: start with an empty model and add always the most significant new variable until no addition has p-valu
e < 0.05.

# X3 -> X1 + X3

# Both methods give the same result.
```

Class exercise

To be solved at the exercise session. Note: the R-script of lecture 10 might prove helpful in solving the below problems.

1. The data set `barro` in the package `quantreg` contains the annual GDP growth rates of several countries along with several explanatory variables. Our objective is to use variable selection to determine which factors are most helpful in predicting the growth rate of GDP.
   a. Visualize the data set.
   b. Fit a standard multiple regression to the data with `y.net` as the response.
   c. Use the function `step` to perform both backward and forward variable selection using the AIC as the criterion.
   d. Do the results of the backward and forward selections agree?
   e. Which variables would you conclude to be the most important in predicting the growth rate of GDP?

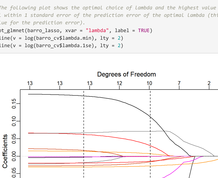
```
library(quantreg)
data(barro)

# a.
# The paired scatter plots reveal complicated relationships, not all of which are linear
# pairs(barro)

# b.
lm_barro <- lm(y.net ~ ., data = barro)
summary(lm_barro)
```

```
##
## Call:
## lm(formula = y.net ~ ., data = barro)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.037876 -0.009417  0.001339  0.009966  0.038882
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0355312  0.052664  -0.671  0.50358
## lgdp2        0.027892   0.005480   5.091 1.15e-13 ***
## mse2         0.015117   0.005125   2.950  0.00370 **
## fse2        -0.004467   0.006238  -0.716  0.47503
## fhe2        -0.023602   0.027751  -0.850  0.39645
## mhe2         0.022279   0.023702   0.940  0.34827
## lexp2        0.067136   0.016310   4.116 6.39e-05 ***
## lintr2      -0.001695   0.005091  -0.333  0.73947
## gedy2       -0.104333   0.118510  -0.880  0.38009
## ly2          0.004770   0.022563   0.211  0.83270
## gcony2      -0.190927   0.038302  -4.985 1.79e-06 ***
## lblakp2     -0.031368   0.004933  -6.358 2.44e-09 ***
## pol2        -0.020233   0.006283  -3.220  0.00158 **
## ttrad2       0.178385   0.039877   4.473 1.53e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01651 on 147 degrees of freedom
## Multiple R-squared:  0.5924,	Adjusted R-squared:  0.5563
## F-statistic: 16.43 on 13 and 147 DF,  p-value: < 2.2e-16
```

```
# The model has good predictive power (R2 ~ 0.60) but, at least based on the p-values, seems to contain several variables wi
th no statistically significant relationship with the response. The variable selection in part c should help get rid of the u
nnecessary ones.

# c.
# The selection codes are not run here due to the enormous amount of produced output.

##### Backward selection
# We start with the full model and at each step drop the variable whose leaving out lowers the AIC most (lower AIC is bette
r).

# step(lm(y.net ~ ., data = barro), direction = "backward")


# The backward selection yields the model:
# y.net ~ lgdp2 + mse2 + lexp2 + lintr2 + ly2 + gcony2 + lblakp2 + pol2 + ttrad2


##### Forward selection
# We start with an empty model and at each step include the variable whose inclusion lowers the AIC most.

# step(lm(y.net ~ 1, data = barro),
#      scope = list(lower = lm(y.net ~ 1, data = barro),
#                   upper = lm(y.net ~ ., data = barro)),
#      direction = "forward")

# The forward selection yields the model:
# y.net ~ lblakp2 + ly2 + lgdp2 + gcony2 + lexp2 + ttrad2 + mse2 + pol2 + lintr2

# The forward selection output also provides us with an order of importance for the variables. The variable which enters the
model first is the most important (the variables are given in the output in the order of importance).

# Thus lblakp2 (Black Market Premium) is the most important variable



# d.
# The results of the backward and forward selection agree. This is a good sign and makes the result more reliable (two strat
egies arrived at the same result).

# e.
# The important variables are:
# Black Market Premium
# Investment/GDP
# Initial Per Capita GDP
# Public Consumption/GDP
# Life Expectancy
# Growth Rate Terms Trade
# Male Secondary Education
# Political Instability
# Human Capital
```

2. We continue the analysis of problem 1.
   a. Fit a LASSO model to the `barro` data set.
   b. Plot the LASSO coefficient profiles. Which variable does LASSO hold the most important?
   c. Use 10-fold cross validation to choose a suitable value for the parameter λ. Which variables are included in the corresponding model?

```
library(glmnet)
library(plotmo)

# a.
barro_lasso <- glmnet(as.matrix(barro[, 2:14]), as.matrix(barro[, 1]), alpha = 1)

# b.
plot_glmnet(barro_lasso, xvar = "lambda", label = TRUE)

# Reading from right to left, the variable (blakp2 (Black Market Premium) enters the model first and is thus the most import
ant. The second most important would be ly2 (Investment/GDP)

# c.
# There is an element of randomness in the results of the cross validation (the division of the data into the folds is done
randomly) and thus we fix the random seed below.
set.seed(08032019)
barro_cv <- cv.glmnet(as.matrix(barro[, 2:14]), as.matrix(barro[, 1]), alpha = 1, nfolds = 10)

# The following plot shows the optimal choice of lambda and the highest value of lambda for which the prediction error is st
ill within 1 standard error of the prediction error of the optimal lambda (this can still be considered a "reasonably good"
value for the prediction error).
plot_glmnet(barro_lasso, xvar = "lambda", label = TRUE)
abline(v = log(barro_cv$lambda.min), lty = 2)
abline(v = log(barro_cv$lambda.1se), lty = 2)
```




```
# To obtain a more narrow set of variables we take the 1 standard-error lambda and arrive at the model
barro_lasso_1se <- glmnet(as.matrix(barro[, 2:14]), as.matrix(barro[, 1]), alpha = 1, lambda = barro_cv$lambda.1se)
coef(barro_lasso_1se)
```

```
## 14 x 1 sparse Matrix of class "dgCMatrix"
##                        s0
## (Intercept)  0.0233671695
## lgdp2       -0.0160720043
## mse2         0.0070822577
## fse2         .
## fhe2         .
## mhe2         .
## lexp2        0.0300585601
## lintr2      -0.0009228819
## gedy2       -0.1199266661
## ly2          0.0711441466
## gcony2      -0.0765576554
## lblakp2     -0.0279256369
## pol2        -0.0149708518
## ttrad2       0.1132610667
```

```
# Thus LASSO retains 10 variables in total. These include the same 9 as obtained with backward and forward selection and add
itionally also Education/GDP.
```

3. **(Optional)** Investigate how ridge regression and LASSO perform in the presence of ``noise'' variables. That is, simulate data where the response depends linearly on a few explanatory variables but include in the model also several explanatory variables (noise) which are independent of the response. Plot then the ridge and LASSO profiles of the variables. Do the profiles of the noise variables perform as expected?

MS-C1620 Statistical Inference

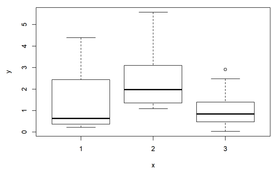
Exercise 12

Homework exercise

To be solved at home before the exercise session.

1. Consider a data set with measurements of the variable `y` for three groups (`x`). Each group has sample size 15. Below are shown boxplots of the groups, along with outputs given by ANOVA and the Kruskal-Wallis test for the data.
   a. What are the conclusions of the two tests?
   b. Which test (if either) would you trust and why?
   c. How would you continue the analysis?

```
boxplot(y ~ x, data = my_data)
```




```
summary(aov(y ~ x, data = my_data))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## x            1   1.13   1.129   0.586  0.448
## Residuals   43  82.89   1.928
```

```
kruskal.test(y ~ x, data = my_data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  y by x
## Kruskal-Wallis chi-squared = 10.185, df = 2, p-value = 0.006142
```

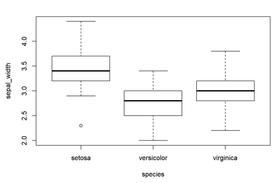
```
# a.
# ANOVA claims that there is no evidence that the expected values of the groups differ.
# K-W claims that the medians of the groups are not all same.

# b.
# The boxplots show evidence that the groups have positively skew distributions, meaning that the normality assumption of ANOVA is unlikely to hold. Moreover, the "replacement" of the normality assumption with a large enough sample size is question able here as we have only 15 obs. per group -> Cannot trust ANOVA.
# K-W requires that the group distributions have the same shape, meaning that the boxplots should look otherwise similar but have possibly different locations in the y-axis. Based on the plot, this seems plausible -> we can trust K-W and thus conclu de that the group medians differ.

# c.
# The analysis could be continued with pair-wise testing using e.g. the two-sample rank test to find out which pairs of grou ps have differing medians (accompanied with a suitable correction, such as the Bonferroni correction).
```

Class exercise

To be solved at the exercise session.

1. A botanist wants to test the hypothesis that the three iris species have equal expected value of `Sepal.Width`.
   a. Visualize the data.
   b. Conduct an analysis of variance.
   c. Are the assumptions of ANOVA satisfied?
   d. If the assumptions are fulfilled, conduct pairwise comparisons using the Bonferroni correction.
   e. State your conclusions.

```
# a.
flowers <- data.frame(sepal_width = iris[, 2], species = iris[, 5])

# The boxplots show that at least the group "Setosa" seems to differ from the others
boxplot(sepal_width ~ species, data = flowers)
```



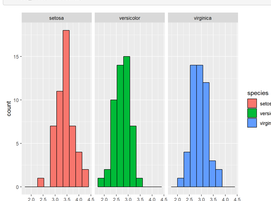

```
# b.
# ANOVA finds differences in the group expected values, p-value < 0.05
# -> Not plausible that the expected values are the same (given that the assumptions of ANOVA hold).
flowers_aov <- aov(sepal_width ~ species, data = flowers)
summary(flowers_aov)
```

```
##              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
```

```
# c.
# Bartlett's test shows no evidence that the variances would differ.
bartlett.test(sepal_width ~ species, data = flowers)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  sepal_width by species
## Bartlett's K-squared = 2.0911, df = 2, p-value = 0.3515
```

```
# Based on the histograms, the group-wise normality assumption seems plausible.
library(ggplot2)
ggplot(flowers, aes(x = sepal_width, fill = species)) +
  geom_histogram(bins = 10, col = "black") +
  facet_wrap(. ~ species)
```

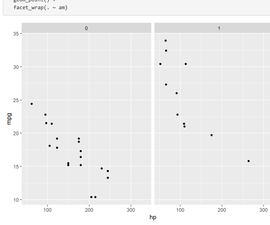



```
# -> we conclude that the assumptions of ANOVA are fulfilled.


# d. & e.:
# Pairwise testing with the Bonferroni correction reveals that all three groups differ
# from each other in terms of expected values.
pairwise.t.test(flowers$sepal_width, flowers$species, p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  flowers$sepal_width and flowers$species
##
##            setosa  versicolor
## versicolor < 2e-16 -
## virginica  1.4e-09 0.0094
##
## P value adjustment method: bonferroni
```

2. The data set `mtcars` has measurements for 32 cars. We investigate the relationship between `mpg` (miles/gallon, the response) and `hp` and `am` (horsepowers and transmission type, the explanatory variables) through an analysis of covariance.
   a. Find a suitable visualization for the data.
   b. Using the function `lm`, fit a regression model with the covariates `hp`, `am` and `hp:am` (the final one is an interaction effect, the product of the two covariates).
   c. Interpret the fitted model (homework problem 10.1.a might prove helpful).

```
#a.
cars <- data.frame(mpg = mtcars$mpg, hp = mtcars$hp, am = mtcars$am)

# Scatter plots of mpg and hp for each level of am
# plot(mpg ~ hp, data = cars[cars$am == "0", ])
# plot(mpg ~ hp, data = cars[cars$am == "1", ])

# Or more succinctly in ggplot
library(ggplot2)

ggplot(cars, aes(x = hp, y = mpg)) +
  geom_point() +
  facet_wrap(. ~ am)
```




```
# Some questions evoked by the plot:
# 1. Is mpg on average higher for am == 1?
# 2. Is the relationship between mpg and hp linear?
# 3. Are the slopes of mpg ~ hp different for different types of transmission?


# b.
cars_lm <- lm(mpg ~ hp*am, data = cars)
summary(cars_lm)
```

```
##
## Call:
## lm(formula = mpg ~ hp * am, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3818 -2.2696  0.1344  1.7058  5.8752
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.6248479  2.1829432  12.197 1.81e-12 ***
## hp          -0.0591370  0.0129493  -4.568 9.02e-05 ***
## am           5.2176534  2.6650931   1.958   0.0603 .
## hp:am        0.0004029  0.0166602   0.024   0.9806
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.961 on 28 degrees of freedom
## Multiple R-squared:  0.782,  Adjusted R-squared:  0.7587
## F-statistic: 33.49 on 3 and 28 DF,  p-value: 2.132e-09
```

```
# c.
# As in homework problem 10.1.a, we write the model separately for am == 0 and am == 1:
#
# am == 0:
# E[mpg] = b0 + b1*hp
#
# am == 1:
# E[mpg] = b0 + b1*hp + b2 + b3*hp = (b0 + b2) + (b1 + b3)*hp
#
# where the b's and the model output above have the following correspondences:
# b0 = "(Intercept)", b1 = "hp", b2 = "am", b3 = "hp:am"

# We interpret the coefficients b1, b2, b3 each in turn:
#
# b3    describes the difference of the hp-slopes for the two transmission types.
#       p-value is almost 1 and we conclude that it is plausible that b3 = 0
#       -> the slopes do not differ from each other (the horsepowers do not affect mpg
#       differently for the two types of transmission).
#
# b2    describes the slope of the group with am == 0 (but also of the group with am ==
#       1, since we now believe that b3 = 0). p-value is below 0.05 so the slope differs
#       significantly from zero -> we conclude that a unit increase in horsepowers
#       lowers mpg by -0.06.
#
# b2    describes the difference of the intercept terms for the two transmission types.
#       p-value >= 0.05 and we conclude that it is plausible that b2 = 0.
#       -> the lines do not differ in their vertical position (that is, even though the
#       points of am == 1 are higher in the plot, we established that there is
#       not enough evidence to show that this effect is not caused by randomness).
```

3. **(Optional)** Consider still the `mtcars` data set but replace the variable `am` with the variable `gear` (and make sure its type is `factor`). Fit the linear regression model `mpg ~ hp + gear` and find out how the function `anova` can be used to test whether all regression coefficients related to `gear` are equal to zero **simultaneously**. *Note that the situation is different from problem 2 as `gear` has three classes (i.e., two coefficients) and thus the p-values from the model only relate to the hypotheses whether the two coefficients can be set to zero individually.*