Classical and Bayesian Statistics

Problems 1

A note on these problem: These are mostly about getting to know how R works. It is *not* the idea that you have to understand or memorize all these commands. You will see the important commands over and over again.

Problem 1.1

(∗∗) a) The following table lists the civilian deaths in WW1 and WW2. We consider the Allied death in WW2. What is *very* problematic about that part of the table?

Hint: Consider the total number of death.

The problematic part about this table is that there is the total number of deaths regardless of the population of the country.
One other problematic thing is that for some countries there are missing data and "an enermous number" enormous compared to what? or "Large number"? what is considered a large number in this case?

CIVILIANS

(a) World War I - Not known

(b) World War II

Allied

United Kingdom	60,595
Belgium	90,000
China	An enormous number
Denmark	Unknown
France	152,000
Netherlands	242,000
Norway	3,638
U.S.S.R.	6,000,000
	6,548,233

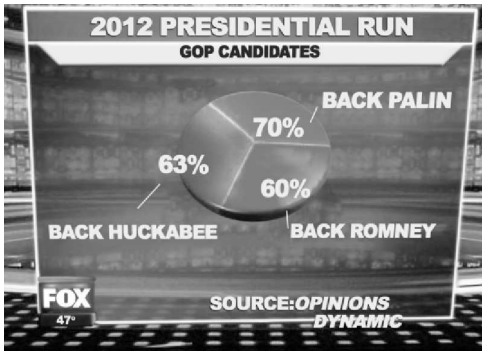
Enemy

Germany	800,000
Austria	125,000
Italy	180,000
Japan	600,000
Poland	5,000,000
Yugoslavia	Large number
	6,705,000

(∗∗) b) In the following figure, a presidential election prediction in the USA is shown in a so-called *pie chart*.

What is problematic about this specific graphic?

The percentages in this pie chart add up to 193% instead of a 100%
as well as, in the cases of this kind of polls I would also like to know how many people were interviewed



(∗∗) c) A little bit of critical thinking[1].

A fictitious lady called Linda is described as follows: Linda is thirty-one years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations.

Which alternative is more probable?

i) Linda is a bank teller.

ii) Linda is a bank teller and is active in the feminist movement.

Explain your answer.

(∗ ∗ ∗) d) A true story from WW II[2].

So here's the question. You don't want your planes to get shot down by enemy fighters, so you armor them. But armor makes the plane heavier, and heavier planes are less maneuverable and use more fuel. Armoring the planes too much is a problem; armoring the planes too little is a problem. Somewhere inbetween there's an optimum.

The military came to the SRG (Statistical research group; a highly influential US think tank in WWII) with some data they thought might be useful to find this optimum. When American planes came back from engagements over Europe, they were covered in bullet holes. But the damage wasn't uniformly distributed across the aircraft. There were more bullet holes in the fuselage, not so many in the engines (see Table 1).

The officers saw an opportunity for efficiency; you can get the same protection with less armor if you concentrate the armor on the places with the greatest

[1] From: *Thinking, fast and slow*; Daniel Kahneman
[2] From: *How not to be wrong*; Jordan Ellenberg

Section of the plane	Bullet holes per square foot
Engine	1.11
Fuselage	1.73
Fuel system	1.55
Rest of the plane	1.8

Table 1: Bullets in plane

need, where the planes are getting hit the most, i.e., the fuselage and asked the SRG But exactly how much more armor belonged on those parts of the plane.

The officers didn't get the answer they expected. They should strengthen the engine they were told. Can you explain why?

Problem 1.2

This task discusses some simple `R` commands. In case of doubt, refer the document Chapter 2 of the lecture notes.

(∗) a) Form a vector `x` with the numbers $4, 2, 1, 3, 3, 5, 7$.

(∗) b) Access with `R` the third value.

(∗) c) Access with `R` the first and fourth value.

(∗) d) Determine the length of the vector `x`.

(∗) e) What does the command `x+2` do? Make a guess first and then execute the command.

(∗) f) What does the command `sum(x+2)` do? Make a guess first and then execute the command.

(∗∗) g) What does the command `x <= 3` do? Make a guess first and then execute the command.

(∗∗) h) What does the command `x[x <= 3]` do? Make a guess first and then execute the command.

(∗) i) What does the command `sort(x)` do? Make a guess first and then execute the command.

(∗ ∗ ∗) j) What does the command `order(x)` do? Make a guess first and then execute the command. Compare the values of `order(x)` with the values of `x`.

(∗∗) k) We want to replace the value of the 4th entry of `x` with the number 8. How do you do this?

Problem 1.3

This task deals with the dataset **`weather.csv`**, which we got to know in the introduction.

In case of doubt, please refer to Chapter 2 of the lecture notes.

(∗) a) Load the dataset and save it under the variable `data`.

(∗) b) Select the value of the second row and third column.

(∗) c) Select the 4th row.

(∗) d) Select the 1st and 4th column. Use the respective column names.

(∗∗) e) Save the above data under the name `data1` and save it as a file under the name **`weather2.csv`**.

(∗∗) f) How can you find out (with `R` of course) which is the name of the 3rd column?

(∗∗) g) We want to replace the columns `Basel` with `Geneva`. How would you proceed?

h) We consider the command

```
data3 <- data [order(data[, 'Zurich']), ]
```

This creates

```
data3 <- data [order(data[, 'Zurich']), ]
data3

##     Luzern Basel Chur Zurich
## Feb      5     6    1      0
## Jan      2     5   -3      4
## Mar     10    11   13      8
## Apr     16    12   14     17
## May     21    23   21     20
## Jun     25    21   23     27
```

(∗∗) i) If you look at the table, what does this command do?

(∗ ∗ ∗) ii) Explain why this command has this effect.

Problem 1.4

The following temperatures are given in degrees Fahrenheit (°F)

$$51.9, \quad 51.8, \quad 51.9, \quad 53$$

(∗) a) Form a vector `fahrenheit` with these values.

(∗) b) Covert these temperatures into degrees Celsius (°C). The conversion formula is

$$C = \frac{5}{9}(F - 32)$$

Form a vector `celsius`.

(∗) c) Given are further temperatures

$$48. \quad 48.2. \quad 48. \quad 48.7$$

Determine the difference to the original temperatures. Again, use vectors.

Problem 1.5

Given are the weight of 6 people (kg)

$$60,\ 72,\ 57,\ 90,\ 95,\ 72$$

and the body height (in m)

$$1.75,\ 1.80,\ 1.65,\ 1.90,\ 1.74,\ 1.91$$

We want to calculate the Body Mass Index (BMI) which is defined by

$$\text{BMI} = \frac{\text{weight}}{\text{height}^2}$$

(∗) a) Create two vectors `weight` and `height` with the respective values.

(∗∗) b) Calculate the BMI of these 6 persons simultaneously. Create a vector `bmi` for this purpose.

Problem 1.6

In task 1.2 we encountered the command `order(...)`, whose output is not obvious at first sight. If we want to know what this command does, we can consult the `R` help function. Unfortunately, this help function is *not* very helpful for beginners. We *can* use `?...` or `help(...)` to query a command, for example `?order` or `help(order)`.

```
help(order)
```

order {base} R Documentation

Ordering Permutation

Description

`order` returns a permutation which rearranges its first argument into ascending or descending order, breaking ties by further arguments. `sort.list` is the same, using only one argument.
See the examples for how to use these functions to sort data frames, etc.

Usage

```
order(..., na.last = TRUE, decreasing = FALSE,
      method = c("auto", "shell", "radix"))

sort.list(x, partial = NULL, na.last = TRUE, decreasing = FALSE,
          method = c("auto", "shell", "quick", "radix"))
```

Arguments

`...`	a sequence of numeric, complex, character or logical vectors, all of the same length, or a classed `R` object.
`x`	an atomic vector.
`partial`	vector of indices for partial sorting. (Non-`NULL` values are not implemented.)
`decreasing`	logical. Should the sort order be increasing or decreasing? For the `"radix"` method, this can be a vector of length equal to the number of arguments in `...`. For the other methods, it must be length one.
`na.last`	for controlling the treatment of `NA`s. If `TRUE`, missing values in the data are put last; if `FALSE`, they are put first; if `NA`, they are removed (see 'Note'.)
`method`	the method to be used: partial matches are allowed. The default (`"auto"`) implies `"radix"` for short numeric vectors, integer vectors, logical vectors and factors. Otherwise, it implies `"shell"`. For details of methods `"shell"`, `"quick"`, and `"radix"`, see the help for `sort`.

This is part of the output (it doesn't get any more understandable below, try it). But the result is usually more confusing than helpful for beginners.

Far more helpful is the use of **`Google`**. If you enter the search term **`r order`**, you will quickly find an explanation that is also useful, for example

`https://www.datacamp.com/community/tutorials/sorting-in-r?utm_source=adwords_ppc&utm_campaignid=898687156&utm_adgroupid=4894725`

In general, `www.datacamp.com` is a treasure trove for `R` and `Python`.

(**) a) A command that occurs frequently is the `seq(...)` command.

Google this command with search words like **`r seq examples`**. What does this command and explain how this command works with the options `by` and `length.out`.

b) Given is the vector

```
x <- c(4, 10, 3, NA, NA, 1, 8)
```

First a remark about the value `NA` (not available). These stand for missing data, which are not available for some reason which happens quite often in statistics.

(**) i) If we try to calculate the mean of `x` (command `mean(x)`), the result is `NA`. Can you explain why this is the case?

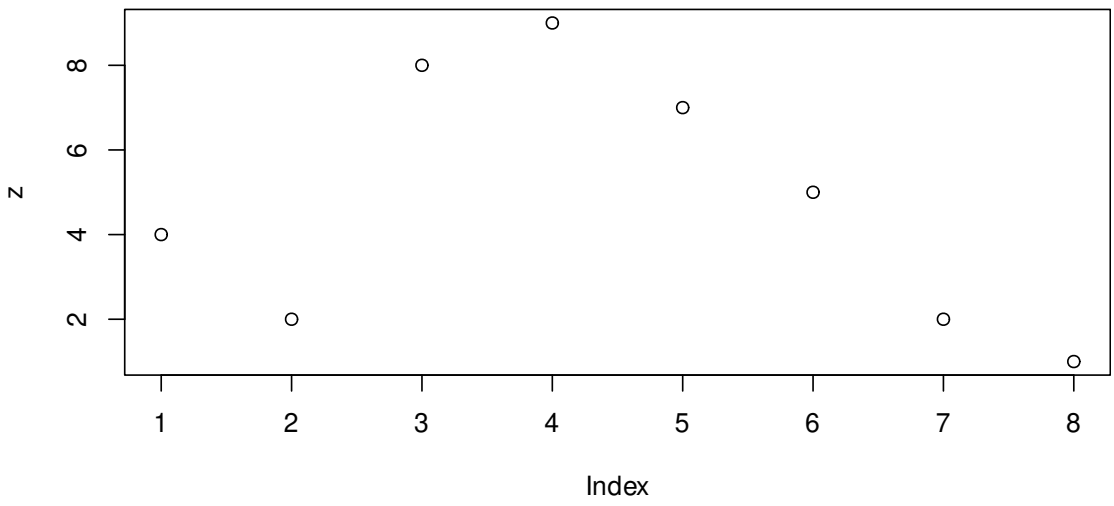
(**) ii) How can you calculate the mean value of all *occurring* values? Again, use **`Google`**.

(**) iii) Apply the commands `sort(...)` and `order(...)` to the list `x`. What do these commands do?

In both of them, the two options `na.last = ...` and `decreasing = ...` (among others) are available, which can be set to `TRUE` (or `T`) or `FALSE` (or `F`). What is the effect of these options?

c) Plots play an important role in statistics. The following plot is very easy to create, but it looks a bit boring and colourless.

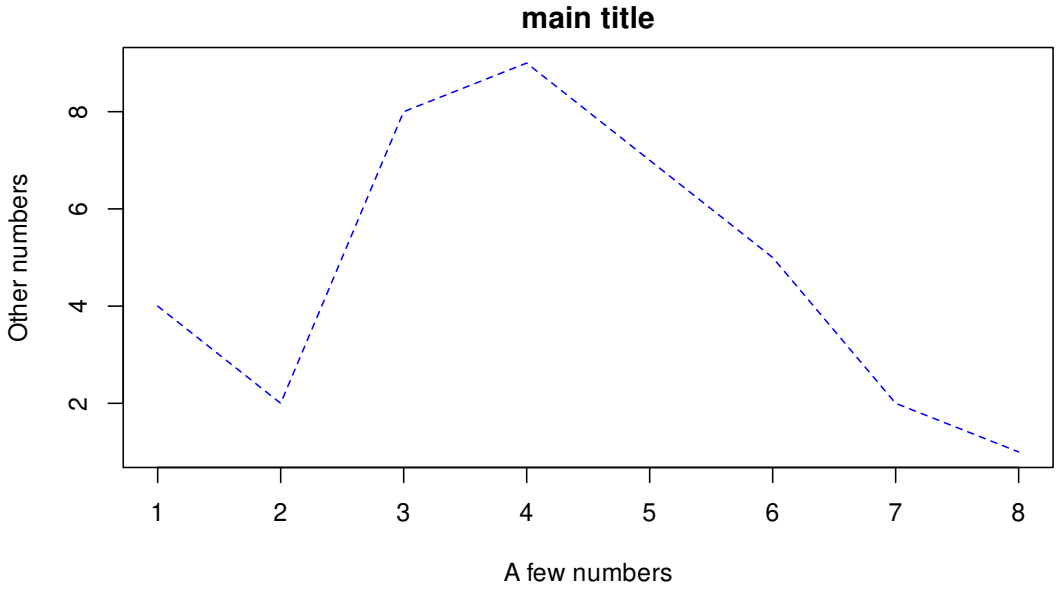
```
z <- c(4, 2, 8, 9, 7, 5, 2, 1)

plot(z)
```



(*) i) Change in the following command the parameters of the options and describe the effect of these options (especially `type` and `lty`, the others should be obvious). Use **`Google`**.

```
plot(z,
     type = "l",
     col = "blue",
     lty = 2,
     main = "main title",
```

```
     xlab = "A few numbers",
     ylab = "Other numbers"

)
```

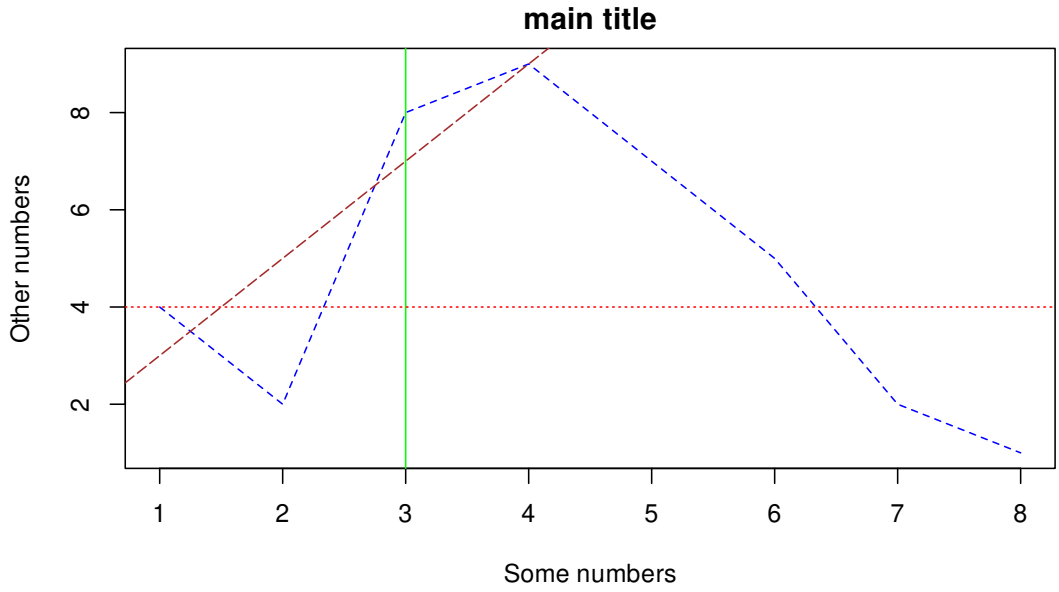


(∗∗∗) ii) Use the command `abline(...)` to add three lines to the graph above (see figure below).

- The vertical straight line $x = 3$, solid, green.
- The horizontal straight line $y = 4$, dotted, red.
- The line $y = 2x + 1$, dashed with long dashes, brown.

```
plot(z,
     type = "l",
     col = "blue",
     lty = 2,
     main = "main title",
     xlab = "A few numbers",
     ylab = "Other numbers

)

abline(...)
abline(...)
abline(...)
```

It is important in this case that the `plot()` and `abline()` are executed together.



Problem 1.7

The dataframe **d.fuel.dat** contains the data of various vehicles from an American investigation of the 1980s. Each row contains the data of one vehicle (one vehicle corresponds to one observation).

(∗) a) Load the file **d.fuel.dat** stored on **ILIAS** with the following R command

```
d_fuel <- read.table(file = "d.fuel.dat" ,header = T, sep = ",")[, -1]
```

The option `sep = ","` is needed, because the columns in the file **d.fuel.dat** are separated by commas. In the file **d.fuel.dat** the lines are numbered and therefore the first column contains the number of the line which we ignore with `[, -1]`

The columns contain the following variables:

weight:	weight in pounds (1 pound = 0.453 59 kg)
mpg:	Range in Miles Per Gallon (1 gallon = 3.789 L; 1 mile = 1.6093 km)
type:	Car type

(∗) b) Select only the fifth row of the dataframe **d.fuel**. Which values are in the fifth row?

(∗) c) Select now the first to fifth observation of the dataframe **d.fuel** By the way, this way you can get a quick overview of the type of dataframe for an unknown dataset.

(∗∗) d) Simultaneously display the 1st to 3rd and the 57th to 60th rows of the dataframe.

(∗) e) Calculate the mean of the ranges of all cars in miles/gallon.

(∗∗) f) Calculate the average value of the range of cars 7 to 22.

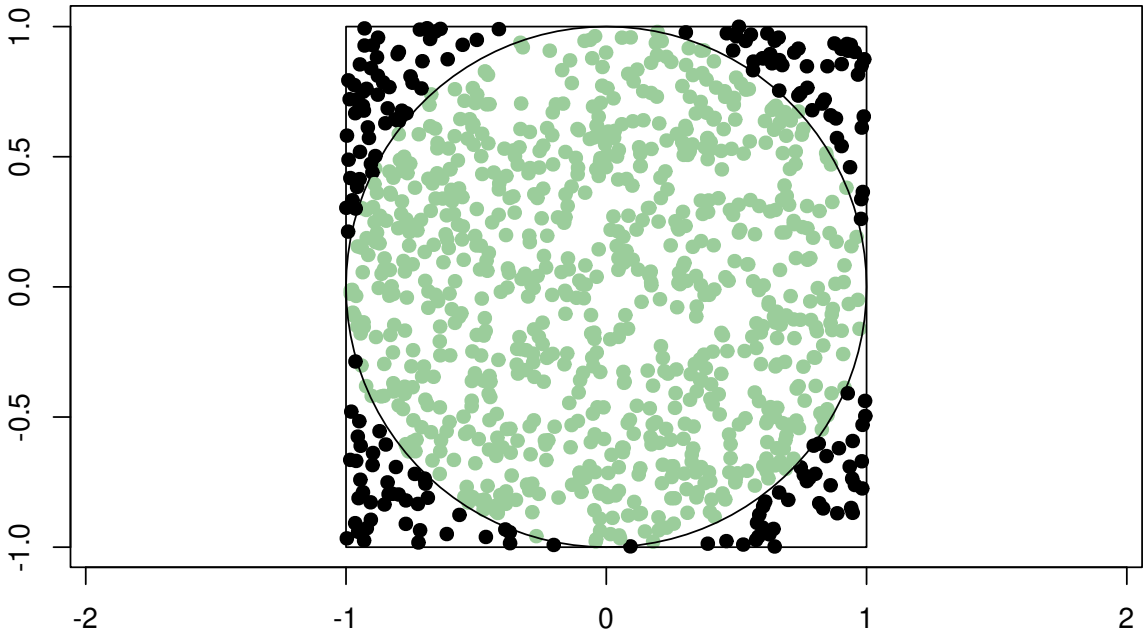
(∗∗) g) Create a new vector `t_kml`, which contains all ranges in km/L, and a vector `t_kg`, which contains all weights in kg.

(∗∗) h) Calculate the mean of the ranges in km/L and that of the vehicle weights in kg.

Problem 1.8

(∗ ∗ ∗) This problem is probably a little bit daunting for beginners in R. If that is the case, you might skip this problem and come back later when you have more experience in R.

We want to explain the code for the approximation of π which is 4 times the ratio of the number of green dots to the number of total dots (see Chapter 1 of the lecture notes).




```
## Approximation of Pi: 3.236
```

The essential part is

```
n <- 1000

x <- runif(n, min = -1, max = 1)
y <- runif(n, min = -1, max = 1)

r_squ <- x^2 + y^2

pi <- 4* sum(r_squ < 1) / n

pi


## [1] 3.224
```

To explain the commands, we choose just 10 dots. The approximation will be bad, but that is not the point. Let $n = 10$ be the total number of points.

```
n <- 10
```

(***) a) The first unknown command is `runif(n, min = -1, max=1)`. Run this command several times.

```
runif(n, min = -1, max=1)

##  [1] -0.4689827 -0.2557522  0.1457067  0.8164156 -0.5966361  0.7967794
##  [7]  0.8893505  0.3215956  0.2582281 -0.8764275
```

What do you observe?

(***) b) `x` contains the x-coordinates of the 10 points and `y` the y-coordinates.

We have to check whether a point lies within the circle. The circle has the equation

$$x^2 + y^2 = 1$$

so a point is within the circle, if

$$x^2 + y^2 < 1$$

```
x <- runif(n, min = -1, max = 1)
y <- runif(n, min = -1, max = 1)

r_squ <- x^2 + y^2

r_squ

##  [1] 1.10167835 0.74990985 0.23192991 0.61456374 0.50800244 0.05190117
##  [7] 1.13658713 1.02321680 0.60425152 0.40985717
```

Can you explain, what

```
r_squ < 1

##  [1] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE
```

does? Hint: Compare this vector with `r_squ`.

(**) c) Explain, what the command

```
sum(r_squ < 1)

## [1] 7
```

does? Hint: Compare the value with `r_squ < 1`.

(**) d) Explain

```
pi <- 4 * sum(r_squ < 1) / n

pi

## [1] 2.8
```

e) Finally the code of the plot above, which is for those who really want to know it right know (not explained).

```
library(plotrix)
n <- 1000
x <- runif(n, min = -1, max = 1)
y <- runif(n, min = -1, max = 1)

r_squ <- x^2 + y^2

plot(x,y,
     col = ifelse(r_squ < 1,"darkseagreen3",  "black"),
     asp = 1,
     pch = 19,
     xlim=c(-1.000,1.0000))
lines(c(-1,-1,1,1,-1),c(-1,1,1,-1,-1))
draw.circle(0,0,1)



cat("Approximation of Pi:", sum(r_squ < 1) / n * 4)
```

Classical and Bayesian Statistic

Sample solution for Problems 1

Solution 1.1

a) The main problem is that the total number of death is accurate to the dead, but only of the *known* numbers. The dead of China with "An enormous number" and Denmark with "Unknown" were not included. Accordingly, the total number cannot be correct, especially not accurate to the dead.

   Another problem is the indication of the numbers. Norway had 3638 death, an exact figure, while the Soviet Union had 6 000 000 death, which must be an estimate.

   Giving the total number *accurate* to the individual dead makes no sense. This is a pretence of accuracy that is not existing.

b) The percentages do not add up to 100 %, but almost 200 %. The reason for this is that the participants in the poll could indicate more than one candidate. However, this creates the problem that the candidates are very difficult to compare with each other.

c) Reading the description of Linda one would expect that the second answer is correct, because it "fits" the description better.

   However, this argument does not hold. The description tells us nothing about Linda's job, so we have to rely on the whole population and there are more bank tellers than banktellers, which are *also* active in the feminist movement.

d) The SRG's insight was simply to ask: where are the *missing* holes? The ones that would have been all over the engine casing, if the damage had been spread *equally* all over the plane (the numbers give are per square foot and should be the same all over the plane)?

   The missing bullet holes were on the missing planes. The reason planes were coming back with fewer hits to the engine is that planes that got hit in the engine weren't coming back. Whereas the large number of planes returning to base with a thoroughly Swiss-cheesed fuselage is pretty strong evidence that hits to the fuselage can (and therefore should) be tolerated.

   If you go the recovery room at the hospital, you'll see a lot more people with bullet holes in their legs than people with bullet holes in their chests. But that's not because people don't get shot in the chest; it's because the people who get shot in the chest don't recover.

In cognitive science, this phenomenon is called the *survival bias* and is everywhere in our thinking. Recently, the Covid vaccination opponents have said that the vaccination does not work because they know someone vaccinated who has contracted Covid-19. No word about the very high proportion of the population where the vaccin *did* work.

Solution 1.2

a) Vectors are created with the command `c(...)`:

```
x <- c(4, 2, 1, 3, 3, 5 ,7)
```

b)
```
x[3]

## [1] 1
```

c)
```
x[c(1, 4)]

## [1] 4 3
```

d)
```
length(x)

## [1] 7
```

e)
```
x+2

## [1] 6 4 3 5 5 7 9
```

Adds 2 to each component of `x`.

f)
```
sum(x+2)

## [1] 39
```

Add up all entries of `x+2`.

g)
```
x <= 3

## [1] FALSE  TRUE  TRUE  TRUE  TRUE FALSE FALSE
```

The command creates a vector of the length of `x`. For all values less than 3 in `x` the "value" of the new vector is `TRUE`, for the others `FALSE`.

h)
```
x[x <= 3]

## [1] 2 1 3 3
```

The construction `x[...]` selects elements from the vector `x`. In this case we select the values of `x` for which `x <= 3` from g) is `TRUE`.

i)
```
sort(x)

## [1] 1 2 3 3 4 5 7
```

The values of x are ordered in ascending order of size.

j)
```
order(x)

## [1] 3 2 4 5 1 6 7

x

## [1] 4 2 1 3 3 5 7
```

At first glance, it is not quite obvious what this command does. For example, the value 6 occurs in `order(x)`, but not in x. If we stare at (or think about) the vectors long enough, we may, perhaps, recognise the following pattern:

- The first value of `order(x)` is 3. If we consider the 3rd entry in x, it is 1.
- The second value of `order(x)` is 2. The 2nd entry of x is 2.
- The third value of `order(x)` is 4. The 3rd entry of x is 3.
- The 4th value of `order(x)` is 5. The 4th entry of x is 3.
- etc. ...

The command `order(x)` specifies the places *where* the values of x in ascending order. Or in other words, it produces the *ranking* of the values of x: The 3rd value is the smallest of x, the 2nd is the second smallest, and so on.

What do we do if we do not recognise this pattern? See task 1.6.

k)
```
x[4] <- 8
x

## [1] 4 2 1 8 3 5 7
```

Solution 1.3

a)
```
setwd("../../../Software_R_Python/R/")
data <- read.csv("weather.csv")
data

##     Luzern Basel Chur Zurich
## Jan      2     5   -3      4
## Feb      5     6    1      0
## Mar     10    11   13      8
## Apr     16    12   14     17
## May     21    23   21     20
## Jun     25    21   23     27
```

Of course your path will be different. For Windows users: You have to replace the \ with /.

b)
```
data [2, 3]

## [1] 1
```

Again: The first value within the square brackets of `data[..., ...]` *always* refers to the rows and the second value to the columns of the dataset.

c)
```
data[4, ]

##     Luzern Basel Chur Zurich
## Apr     16    12   14     17
```

d)
```
data[, c("Luzern", "Zurich")]

##     Luzern Zurich
## Jan      2      4
## Feb      5      0
## Mar     10      8
## Apr     16     17
## May     21     20
## Jun     25     27
```

e)
```
data1 <- data[, c("Luzern", "Zurich")]
write.csv(data1, "weather2.csv", row.names=F)

data2 <- read.csv("weather2.csv")
data2

##   Luzern Zurich
## 1      2      4
## 2      5      0
## 3     10      8
## 4     16     17
## 5     21     20
## 6     25     27
```

f)
```
colnames(data)[3]

## [1] "Chur"
```

The command `colnames(data)` creates a vector with the column names of the dataset `data` and `...[3]` selects the third entry.

g)
```
colnames(data)[2] <- "Geneva"
data

##     Luzern Geneva Chur Zurich
## Jan      2      5   -3      4
## Feb      5      6    1      0
## Mar     10     11   13      8
## Apr     16     12   14     17
```

```
## May     21     23   21     20
## Jun     25     21   23     27
```

If we don't want to look up columns, the following is a good option.

```
data <- read.csv("../../../Software_R_Python/R/weather.csv")
data

##     Luzern Basel Chur Zurich
## Jan      2     5   -3      4
## Feb      5     6    1      0
## Mar     10    11   13      8
## Apr     16    12   14     17
## May     21    23   21     20
## Jun     25    21   23     27

colnames(data)[which("Basel" == colnames(data))] <- "Geneva"
data

##     Luzern Geneva Chur Zurich
## Jan      2      5   -3      4
## Feb      5      6    1      0
## Mar     10     11   13      8
## Apr     16     12   14     17
## May     21     23   21     20
## Jun     25     21   23     27
```

All `Basel` entries would be replaced by `Geneva`.

`"Basel" == colnames(data)` produces a vector with `TRUE` and `FALSE` entries. At each position where "Basel" occurs, `TRUE` is set, otherwise `FALSE`.

The command `which("Basel" == colnames(data))` determines all places where `Basel` occurs in `colnames(data)`. These are the locations where `Basel` is replaced by `Geneva`.

`==` is the logical equal and not an assignment. We only check if the values are equal.

Translated with www.DeepL.com/Translator (free version)

h)
```
data3 <- data [order(data[, 'Zurich']), ]
data3

##     Luzern Geneva Chur Zurich
## Feb      5      6    1      0
## Jan      2      5   -3      4
## Mar     10     11   13      8
## Apr     16     12   14     17
## May     21     23   21     20
## Jun     25     21   23     27
```

i) i) The table is sorted in ascending order according to the values of Zurich.

ii) The first entry for `data` contains `order(data[, 'Zurich'])`. This specifies the order of the column `Zurich`.

```
order(data[, 'Zurich'])

## [1] 2 1 3 4 5 6
```

Thus the rows of `data` are ordered by this command.

Solution 1.4

a) Vector `fahrenheit`:

```
fahrenheit <- c(51.9, 51.8, 51.9, 53)

fahrenheit

## [1] 51.9 51.8 51.9 53.0
```

b) Temperatures in degrees Celsius (°C):

```
celsius <- 5/9 * (fahrenheit - 32)

celsius

## [1] 11.05556 11.00000 11.05556 11.66667
```

c) Other temperatures:

```
fahrenheit_2 <- c(48, 48.2, 48, 48.7)

fahrenheit_3 <- fahrenheit - fahrenheit_2

fahrenheit_3

## [1] 3.9 3.6 3.9 4.3
```

Solution 1.5

a)
```
weight <- c(60, 72, 57, 90, 95, 72)
height <- c(1.75, 1.80, 1.65, 1.90, 1.74, 1.91)
```

b) In `R` we achieve this through

```
bmi <- weight / height^2

bmi

## [1] 19.59184 22.22222 20.93664 24.93075 31.37799 19.73630
```

We have calculated the BMI for all 6 people in one go!

Solution 1.6

a) The command `seq(...)` forms a sequence of numbers starting with the first number (`from = ...`) and ending with the second number (`to = ...`), if possible. The increment is indicated by `by = ...`.

```
seq(from = 3, to = 10, by = 2)

## [1] 3 5 7 9
```

or:

```
seq(3, 10, 2)

## [1] 3 5 7 9
```

The number 10 is not included, because it does not appear at all in the enumeration.

The option `length.out` specifies how *many* numbers (instead of the increment) are formed between the start and the end of the sequence, all in equal distance from one to the next.

```
seq(from = 3, to = 10, length.out = 10)

##  [1]  3.000000  3.777778  4.555556  5.333333  6.111111  6.888889
##  [7]  7.666667  8.444444  9.222222 10.000000
```

Or

```
seq(3, 10, length.out = 10)

##  [1]  3.000000  3.777778  4.555556  5.333333  6.111111  6.888889
##  [7]  7.666667  8.444444  9.222222 10.000000
```

The options `by` and `length.out` do *not* "like" each other, since the information is usually contradictory and an error message occurs

```
seq(from = 3, to = 10, length.out = 10, by = 2)

## Error in seq.default(from = 3, to = 10, length.out = 10, by = 2):  too many arguments
```

Remark:

Although the command

```
seq(3, 10, 2)

## [1] 3 5 7 9
```

is perfectly acceptable, this notation is *not* recommended. First of all, if you use this notation, you have to stick to the order of the arguments, in this case `from`, `to` and `by`. Secondly, it makes the code far less readable.

b) i) The `mean` command does not make much sense in this case, because `R` tries to take the mean from *all* values and of course this does not work with the `NA`'s.

ii) However, we can use the option `na.rm = TRUE` (default is `FALSE`) will cause the `NA`'s to be ignored (`.rm` stands for *remove*).

```
mean(x, na.rm = TRUE)

## [1] 5.2
```

iii) The `sort` command is the easier one of the two to explain.

```
sort(x)

## [1]  1  3  4  8 10
```

This command sorts (as the name suggests) the existing numbers in ascending order.

But if we want to sort in descending order, we use the option `decreasing = TRUE` (default is `FALSE`).

```
sort(x, decreasing = TRUE)

## [1] 10  8  4  3  1
```

However, the `NA` values were ignored. If we want to include them, we select the option `na.last = TRUE`.

```
sort(x, decreasing = TRUE, na.last = TRUE)

## [1] 10  8  4  3  1 NA NA
```

If we want the `NA`'s at the beginning, we set this option `FALSE`.

```
sort(x, decreasing = TRUE, na.last = FALSE)

## [1] NA NA 10  8  4  3  1
```

We will now take look at the `order` command.

```
order(x)

## [1] 6 3 1 7 2 4 5
```

Let's look at the original list

```
x

## [1]  4 10  3 NA NA  1  8
```

We see that the 6th number is the smallest, the 3rd number the second smallest, and so on. The `NA`'s (4th and 5th number) are at the end.

The options `decreasing = ...` and `na.last = ...` work here in the same way as the `sort` command.

c) i) The options `main = "..."`, `col = "..."`, `xlab = "..."` and `ylab = "..."` should be obvious.

The option `type = "..."` specifies the line type. See also

`https://www.dummies.com/programming/r/how-to-create-different-plot-types-in-r/`

The option `lty = "..."` specifies the line type for "solid" lines. See also

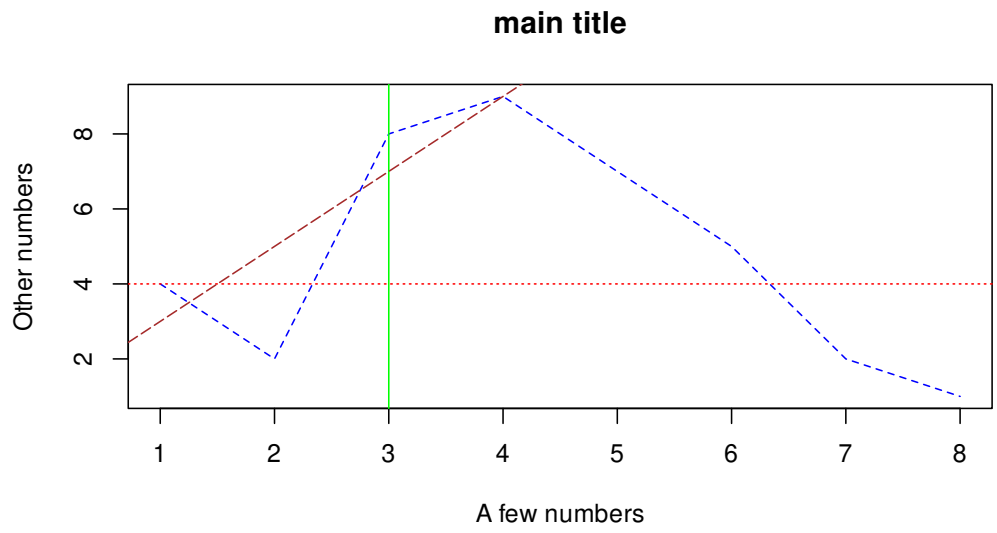
`http://www.sthda.com/english/wiki/line-types-in-r-lty`

For the color palette in `R` see

`http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf`

The code:

```
plot(z,
     type = "l",
     col = "blue",
     lty = 2,
     main = "main title",
     xlab = "A few numbers",
     ylab = "Other numbers")

abline(v = 3, col = "green", lty = 1)
abline(h = 4, col = "red", lty = 3)
abline(a = 1, b = 2, col = "brown", lty = 5)
```



The argument `v = 3` in `abline(...)` stands for *vertical*. `R` draws a vertical line at $x = 3$.

Similarly, the argument `h = 4` in `abline(...)` stands for *horizontal*. `R` draws a horizontal line at $y = 4$.

Now it get's slightly more complicated. The arguments `a = 1, b = 2` in `abline(...)` stand for $a = 1$ the intercept and $b = 2$ the slope of the line, i.e. a line with the equation $y = 1 + 2x = a + bx$.

You might remember from your school days that the equation of a line has the form

$$y = ax + b$$

with a the slope and b the intercept. However, statisticians often use different notations to the ones of the mathematicians and for them the equation of a line has the form

$$y = a + bx$$

with a the intercept and b the slope.

Solution 1.7

a) See task definition.

To view the data in the data frame, type the name of the object

```
d_fuel
```

```
##    weight mpg    type
## 1    2560  33   Small
## 2    2345  33   Small
## 3    1845  37   Small
## 4    2260  32   Small
## 5    2440  32   Small
## 6    2285  26   Small
## 7    2275  33   Small
## 8    2350  28   Small
## 9    2295  25   Small
## 10   1900  34   Small
## 11   2390  29   Small
## 12   2075  35   Small
## 13   2330  26   Small
## 14   3320  20  Sporty
## 15   2885  27  Sporty
## 16   3310  19  Sporty
## 17   2695  30  Sporty
## 18   2170  33  Sporty
## 19   2710  27  Sporty
## 20   2775  24  Sporty
```

```
## 21   2840  26  Sporty
## 22   2485  28  Sporty
## 23   2670  27 Compact
## 24   2640  23 Compact
## 25   2655  26 Compact
## 26   3065  25 Compact
## 27   2750  24 Compact
## 28   2920  26 Compact
## 29   2780  24 Compact
## 30   2745  25 Compact
## 31   3110  21 Compact
## 32   2920  21 Compact
## 33   2645  23 Compact
## 34   2575  24 Compact
## 35   2935  23 Compact
## 36   2920  27 Compact
## 37   2985  23 Compact
## 38   3265  20  Medium
## 39   2880  21  Medium
## 40   2975  22  Medium
## 41   3450  22  Medium
## 42   3145  22  Medium
## 43   3190  22  Medium
## 44   3610  23  Medium
## 45   2885  23  Medium
## 46   3480  21  Medium
## 47   3200  22  Medium
## 48   2765  21  Medium
## 49   3220  21  Medium
## 50   3480  23  Medium
## 51   3325  23   Large
## 52   3855  18   Large
## 53   3850  20   Large
## 54   3195  18     Van
## 55   3735  18     Van
## 56   3665  18     Van
## 57   3735  19     Van
## 58   3415  20     Van
## 59   3185  20     Van
## 60   3690  19     Van
```

b) Select the fifth row:

```
d_fuel[5, ]

##   weight mpg  type
## 5   2440  32 Small
```

c) Select the 1st to 5th observation:

```
d_fuel[1:5, ]

##   weight mpg  type
## 1   2560  33 Small
## 2   2345  33 Small
## 3   1845  37 Small
## 4   2260  32 Small
## 5   2440  32 Small
```

Alternatively you can get an overview with the help of the R function `head(...)`

```
head(d_fuel)

##   weight mpg  type
## 1   2560  33 Small
## 2   2345  33 Small
## 3   1845  37 Small
## 4   2260  32 Small
## 5   2440  32 Small
## 6   2285  26 Small
```

d) Select the 1st to 3rd and 57th to 60th observation:

```
d_fuel[c(1:3, 57:60), ]

##    weight mpg  type
## 1    2560  33 Small
## 2    2345  33 Small
## 3    1845  37 Small
## 57   3735  19   Van
## 58   3415  20   Van
## 59   3185  20   Van
## 60   3690  19   Van
```

e) The values of the ranges are in the second column, which is called `mpg`.

```
colnames(d_fuel)

## [1] "weight" "mpg"    "type"
```

There are several possibilities to calculate the mean value, which differ in the way of how we select the columdat:

```
mean(d_fuel[, 2])

## [1] 24.58333

mean(d_fuel[, "mpg"])

## [1] 24.58333

mean(d_fuel$mpg)

## [1] 24.58333
```

f) Again, there are different possibilities. One of them is:

```
mean(d_fuel[7:22, "mpg"])

## [1] 27.75
```

g) Conversion of miles per gallon to kilometers per liter and pounds to kilograms:

```
t_kml <- d_fuel[, "mpg"] * 1.6093 / 3.789
t_kg <- d_fuel[, "weight"] * 0.45359
```

h) Mean value of range and weight:

```
mean(t_kml)

## [1] 10.44127

mean(t_kg)

## [1] 1315.789
```

Solution 1.8

a) Let us run the command several times

```
runif(n, min = -1, max = 1)

##  [1] -0.03583977  0.19913165 -0.01291739 -0.62756480  0.65474664
##  [6]  0.33693348  0.58847972 -0.78411275  0.44742189 -0.17745114

runif(n, min = -1, max = 1)

##  [1]  0.64189259  0.29412039  0.56586552  0.10607262  0.05943916
##  [6]  0.57871246 -0.95333760 -0.04553987  0.46462748  0.38546311

runif(n, min = -1, max = 1)

##  [1] -0.04476076  0.72241895 -0.12380579 -0.51040545 -0.85864191
##  [6] -0.80106768 -0.36745659  0.03726853  0.32401015 -0.18633963
```

The command returns 10 numbers between -1 and 1. These numbers are different each time we execute the command, so we can assume, correctly, that these are random numbers, the `r` (random) of `runif`. The 10 numbers are more less evenly distributed in the range from -1 and 1, the `unif` (uniform) in `runif`.

b) `r_squ < 1` returns a vector with so-called boolean values `TRUE` and `FALSE`.

`R` checks all components of `r_squ` whether they are less than 1. If that is the case, the value of this component is `TRUE`, else `FALSE`.

In other words, all entries `TRUE` are the points in the circle, the entries `FALSE` are the points on or outside the circle.

c) `sum(r_squ < 1)` counts the numbers of `TRUE` in `r_squ < 1`, i.e. the number of points within the circle.

Internally, `R` stores `FALSE` as 0 and `TRUE` as 1. The `sum(...)` command adds up all the components in a vector, so the result in this case is the number of ones, i.e. `TRUE`.

d) `4 * sum(r_squ < 1) / n` is the final step. `sum(r_squ < 1) / n` is the proportion of the number of points in the circle to the total number of points. This ratio times 4 gives an approximation of π.

Classical and Bayesian Statistics

Problems 2

Problem 2.1

In the Wikipedia article

https://en.wikipedia.org/wiki/Heights_of_presidents_and_presidential_candidates_of_the_United_States

the body height of US presidents and their challengers in the presdential elections are listed (see Table 1). It was mentioned that the taller candidate typically wins the elections.

In this exercise, we examine the data of the presidential elections since they are broadcast on television.

Year	Winner	Height	Opponent	Height
2020	Joe Biden	183 cm	Donald Trump	191 cm
2016	Donald Trump	191 cm	Hillary Clinton	165 cm
2012	Barack Obama	185 cm	Mitt Romney	187 cm
2008	Barack Obama	185 cm	John McCain	175 cm
2004	George W. Bush	182 cm	John Kerry	193 cm
2000	George W. Bush	182 cm	Al Gore	185 cm
1996	Bill Clinton	188 cm	Bob Dole	187 cm
1992	Bill Clinton	188 cm	George H. W. Bush	188 cm
1988	George H. W. Bush	188 cm	Michael Dukakis	173 cm
1984	Ronald Reagan	185 cm	Walter Mondale	180 cm
1980	Ronald Reagan	185 cm	Jimmy Carter	177 cm
1976	Jimmy Carter	177 cm	Gerald Ford	183 cm
1972	Richard Nixon	182 cm	George McGovern	185 cm
1968	Richard Nixon	182 cm	Hubert Humphrey	180 cm
1964	Lyndon B. Johnson	193 cm	Barry Goldwater	180 cm
1960	John F. Kennedy	183 cm	Richard Nixon	182 cm
1956	Dwight D. Eisenhower	179 cm	Adlai Stevenson	178 cm
1952	Dwight D. Eisenhower	179 cm	Adlai Stevenson	178 cm
1948	Harry S. Truman	175 cm	Thomas Dewey	173 cm

Table 1: Body sizes of presidents and their challengers since 1948

We create two vectors for the corresponding body heights

```
winner <- c(183, 191, 185, 185, 182, 182, 188, 188, 188, 185, 185, 177,
            182, 182, 193, 183, 179, 179, 175)
opponent <- c(191, 165, 187, 175, 193, 185, 187, 188, 173, 180, 177, 183,
              185, 180, 180, 182, 178, 178, 173)
```

a) Determine the length of the two vectors. This way you can check if there are the same number of entries in both vectors.

b) Determine the entries 6. to 10. entry of the vector `winner`. Use the square brackets notation `winner[...]` for this purpose.

c) Determine the 3rd, 5th and 10th to 12th entry.

d) The Washington Post found out that the entries for Bill Clinton (7th and 8th entry) are too small. He has a height of 189 cm. Change the entries in the vector `winner` accordingly and output the new vector again.

e) The claim is that the winners are taller than their challengers. Check this by comparing the mean values of the vectors.

f) Determine the average height *differences*.

g) Determine the variance s^2 and the standard deviation s of the vector `winner`.

h) Determine these values without using the implemented functions (for practicing the handling of `R`). The variance is defined by

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n - 1}$$

Problem 2.2

In a class, the following grades were achieved in a statistics test:

4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9, 6, 4, 3.7, 5, 5.2, 4.5, 3.6, 5, 6, 2.8, 3.3, 5.5, 4.2, 4.9, 5.1

a) Alter three grades in the dataset such that the median remains the same, but the mean value decreases as significantly as possible.

   Use the `sort(...)` command to do this.

b) Create a joint box plot for the two data sets. What do you recognize?

Problem 2.3

From our own experience, we have the impression that in married couples the husband is generally older than his wife. we Want to examine statistically whether this is true.

In a study from the UK, the age (in years) and the body height (in cm) of 170 married couples were collected.

(∗) a) Load the file **husband_wife.csv**. Assign it to a variable.

(∗∗) b) Execute the **summary(...)** command. Explain what the command does and interpret the values for the age of husband and wife.

(∗∗) c) Create a box plot of the *difference* of age between husbands and wives.

(∗∗) d) Interpret the median and quartiles in the box plot. What can you say about the outliers?

Problem 2.4

This task is about getting to know more **R** commands and practicing the use of **R**.

We will use the **InsectSprays** data set that is already contained in **R**.

```
head(InsectSprays)


##   count spray
## 1    10     A
## 2     7     A
## 3    20     A
## 4    14     A
## 5    14     A
## 6    12     A
```

Six different insect sprays were used, which were sprayed on different fields. Then the number of insects was counted that were on the respective field after spraying. (Beall, G., (1942) The Transformation of data from entomological field experiments, Biometrika, 29, 243-262.)

a) First we want to determine the average values of the individual sprays. For this purpose we use the **R** command **tapply(...)**

```
tapply(InsectSprays[, "count"], InsectSprays[, "spray"], FUN = mean)

##         A         B         C         D         E         F
## 14.500000 15.333333  2.083333  4.916667  3.500000 16.666667
```

This command applies (apply) the function **FUN**, in this case **mean** to the column **count** (**InsectSprays[, "count"]**). The average are taken ordered by the column **spray** (**InsectSprays[, "spray"]**). This means that the average values for **count** are calculated separately for the sprays $A, B, \ldots, F$.

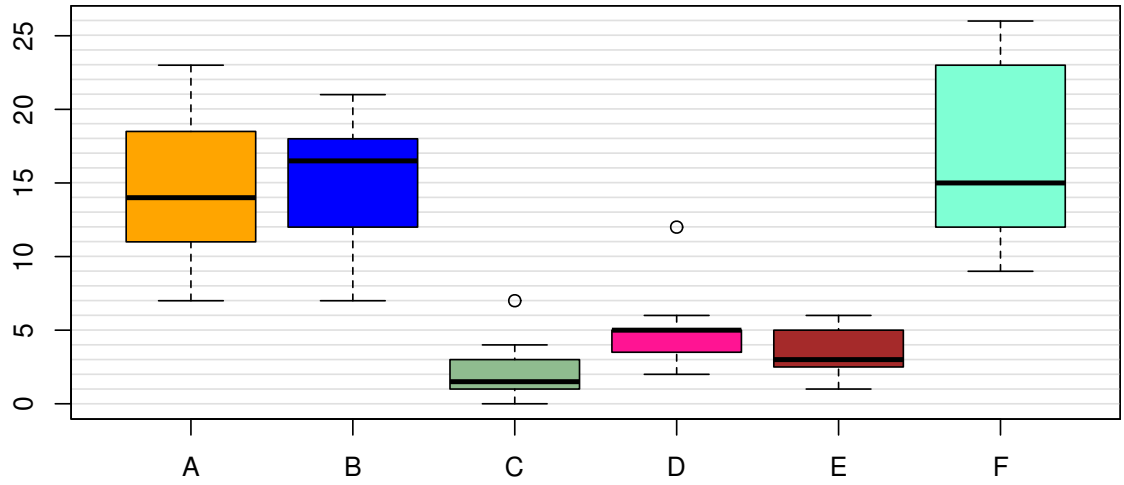
The mean values are very different. The sprays *C*, *D* and *E* seem to be much more efficient than the sprays *A*, *B* and *F*.

The following notation is somewhat easier:

```
tapply(InsectSprays$count, InsectSprays$spray, mean)

##         A         B         C         D         E         F
## 14.500000 15.333333  2.083333  4.916667  3.500000 16.666667
```

b) Now we want to make a box plot of the data. Since the data is ordered by the column **spray**, **R** requires the input **boxplot(y ~ x)**, where **y** are the values of which **R** is to take the box plot and **x** are the names by which the values are to be ordered.

```
plot(NULL,xlim=c(0,10),ylim=c(0,27),xaxt="n", yaxt="n")
for (i in 0:26){
lines(c(-1,11),c(1.04*i,1.04*i),col="gray88")
}
par(new=T)
boxplot(count ~ spray,
        data = InsectSprays,
        col=c("orange", "blue", "darkseagreen", "deeppink",
              "brown", "aquamarine")
)
```



Again, it is obvious that the sprays *C*, *D* and *E* appear to be much more efficient than the sprays *A*, *B* and *F*.

Problem 2.5

In the file **Diet.csv** 76 persons are listed, who each followed one of the diets 1,2 or 3 for 6 weeks.

```
diet <- read.csv("../../../Themen/Varianzanalyse/Uebungen_de/Daten/Diet.csv")

head(diet)
```

```
##   Person gender Age Height pre.weight Diet weight6weeks
## 1     25      0  41    171         60    2         60.0
## 2     26      0  32    174        103    2        103.0
## 3      1      0  22    159         58    1         54.2
## 4      2      0  46    192         60    1         54.0
## 5      3      0  55    170         64    1         63.3
## 6      4      0  33    171         64    1         61.1
```

The file shows the weight **pre.weight** before taking the diet and the weight after 6 weeks **weight6weeks**. We are interested in the weight loss. Therefore we add a column **weight.loss** to the file. This is done as the following way:

```
diet$weight.loss <- diet$weight6weeks - diet$pre.weight

head(diet)
```

```
##   Person gender Age Height pre.weight Diet weight6weeks weight.loss
## 1     25      0  41    171         60    2         60.0         0.0
## 2     26      0  32    174        103    2        103.0         0.0
## 3      1      0  22    159         58    1         54.2        -3.8
## 4      2      0  46    192         60    1         54.0        -6.0
## 5      3      0  55    170         64    1         63.3        -0.7
## 6      4      0  33    171         64    1         61.1        -2.9
```

R recognises **diet$weight.loss** automatically as a new column and adds it at the end of the dataset.

Now perform the subtasks in the task before for **weight.loss** and **Diet**. Interpret the results in each case.

Classical and Bayesian Statistic

Sample solution for Problems 2

Solution 2.1

a) We create two vectors for the corresponding body heights

```
winner <- c(183, 191, 185, 185, 182, 182, 188, 188, 188, 185, 185, 177,
            182, 182, 193, 183, 179, 179, 175)
opponent <- c(191, 165, 187, 175, 193, 185, 187, 188, 173, 180, 177, 183,
              185, 180, 180, 182, 178, 178, 173)
```

a) Length of the two vectors

```
length(winner)

## [1] 19

length(opponent)

## [1] 19
```

b) 6th to 10th entries of the vector `winner`

```
winner [6:10]

## [1] 182 188 188 188 185
```

c) 3rd, 5th and 10th to 12th entries.

```
winner [c(3, 5, 10:12)]

## [1] 185 182 185 185 177
```

d) Change 7th and 7th entries:

```
winner[7] <- 189
winner[8] <- 189
#or winner[c(7,8)] <- 189

winner

##  [1] 183 191 185 185 182 182 189 189 188 185 185 177 182 182 193 183
## [17] 179 179 175
```

e) Compare the mean values of the vectors

```
mean(winner)

## [1] 183.8947

mean(opponent)

## [1] 181.0526
```

The winner of the election is thus on average about 3.16 cm taller.

f) Average of height differences:

```
diff <- winner - opponent

mean(diff)

## [1] 2.842105
```

g) Variance s^2 and the standard deviation s of the vector `winner`.

```
var(winner)

## [1] 22.09942

sd(winner)

## [1] 4.701002
```

h) Without using the implemented functions.

```
winner_var <- sum((winner - mean(winner))^2) / (length(winner)-1)

winner_var

## [1] 22.09942

winner_sd <- sqrt(winner_var)

winner_sd

## [1] 4.701002
```

Solution 2.2

a) The original dataset have the following values:

```
grades_1 <- c(4.2,2.3,5.6,4.5,4.8,3.9,5.9,2.4,5.9,
              6,4,3.7,5,5.2,4.5,3.6,5,6,
              2.8,3.3,5.5,4.2,4.9,5.1)
```

For the median and mean we obtain:

```
median(grades_1)

## [1] 4.65

mean(grades_1)

## [1] 4.5125
```

First we reorder the data values according to their size:

```
sort(grades_1)
```

```
##  [1] 2.3 2.4 2.8 3.3 3.6 3.7 3.9 4.0 4.2 4.2 4.5 4.5 4.8 4.9 5.0 5.0
## [17] 5.1 5.2 5.5 5.6 5.9 5.9 6.0 6.0
```

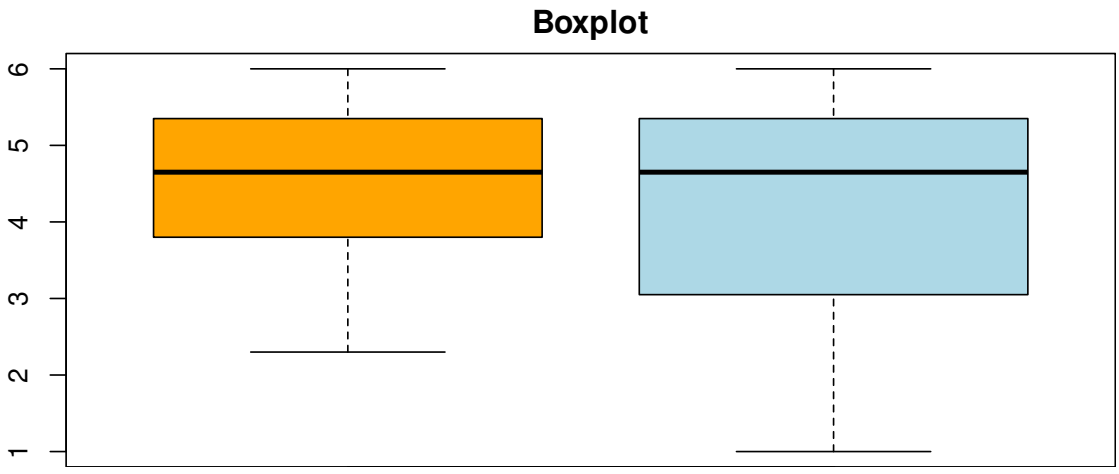
Since the number of grades is even, the median is formed from the average of $x_{(12)}$ and $x_{(13)}$. Thus, if we change grades smaller than $x_{(12)}$, the median will not change. Accordingly, we change the grade values of $x_{(9)}, x_{(10)}, x_{(11)}$ in the most extreme way, namely to 1. This leaves the median unchanged, but decreases the average value as much as possible.

$$2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9, 6, 4, 3.7, 5, 5.2, 1, 3.6, 5, 6, 2.8, 3.3, 5.5, 1, 4.9, 5.1$$

```
grades_2 <- sort(grades_1)

grades_2[c(9,10,11)] <- 1

median(grades_2)

## [1] 4.65

mean(grades_2)

## [1] 4.1
```

b) The boxplots

```
par(mar=c(0,2,2,0))
boxplot(grades_1,
        grades_2,
        main = "Boxplot",
        col = c("orange", "lightblue")
)
```



In fact, the median remains the same. The box becomes wider as extreme values

are added. The median and the upper quartiles stay the same since we didn't change values in this range.

Solution 2.3

a) Load the file

```
hw <- read.csv(".../husband_wife.csv")
```

Use `head(...)` to check if the file was read correctly:

```
head(hw)

##   age.husband height.husband age.wife height.wife
## 1          49            180       43         159
## 2          25            184       28         156
## 3          40            165       30         162
## 4          52            177       57         154
## 5          58            161       52         142
## 6          32            169       27         166
```

b)

```
summary(hw)

##   age.husband    height.husband     age.wife      height.wife
##  Min.   :20.00   Min.   :155.0   Min.   :18.00   Min.   :141.0
##  1st Qu.:33.00   1st Qu.:169.0   1st Qu.:32.00   1st Qu.:156.0
##  Median :43.50   Median :172.0   Median :41.00   Median :160.0
##  Mean   :42.92   Mean   :172.8   Mean   :40.68   Mean   :160.3
##  3rd Qu.:53.00   3rd Qu.:177.0   3rd Qu.:50.00   3rd Qu.:165.0
##  Max.   :64.00   Max.   :190.0   Max.   :64.00   Max.   :176.0
```

With the command `summary()` we get a short statistical overview of the data. It lists the smallest value (`min.`), the lower quartile (`1st Qu.`), the median (`Median`), the mean value (`Mean`), the upper quartile (`3rd Qu.`) and the maximum value (`Max.`) of the data.

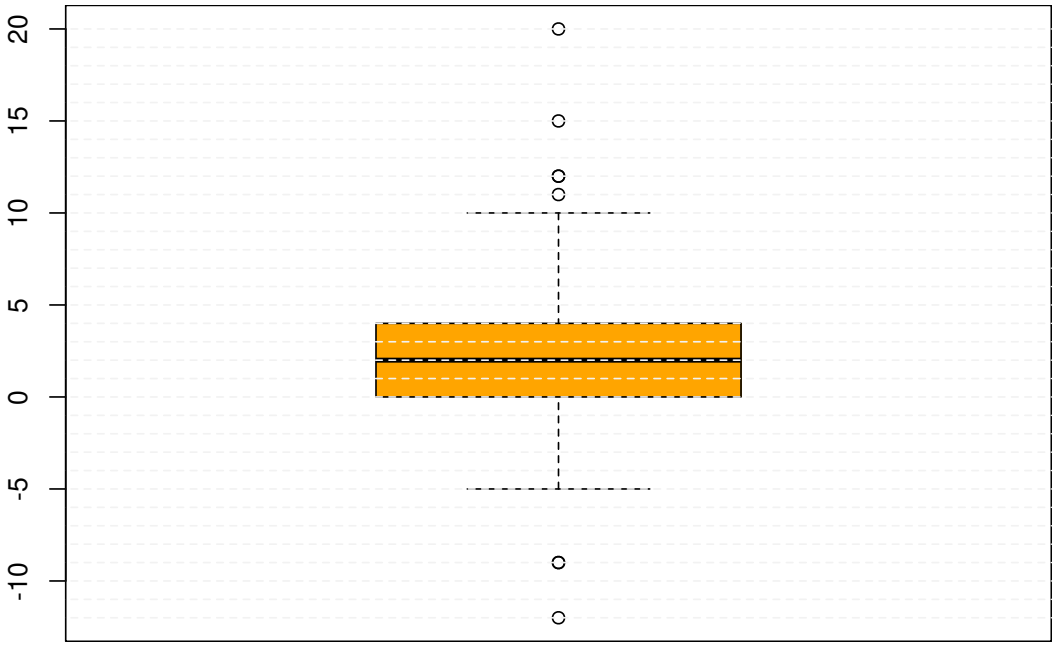
For the 170 husbands, 20 is the lowest age and 64 is the highest. The average age is almost 43 years. The lower quartile is 33 years, so 25 % of the husbands are 33 or younger and 75 % are 33 or older. The median is 43.5 years, so 50 % of the husbands are 43.5 or younger and 50 % are 43.5 or older. The upper quartile is 53 years, so 75 % of the husbands are 53 or younger and 25 % are 53 or older.

The key figures for the wives can be interpreted analogously. However, a quick glance shows that the women tend to be a little younger than the men. This does *not* mean that the husbands are generally older than their wives.

c) Code:

```
age_man <- hw[, 1]
age_woman <- hw[, 3]
```

```r
boxplot(age_man-age_woman, col="orange")
for (i in  -12:20){
lines(c(-2,2),c(i,i),col="gray95",lty="dashed")
}
```



d) The median is about 2, so the age difference is 2 or less for the half of the married couples and 2 or greater for the other half.

The lower quartile is at about 0, i.e. for 25 % of all couples, the wife is older than her husband.

The upper quartile is at 4, i.e. 25 % of all investigated married couples the husband is 4 or more years older than his wife.

The "middle" half of all married couples has an age difference (husband older than his wife) between 0 and 4 years.

The maximum difference is 20 years and the minimum is about -12. In the latter case the wife is about 12 years older than her husband.

Solution 2.4

Solution 2.5

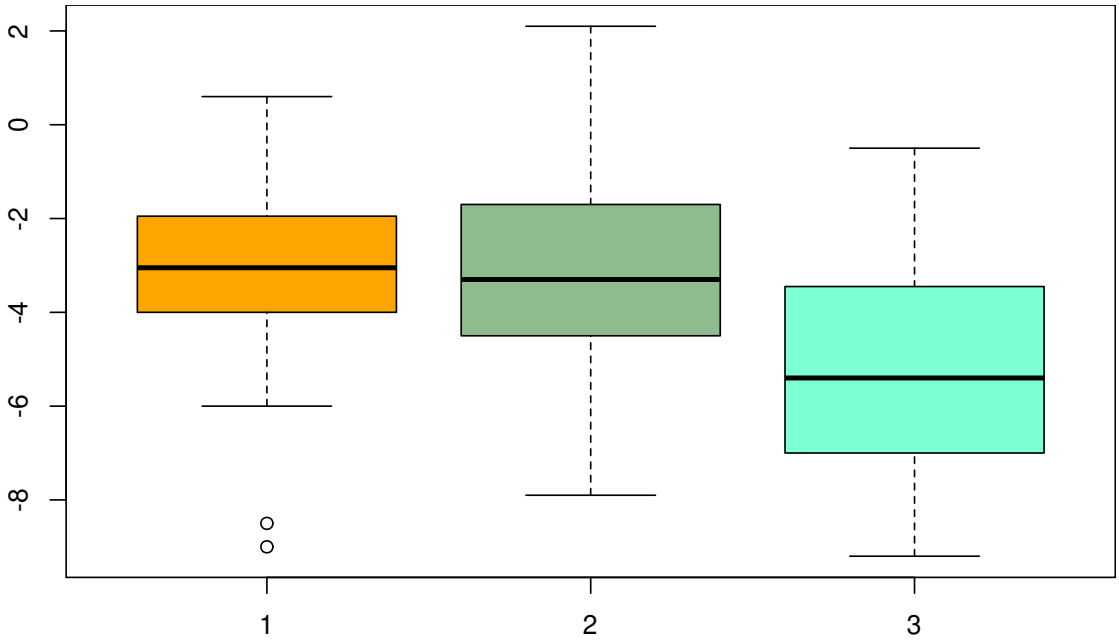
a) Group means:

```r
tapply(diet$weight.loss, diet$Diet, mean)

##         1         2         3
## -3.300000 -3.025926 -5.148148
```

Diets 1 and 2 lead to an average weight loss of about 3 kilograms. For Diet 3 on the other hand, the average weight loss is 5 kilograms and seems to be more efficient than the other two.

b) Graphical representation by a boxplot:

```
boxplot(weight.loss ~ Diet,
        data = diet,
        col = c("orange", "darkseagreen", "aquamarine")

)
```



Boxplot confirms assumption from subtask a).

Classical and Bayesian Statistics

Problems 3

Problem 3.1

The Old Faithful geyser in the Yellowstone National Park is one of the most famous hot springs in the world. Of great interest to spectators and the National Park Service is the time between two eruptions and the duration of the eruption.

The dataset of 272 the measurements is already included in **R** from 1.8.1978 columns, **eruptions** and **waiting**.

a) Draw histograms of the time interval between two eruptions:

```
# Read in data record
geyser <- faithful
head(geyser)

# 4 Graphics in the graphics window
par(mfrow = c(2,2))

hist(geyser[, "waiting"])
hist(geyser[, "waiting"], breaks = 20)
hist(geyser[, "waiting"], breaks = seq(41, 96, by = 11))
```

What do you see? What is the difference between the three histograms?

Remark: If you specify the number of classes with **breaks = 20**, **R** interpretes this value as suggestion and can be changed internally under circumstances.

b) Draw histograms (vary number of classes) of the eruption duration:

```
hist(geyser[, "eruptions"])
```

What do you see? Compare with the first subtask.

Problem 3.2

What's wrong with the following statements? Discuss.

(∗) a) The probabilities of a rigged (biased) coin were determined as $P(\text{heads}) = 0.32$ and $P(\text{tails}) = 0.73$.

(∗) b) The probability of winning in a lottery is $-3 \cdot 10^{-6}$.

(∗∗) c) A survey investigated the following events:

S: The interviewed person is pregnant.

M: The interviewed person is male.

It was found that $P(S) = 0.1$, $P(M) = 0.5$ and $P(S \cup M) = 0.7$

Problem 3.3

In a random experiment a red and a blue die are thrown simultaneously. We assume that both die are "fair" (unbiased), i.e. the numbers 1 to 6 of a die occur with the same probability.

(∗) a) Describe the sample space in the form of elementary events.

(∗) b) What is the probability of a single elementary event occuring?

(∗) c) Calculate the probability that the event E_1 "The sum of the eyes is 7" occurs.

(∗) d) What is the probability that the event E_2 "The eye sum is less than 4" occurs?

(∗) e) Calculate $P(E_3)$ for the event E_3 "Both numbers are odd".

(∗∗) f) Calculate $P(E_2 \cup E_3)$.

Problem 3.4

The events A and B are independent with probabilities $P(A) = 3/4$ and $P(B) = 2/3$. Calculate the probabilities of the following events. Make suitable freehand sketches using Venn diagramms.

(∗) a) Both events occur.

(∗∗) b) At least one of the two events occurs.

(∗∗) c) At most one of the two events occurs.

(∗∗) d) None of the two events occurs.

(∗∗) e) Exact one of the events occurs.

Problem 3.5

$(**)$ The collapse of a building in Tokyo can be caused by two independent events.

- E_1: Strong earthquake
- E_2: Big typhoon

The annual probabilities of these two events are $P(E_1) = 0.04$ and $P(E_2) = 0.08$.

Calculate the annual probability of the building collapsing.

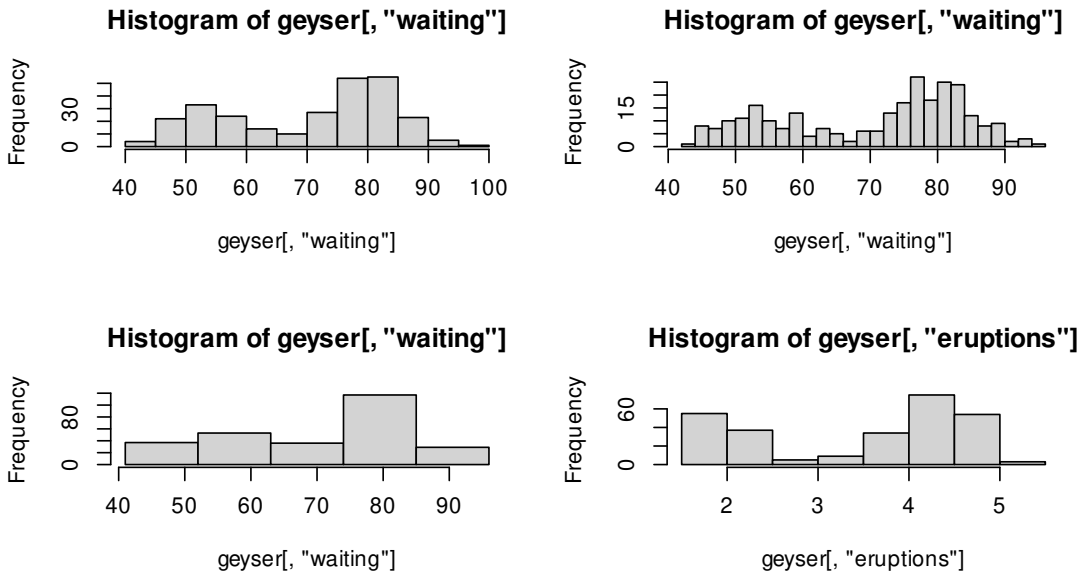
Classical and Bayesian Statistic

Sample solution for Problems 3

Solution 3.1

a)

```
# Read in data record
geyser <- faithful
head(geyser)

##   eruptions waiting
## 1     3.600      79
## 2     1.800      54
## 3     3.333      74
## 4     2.283      62
## 5     4.533      85
## 6     2.883      55

par(mfrow = c(2,2)) # 4 Graphics in the graphics window
# Draw histograms
hist(geyser[, "waiting"])
hist(geyser[, "waiting"], breaks = 20)
hist(geyser[, "waiting"], breaks = seq(41, 96, by = 11))
hist(geyser[, "eruptions"])
```



The first three histograms in the figure show the intervals between two eruptions of Old Faithful. It is noticeable that time spans of 55 minutes but also

between 70 and 85 minutes occur more frequently than other intervals. Such a distribution with two peaks is also called *bimodal*.

If the bar widths are chosen inappropriately, this special feature of the geyser data is not visible. This happened in the third histogram. The example illustrates that the correct choice of bar widths must be well-considered.

b) Finally, the fourth histogram shows the characteristics of different eruption durations. Two peaks are clearly visible: Either the eruption is over after 1.5-2 minutes, or it lasts at least three and a half minutes.

But whether the duration of an eruption has something to do with the duration of the preceding time interval between two eruptions (in other words: whether the peaks of the histogram from subtask b) correspond to the peaks of the histograms from subtask a)) cannot be said on the basis of these histograms.

Solution 3.2

a) Because "head" and "tail" are the only possible outcomes, the probabilities have to add up to 1. This is not the case:

$$P(\Omega) = P(\text{Head}) + P(\text{Tail}) = 1.05$$

Axiom 2 is violated.

b) The probability is less than zero. Axiom 1 is violated.

c) Because $S \cap M = \{\}$ it follows $P(S) + P(M) = P(S \cup M)$ because of Axiom 3. But this is not the case.

Solution 3.3

a) $\Omega = \{(1,1),(1,2),\ldots,(1,6),(2,1),(2,2),\ldots,(2,6),\ldots,(6,6)\}$, $|\Omega| = 36$

b) P(elementary event) = $\frac{1}{|\Omega|} = \frac{1}{36}$

c)
- $E_1 = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$
- Number of favorable elementary events: $|E_1| = 6$
- Number of possibe elemetary events: $|\Omega| = 36$
- Probability of E_1 occuring:

$$P(E_1) = \frac{|E_1|}{|\Omega|} = \frac{6}{36} = \frac{1}{6}$$

d) $E_2 = \{(1,1),(2,1),(1,2)\}$:

$$P(E_2) = \frac{|E_2|}{|\Omega|} = \frac{3}{36} = \frac{1}{12}$$

e) $E_3 = \{(1,1),(1,3),(1,5),(3,1),(3,3),(3,5),(5,1),(5,3),(5,5)\}$:

$$P(E_3) = \frac{|E_3|}{|\Omega|} = \frac{9}{36} = \frac{1}{4}$$

f) With the addition theorem:

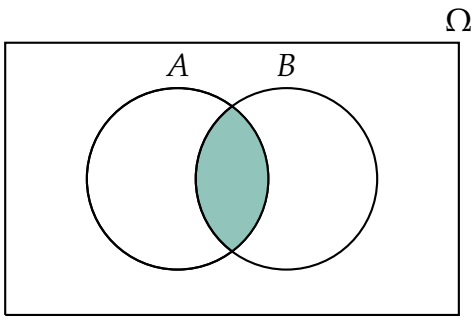
$$\begin{aligned} P(E_2 \cup E_3) &= P(E_2) + P(E_3) - P(E_2 \cap E_3) \\ &= P(E_2) + P(E_3) - P(\{(1,1)\}) \\ &= \frac{3}{36} + \frac{9}{36} - \frac{1}{36} \\ &= \frac{11}{36} \end{aligned}$$

Be careful: The rule $P(E_2 \cap E_3) = P(E_2)P(E_3)$ does not apply in this case as the events E_2 and E_3 are *not* independent.

Solution 3.4

```
A <- 3/4
B <- 2/3
```

a)



$P(\text{both events}) = P(A \cap B) = P(A) \cdot P(B) = \frac{3}{4} \cdot \frac{2}{3} =$

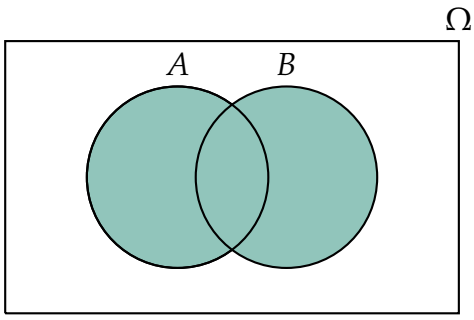
```
library(MASS)

##
## Attaching package:  'MASS'
```

```
## The following object is masked _by_ '.GlobalEnv':
##
##     geyser
```

```
fractions(A * B)
```

```
## [1] 1/2
```

b)

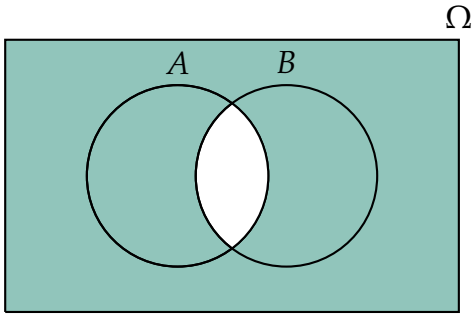


$$\begin{aligned} P(\text{at least one}) &= P(A \cup B) \\ &= P(A) + P(B) - P(A \cap B) \\ &= P(A) + P(B) - P(A) \cdot P(B) \\ &= \end{aligned}$$

```
fractions(A + B - A*B)
```

```
## [1] 11/12
```

c)

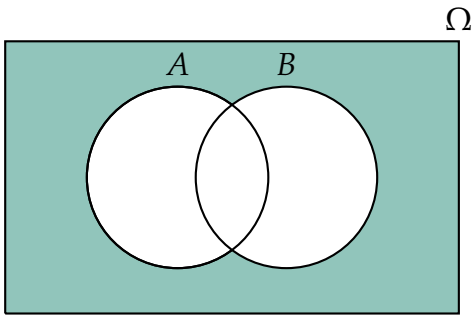


$$P(\text{at most one}) = 1 - P(A \cap B) = 1 - P(A) \cdot P(B)$$

```
fractions(1 - A*B)
```

```
## [1] 1/2
```

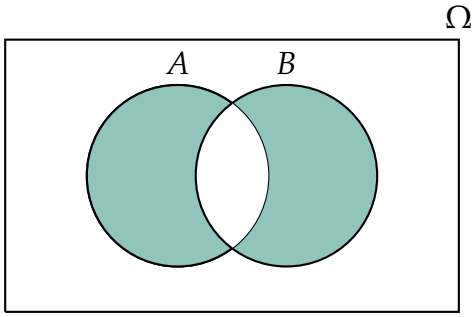
d)



$$\begin{aligned} P(\text{no event}) &= P(\overline{A \cup B}) \\ &= 1 - P(A \cup B) \\ &= 1 - (P(A) + P(B) - P(A) \cdot P(B)) \\ &= \end{aligned}$$

```
fractions(1-(A + B - A*B))

## [1] 1/12
```

e)



$$\begin{aligned} P(\text{exactly one event}) &= P(A \cup B) - P(A \cap B) \\ &= P(A) + P(B) - 2P(A) \cdot P(B) \\ &= \end{aligned}$$

```
fractions(A + B - 2*A*B)

## [1] 5/12
```

Solution 3.5

The annual probability of collapse $E_1 \cup E_2$ can be calculated as follows

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) = 0.04 + 0.08 - 0.04 \cdot 0.08 = 0.1168$$

where $P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$, since E_1 and E_2 are independent.

Classical and Bayesian Statistics

Problems 4

Problem 4.1

(∗) Calculate p_2 so that the table below becomes a probability distribution.

x_k	-5	-4	1	3	6
p_k	0.3	p_2	0.1	0.2	0.3

Problem 4.2

The random variable X describes the number of household members in a sample and has the distribution:

k	1	2	3	4	5
$P(X = k)$	0.4	0.2	0.2	0.1	0.1

Describe the looked for probabilities in b) – e) in the form $P(\ldots)$. For example: $P(X \leq 5)$ or $P(3 \leq X \leq 5)$.

(∗) a) Does the table describe a probability distribution? Justify your answer.

(∗) b) Calculate the probability that a randomly selected household has between 2 and 4 members.

(∗∗) c) Calculate the probability that a randomly selected household has more than 2 members.

(∗) d) Calculate the probability that a randomly selected household has no more than 4 members.

(∗∗) e) Calculate the probability that a randomly selected household has more than one member.

Problem 4.3

A multiple-choice test consists of 15 questions, each with 5 possible answers, of which

exactly one is correct. The probability of answering a question correctly is therefore 0.2. The probability distribution function is given by

k	8	9	10	11	12	13	14	15
$P(X \leq k)$	0.711	0.939	0.969	0.982	0.989	0.992	0.999	1

Note: These are the *cumulated* probabilities $P(X \leq k)$ and *not* $P(X = k)$.

Describe the looked for probabilities again in the form $P(\ldots)$.

(∗) a) The probability that at most 13 questions are answered correctly.

(∗∗) b) The probability that at least 10 questions are correct.

(∗∗) c) The probability that exactly 15 questions are answered correctly.

(∗∗) d) The probability that between 9 and 12 questions are answered correctly.

Problem 4.4

We toss a coin three times. The random variable X indicates how many times "head" are tossed.

(∗∗) a) Set up the probability distribution of X as a table.

(∗) b) Calculate the probability that exactly 2 heads are tossed.

(∗∗) c) Calculate the probability that at least 2 heads are tossed.

(∗∗) d) Calculate the probability that no more than 1 head is tossed.

Problem 4.5

(∗∗) Calculate the expected value of the following probability distribution. Use R.

x_k	-5	-4	1	3	6
p_k	0.3	p_2	0.1	0.2	0.3

Problem 4.6

We roll a blue and a red dice together.

(∗∗) a) Determine the probability distribution of the eye sum cast.

(**) b) Calculate the expected value and the standard deviation. Interpret these values.

Use `R` by creating two vectors `x` and `p`, multiplying the two and using the command `sum(...)`.

Classical and Bayesian Statistic

Sample solution for Problems 4

Solution 4.1

First we need to determine the value for p_2. Since the sum *has to be* equal 1, it follows that

$$p_2 = 1 - 0.3 - 0.1 - 0.2 - 0.3 = 0.1$$

Solution 4.2

a) Yes, because the probabilities add up to 1:

$$0.4 + 0.2 + 0.2 + 0.1 + 0.1 = 1$$

b) Wanted:

$$P(2 \leq X \leq 4) = P(X = 2) + P(X = 3) + P(X = 4) = 0.2 + 0.2 + 0,1 = 0.5$$

c) Sought:

$$P(X > 2) = P(X \geq 3) = P(X = 3) + P(X = 4) + P(X = 5) = 0.2 + 0.1 + 0.1 = 0.4$$

Or:

$$P(X > 2) = 1 - P(X \leq 2) = 1 - (P(X = 1) + P(X = 2)) = 1 - 0.4 - 0.2 = 0.4$$

d) Wanted:

$$P(X \leq 4) = 1 - P(X = 5) = 1 - 0.1 = 0.9$$

e) Sought:

$$P(X \geq 2) = 1 - P(X = 1) = 1 - 0.4 = 0.6$$

Solution 4.3

a) Sought:

$$P(X \leq 13) = 0.992$$

Can be read directly from the table.

b) Sought:

$$P(X \geq 10) = 1 - P(X \leq 9) = 1 - 0.939 = 0.061$$

c) Sought:

$$P(X = 15) = P(X \leq 15) - P(X \leq 14) = 1 - 0.999 = 0.001$$

d) Sought:

$$P(9 \leq X \leq 12) = P(X \leq 12) - P(X \leq 8) = 0.989 - 0.711 = 0.278$$

Solution 4.4

a) The sample space is (T: tails, H: heads)

$$\Omega = \{TTT, TTH, THT, HTT, THH, HTH, HHT, HHH\}$$

It follows that

$$P(X = 0) = \frac{1}{8}, \qquad P(X = 1) = \frac{3}{8}, \qquad P(X = 2) = \frac{3}{8}, \qquad P(X = 3) = \frac{1}{8}$$

So:

k	0	1	2	3
$P(X = k)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

b) Sought:

$$P(X = 2) = \frac{3}{8}$$

c) Sought:

$$P(X \geq 2) = P(X = 2) + P(X = 3) = \frac{3}{8} + \frac{1}{8} = \frac{4}{8} = \frac{1}{2}$$

d) Sought:

$$P(X \leq 1) = P(X = 0) + P(X = 1) = \frac{1}{8} + \frac{3}{8} = \frac{4}{8} = \frac{1}{2}$$

Solution 4.5

See Problem 4.1

$$p_2 = 0.1$$

The expected value is determined by

$$\begin{aligned} \mathrm{E}(X) &= x_1 p_1 + x_2 p_2 + \ldots + x_5 p_5 \\ &= -5 \cdot 0.3 + (-4) \cdot 0.1 + \ldots + 6 \cdot 0.3 = 0.6 \end{aligned}$$

We use R:

```
x <- c(-5, -4, 1, 3, 6)
p <- c(0.3, 0.1, 0.1, 0.2, 0.3)

sum(x*p)


## [1] 0.6
```

Solution 4.6

a) Let X be the random variable for the eye sum thrown. Then

$$\Omega = \{2, 3, 4, \ldots, 12\}$$

This results in the following table for the probability distribution:

x_i	elementary event	abs. frequency	p_i
2	11	1	$\frac{1}{36}$
3	12,21	2	$\frac{2}{36}$
4	13,22,31	3	$\frac{3}{36}$
5	14,23,32,41	4	$\frac{4}{36}$
6	15,24,33,42,51	5	$\frac{5}{36}$
7	16,25,34,43,52,61	6	$\frac{6}{36}$
8	26,35,44,53,62	5	$\frac{5}{36}$
9	36,45,54,63	4	$\frac{4}{36}$
10	46,55,64	3	$\frac{3}{36}$
11	56,65	2	$\frac{2}{36}$
12	66	1	$\frac{1}{36}$

b) The expected value is determined by

$$\begin{aligned} \mathrm{E}(X) &= x_1 p_1 + x_2 p_2 + \ldots + x_{11} p_{11} \\ &= 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + \ldots + 12 \cdot \frac{1}{36} \\ &= 7 \end{aligned}$$

We use R for the calculation:

```
x <- 2:12
p <- c(1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1) / 36
```

```
E <- sum(x*p)
E

## [1] 7
```

If we roll the two dice a lot of times, the rolled mean of the eye sum is about 7. This was to be expected because the table above is symmetrical.

We calculate the standard deviation.

```
E <- sum(x*p)
var_X <- sum((x-E)^2*p)
var_X

## [1] 5.833333

sigma <- sqrt(var_X)
sigma

## [1] 2.415229
```

The standard deviation is 2.42. If we roll the two dice a lot of times, we deviate on „average“ 2.42 from the expected value 7.

Classical and Bayesian Statistics

Problems 5

Problem 5.1

For the height of 18-20 year old men the mean value is 1.80 m with a standard deviation of 7.4 cm. The body height can be considered as normally distributed.

Make a sketch for each of the following probabilities.

(∗∗) a) What is the probability that a randomly selected man in this age group is taller than 1.85 m?

(∗∗) b) What is the probability that a randomly selected man in this age group has a height between 1.70 m and 1.80 m?

(∗∗) c) In what symmetrical range around the mean are the heights of 50 % of the body heights?

(∗∗) d) How tall must a man be to be among the 5 % of the tallest men?

Problem 5.2

In one place there are several carp ponds. The mass of the carp is normally distributed with the expected value $\mu = 4$ kg and the standard deviation 1.25 kg.

(∗) a) What is the probability of catching a carp that is at most 2.5 kg?

(∗) b) What is the probability of catching a carp that weighs at least 5 kg?

(∗∗) c) What percentage of all carp weigh between 3 kg and 4.5 kg?

(∗∗) d) The Fishing Association wants to offer a prize for the heaviest carp.

What is the minimum weight required to have a probability of 2 % of getting the prize?

Problem 5.3

(∗∗) A cigarette manufacturer pretends that the nicotine content in a cigarette is on average 2.2 mg with standard deviation ofz 0.3 mg. However, for a sample of 100 randomly selected cigarettes, the sample mean is 3.1 mg.

If the cigarette manufacturer's statement is true, what is the probability that the sample mean reaches a value of 3.1 mg or more? Interpret your result.

Problem 5.4

The time a passenger spends at an airport check-in counter is a random variable with mean value 8.2 minutes and standard deviation 6 minutes. We randomly observe 36 passengers.

(∗) a) Calculate the probability that the average waiting time of these passengers is less than 10 minutes. Interpret your result.

(∗) b) Calculate the probability that the average waiting time of these passengers is between 5 and 10 minutes. Interpret your result.

(∗) c) Calculate the probability that the average waiting time of these passengers is more than 20 minutes. Interpret your result.

(∗ ∗ ∗) d) All of us have probably already had the experience of a longer waiting time at a check-in counter. Why is the probability of c) then so small?

(∗∗) e) Does the i.i.d. assumption hold here at all?

Problem 5.5

A lecturer knows from experience that the average score in an exam is 77 points with a standard deviation of 15 points. This semester the lecturer will teach two courses: one has 25 participants, the other 64.

(∗∗) a) What is the probability that the approximate average examination result in the course with 25 participants is between 72 and 82 points?

(∗) b) Repeat the calculation from part a) for the course with 64 participants. Compare and interpret the results a) and b).

Problem 5.6

(∗∗) A wine merchant claims that the wine bottles he fills contain 70 centiliters. However, a sceptical consumer suspects that the wine dealer is bottling too little wine and wants to verify this claim. He therefore buys 12 bottles of wine and measures their contents. He finds:

$$71, 69, 67, 68, 73, 72, 71, 71, 68, 72, 69, 72 \quad \text{(in centiliters)}$$

First assume that the standard deviation of the filling is known in advance. It is $\sigma = 1.5$ centiliters.

Perform the (one-sided; in which direction?) hypothesis test at the 5 % significance level. Formulate or calculate *explicitly*

- the model assumptions, H_0, H_A
- the rejection range
- the value of the test statistics and the test result
- the p value

Formulate the conclusion for the critical consumer in one sentence.

Problem 5.7

A bakery states that the rolls it produces have a minimum weight of 50 g with known standard deviation $\sigma = 3$ g. The weights are normally distributed.

A statistics student who is suspicious and suspects that the rolls are too light buys 16 rolls at the bakery and weighs all the rolls. He gets the following values (in g):

$$46,\ 48,\ 52,\ 49,\ 46,\ 51,\ 52,\ 47,\ 49,\ 44,\ 48,\ 51,\ 49,\ 50,\ 53,\ 47$$

(∗∗) a) Formulate the null and alternative hypothesis and carry out a hypothesis test at the 5 % significance level.

(∗∗) b) The student is concerned about the small sample size of 16 in his experiment. Therefore, he examines the weight of the rolls again, this time for 100 rolls. He gets the same average in the sample as for the 16 rolls in a).

Is the test decision the same as in a)? Justify your the answer.

Short answers to selected problems

A 5.1:

a) 25 %, 41.2 %

b) $[175, 185]$

c) At least 192.2 cm

A 5.2:

a) 0.114

b) 0.212

c) 44.4 %

d) 6.57 kg

A 5.3: ≈ 0

A 5.5:

a) 0.9044193

b) 0.9923392

Classical and Bayesian Statistic

Sample solution for Problems 5

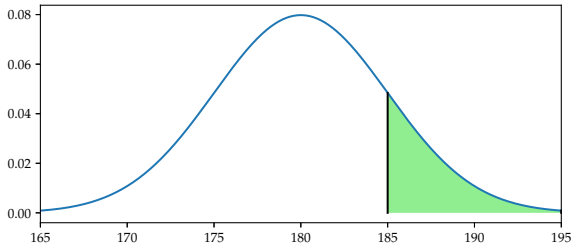
Solution 5.1

The random variable X denotes the body length of a randomly selected person. The distribution of X normally distributed:

$$X \sim \mathcal{N}(1.8, 0.074^2)$$

a) Sought is the probability $P(X \geq 1.85)$ and we obtain

$$P(X \geq 1.85) = 0.2496$$

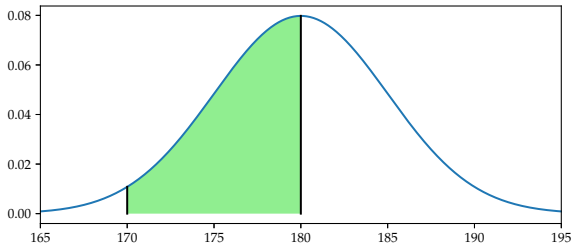


So about 25 % of the 18-20 year old men are taller than 1.85 m.

```
1 - pnorm(q = 1.85, mean = 1.80, sd = 0.074)

## [1] 0.2496233
```

b) Sought is the probability $P(1.70 \leq X \leq 1.80)$ and we get

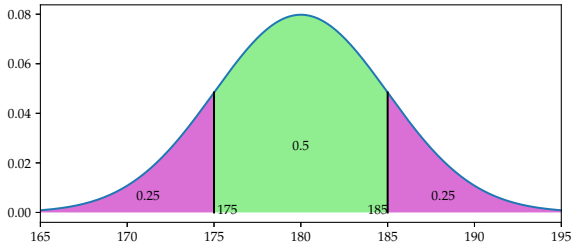
$$P(1.70 \leq X \leq 1.80) = 0.2496$$



So about 41 % of the 18-20 year old men are between 1.70 m and 1.80 m.

```
pnorm(q = 1.80, mean = 1.80, sd = 0.074) - pnorm(1.70, 1.80, 0.074)

## [1] 0.4117085
```

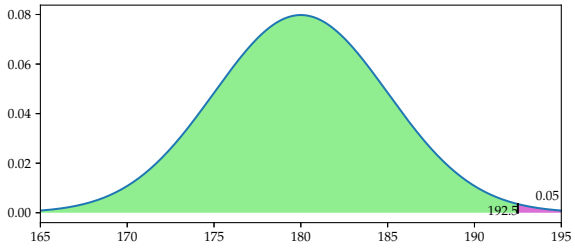
c) Sought are the quantiles $q_{0.25}$ and $q_{0.75}$ (these are just the lower and upper quartiles):




```
qnorm(p = c(0.25,0.75), mean = 1.80, sd = 0.074)

## [1] 1.750088 1.849912
```

That is, 50 % of the men are between 1.75 m and 1.85 m tall.

d) Sought is the quantile $q_{0.95}$:




```
qnorm(p = 0.95, mean = 1.80, sd = 0.074)

## [1] 1.921719
```

That is, 5 % of the men are taller than 1.92 m.

Solution 5.2

The random variable X denotes the weight of the carp. X is distributed as follows

$$X \sim \mathcal{N}(4, 1.25^2)$$

a) Sought is the probability $P(X \leq 2.5)$ and we obtain

$$P(X \leq 2.5) = 0.115$$

About 11 % of the carp weigh less than 2.5 kg.

```
pnorm(q = 2.5, mean = 4, sd = 1.25)
```

```
## [1] 0.1150697
```

b) Sought is the probability $P(X \geq 5)$ and we get

$$P(X \geq 5) = 0.212$$

About 21 % of the carp weigh more than 5 kg.

```
1 - pnorm(q = 5, mean = 4, sd = 1.25)

## [1] 0.2118554
```

c) Sought is the probability

$$P(3 \leq X \leq 4.5) = 0.4436$$

About 44 % of the carp weigh between 3 kg and 4.5 kg.

```
pnorm(q = 4.5, mean = 4, sd = 1.25) - pnorm(3, 4, 1.25)

## [1] 0.4435663
```

d) The quantile $q_{0.98}$ is sought

```
qnorm(p = 0.98, mean = 4, sd = 1.25)

## [1] 6.567186
```

At 2 % of winning the prize, you have to catch a carp that weighs 6.57 kg or more.

Solution 5.3

Let X_i denote the random variable of the nicotine content in the i-th cigarette. We know that $\mu = 2.2$ and $\sigma_X = 0.3$.

We consider the average nicotine content $\overline{X}_{100}$, which according to the Central Limit Theorem (CLT) is approximately normally distributed as follows:

$$\overline{X}_{100} \sim \mathcal{N}\left(\mu, \frac{\sigma_X^2}{n}\right) = \mathcal{N}\left(2.2, \frac{0.3^2}{100}\right) = \mathcal{N}(2.2, 0.0009)$$

We are looking for $P(\overline{X}_{100} \geq 3.1)$ and get

$$P(\overline{X}_{100} \geq 3.1) \approx 0$$

```
1 - pnorm(q = 3.1, mean = 2.2, sd = 0.3/sqrt(100))
```

```
## [1] 0
```

This probability is practically 0, that means that the mean 3.1 mg is extremely unlikely, *assuming* that the manufacturer's information of the average 2.2 mg is true. That indicates that something with the value 2.2 mg is dubious.

Note that even though R returns 0, the probability is *never* exactly 0.

Solution 5.4

Let X_i denote the random variable of the waiting time for the i-th passenger (in minutes). We know that $\mu = 8.2$ and $\sigma_X = 6$.

We consider the average waiting time $\overline{X}_{36}$, which according to the CLT is approximately distributed as

$$\overline{X}_{36} \sim \mathcal{N}\left(\mu, \frac{\sigma_X^2}{n}\right) = \mathcal{N}\left(8.2, \frac{6^2}{36}\right) = \mathcal{N}(8.2, 1)$$

a) Sought is $P(\overline{X}_{36} \leq 10)$ and we get

$$P(\overline{X}_{36} \leq 10) = 0.9640697$$

```
pnorm(q = 10, mean = 8.2, sd = 1)

## [1] 0.9640697
```

The probability of the average waiting time of less than 10 minutes for the these 36 passengers is quite high. That means this group has to wait for *more* than 10 minutes is quite low. So, they unlikely have to wait for more than 10 minutes *on average*.

b) Sought is $P(5 \leq \overline{X}_{36} \leq 10)$ and we calculate

$$P(5 \leq \overline{X}_{36} \leq 10) = 0.9633825$$

```
pnorm(q = 10, mean = 8.2, sd = 1) - pnorm(q = 5, 8.2, 1)

## [1] 0.9633825
```

The difference to b) is very small. This means that the probability that the average waiting time is below 5 minutes is very small.

c) Sought is $P(\overline{X}_{36} \geq 20)$ and we get

$$P(\overline{X}_{36} \geq 20) \approx 0$$

```
1 - pnorm(q = 20, mean = 8.2, sd = 1)

## [1] 0
```

The probability that the average waiting time is longer than 20 minutes is *very* small. It is almost impossible to wait more than 20 minutes on average.

d) The probability that *you* can wait more than 20 minutes is of course much higher.

The probability in c) describes the probability that 36 randomly chosen people waited *on average* more than 20 minutes and that probability is almost 0.

The probability that a lot of people have to wait more than 20 minutes is *on average* less than the probability that *one* person has to wait more than 20 minutes.

e) This is a tricky one. If you consider a large airport and choose passengers randomly from *any* check-in counter at *any* time of the day, then the i.i.d. assumption seems to be justified.

If you pick *one* check-in counter at random and choose 36 passengers from *that* counter then the assumption is not justified. If there is already a long queue then *most* passenger wait longer than average.

The same applies if you choose a *specific* time during the day, when the airport is unusually busy. The waiting time for *most* passengers would be longer.

Solution 5.5

Let X_i denote the random variable for the score of the i-th student. We know that $\mu = 77$ and $\sigma_X = 15$.

a) We consider the average number of points $\overline{X}_{25}$, which is according to the CLT approximately distributed as

$$\overline{X}_{25} \sim \mathcal{N}\left(\mu, \frac{\sigma_X^2}{n}\right) = \mathcal{N}\left(77, \frac{15^2}{25}\right) = \mathcal{N}(7.7, 9)$$

We are looking for $P(72 \leq \overline{X}_{25} \leq 82)$ and obtain

$$P(72 \leq \overline{X}_{25} \leq 82) = 0.9044193$$

```
pnorm(q = 82, mean = 77, sd = 15/sqrt(25)) - pnorm(72, 77, 15/sqrt(25))

## [1] 0.9044193
```

b) We consider the average number of points $\overline{X}_{64}$, which is according to the CLT approximately distributed as

$$\overline{X}_{64} \sim \mathcal{N}\left(\mu, \frac{\sigma_X^2}{n}\right) = \mathcal{N}\left(77, \frac{15^2}{64}\right) = \mathcal{N}(7.7, 225/64)$$

We are looking for $P(72 \leq \overline{X}_{64} \leq 82)$ and get

$$P(72 \leq \overline{X}_{64} \leq 82) = 0.9923392$$

```
pnorm(q = 82, mean = 77, sd = 15/sqrt(64)) - pnorm(72, 77, 15/sqrt(64))

## [1] 0.9923392
```

This probability is larger then in a). The reason is that there are more students in the class and the larger sample reduces randomness *on average*.

Solution 5.6

Let X_i denote the content (in centiliters) of the i-th wine bottle for $i = 1, \ldots, n = 12$.

a) *Model*:

$$X_1, \ldots, X_{12} \text{ i.i.d. } \sim \mathcal{N}(\mu, \sigma^2), \quad \sigma^2 = 1.5^2 \text{ known}$$

b) *Null hypothesis*:

$$H_0: \quad \mu = \mu_0 = 70$$

Alternative hypothesis:

$$H_A: \quad \mu < \mu_0$$

c) *Test statistics*:

$$\overline{X}_n$$

Distribution of test statistics under H_0 being true:

$$\overline{X}_n \sim \mathcal{N}\left(70, \frac{1.5^2}{12}\right)$$

d) *Significance level*:

$$\alpha = 5\,\%$$

e) *Rejection range for the test statistics*:

$$K = (-\infty, 69.29].$$

```
wine <- c(71, 69, 67, 68, 73, 72, 71, 71, 68, 72, 69, 72)

qnorm(p = 0.05, mean = 70, sd = 1.5 / sqrt(length(wine)))

## [1] 69.28776
```

f) *Test decision*:

The observed mean is

$$\overline{x}_n = 70.25$$

```
mean(wine)

## [1] 70.25
```

Thus $\overline{x}_n \notin K$ and H_0 is not rejected. It is therefore quite plausible that the wine merchant is bottling the wine correctly.

g) p-value

```
pnorm(q = 70.25, mean = 70, sd = 1.5 / sqrt(length(wine)))

## [1] 0.7181486
```

The p-value of 0.718 is much higher than the significance level of 0.05 and therefore the test decision is, of course, the same as using the rejection range.

The test was not really necessary, as the mean v of the data is already greater than 70 and we are assuming that the mean is *less* than 70.

Solution 5.7

```
x <- c(46, 48, 52, 49, 46, 51, 52, 47, 49, 44, 48, 51, 49, 50, 53, 47)
```

a) Null hypothesis:

$$H_0: \quad \mu = 50$$

Alternative hypothesis

$$H_A: \quad \mu < 50$$

```
pnorm(mean(x), 50, 3 / sqrt(16))

## [1] 0.0668072
```

The p-value of 0.067 is just above the significance level and thus the null hypothesis H_0 is *not* rejected. The average of the values *is* lower than 50 g, but not significantly lower.

b) The p-value becomes *very* much smaller

```
pnorm(mean(x), 50, 3 / sqrt(100))

## [1] 8.841729e-05
```

and is well below the significance level. In this case, the null hypothesis would be clearly rejected. The value 50 g is statistically significantly wrong.

The reason *why* the p-value becomes much smaller is that with an increasing number of measurements the certainty of where the true mean lies *increases*. Thus, for the mean 48.875 g is quite possible for a small number of rolls, provided that $\mu = 50$ g, the null hypothesis, is correct.

For a large number of measurements the *same* mean 48.875 g is already rather unlikely, always assuming that null hypothesis $\mu = 50$ g is correct.

Classical and Bayesian Statistics

Problems 6

Problem 6.1

(∗∗) A wine merchant claims that the wine bottles he fills contain 70 centiliters. However, a sceptical consumer suspects that the wine dealer is bottling too little wine and wants to verify this claim. He therefore buys 12 bottles of wine and measures their contents. He finds:

$$71, 69, 67, 68, 73, 72, 71, 71, 68, 72, 69, 72 \quad \text{(in centiliters)}$$

First assume that the standard deviation of the filling is known in advance. It is $\sigma = 1.5$ centiliters.

Perform the (one-sided; in which direction?) hypothesis test at the 5 % significance level. Formulate or calculate *explicitly*

- the model assumptions, H_0, H_A
- the rejection range
- the value of the test statistics and the test result
- the p value

Formulate the conclusion for the critical consumer in one sentence.

Problem 6.2

A bakery states that the rolls it produces have a minimum weight of 50 g with known standard deviation $\sigma = 3$ g. The weights are normally distributed.

A statistics student who is suspicious and suspects that the rolls are too light buys 16 rolls at the bakery and weighs all the rolls. He gets the following values (in g):

$$46,\ 48,\ 52,\ 49,\ 46,\ 51,\ 52,\ 47,\ 49,\ 44,\ 48,\ 51,\ 49,\ 50,\ 53,\ 47$$

(∗∗) a) Formulate the null and alternative hypothesis and carry out a hypothesis test at the 5 % significance level.

(∗∗) b) The student is concerned about the small sample size of 16 in his experiment. Therefore, he examines the weight of the rolls again, this time for 100 rolls. He gets the same average in the sample as for the 16 rolls in a).

Is the test decision the same as in a)? Justify your the answer.

Problem 6.3

We continue Problem 6.1. The standard deviation of the deviations is not known. Thus we perform a t-test.

(∗) a) Perform the test and make the test decision with the p value.

```
t.test(...)
```

(∗∗) b) What is different compared to task 6.4?

(∗∗) c) Determine and interpret the confidence interval and carry out the test decision with the confidence interval.

Problem 6.4

(See Problem 6.2) A bakery states that the rolls it produces have a minimum weight of 50 g with known standard deviation $\sigma = 3$ g. The weights are normally distributed.

A statistics student who is suspicious and suspects that the rolls are too light buys 16 rolls at the bakery and weighs all the rolls. He gets the following values (in g):

$$46,\ 48,\ 52,\ 49,\ 46,\ 51,\ 52,\ 47,\ 49,\ 44,\ 48,\ 51,\ 49,\ 50,\ 53,\ 47$$

(∗∗) a) The student is now also suspicious of the known standard deviation and only wants to rely on the given data. How does he proceed? Carry out the hypothesis test.

(∗∗) b) Make the test decision with the confidence interval.

Problem 6.5

Below you will find several examples of comparisons of two samples. Give *short* answers to the following questions for each example:

- Is the sample paired or unpaired? Justify your answer!
- Is the test to be performed one-sided or two-sided? Justify your answer!

- What is the null hypothesis in words?
- What is the alternative hypothesis in words?

(∗∗) a) One experiment was to investigate the effect of cigarette smoking on platelet aggregations. Blood samples were taken from 11 people before and after smoking a cigarette, and the amount of platelet accumulation was measured. We are interested in whether platelet accumulation is increased by smoking.

(∗∗) b) The next data are from a study by Charles Darwin on cross-pollination and self-insemination. 15 pairs of seedlings of the same age, one produced by self-insemination and one by cross-pollination, were bred. Both parts of each pair had almost identical conditions. The aim was to see whether the cross-pollinated plants had more vitality than the self-pollinated ones, i.e. whether they grew bigger. The height of each plant was measured after a fixed period of time.

(∗∗) c) Does the calcium content in the diet affect the systolic blood pressure? To test this question, a trial group of 10 men were given calcium supplementation for 12 weeks. A control group of 11 men were given a placebo.

(∗∗) d) One experiment investigated whether mice had different levels of iron intake in two forms (Fe^{2+} and Fe^{3+}). For this purpose 36 mice were divided into two groups of 18 each and one group was "fed" Fe^{2+} and the other Fe^{3+}. As the iron was radioactively marked, both the initial concentration and the concentration could be measured some time later. From this, the proportion of iron absorbed was calculated for each mouse.

Problem 6.6

Two depth gauges measure the following values for the depth of lakes at 9 different locations:

Gauge A	120	265	157	187	219	288	156	205	163
Gauge B	127	281	160	185	220	298	167	203	171

We assume that the measurements are normally distributed.

Earlier studies show that gauge B systematically measures larger values than gauge A. Do the readings confirm this assumption or is a random fluctuation plausible as an explanation?

(∗∗) a) Are the samples paired or independent?

(∗∗) b) Do we perform a one- or two-sided test? Justify your answer.

(∗∗) c) Perform a t-test at significance level of $\alpha = 0.05$. Formulate explicitly the model assumptions, the null hypothesis, the alternative hypothesis, and the test result.

Problem 6.7

The following table shows the jaw lengths of 10 male and 10 female golden jackals:

Male x_i	120	107	110	116	114	111	113	117	114	112
Female y_j	110	111	107	108	110	105	107	106	111	111

We want to investigate whether the male and female jackals have different jaw lengths.

(∗∗) a) Are the samples paired or unpaired? Justify your answer.

(∗) b) Formulate null and alternative hypothesis.

(∗∗) c) Perform a t test with the help of R. Interpret the p-value and the resulting test decision.

```
 # View record
jackals <- read.table(file="../../../Themen/Statistik_Messdaten/Uebungen_de/Daten/jackals.txt",
                      header=TRUE) # Read in data record
head(jackals)
# Perform t-test
t.test(jackals[, "M"], jackals[, "F"])
```

d) Perform a Wilcoxon test with the help of R. Again, interpret the p-value and make the test decision.

```
# Perform Wilcoxon test
wilcox.test(jackals[, "M"], jackals[, "F"],)
```

e) If the results of the two tests are different, which one would you rather trust? Why?

Problem 6.8

A U.S. magazine, Consumer Reports, conducted an investigation into the calorie and salt content of various hot dog brands. There were three different types of hot dogs: beef, "meat" (beef, pork, mixed poultry) and poultry.

The results below list the calorie content of different brands of beef and poultry hot dogs.

Beef hot dog:

186, 181, 176, 149, 184, 190, 158, 139, 175, 148, 152, 111, 141, 153, 190, 157, 131, 149, 135, 132

Poultry hot dog:

129, 132, 102, 106, 94, 102, 87, 99, 170, 113, 135, 142, 86, 143, 152, 146, 144

Do the two types of hot dog have a significantly different calorie content? We answer this question with a hypothesis test.

(∗) a) Is it a paired or unpaired test? Justify your answer.

(∗∗) b) Is it a one- or two-sided test? Justify your answer.

(∗) c) Formulate null and alternative hypothesis.

(∗) d) Calculate the means of the two datasets. What is your assumption?

(∗∗) e) Which test would you choose, a t-test or Wilcoxon-test? Justify your answer.

(∗∗) f) Perform the corresponding test with R. Interpret the p-value.

Problem 6.9

In the year 2013, within the framework of a international cooperation under the leadership of EAWAG in Dübendorf, concentrations of illegal substances in waste water from 42 European cities during one week were investigated (Ort C. et all, *Spatial differences and temporal changes in illicit drug use in Europe quantified by wastewater analysis*, Addiction 2014 Aug).

The median concentrations of ecstasy (MDMA) in waste water were measured on 7 consecutive days (6-12 March) in addition to other substances. On the basis of this study, a widely read Swiss free newspaper stated that a lot more drugs are consumed in Zurich than elsewhere.

The following table shows for the cities of Zurich and Basel the quantities of MDMA that were extracted on the days of the week - the values can be found in the file *mdma.txt*. The values are in mg per 1000 inhabitants per day.

Weekdays	Wed	Thu	Fri	Sat	Sun	Mon	Tue
Zurich	16.3	12.7	14.0	53.3	117	62.6	27.6
Basel	10.4	8.91	11.7	29.9	46.3	25.0	29.4

Assume that the daily differences D_i between the quantities of MDMA extracted per thousand inhabitants in the wastewater of Zurich and Basel are independently normally distributed with expected value μ_D and standard deviation σ_D.

Hint:

```
... <- read.table("...", header = TRUE)
```

(∗∗) a) Estimate (calculate) from the data the mean and standard deviation of the differences, i.e. $\widehat{\mu}_D$ and $\widehat{\sigma}_D$.

(∗) b) Are the samples paired or unpaired? Justify your answer.

(∗) c) Formulate the null hypothesis and the alternative hypothesis if you want to check the statement of the said free newspaper.

(∗∗) d) Perform a statistical test with the help of **R** on the significance level 5 %, assuming that the data are normally distributed.

What is your test decision?

(∗∗) e) Specify the 95 % confidence interval for the differences D_i (using **R**).

How do you interpret this confidence interval?

(∗∗) f) Now perform a statistical test with the help of **R** at significance level of 5 %, assuming that the data are not normally distributed. What is your conclusion?

Problem 6.10

(Continuation of Problem 2.3)

From our own experience, we have the impression that in married couples the husband tends to be older than his wife. Now we want to examine with a hypothesis test whether this is the case.

The data set **husband_wife.csv** contains the values of height and age for men and women of 170 British married couples. The height of husbands and wives is given in cm and the age in years.

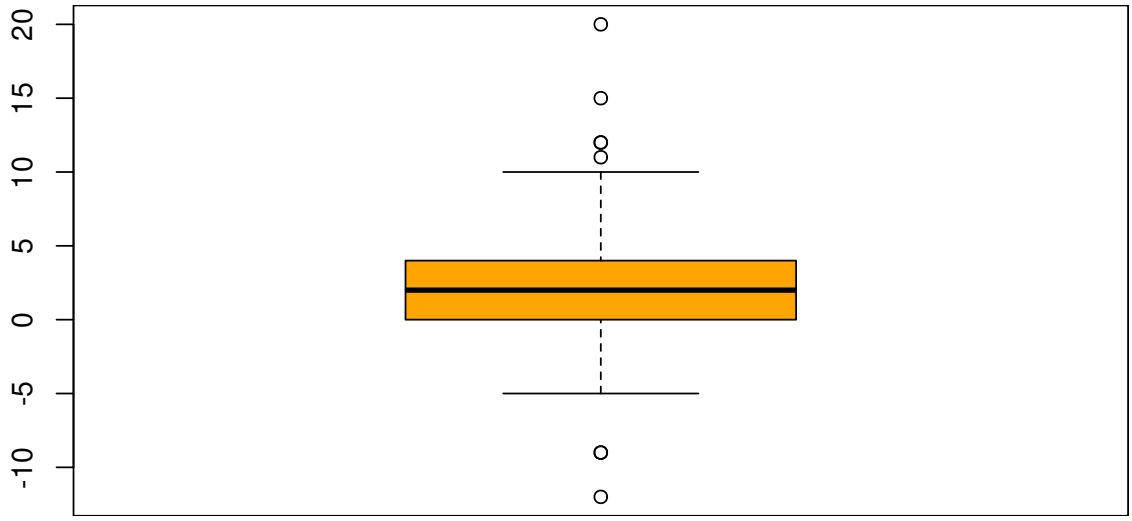
Note:

```
mf <- read.csv(".../husband_wife.csv")
```

The ... represent the path where the file was saved.

In Problem 2.3, we have already seen the box plot for the age difference of the married couples.

```
mf <- read.csv("../../../Themen/Deskriptive_Statistik/Uebungen_en/Daten/husband_wife.cs
head(mf)


##   age.husband height.husband age.wife height.wife
## 1          49            180       43         159
## 2          25            184       28         156
## 3          40            165       30         162
## 4          52            177       57         154
## 5          58            161       52         142
## 6          32            169       27         166


diff <- mf$age.husband - mf$age.wife

boxplot(diff, col = "orange")
```



Remark R: The expression

```
mf$age.husband
```

is equivalent to

```
mf[, "age.husband"]
```

In about 50 % of the married couples the age difference is between 0 and about 5 years. In about 25 % of the married couples the wife is older than the husband. Our

assumption seems to be correct, but the question still remains whether the difference is statistically significant.

a) We want to investigate our assumption that husbands are more likely to be older their wives with a hypothesis test.

(∗) i) Do you choose a paired or unpaired test? Justify your answer.

(∗) ii) Do you choose a one- or two-sided test? Justify your answer.

(∗∗) iii) We assume normal distribution of the data. Carry out a hypothesis test at a significance level of 5 %.

Formulate the null and alternative hypothesis and perform the test and make the test decision.

Make the test decision with the confidence interval.

(∗∗) iv) If you do not assume a normal distribution, which test do you choose? Carry out this test and interpret the result.

b) We are studying the differences in height between women and men. In general, men are larger than women. In England, according to Wikipedia, men are on average 13 cm taller than women.

(∗) i) Which test is appropriate here (one- or two-sided, paired or unpaired)? Justify your answer.

(∗) ii) Formulate the null and alternative hypothesis.

(∗∗) iii) Is the statement that on average the men are 13 cm taller than the women statistically significantly refuted by our data set at a significance level of 5 %? Perform the test and interpret the result. We assume that the body heights are normally distributed.

Make the test decision with the confidence interval.

Problem 6.11

The body temperature of 10 patients is measured at the time of administration of a

drug (T_1) and 2 hours later (T_2). The aim is to test with a hypothesis test whether this drug has a fever-lowering effect.

Patient-Nr.	1	2	3	4	5	6	7	8	9	10
Temp. 1 in °C	39.1	39.3	38.9	40.6	39.5	38.4	38.6	39.0	38.6	39.2
Temp. 2 in °C	38.1	38.3	38.8	37.8	38.2	37.3	37.6	37.8	37.4	38.1

(∗) a) Is it a paired or unpaired test? Justify your answer.

(∗∗) b) Is it a one- or two-sided test? Justify your answer.

(∗) c) Formulate the null and alternative hypothesis.

(∗∗) d) Assume that the data are normally distributed. Which test do you choose? Carry out the test with R at significance level 5 %. What is your conclusion?

(∗∗) e) If we cannot assume that the data are normally distributed, which test do you choose? Perform this at significance level 5 %.

(∗∗) f) Explain the difference of the p-values in subtasks d) and e).

Problem 6.12

Consider a one-sided t-test of $H_0 : \mu = 0$ against $H_A : \mu > 0$ at the significance level of 0.05.

Although the observed n data points have an empirical mean greater than 0, the calculations show that the null hypothesis is not rejected.

Decide whether the following statements are *true or false*.

Hint: Make useful sketches including all relevant information.

(∗ ∗ ∗) a) We reject H_0 for no level $\alpha < 0.05$.

(∗ ∗ ∗) b) There is a level $\alpha < 1$ where we discard H_0.

(∗ ∗ ∗) c) The p-value is strictly smaller than 0.5.

(∗ ∗ ∗) d) If we perform a two-sided test at the level 0.05 instead of a one-sided test, we do not discard H_0.

(∗ ∗ ∗) e) If we copy the data more and more often (i.e. we look at each data point k times, so that we obtain a total of $k \cdot n$ data points with the same mean as for n data points), we discard H_0 for a large k at significance level of 0.05.

Classical and Bayesian Statistic

Sample solution for Problems 6

Solution 6.1

Let X_i denote the content (in centiliters) of the i-th wine bottle for $i = 1, \ldots, n = 12$.

a) *Model*:

$$X_1, \ldots, X_{12} \text{ i.i.d. } \sim \mathcal{N}(\mu, \sigma^2), \quad \sigma^2 = 1.5^2 \text{ known}$$

b) *Null hypothesis*:

$$H_0: \quad \mu = \mu_0 = 70$$

Alternative hypothesis:

$$H_A: \quad \mu < \mu_0$$

c) *Test statistics*:

$$\overline{X}_n$$

Distribution of test statistics under H_0 being true:

$$\overline{X}_n \sim \mathcal{N}\left(70, \frac{1.5^2}{12}\right)$$

d) *Significance level*:

$$\alpha = 5\,\%$$

e) *Rejection range for the test statistics*:

$$K = (-\infty, 69.29].$$

```
wine <- c(71, 69, 67, 68, 73, 72, 71, 71, 68, 72, 69, 72)

qnorm(p = 0.05, mean = 70, sd = 1.5 / sqrt(length(wine)))

## [1] 69.28776
```

f) *Test decision*:

The observed mean is

$$\overline{x}_n = 70.25$$

```
mean(wine)

## [1] 70.25
```

Thus $\overline{x}_n \notin K$ and H_0 is not rejected. It is therefore quite plausible that the wine merchant is bottling the wine correctly.

g) p-value

```
pnorm(q = 70.25, mean = 70, sd = 1.5 / sqrt(length(wine)))

## [1] 0.7181486
```

The p-value of 0.718 is much higher than the significance level of 0.05 and therefore the test decision is, of course, the same as using the rejection range.

The test was not really necessary, as the mean v of the data is already greater than 70 and we are assuming that the mean is *less* than 70.

Solution 6.2

```
x <- c(46, 48, 52, 49, 46, 51, 52, 47, 49, 44, 48, 51, 49, 50, 53, 47)
```

a) Null hypothesis:

$$H_0: \quad \mu = 50$$

Alternative hypothesis

$$H_A: \quad \mu < 50$$

```
pnorm(mean(x), 50, 3 / sqrt(16))

## [1] 0.0668072
```

The p-value of 0.067 is just above the significance level and thus the null hypothesis H_0 is *not* rejected. The average of the values *is* lower than 50 g, but not significantly lower.

b) The p-value becomes *very* much smaller

```
pnorm(mean(x), 50, 3 / sqrt(100))

## [1] 8.841729e-05
```

and is well below the significance level. In this case, the null hypothesis would be clearly rejected. The value 50 g is statistically significantly wrong.

The reason *why* the p-value becomes much smaller is that with an increasing number of measurements the certainty of where the true mean lies *increases*.

Thus, for the mean 48.875 g is quite possible for a small number of rolls, provided that $\mu = 50$ g, the null hypothesis, is correct.

For a large number of measurements the *same* mean 48.875 g is already rather unlikely, always assuming that null hypothesis $\mu = 50$ g is correct.

Solution 6.3

a) Test:

```
wine <- c(71, 69, 67, 68, 73, 72, 71, 71, 68, 72, 69, 72)

t.test(wine, mu = 70, alternative = "less")

##
##  One Sample t-test
##
## data:  wine
## t = 0.44189, df = 11, p-value = 0.6664
## alternative hypothesis: true mean is less than 70
## 95 percent confidence interval:
##      -Inf 71.26603
## sample estimates:
## mean of x
##     70.25
```

The p-value of 0.666 is significantly higher than the significance level of 0.05 and thus, the null hypothesis is *not* rejected as well. The observations *fit* the null hypothesis value 70 ml.

b) The p-value is smaller than the p-value of task 6.1, because with unknown standard deviation for the t test an additional uncertainty is added.

c) The confidence interval is

$$(-\infty,\ 71.27)$$

With a probability of 95 % the true mean value lies in this interval. Since $\mu = 70$ lies in this interval, the null hypothesis is *not* rejected.

However, the decision test is already clear beforehand. The test is left-tailed, but $\overline{x}_{12} = 70.25$ is already *greater* than 70.

Solution 6.4

```
x <- c(46, 48, 52, 49, 46, 51, 52, 47, 49, 44, 48, 51, 49, 50, 53, 47)
```

a) t-test

```
t.test(x, mu = 50, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  x
## t = -1.7811, df = 15, p-value = 0.04758
## alternative hypothesis: true mean is less than 50
## 95 percent confidence interval:
##      -Inf 49.98228
## sample estimates:
## mean of x
##    48.875
```

The p-value of 0.047 is slightly below the significance level of 0.05 and thus the null hypothesis H_0 is rejected (just). The value 50 g is statistically significantly not correct.

Note that in a) the null hypothesis is not rejected, but here it is. So it may well make a difference which test we use.

b) The confidence interval is

$$(-\infty, 49.98)$$

Thus the true mean lies with a probability of 95 % in this interval. But since the 50 g does *not* lie in this interval, the null hypothesis is rejected as well.

Solution 6.5

a) *Paired sample*: Each platelet count *before* smoking corresponds to the platelet count of the same person *after* smoking.

One-sided test: We do not want to know whether the platelet count has *changed*, but whether it has *increased*.

H_0: Smoking has no influence on the accumulation of platelets. ($\mu_\text{S} = \mu_\text{NS}$)

H_A: Smoking increases the accumulation of platelets. ($\mu_S > \mu_{NS}$)

b) *Paired sample*: To each height of a self-pollinated seedling belongs the height of the cross-pollinated "partner".

One-sided test: We do not want to know whether the heights *differ*, but whether the cross-pollinated seedlings become *larger* than the self-pollinated ones.

H_0: The heights do not differ.($\mu_c = \mu_s$)

H_A: Cross-pollinated seedlings become larger than self-pollinated ones. ($\mu_c > \mu_s$)

c) *Unpaired sample*: Unequal numbers in the groups. One blood pressure measurement from the experimental group does not correspond to a specific one from the control group.

Two-sided test: We just want to know if the calcium has an effect on the blood pressure, *no matter* if the blood pressure is higher or lower.

H_0: Calcium has no effect on blood pressure. ($\mu_{\text{Calcium}} = \mu_{\text{Contrast}}$)

H_A: Calcium has an effect on blood pressure. ($\mu_{\text{Calcium}} \neq \mu_{\text{Contrast}}$)

d) *Unpaired sample*: The numbers in the two groups need not be of equal size. The iron measurement of a "Fe^{2+}-mouse" does not correspond to a specific measurement of a "Fe^{3+}-mouse".

Two-sided test: We just want to know if the mice absorb the different forms of iron *differently*.

H_0: Iron absorption is independent of iron variety. ($\mu_2 = \mu_3$)

H_A: The iron absorption depends on the iron variety. ($\mu_2 \neq \mu_3$)

Solution 6.6

a) These are *paired* samples. Measurements are taken at the same location with both gauges.

b) It is assumed that gauge B has the higher values than gauge A. So we perform a one-sided test.

c)
- *Model:*
$$D_1, \ldots, D_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2),$$
σ is estimated by $\widehat{\sigma}$ and $d_i = x_A - x_B$.
- *Null hypothesis*
$$H_0: \quad \mu_D = \mu_0 = 0$$
Alternative hypothesis:
$$H_A: \ \mu_D < \mu_0$$
- *Significance level:*
$$\alpha = 5\,\%$$
- *p-value*
$$0.01168$$

```r
A <- c(120, 265, 157, 187, 219, 288, 156, 205, 163)
B <- c(127, 281, 160, 185, 220, 298, 167, 203, 171)

t.test(A, B, paired = TRUE, alternative = "less")

##
##  Paired t-test
##
## data:  A and B
## t = -2.7955, df = 8, p-value = 0.01168
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -1.93449
## sample estimates:
## mean of the differences
##               -5.777778
```

- *Test decision*

  The p-value is less than 0.05 and thus the null hypothesis is rejected. The gauge B produces indeed statistically significantly larger values than gauge A.

Solution 6.7

Load the dataset

```r
jackals <- read.table(file="./Data/jackals.txt",header=TRUE)
head(jackals)


##     M   F
## 1 120 110
## 2 107 111
## 3 110 107
## 4 116 108
## 5 114 110
## 6 111 105
```

a) The samples are unpaired, as the individual males do not correspond to a specific female. The numbers in the two samples need not be the same.

b) We introduce the following terms:

- X_i: ith value of the jaw length of the males, $i = 1, \ldots, n = 10$
- Y_j: jth value of the jaw length of the female, $j = 1, \ldots, m = 10$

Model:

$$X_i \text{ i.i.d. } \mathcal{N}(\mu_M, \sigma_M^2), \qquad Y_i \text{ i.i.d. } \mathcal{N}(\mu_F, \sigma_F^2)$$

Null hypothesis:

$$H_0: \quad \mu_M = \mu_F$$

Alternative hypothesis:

$$H_A: \quad \mu_M \neq \mu_F$$

c) The R output for the t-test:

```
t.test(jackals[, "M"], jackals[, "F"])

##
##  Welch Two Sample t-test
##
## data:  jackals[, "M"] and jackals[, "F"]
## t = 3.4843, df = 14.894, p-value = 0.00336
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.861895 7.738105
## sample estimates:
## mean of x mean of y
##     113.4     108.6
```

The p-value is $0.0034 < 0.05$, hence the null hypothesis is rejected. The male and female jackals have statistically significantly different jaw length.

d) The R output for the Wilcoxon-test looks:

```
wilcox.test(jackals[, "M"], jackals[, "F"])

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  jackals[, "M"] and jackals[, "F"]
## W = 87.5, p-value = 0.004845
## alternative hypothesis: true location shift is not equal to 0
```

The p-value is $0.0048 < 0.05$, hence the null hypothesis is rejected in this test as well.

e) The result of the Wilcoxon-test is more trustworthy because, unlike the t-test, it does not assume that the data are normally distributed and we cannot verifiy this condition in any way.

However, the very different standard deviations in the two groups may be problematic for both tests.

Solution 6.8

a) We cannot unambiguously assign the observations of one data set to the values of the other data set. So it is an unpaired test. Moreover, the data sets have different lengths.

b) No preference a priori between poultry and beef hot dogs is evident, i.e. we perform a two-sided test.

c) As this is an unpaired test, the means μ_{Beef} and μ_{Poultry} are compared.

Null hypothesis (no difference in calorie content)

$$H_0: \quad \mu_{\text{Beef}} = \mu_{\text{Poultry}}$$

Alternative hypothesis (difference in calorie content)

$$H_A: \quad \mu_{\text{Beef}} \neq \mu_{\text{Poultry}}$$

d) R output:

```
beef <- c(186, 181, 176, 149, 184, 190, 158, 139, 175, 148, 152, 111,
          141, 153, 190, 157, 131, 149, 135, 132)

poultry <- c(129, 132, 102, 106, 94, 102, 87, 99, 170, 113, 135, 142,
       86, 143, 152, 146, 144)

mean(beef)

## [1] 156.85

mean(poultry)

## [1] 122.4706
```

The calorie content of beef hot dogs seems to be much higher than that of poultry hot dogs. The null hypothesis should be rejected.

Note that is *not* good practice to decide from the data whether to perform a one- or two-sided test. We have to make this decision *before* collecting data.

We *could* argue that we already "know" ("studies has shown that. . ."), that beef is more fatty than poultry and *then* a one-sided test would be appropriate. But this information comes from *outside* the data sets.

e) Since there is no indication whether the data are normally distributed, we choose a Wilcoxon test as a precautionary measure.

f) R output:

```
wilcox.test(beef, poultry, paired = FALSE)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  beef and poultry
## W = 285.5, p-value = 0.0004549
## alternative hypothesis: true location shift is not equal to 0
```

The p-value is 0.00046 and thus far below the significance level of 0.05. Hence, the null hypothesis that the two types of hot dog have the same statistically significant calorie content is *rejected*.

Now we can argue based on d) that beef hot dogs has statistically significantly more calories than poultry hot dogs.

Solution 6.9

Load the file:

```
mdma <- read.table("./Data/mdma.txt", header = TRUE)
head(mdma)


##   Zurich Basel
## 1   16.3 10.40
## 2   12.7  8.91
## 3   14.0 11.70
## 4   53.3 29.90
## 5  117.0 46.30
## 6   62.6 25.00
```

a) We estimate with R the mean value and the standard deviation as follows:

```
d <- mdma$Zurich - mdma$Basel

mean(d)

## [1] 20.27

sd(d)

## [1] 26.2723
```

We see that the standard deviation is very large compared to the mean. This could be an indication that the data are problematic for a hypothesis test.

b) We can consider the *days* as an experiment units, then these are paired samples because we have two observations per day.

However, it could also be argued that the *cities* are experiment units. In this case the samples are considered unpaired.

c) The null hypothesis is that there is no difference between the two cities in terms of the quantity of MDMA extracted, i.e.

$$H_0: \quad \mu_D = \mu_0 = 0$$

The alternative hypothesis is

$$H_A: \quad \mu_D \neq \mu_0 = 0$$

Note that we *cannot* decide on a one-sided test because of the claim of the free newspaper, namely that more drugs are consumed in Zurich and therefore more MDMA is extracted, therefore

$$\mu_D > \mu_0 = 0$$

That statement is based on the *same* data which is not good practice (see DoE).

d) Output:

```
t.test(mdma$Zurich,
       mdma$Basel,
       paired = TRUE)

##
##  Paired t-test
##
## data:  mdma$Zurich and mdma$Basel
## t = 2.0413, df = 6, p-value = 0.08729
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.027829 44.567829
## sample estimates:
## mean of the differences
##                   20.27
```

From the R output we can see that the p-value is 0.08729 and therefore greater than $\alpha = 0.05$. So at 5 % level of significance we do not reject the null hypothesis. Hence, there is not significantly more ecstasy is consumed in Zurich than in Basel. The claim of the newspaper is not valid.

If we interpret the samples as unpaired, then

```
t.test(mdma$Zurich,
       mdma$Basel,
       paired = FALSE)

##
##  Welch Two Sample t-test
##
## data:  mdma$Zurich and mdma$Basel
## t = 1.3273, df = 7.5245, p-value = 0.2233
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -15.33677  55.87677
## sample estimates:
## mean of x mean of y
##  43.35714  23.08714
```

We have the same test decision.

e) From the R output for a paired test, we can see that the 95 % confidence interval is given by:

$$[-4.028,\ 44.57]$$

With 95 % probability, the true difference between the values of MDMA per thousand inhabitants in Zurich and Basel lies in the interval $[-4.028,\ 44.57]$.

Because 0 is contained in the 95 % confidence interval, we cannot reject the null hypothesis at the 5 % significance level.

f) Output:

```
wilcox.test(mdma$Zurich,
            mdma$Basel,
            paired = TRUE)

##
##  Wilcoxon signed rank exact test
##
## data:  mdma$Zurich and mdma$Basel
## V = 27, p-value = 0.03125
## alternative hypothesis: true location shift is not equal to 0
```

In this case we reject the null hypothesis, because the p-value is 0.0312 and thus lower than the significance level 5 %.

The different results for t-test and Wilcoxon-Test are probably because of the huge standard deviation compared to the mean. The normal distribution assumption may not be satisfied, so the Wilcoxon-Test is more trustworthy and the claim of the newspaper is correct.

Solution 6.10

a) i) It is a paired test. For each test unit (married couple) there are two associated measurements (age husband, age wife).

ii) We are not sure whether the husbands are really older than their wives. It is simply our *impression* and *not* a fact. So perform do a two-sided test.

iii) Let D denote the age difference between husband and wife.

Null hypothesis

$$H_0: \quad \mu_D = 0$$

Alternative hypothesis:

$$H_0: \quad \mu_D \neq 0$$

```
t.test(diff)

##
##  One Sample t-test
##
## data:  diff
## t = 7.1518, df = 169, p-value = 2.474e-11
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.618286 2.852302
## sample estimates:
## mean of x
##  2.235294
```

or

```
t.test(mf$age.husband,
       mf$age.wife,
       paired = TRUE)

##
##  Paired t-test
##
## data:  mf$age.husband and mf$age.wife
## t = 7.1518, df = 169, p-value = 2.474e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.618286 2.852302
## sample estimates:
## mean of the differences
##                2.235294
```

The p-value is far below the significance level of 5 % and thus the null hypothesis is rejected. The husbands are statistically significantly older than their wives.

The confidence interval is

$$(1.61,\ 2.85)$$

With a probability of 95 % the true mean lies in this interval. The null hypothesis was $\mu_D = 0$, i.e. no age difference. This value does *not* lie in the confidence interval and therefore the null hypothesis is rejected here as well. There is a statistically significant age difference within the married couples.

iv) Wilcoxon-test:

```
wilcox.test(diff)

##
##  Wilcoxon signed rank test with continuity correction
##
## data:  diff
## V = 9460, p-value = 3.977e-12
## alternative hypothesis: true location is not equal to 0
```

or

```
wilcox.test(mf$age.husband,
            mf$age.wife,
            paired = TRUE)

##
##  Wilcoxon signed rank test with continuity correction
##
## data:  mf$age.husband and mf$age.wife
## V = 9460, p-value = 3.977e-12
## alternative hypothesis: true location shift is not equal to 0
```

Again, the p-value is far below the significance level of 5 % and thus the null hypothesis is rejected. The husbands are statistically significantly older than their wives.

b) i) In this case only the average heights of men and women are compared, so it is an unpaired test. Since we do not know whether the deviation is upwards or downwards from 13 cm, we again perform a two-sided test.

Note that it is *not* the question whether men are taller than women. The question whether men are on average 13 cm taller than women.

ii) Let μ_W be the average height of the women and μ_M be the average height of the men. The null hypothesis is

$$H_0: \quad \mu_W = \mu_M - 13$$

and the alternative hypothesis is

$$H_A: \quad \mu_W \neq \mu_M - 13$$

iii) We assume normal distribution of body heights:

```
t.test(mf$height.husband - 13,
       mf$height.wife,
       paired = FALSE)

##
##  Welch Two Sample t-test
##
## data:  mf$height.husband - 13 and mf$height.wife
## t = -0.63293, df = 336.53, p-value = 0.5272
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.812281  0.929928
## sample estimates:
## mean of x mean of y
##  159.8471  160.2882
```

or

```
t.test(mf$height.husband,
       mf$height.wife,
       mu = 13,
       paired = FALSE)

##
##  Welch Two Sample t-test
##
## data:  mf$height.husband and mf$height.wife
## t = -0.63293, df = 336.53, p-value = 0.5272
## alternative hypothesis: true difference in means is not equal to 13
## 95 percent confidence interval:
##  11.18772 13.92993
## sample estimates:
## mean of x mean of y
##  172.8471  160.2882
```

The p-value is far greater than the significance level and thus the null hypothesis is not rejected. The data do *not* refute the difference in height of 13 cm statistically significant.

The confidence interval is

$$(-1.81,\ 0.92)$$

With a probability of 95 % the true mean lies in this interval. The null hypothesis was $\mu_W = \mu_M - 13$, i.e. no statistically significant deviation from the

height difference of 13 cm. This value 0 lies in the confidence interval and thus the null hypothesis is *not* rejected. The data do *not* refute the difference in height of 13 cm in a statistically significant way.

If we do not assume normal distribution, we choose an unpaired Wilcoxon-test (Mann-Whitney-U).

```
wilcox.test(mf$height.husband - 13, mf$height.wife, paired = FALSE)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  mf$height.husband - 13 and mf$height.wife
## W = 13760, p-value = 0.4461
## alternative hypothesis: true location shift is not equal to 0
```

Again, the null hypothesis is clearly *not* rejected.

Solution 6.11

a) It is a paired test, as two measurements were taken on one test unit (patient).

b) We want to test the fever-lowering effectiveness. For this purpose we calculate the average of the μ_D of the differences D_i (Temp. 1 − Temp. 2). In order to be able to prove the effectiveness

$$\mu_D > 0$$

Note that we are only interested in the fever-*lowering* and *not* the fever-rising effect. Therefore a one-sided test.

c) Null hypothesis (drug has no effect)

$$H_0: \quad \mu_D = 0$$

Alternative hypothesis (drug is fever-lowering)

$$H_A: \quad \mu_D > 0$$

d) R output:

```
t_1 <- c(39.1, 39.3, 38.9, 40.6, 39.5, 38.4, 38.6, 39.0, 38.6, 39.2)
t_2 <- c(38.1, 38.3, 38.8, 37.8, 38.2, 37.3, 37.6, 37.8, 37.4, 38.1)

t.test(t_1, t_2, paired=T, alternative="greater")

##
##  Paired t-test
##
## data:  t_1 and t_2
## t = 5.6569, df = 9, p-value = 0.0001554
```

```
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.7976252       Inf
## sample estimates:
## mean of the differences
##                    1.18
```

The p-value 0.0001554 is less than 0.05 here. Therefore the difference is statistically significant. We can therefore assume that the drug is antipyretic (lowering the fever).

e) R output:

```
wilcox.test(t_1, t_2, paired=T, alternative="greater")

##
##  Wilcoxon signed rank test with continuity correction
##
## data:  t_1 and t_2
## V = 55, p-value = 0.002865
## alternative hypothesis: true location shift is greater than 0
```

The p-value 0.002865 is less than 0.05. Therefore the difference is statistically significant. We can therefore assume that the drug is antipyretic.

f) The p-value of the Wilcoxon-test is greater than the p-value of the t-test. Since the Wilcoxon-test assumes less (no normal distribution) than the t-test, there is an additional uncertainty. The null hypothesis is less strongly rejected.

However, the t-test suggests that it is more "precise". This is true, but only if the data are normally distributed, which is often not known.

Therefore the Wilcoxon test is often preferable to the t-test.

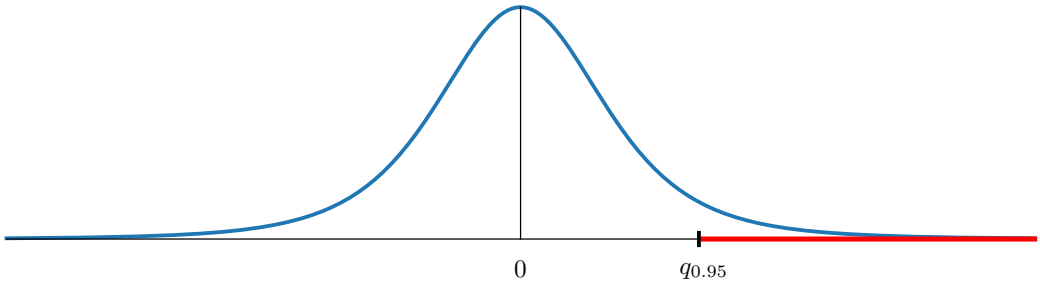
Solution 6.12

This task is not easy and we have to read the text *very* carefully:

Consider a one-sided t-test of $H_0 : \mu = 0$ *against* $H_A : \mu > 0$ *against* H_A *at the level of* 0.05.

We draw a t-distribution with $\mu = 0$ and degree of freedom 4 (these assumptions are

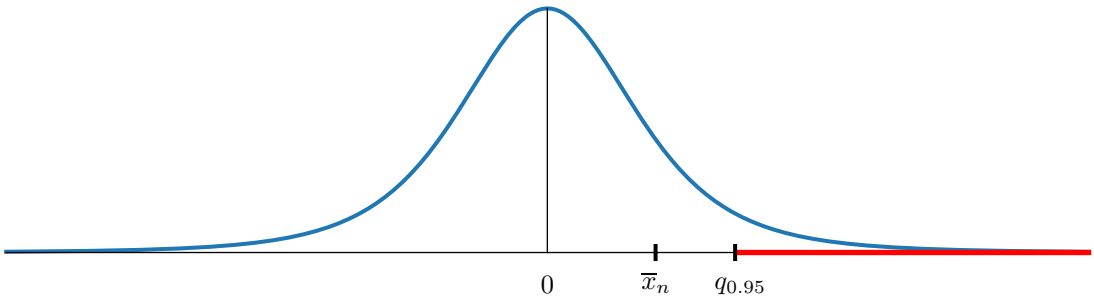
not relevant) and with the rejection zone (right-tailed test).



Although the observed n data points have an empirical mean greater than zero, the calculations show that the null hypothesis is not rejected.

What does this mean?

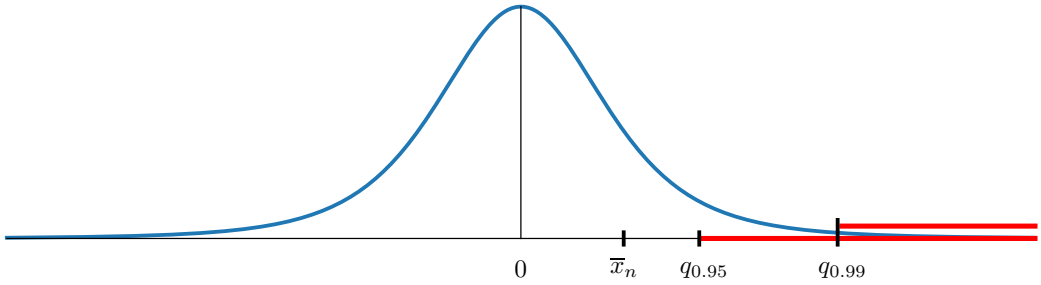
- First of all, the mean is $\overline{x}_n > 0$ and is to the right of 0 on the x-axis.
- H_0 is *not* rejected, hence $\overline{x}_n$ is not in the rejection zone.
- $\overline{x}_n$ lies somewhere between 0 and the boundary of the rejection zone.



a) *We discard H_0 for no level $\alpha < 0.05$.*

What does that mean? Our significance level is $\alpha = 0.05$. Now we choose a significance level α^* that is *less* than 0.05, for example $\alpha^* = 0.01$. The boundary of the rejection zone moves to the right from $q_{0.95}$ to $q_{0.99}$ and the rejection zone becomes smaller. This means that $\overline{x}_n$ is still *not* in the rejection zone (see Figure

below).



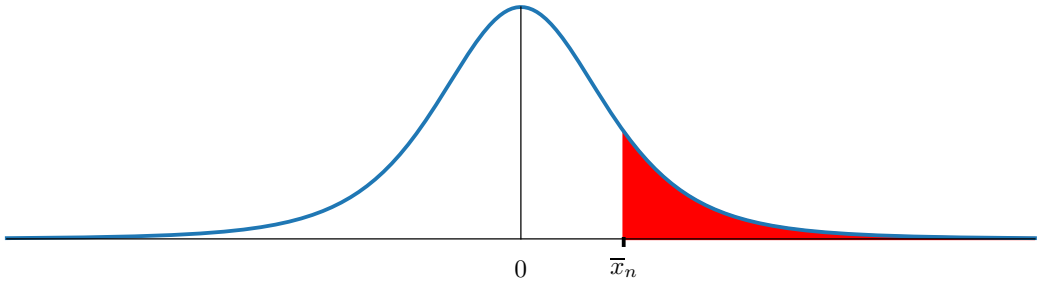
So for all $\alpha < 0.05$ the null hypothesis is not rejected, because $\overline{x}_n$ can never be in the rejection zone. So the statement is correct.

b) *There is a level $\alpha < 1$ where H_0 is rejected.*

The reasoning is the same as in a) but in the opposite direction: Let $\alpha = 0.05$ and we *increase* α, then the boundary of the rejection zone moves to the left and for some α, $\overline{x}_n$ is in the rejection zone and the null hypothesis is rejected. So the statement is correct.

c) *The p-value is strictly smaller than 0.5.*

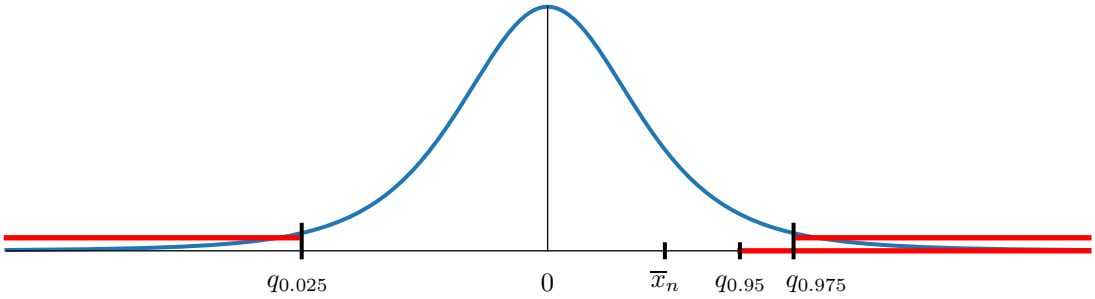
Now what is the p-value? This is the probability of observing a certain value (here $\overline{x}_n$) or a more extreme value towards the alternative hypothesis. We can represent probabilities as areas in continuous distributions and this is the area under the curve to the *right* of $\overline{x}_n$ (see Figure below).



Now the total area under the curve is 1, the area to the right of 0 is 0.5 and then the red area (p-value) must be less than 0.5. So the statement is correct.

d) *If we perform a two-sided test on significance level of 0.05 instead of a one-sided test, we do not discard H_0.*

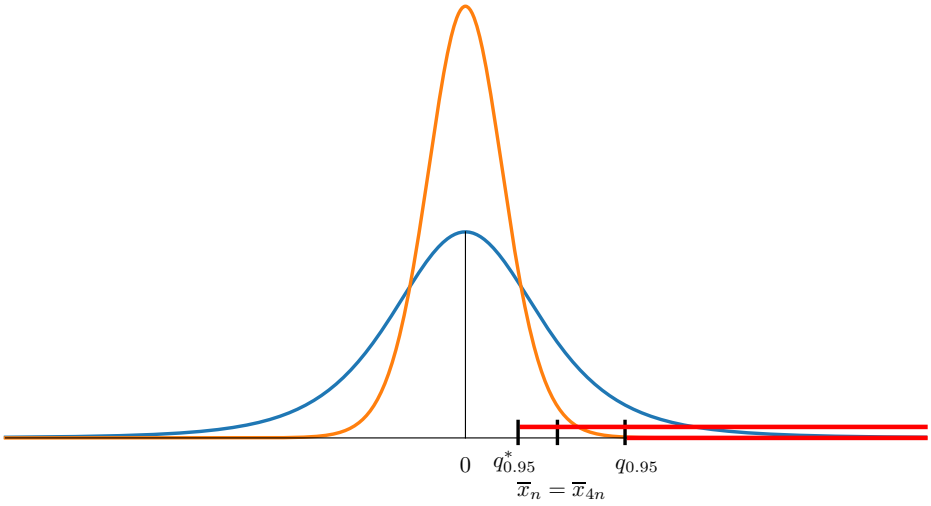
Similar answer as in a): If we switch from a one-sided to a two-sided test, the rejection zone on the right-hand side becomes smaller. Since $\overline{x}_n$ is not in the rejection zone of the one-sided test, it cannot be in the rejection zone of the two-sided test. So the statement is correct.



e) *If we copy the data more and more often (i.e. we look at each data point k times, so that we obtain a total of $k \cdot n$ data points with the same mean as for n data points), we discard H_0 for a large k at significance level of 0.05.*

For example, if we quadruple the data points ($k = 4$), then the mean remains the same and the standard deviation *decreases* by a factor $\sqrt{k} = 2$.

Hence $\overline{x}_n = \overline{x}_{nk}$, but the number of observations increases. That means the curve becomes narrower and the boundary of the rejection zone moves to the left. For a sufficiently large k, the mean $\overline{x}_n$ lies in the rejection zone and the null hypothesis is rejected. So the statement is correct.



Classical and Bayesian Statistics

Problems 7

Problem 7.1

Below you will find several examples of comparisons of two samples. Give *short* answers to the following questions for each example:

- Is the sample paired or unpaired? Justify your answer!
- Is the test to be performed one-sided or two-sided? Justify your answer!
- What is the null hypothesis in words?
- What is the alternative hypothesis in words?

(∗∗) a) One experiment was to investigate the effect of cigarette smoking on platelet aggregations. Blood samples were taken from 11 people before and after smoking a cigarette, and the amount of platelet accumulation was measured. We are interested in whether platelet accumulation is increased by smoking.

(∗∗) b) The next data are from a study by Charles Darwin on cross-pollination and self-insemination. 15 pairs of seedlings of the same age, one produced by self-insemination and one by cross-pollination, were bred. Both parts of each pair had almost identical conditions. The aim was to see whether the cross-pollinated plants had more vitality than the self-pollinated ones, i.e. whether they grew bigger. The height of each plant was measured after a fixed period of time.

(∗∗) c) Does the calcium content in the diet affect the systolic blood pressure? To test this question, a trial group of 10 men were given calcium supplementation for 12 weeks. A control group of 11 men were given a placebo.

(∗∗) d) One experiment investigated whether mice had different levels of iron intake in two forms (Fe^{2+} and Fe^{3+}). For this purpose 36 mice were divided into two groups of 18 each and one group was "fed" Fe^{2+} and the other Fe^{3+}. As the iron was radioactively marked, both the initial concentration and the concentration could be measured some time later. From this, the proportion of iron absorbed was calculated for each mouse.

Problem 7.2

Two depth gauges measure the following values for the depth of lakes at 9 different locations:

Gauge A	120	265	157	187	219	288	156	205	163
Gauge B	127	281	160	185	220	298	167	203	171

We assume that the measurements are normally distributed.

Earlier studies show that gauge B systematically measures larger values than gauge A. Do the readings confirm this assumption or is a random fluctuation plausible as an explanation?

(∗∗) a) Are the samples paired or independent?

(∗∗) b) Do we perform a one- or two-sided test? Justify your answer.

(∗∗) c) Perform a t-test at significance level of $\alpha = 0.05$. Formulate explicitly the model assumptions, the null hypothesis, the alternative hypothesis, and the test result.

Problem 7.3

The following table shows the jaw lengths of 10 male and 10 female golden jackals:

Male x_i	120	107	110	116	114	111	113	117	114	112
Female y_j	110	111	107	108	110	105	107	106	111	111

We want to investigate whether the male and female jackals have different jaw lengths.

(∗∗) a) Are the samples paired or unpaired? Justify your answer.

(∗) b) Formulate null and alternative hypothesis.

(∗∗) c) Perform a t test with the help of R. Interpret the p-value and the resulting test decision.

```r
 # View record
jackals <- read.table(file="../../../Themen/Statistik_Messdaten/Uebungen_de/Daten/jackals.txt",
                      header=TRUE) # Read in data record
head(jackals)
# Perform t-test
t.test(jackals[, "M"], jackals[, "F"])
```

d) Perform a Wilcoxon test with the help of R. Again, interpret the p-value and make the test decision.

```
# Perform Wilcoxon test
wilcox.test(jackals[, "M"], jackals[, "F"],)
```

e) If the results of the two tests are different, which one would you rather trust? Why?

Problem 7.4

A U.S. magazine, Consumer Reports, conducted an investigation into the calorie and salt content of various hot dog brands. There were three different types of hot dogs: beef, "meat" (beef, pork, mixed poultry) and poultry.

The results below list the calorie content of different brands of beef and poultry hot dogs.

Beef hot dog:

186, 181, 176, 149, 184, 190, 158, 139, 175, 148, 152, 111, 141, 153, 190, 157, 131, 149, 135, 132

Poultry hot dog:

129, 132, 102, 106, 94, 102, 87, 99, 170, 113, 135, 142, 86, 143, 152, 146, 144

Do the two types of hot dog have a significantly different calorie content? We answer this question with a hypothesis test.

(∗) a) Is it a paired or unpaired test? Justify your answer.

(∗∗) b) Is it a one- or two-sided test? Justify your answer.

(∗) c) Formulate null and alternative hypothesis.

(∗) d) Calculate the means of the two datasets. What is your assumption?

(∗∗) e) Which test would you choose, a t-test or Wilcoxon-test? Justify your answer.

(∗∗) f) Perform the corresponding test with R. Interpret the p-value.

Problem 7.5

In the year 2013, within the framework of a international cooperation under the leadership of EAWAG in Dübendorf, concentrations of illegal substances in waste water from 42 European cities during one week were investigated (Ort C. et all, *Spatial differences and temporal changes in illicit drug use in Europe quantified by wastewater analysis*, Addiction 2014 Aug).

The median concentrations of ecstasy (MDMA) in waste water were measured on 7 consecutive days (6-12 March) in addition to other substances. On the basis of this study, a widely read Swiss free newspaper stated that a lot more drugs are consumed in Zurich than elsewhere.

The following table shows for the cities of Zurich and Basel the quantities of MDMA that were extracted on the days of the week - the values can be found in the file *mdma.txt*. The values are in mg per 1000 inhabitants per day.

Weekdays	Wed	Thu	Fri	Sat	Sun	Mon	Tue
Zurich	16.3	12.7	14.0	53.3	117	62.6	27.6
Basel	10.4	8.91	11.7	29.9	46.3	25.0	29.4

Assume that the daily differences D_i between the quantities of MDMA extracted per thousand inhabitants in the wastewater of Zurich and Basel are independently normally distributed with expected value μ_D and standard deviation σ_D.

Hint:

```
... <- read.table("...", header = TRUE)
```

(∗∗) a) Estimate (calculate) from the data the mean and standard deviation of the differences, i.e. $\widehat{\mu}_D$ and $\widehat{\sigma}_D$.

(∗) b) Are the samples paired or unpaired? Justify your answer.

(∗) c) Formulate the null hypothesis and the alternative hypothesis if you want to check the statement of the said free newspaper.

(∗∗) d) Perform a statistical test with the help of R on the significance level 5 %, assuming that the data are normally distributed.

What is your test decision?

(**) e) Specify the 95 % confidence interval for the differences D_i (using R).

How do you interpret this confidence interval?

(**) f) Now perform a statistical test with the help of R at significance level of 5 %, assuming that the data are not normally distributed. What is your conclusion?

Problem 7.6

(Continuation of Problem 2.3)

From our own experience, we have the impression that in married couples the husband tends to be older than his wife. Now we want to examine with a hypothesis test whether this is the case.

The data set **husband_wife.csv** contains the values of height and age for men and women of 170 British married couples. The height of husbands and wives is given in cm and the age in years.

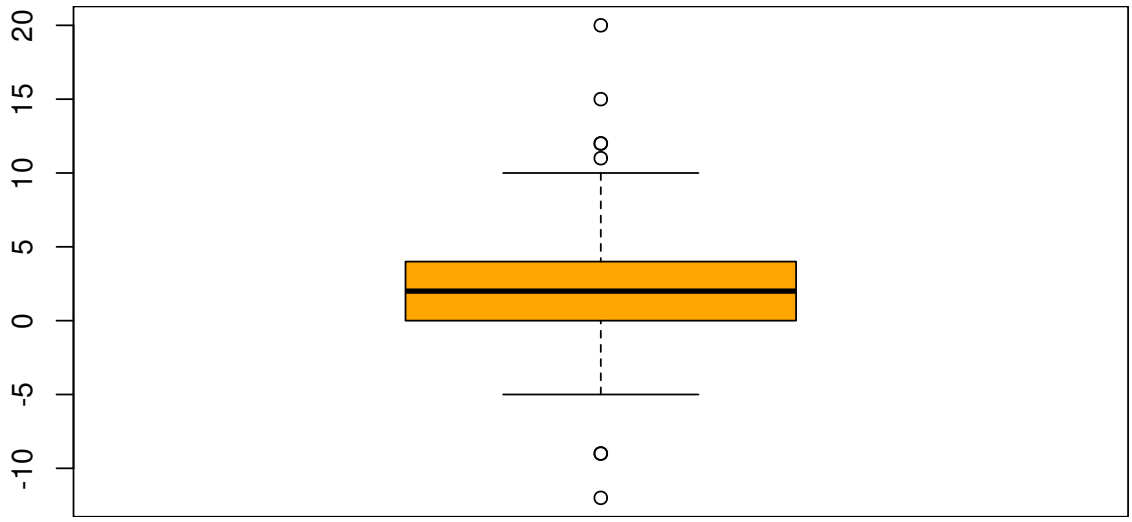
Note:

```
mf <- read.csv(".../husband_wife.csv")
```

The ... represent the path where the file was saved.

In Problem 2.3, we have already seen the box plot for the age difference of the married couples.

```
mf <- read.csv("../../../Themen/Deskriptive_Statistik/Uebungen_en/Daten/husband_wife.cs
head(mf)


##   age.husband height.husband age.wife height.wife
## 1          49            180       43         159
## 2          25            184       28         156
## 3          40            165       30         162
## 4          52            177       57         154
## 5          58            161       52         142
## 6          32            169       27         166


diff <- mf$age.husband - mf$age.wife

boxplot(diff, col = "orange")
```



Remark R: The expression

```
mf$age.husband
```

is equivalent to

```
mf[, "age.husband"]
```

In about 50 % of the married couples the age difference is between 0 and about 5 years. In about 25 % of the married couples the wife is older than the husband. Our assumption seems to be correct, but the question still remains whether the difference is statistically significant.

a) We want to investigate our assumption that husbands are more likely to be older their wives with a hypothesis test.

(∗) i) Do you choose a paired or unpaired test? Justify your answer.

(∗) ii) Do you choose a one- or two-sided test? Justify your answer.

(∗∗) iii) We assume normal distribution of the data. Carry out a hypothesis test at a significance level of 5 %.

Formulate the null and alternative hypothesis and perform the test and make the test decision.

Make the test decision with the confidence interval.

(∗∗) iv) If you do not assume a normal distribution, which test do you choose? Carry out this test and interpret the result.

b) We are studying the differences in height between women and men. In general, men are larger than women. In England, according to Wikipedia, men are on average 13 cm taller than women.

(∗) i) Which test is appropriate here (one- or two-sided, paired or unpaired)? Justify your answer.

(∗) ii) Formulate the null and alternative hypothesis.

(∗∗) iii) Is the statement that on average the men are 13 cm taller than the women statistically significantly refuted by our data set at a significance level of 5 %? Perform the test and interpret the result. We assume that the body heights are normally distributed.

Make the test decision with the confidence interval.

Problem 7.7

The body temperature of 10 patients is measured at the time of administration of a drug (T_1) and 2 hours later (T_2). The aim is to test with a hypothesis test whether this drug has a fever-lowering effect.

Patient-Nr.	1	2	3	4	5	6	7	8	9	10
Temp. 1 in °C	39.1	39.3	38.9	40.6	39.5	38.4	38.6	39.0	38.6	39.2
Temp. 2 in °C	38.1	38.3	38.8	37.8	38.2	37.3	37.6	37.8	37.4	38.1

(∗) a) Is it a paired or unpaired test? Justify your answer.

(∗∗) b) Is it a one- or two-sided test? Justify your answer.

(∗) c) Formulate the null and alternative hypothesis.

(∗∗) d) Assume that the data are normally distributed. Which test do you choose? Carry out the test with R at significance level 5 %. What is your conclusion?

(∗∗) e) If we cannot assume that the data are normally distributed, which test do you choose? Perform this at significance level 5 %.

(∗∗) f) Explain the difference of the p-values in subtasks d) and e).

Problem 7.8

Consider a one-sided t-test of $H_0 : \mu = 0$ against $H_A : \mu > 0$ at the significance level of 0.05.

Although the observed n data points have an empirical mean greater than 0, the calculations show that the null hypothesis is not rejected.

Decide whether the following statements are *true or false*.

Hint: Make useful sketches including all relevant information.

$(***)$ a) We reject H_0 for no level $\alpha < 0.05$.

$(***)$ b) There is a level $\alpha < 1$ where we discard H_0.

$(***)$ c) The p-value is strictly smaller than 0.5.

$(***)$ d) If we perform a two-sided test at the level 0.05 instead of a one-sided test, we do not discard H_0.

$(***)$ e) If we copy the data more and more often (i.e. we look at each data point k times, so that we obtain a total of $k \cdot n$ data points with the same mean as for n data points), we discard H_0 for a large k at significance level of 0.05.

Classical and Bayesian Statistic

Sample solution for Problems 7

Solution 7.1

a) *Paired sample*: Each platelet count *before* smoking corresponds to the platelet count of the same person *after* smoking.

One-sided test: We do not want to know whether the platelet count has *changed*, but whether it has *increased*.

H_0: Smoking has no influence on the accumulation of platelets. ($\mu_S = \mu_{NS}$)

H_A: Smoking increases the accumulation of platelets. ($\mu_S > \mu_{NS}$)

b) *Paired sample*: To each height of a self-pollinated seedling belongs the height of the cross-pollinated "partner".

One-sided test: We do not want to know whether the heights *differ*, but whether the cross-pollinated seedlings become *larger* than the self-pollinated ones.

H_0: The heights do not differ.($\mu_c = \mu_s$)

H_A: Cross-pollinated seedlings become larger than self-pollinated ones. ($\mu_c > \mu_s$)

c) *Unpaired sample*: Unequal numbers in the groups. One blood pressure measurement from the experimental group does not correspond to a specific one from the control group.

Two-sided test: We just want to know if the calcium has an effect on the blood pressure, *no matter* if the blood pressure is higher or lower.

H_0: Calcium has no effect on blood pressure. ($\mu_{\text{Calcium}} = \mu_{\text{Contrast}}$)

H_A: Calcium has an effect on blood pressure. ($\mu_{\text{Calcium}} \neq \mu_{\text{Contrast}}$)

d) *Unpaired sample*: The numbers in the two groups need not be of equal size. The iron measurement of a "Fe^{2+}-mouse" does not correspond to a specific measurement of a "Fe^{3+}-mouse".

Two-sided test: We just want to know if the mice absorb the different forms of iron *differently*.

H_0: Iron absorption is independent of iron variety. ($\mu_2 = \mu_3$)

H_A: The iron absorption depends on the iron variety. ($\mu_2 \neq \mu_3$)

Solution 7.2

a) These are *paired* samples. Measurements are taken at the same location with both gauges.

b) It is assumed that gauge B has the higher values than gauge A. So we perform a one-sided test.

c)
- *Model:*

$$D_1, \ldots, D_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2),$$

σ is estimated by $\widehat{\sigma}$ and $d_i = x_A - x_B$.

- *Null hypothesis*

$$H_0: \quad \mu_D = \mu_0 = 0$$

Alternative hypothesis:

$$H_A: \ \mu_D < \mu_0$$

- *Significance level:*

$$\alpha = 5\,\%$$

- *p-value*

0.01168

```
A <- c(120, 265, 157, 187, 219, 288, 156, 205, 163)
B <- c(127, 281, 160, 185, 220, 298, 167, 203, 171)

t.test(A, B, paired = TRUE, alternative = "less")

##
##  Paired t-test
##
## data:  A and B
## t = -2.7955, df = 8, p-value = 0.01168
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -1.93449
## sample estimates:
## mean of the differences
##               -5.777778
```

- *Test decision*

The p-value is less than 0.05 and thus the null hypothesis is rejected. The gauge B produces indeed statistically significantly larger values than gauge A.

Solution 7.3

Load the dataset

```
jackals <- read.table(file="./Data/jackals.txt",header=TRUE)
head(jackals)


##     M   F
## 1 120 110
## 2 107 111
## 3 110 107
## 4 116 108
## 5 114 110
## 6 111 105
```

a) The samples are unpaired, as the individual males do not correspond to a specific female. The numbers in the two samples need not be the same.

b) We introduce the following terms:

- X_i: ith value of the jaw length of the males, $i = 1, \ldots, n = 10$
- Y_j: jth value of the jaw length of the female, $j = 1, \ldots, m = 10$

Model:

$$X_i \text{ i.i.d. } \mathcal{N}(\mu_M, \sigma_M^2), \qquad Y_i \text{ i.i.d. } \mathcal{N}(\mu_F, \sigma_F^2)$$

Null hypothesis:

$$H_0: \quad \mu_M = \mu_F$$

Alternative hypothesis:

$$H_A: \quad \mu_M \neq \mu_F$$

c) The R output for the t-test:

```
t.test(jackals[, "M"], jackals[, "F"])

##
##  Welch Two Sample t-test
##
## data:  jackals[, "M"] and jackals[, "F"]
## t = 3.4843, df = 14.894, p-value = 0.00336
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.861895 7.738105
## sample estimates:
## mean of x mean of y
##     113.4     108.6
```

The p-value is $0.0034 < 0.05$, hence the null hypothesis is rejected. The male and female jackals have statistically significantly different jaw length.

d) The R output for the Wilcoxon-test looks:

```
wilcox.test(jackals[, "M"], jackals[, "F"])

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  jackals[, "M"] and jackals[, "F"]
## W = 87.5, p-value = 0.004845
## alternative hypothesis: true location shift is not equal to 0
```

The p-value is $0.0048 < 0.05$, hence the null hypothesis is rejected in this test as well.

e) The result of the Wilcoxon-test is more trustworthy because, unlike the t-test, it does not assume that the data are normally distributed and we cannot verifiy this condition in any way.

However, the very different standard deviations in the two groups may be problematic for both tests.

Solution 7.4

a) We cannot unambiguously assign the observations of one data set to the values of the other data set. So it is an unpaired test. Moreover, the data sets have different lengths.

b) No preference a priori between poultry and beef hot dogs is evident, i.e. we perform a two-sided test.

c) As this is an unpaired test, the means μ_{Beef} and μ_{Poultry} are compared.

Null hypothesis (no difference in calorie content)

$$H_0: \quad \mu_{\text{Beef}} = \mu_{\text{Poultry}}$$

Alternative hypothesis (difference in calorie content)

$$H_A: \quad \mu_{\text{Beef}} \neq \mu_{\text{Poultry}}$$

d) R output:

```
beef <- c(186, 181, 176, 149, 184, 190, 158, 139, 175, 148, 152, 111,
          141, 153, 190, 157, 131, 149, 135, 132)

poultry <- c(129, 132, 102, 106, 94, 102, 87, 99, 170, 113, 135, 142,
       86, 143, 152, 146, 144)

mean(beef)

## [1] 156.85

mean(poultry)

## [1] 122.4706
```

The calorie content of beef hot dogs seems to be much higher than that of poultry hot dogs. The null hypothesis should be rejected.

Note that is *not* good practice to decide from the data whether to perform a one- or two-sided test. We have to make this decision *before* collecting data.

We *could* argue that we already "know" ("studies has shown that. . ."), that beef is more fatty than poultry and *then* a one-sided test would be appropriate. But this information comes from *outside* the data sets.

e) Since there is no indication whether the data are normally distributed, we choose a Wilcoxon test as a precautionary measure.

f) R output:

```
wilcox.test(beef, poultry, paired = FALSE)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  beef and poultry
## W = 285.5, p-value = 0.0004549
## alternative hypothesis: true location shift is not equal to 0
```

The p-value is 0.00046 and thus far below the significance level of 0.05. Hence, the null hypothesis that the two types of hot dog have the same statistically significant calorie content is *rejected*.

Now we can argue based on d) that beef hot dogs has statistically significantly more calories than poultry hot dogs.

Solution 7.5

Load the file:

```
mdma <- read.table("./Data/mdma.txt", header = TRUE)
head(mdma)


##   Zurich Basel
## 1   16.3 10.40
## 2   12.7  8.91
## 3   14.0 11.70
## 4   53.3 29.90
## 5  117.0 46.30
## 6   62.6 25.00
```

a) We estimate with R the mean value and the standard deviation as follows:

```
d <- mdma$Zurich - mdma$Basel

mean(d)

## [1] 20.27

sd(d)

## [1] 26.2723
```

We see that the standard deviation is very large compared to the mean. This could be an indication that the data are problematic for a hypothesis test.

b) We can consider the *days* as an experiment units, then these are paired samples because we have two observations per day.

However, it could also be argued that the *cities* are experiment units. In this case the samples are considered unpaired.

c) The null hypothesis is that there is no difference between the two cities in terms of the quantity of MDMA extracted, i.e.

$$H_0: \quad \mu_D = \mu_0 = 0$$

The alternative hypothesis is

$$H_A: \quad \mu_D \neq \mu_0 = 0$$

Note that we *cannot* decide on a one-sided test because of the claim of the free newspaper, namely that more drugs are consumed in Zurich and therefore more MDMA is extracted, therefore

$$\mu_D > \mu_0 = 0$$

That statement is based on the *same* data which is not good practice (see DoE).

d) Output:

```
t.test(mdma$Zurich,
       mdma$Basel,
       paired = TRUE)

##
##  Paired t-test
##
## data:  mdma$Zurich and mdma$Basel
## t = 2.0413, df = 6, p-value = 0.08729
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.027829 44.567829
## sample estimates:
## mean of the differences
##                   20.27
```

From the R output we can see that the p-value is 0.08729 and therefore greater than $\alpha = 0.05$. So at 5 % level of significance we do not reject the null hypothesis. Hence, there is not significantly more ecstasy is consumed in Zurich than in Basel. The claim of the newspaper is not valid.

If we interpret the samples as unpaired, then

```
t.test(mdma$Zurich,
       mdma$Basel,
       paired = FALSE)

##
##  Welch Two Sample t-test
##
## data:  mdma$Zurich and mdma$Basel
## t = 1.3273, df = 7.5245, p-value = 0.2233
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -15.33677  55.87677
## sample estimates:
## mean of x mean of y
##  43.35714  23.08714
```

We have the same test decision.

e) From the R output for a paired test, we can see that the 95 % confidence interval is given by:

$$[-4.028,\ 44.57]$$

With 95 % probability, the true difference between the values of MDMA per thousand inhabitants in Zurich and Basel lies in the interval $[-4.028,\ 44.57]$.

Because 0 is contained in the 95 % confidence interval, we cannot reject the null hypothesis at the 5 % significance level.

f) Output:

```
wilcox.test(mdma$Zurich,
            mdma$Basel,
            paired = TRUE)

##
##  Wilcoxon signed rank exact test
##
## data:  mdma$Zurich and mdma$Basel
## V = 27, p-value = 0.03125
## alternative hypothesis: true location shift is not equal to 0
```

In this case we reject the null hypothesis, because the p-value is 0.0312 and thus lower than the significance level 5 %.

The different results for t-test and Wilcoxon-Test are probably because of the huge standard deviation compared to the mean. The normal distribution assumption may not be satisfied, so the Wilcoxon-Test is more trustworthy and the claim of the newspaper is correct.

Solution 7.6

a) i) It is a paired test. For each test unit (married couple) there are two associated measurements (age husband, age wife).

ii) We are not sure whether the husbands are really older than their wives. It is simply our *impression* and *not* a fact. So perform do a two-sided test.

iii) Let D denote the age difference between husband and wife.

Null hypothesis

$$H_0: \quad \mu_D = 0$$

Alternative hypothesis:

$$H_0: \quad \mu_D \neq 0$$

```
t.test(diff)

##
##  One Sample t-test
##
## data:  diff
## t = 7.1518, df = 169, p-value = 2.474e-11
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.618286 2.852302
## sample estimates:
## mean of x
##  2.235294
```

or

```
t.test(mf$age.husband,
       mf$age.wife,
       paired = TRUE)

##
##  Paired t-test
##
## data:  mf$age.husband and mf$age.wife
## t = 7.1518, df = 169, p-value = 2.474e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.618286 2.852302
## sample estimates:
## mean of the differences
##                2.235294
```

The p-value is far below the significance level of 5 % and thus the null hypothesis is rejected. The husbands are statistically significantly older than their wives.

The confidence interval is

$$(1.61,\ 2.85)$$

With a probability of 95 % the true mean lies in this interval. The null hypothesis was $\mu_D = 0$, i.e. no age difference. This value does *not* lie in the confidence interval and therefore the null hypothesis is rejected here as well. There is a statistically significant age difference within the married couples.

iv) Wilcoxon-test:

```
wilcox.test(diff)

##
##  Wilcoxon signed rank test with continuity correction
##
## data:  diff
## V = 9460, p-value = 3.977e-12
## alternative hypothesis: true location is not equal to 0
```

or

```
wilcox.test(mf$age.husband,
            mf$age.wife,
            paired = TRUE)

##
##  Wilcoxon signed rank test with continuity correction
##
## data:  mf$age.husband and mf$age.wife
## V = 9460, p-value = 3.977e-12
## alternative hypothesis: true location shift is not equal to 0
```

Again, the p-value is far below the significance level of 5 % and thus the null hypothesis is rejected. The husbands are statistically significantly older than their wives.

b) i) In this case only the average heights of men and women are compared, so it is an unpaired test. Since we do not know whether the deviation is upwards or downwards from 13 cm, we again perform a two-sided test.

Note that it is *not* the question whether men are taller than women. The question whether men are on average 13 cm taller than women.

ii) Let μ_W be the average height of the women and μ_M be the average height of the men. The null hypothesis is

$$H_0: \quad \mu_W = \mu_M - 13$$

and the alternative hypothesis is

$$H_A: \quad \mu_W \neq \mu_M - 13$$

iii) We assume normal distribution of body heights:

```
t.test(mf$height.husband - 13,
       mf$height.wife,
       paired = FALSE)

##
##  Welch Two Sample t-test
##
## data:  mf$height.husband - 13 and mf$height.wife
## t = -0.63293, df = 336.53, p-value = 0.5272
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.812281  0.929928
## sample estimates:
## mean of x mean of y
##  159.8471  160.2882
```

or

```
t.test(mf$height.husband,
       mf$height.wife,
       mu = 13,
       paired = FALSE)

##
##  Welch Two Sample t-test
##
## data:  mf$height.husband and mf$height.wife
## t = -0.63293, df = 336.53, p-value = 0.5272
## alternative hypothesis: true difference in means is not equal to 13
## 95 percent confidence interval:
##  11.18772 13.92993
## sample estimates:
## mean of x mean of y
##  172.8471  160.2882
```

The p-value is far greater than the significance level and thus the null hypothesis is not rejected. The data do *not* refute the difference in height of 13 cm statistically significant.

The confidence interval is

$$(-1.81,\ 0.92)$$

With a probability of 95 % the true mean lies in this interval. The null hypothesis was $\mu_W = \mu_M - 13$, i.e. no statistically significant deviation from the

height difference of 13 cm. This value 0 lies in the confidence interval and thus the null hypothesis is *not* rejected. The data do *not* refute the difference in height of 13 cm in a statistically significant way.

If we do not assume normal distribution, we choose an unpaired Wilcoxon-test (Mann-Whitney-U).

```
wilcox.test(mf$height.husband - 13, mf$height.wife, paired = FALSE)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  mf$height.husband - 13 and mf$height.wife
## W = 13760, p-value = 0.4461
## alternative hypothesis: true location shift is not equal to 0
```

Again, the null hypothesis is clearly *not* rejected.

Solution 7.7

a) It is a paired test, as two measurements were taken on one test unit (patient).

b) We want to test the fever-lowering effectiveness. For this purpose we calculate the average of the μ_D of the differences D_i (Temp. 1 − Temp. 2). In order to be able to prove the effectiveness

$$\mu_D > 0$$

Note that we are only interested in the fever-*lowering* and *not* the fever-rising effect. Therefore a one-sided test.

c) Null hypothesis (drug has no effect)

$$H_0: \quad \mu_D = 0$$

Alternative hypothesis (drug is fever-lowering)

$$H_A: \quad \mu_D > 0$$

d) R output:

```
t_1 <- c(39.1, 39.3, 38.9, 40.6, 39.5, 38.4, 38.6, 39.0, 38.6, 39.2)
t_2 <- c(38.1, 38.3, 38.8, 37.8, 38.2, 37.3, 37.6, 37.8, 37.4, 38.1)

t.test(t_1, t_2, paired=T, alternative="greater")

##
##  Paired t-test
##
## data:  t_1 and t_2
## t = 5.6569, df = 9, p-value = 0.0001554
```

```
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.7976252       Inf
## sample estimates:
## mean of the differences
##                    1.18
```

The p-value 0.0001554 is less than 0.05 here. Therefore the difference is statistically significant. We can therefore assume that the drug is antipyretic (lowering the fever).

e) R output:

```
wilcox.test(t_1, t_2, paired=T, alternative="greater")

##
##  Wilcoxon signed rank test with continuity correction
##
## data:  t_1 and t_2
## V = 55, p-value = 0.002865
## alternative hypothesis: true location shift is greater than 0
```

The p-value 0.002865 is less than 0.05. Therefore the difference is statistically significant. We can therefore assume that the drug is antipyretic.

f) The p-value of the Wilcoxon-test is greater than the p-value of the t-test. Since the Wilcoxon-test assumes less (no normal distribution) than the t-test, there is an additional uncertainty. The null hypothesis is less strongly rejected.

However, the t-test suggests that it is more "precise". This is true, but only if the data are normally distributed, which is often not known.

Therefore the Wilcoxon test is often preferable to the t-test.

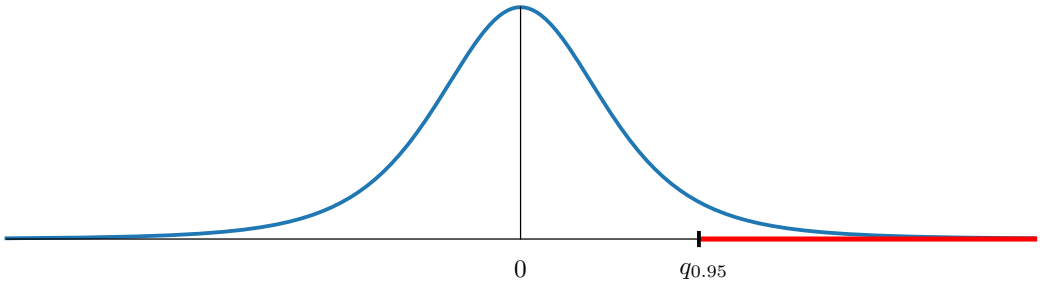
Solution 7.8

This task is not easy and we have to read the text *very* carefully:

Consider a one-sided t-test of $H_0 : \mu = 0$ *against* $H_A : \mu > 0$ *against* H_A *at the level of* 0.05.

We draw a t-distribution with $\mu = 0$ and degree of freedom 4 (these assumptions are

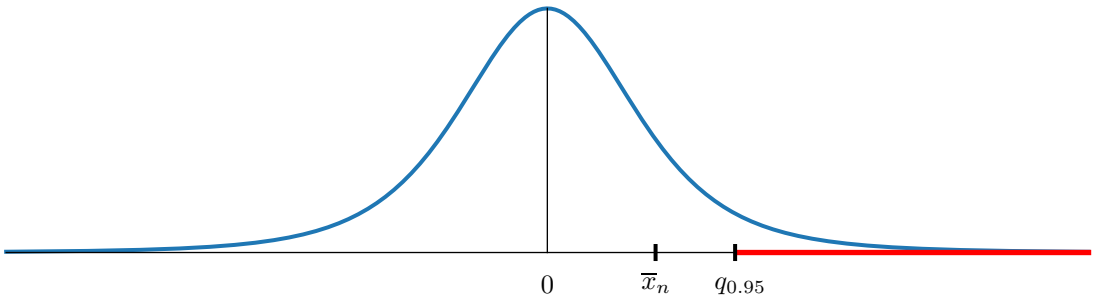
not relevant) and with the rejection zone (right-tailed test).



Although the observed n data points have an empirical mean greater than zero, the calculations show that the null hypothesis is not rejected.

What does this mean?

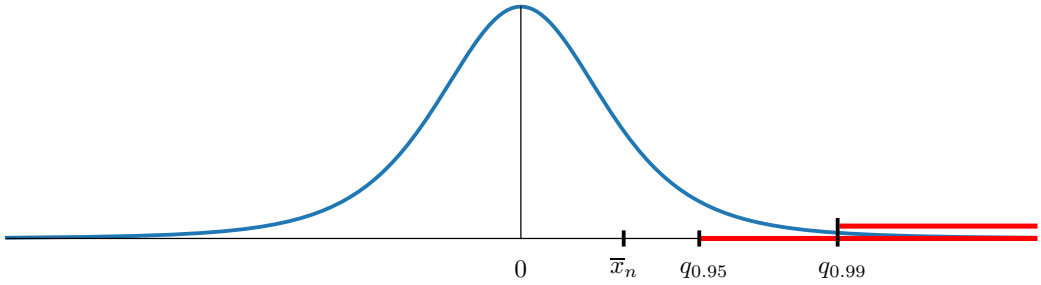
- First of all, the mean is $\overline{x}_n > 0$ and is to the right of 0 on the x-axis.
- H_0 is *not* rejected, hence $\overline{x}_n$ is not in the rejection zone.
- $\overline{x}_n$ lies somewhere between 0 and the boundary of the rejection zone.



a) *We discard H_0 for no level $\alpha < 0.05$.*

What does that mean? Our significance level is $\alpha = 0.05$. Now we choose a significance level α^* that is *less* than 0.05, for example $\alpha^* = 0.01$. The boundary of the rejection zone moves to the right from $q_{0.95}$ to $q_{0.99}$ and the rejection zone becomes smaller. This means that $\overline{x}_n$ is still *not* in the rejection zone (see Figure

below).



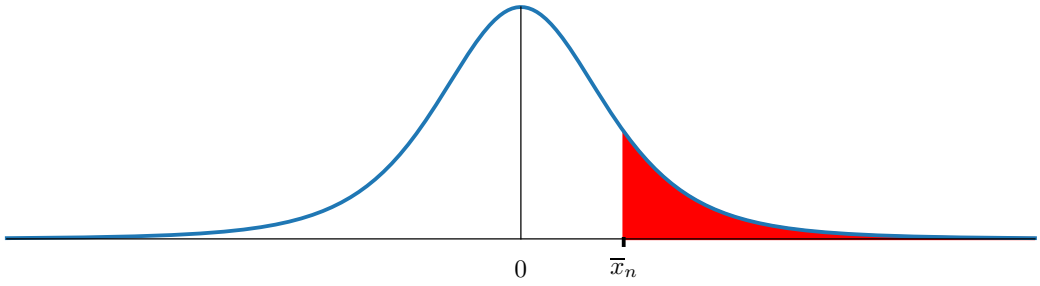
So for all $\alpha < 0.05$ the null hypothesis is not rejected, because $\overline{x}_n$ can never be in the rejection zone. So the statement is correct.

b) *There is a level $\alpha < 1$ where H_0 is rejected.*

The reasoning is the same as in a) but in the opposite direction: Let $\alpha = 0.05$ and we *increase* α, then the boundary of the rejection zone moves to the left and for some α, $\overline{x}_n$ is in the rejection zone and the null hypothesis is rejected. So the statement is correct.

c) *The p-value is strictly smaller than 0.5.*

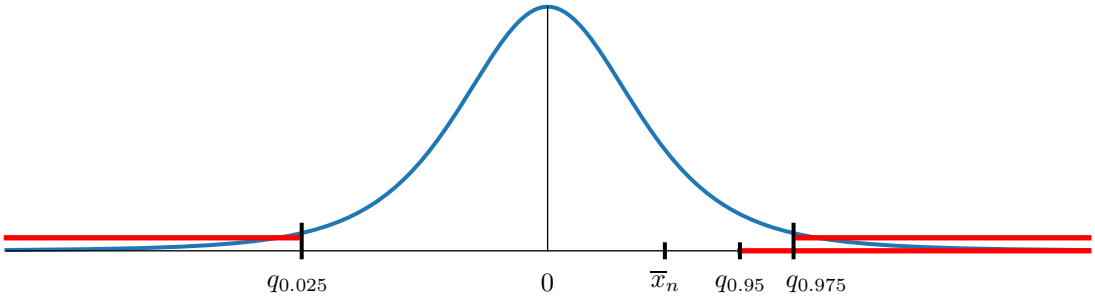
Now what is the p-value? This is the probability of observing a certain value (here $\overline{x}_n$) or a more extreme value towards the alternative hypothesis. We can represent probabilities as areas in continuous distributions and this is the area under the curve to the *right* of $\overline{x}_n$ (see Figure below).



Now the total area under the curve is 1, the area to the right of 0 is 0.5 and then the red area (p-value) must be less than 0.5. So the statement is correct.

d) *If we perform a two-sided test on significance level of 0.05 instead of a one-sided test, we do not discard H_0.*

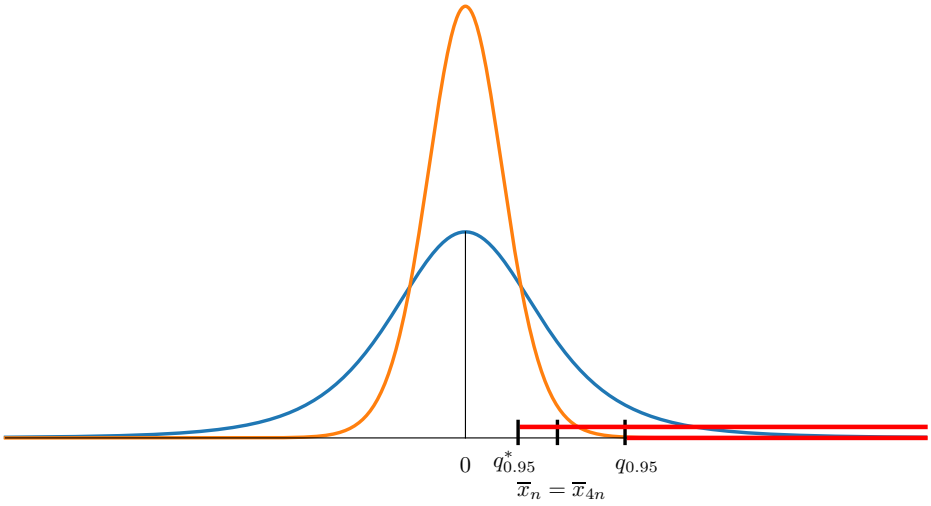
Similar answer as in a): If we switch from a one-sided to a two-sided test, the rejection zone on the right-hand side becomes smaller. Since $\overline{x}_n$ is not in the rejection zone of the one-sided test, it cannot be in the rejection zone of the two-sided test. So the statement is correct.



e) *If we copy the data more and more often (i.e. we look at each data point k times, so that we obtain a total of $k \cdot n$ data points with the same mean as for n data points), we discard H_0 for a large k at significance level of 0.05.*

For example, if we quadruple the data points ($k = 4$), then the mean remains the same and the standard deviation *decreases* by a factor $\sqrt{k} = 2$.

Hence $\overline{x}_n = \overline{x}_{nk}$, but the number of observations increases. That means the curve becomes narrower and the boundary of the rejection zone moves to the left. For a sufficiently large k, the mean $\overline{x}_n$ lies in the rejection zone and the null hypothesis is rejected. So the statement is correct.



Classical and Bayesian Statistics

Problems 8

Problem 8.1

We look at a study conducted in the United States in 1979 (National Longitudinal Study of Youth, NLSY79): of 2584 Americans in 1981, the intelligence quotient (as per the AFQT - armed forces qualifying test score) was measured; in 2006, the same people were asked about their annual income in 2005 and the number of years of schooling.

We are naturally interested here in whether a high IQ or a long school education leads to a higher income.

In the file **income.dat** on ILIAS you will find the data set with the income, the number of years of completed school education and the intelligence quotients of 2584 Americans.

(∗) a) Read the data set **income.dat**.

```
... <- read.table(..., header = TRUE)
```

(∗∗) b) Generate scatter plots with the corresponding regression lines showing income versus number of years of schooling and income versus intelligence quotient. What do you find?

(∗∗) c) Determine the parameters a and b of the linear model $y = a + bx$, where y is the income and x is the number of years of schooling.

How do you interpret the parameters a and b?

(∗∗) d) Calculate the correlation between income and number of years of schooling. How appropriate is a regression model for this data set?

Problem 8.2

In this assignment we look at 4 data sets constructed by the statistician Francis John Anscombe (13 May 1918 – 17 October 2001). In each of the records there is a response variable y and an predictor x.

a) The file is already included in **R**.

```
head(anscombe)

##   x1 x2 x3 x4   y1   y2    y3   y4
## 1 10 10 10  8 8.04 9.14  7.46 6.58
## 2  8  8  8  8 6.95 8.14  6.77 5.76
## 3 13 13 13  8 7.58 8.74 12.74 7.71
## 4  9  9  9  8 8.81 8.77  7.11 8.84
## 5 11 11 11  8 8.33 9.26  7.81 8.47
## 6 14 14 14  8 9.96 8.10  8.84 7.04
```

(∗) b) Produce a scatter plot each of the 4 data sets, draw the regression line and comment on the results.

```
plot(anscombe$x1, anscombe$y1)
reg <- lm(anscombe$y1 ~ anscombe$x1)
abline(reg)
```

With `par(mfrow=c(2,2))` the graphics window is divided so that all 4 panes fit next to each other.

(∗) c) Compare a and b each, where $y = a + bx$.

```
lm(y1 ~ x1, data = anscombe) # or

lm(anscombe$y1 ~ anscombe$x1)
```

d) Determine the correlation coefficients. What stands out?

Problem 8.3

In this task we use the data set **Auto**, which is contained in the library **ISLR**.

```
library(ISLR)
```

If an error message appears, the library must be installed first (this only needs to be done once):

```
install.packages("ISLR")
```

(∗∗) a) Investigate the data record with `head(Auto)` and `?Auto` or `help(Auto)`.

(∗) b) Adjust the model to a simple linear regression with `mpg` as the target variable and `horsepower` as the predictor.

c) Use the `lm()` command to perform this regression.

Use the `summary()` command to output the results. Comment on it:

(**) i) Is there a connection between the target variable and the predictor?

(**) ii) How do you interpret the coefficients for `(intercept)` and `horsepower`? Is the correlation positive or negative?

(**) iii) How do you determine the confidence intervals (with `confint()`) and interpret them?

(**) iv) Interpret the R^2 value.

(**) d) Plot response variable and the predictor with the regression line (`abline`). How do you interpret this plot compared to the `summary()` output.

Problem 8.4

The `MASS` library contains the `Boston` data set, which records `medv` (median house value) for 506 neighborhoods around Boston. We will seek to predict `medv` using 13 predictors such as `rm` (average number of rooms per house), `age` (average age of houses), and `lstat` (percent of households with low socioeconomic status).

(**) a) To find out more about the data set, we can type `?Boston`.

(**) b) Which column names are available?

(*) c) Use the `attach(...)`-command to let `R` recognize the column names of the data set `Boston`.

d) We will start by using the `lm()` function to fit a simple linear regression model, with `medv` as the response and `lstat` as the predictor.

(*) i) Define the simple regression model using the two variables above.

(**) ii) The basic syntax is `lm(y~x,data)` , where `y` is the response, `x` is the predictor, and data is the data set in which these two variables are kept.

```
lm.fit <- lm(...)
summary(lm.fit)
```

(*) e) We can use the `names(...)` function in order to find out what other pieces of information are stored in `lm.fit`.

(**) f) Although we can extract these quantities by name — e.g. `lm.fit$coefficients` — it is safer to use the extractor functions like `coef(...)` to access them.

Interpret these values and the corresponding p–values in the summary above.

(∗∗) g) In order to obtain a confidence interval for the coefficient estimates, we can use the `confint(...)` command.

Give an interpretation of these values.

(∗∗) h) We will now plot `medv` and `lstat` along with the least squares regression line using the `plot(...)` and `abline()` functions (see exercise sheet 2).

Use `lty = ...`, `pch = ...` and `col = ...` to make the plot look nicer.

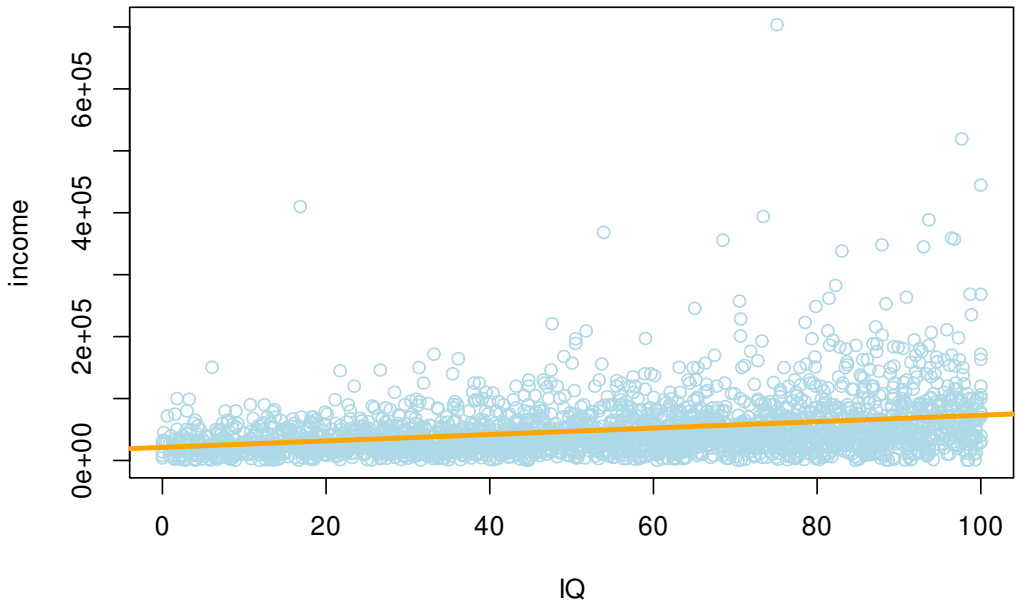
(∗∗) i) Interpret the R^2 value in the `summary`–output above.

Classical and Bayesian Statistic

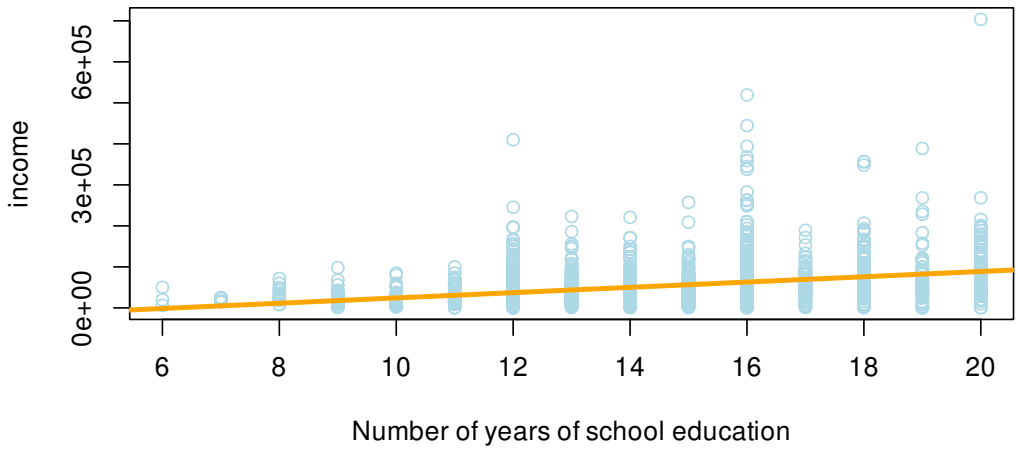
Sample solution for Problems 8

Solution 8.1

a)
```
income <- read.table(file="./Daten/income.dat", header=TRUE)
head(income)

##     AFQT Educ Income2005
## 1  6.841   12       5500
## 2 99.393   16      65000
## 3 47.412   12      19000
## 4 44.022   14      36000
## 5 59.683   14      65000
## 6 72.313   16       8000

iq <- income[,1]

number_years_of_school <- income[, 2]

income <- income[,3]

plot(iq,
     income,
     type = "p",
     xlab = "IQ",
     ylab = "income",
     col = "light blue"

)

abline(lm(income ~ iq),
       col = "orange",
       lwd = 3)
```




```r
plot(number_years_of_school,
     income,
     type="p",
     xlab = "Number of years of school education",
     ylab="income",
     col = "light blue"

)

abline(lm(income ~ number_years_of_school),
       col = "orange",
       lwd = 3)
```



In both cases, the regression line is very flat and the points scatter quite a bit around the regression line.

b) With R we calculate for a and b

```
lm(income ~ number_years_of_school)

##
## Call:
## lm(formula = income ~ number_years_of_school)
##
## Coefficients:
##             (Intercept)  number_years_of_school
##                  -40200                    6451
```

So we find the values $a = -40'200$ and $b = 6451$ für the case of income versus number of years of schooling (and $a = 21'182$ and $b = 518.68$ für the case of income versus intelligence quotient). Thus, every additional year of schooling is accompanied by an annual increase in income of 6451 USD. But be careful: someone without education would have an income of $-40'200$ USD. Of course this makes no sense. Whenever extrapolating into areas where no data points were available, caution should be exercised in interpretation.

c) For the *empirical correlation* we get

```
cor(number_years_of_school, income)

## [1] 0.3456474
```

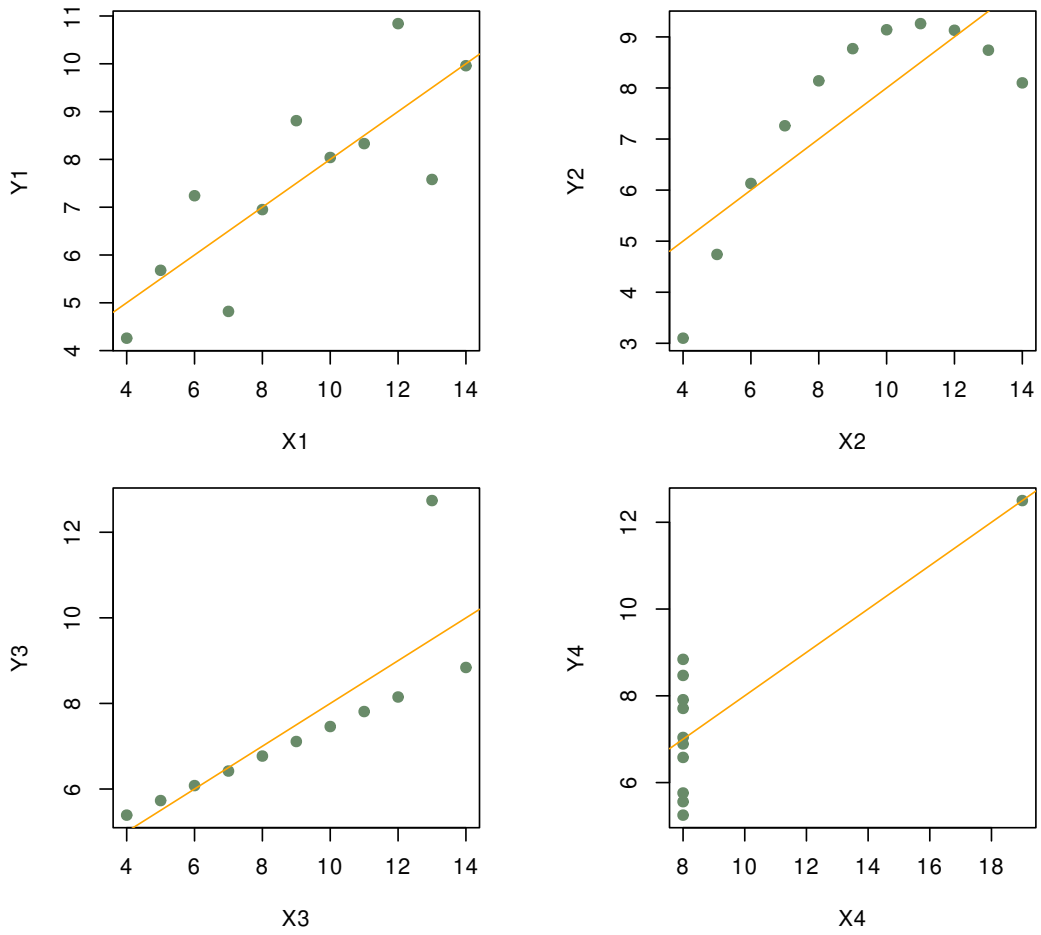
As the correlation coefficient is relatively small, a model based on a linear relationship between income and number of years of schooling does not seems to be appropriate.

Solution 8.2

a) If we look at the four scatter plots, we see that only in the first case a linear regression is correct. In the second case, the relationship between X and Y is not linear but quadratic. In the third case there is an outlier which strongly influences the estimated parameters. In the fourth case, the regression line is determined by a single point.

```
data(anscombe)
reg <- lm(anscombe$y1 ~ anscombe$x1)
reg2 <- lm(anscombe$y2 ~ anscombe$x2)
reg3 <- lm(anscombe$y3 ~ anscombe$x3)
reg4 <- lm(anscombe$y4 ~ anscombe$x4)
par(mfrow=c(2,2))
plot(anscombe$x1, anscombe$y1, ylab = "Y1", xlab = "X1", col="darkseagreen4", pch=
abline(reg, col = "orange")
plot(anscombe$x2, anscombe$y2, ylab = "Y2", xlab = "X2", col="darkseagreen4", pch=
abline(reg2, col = "orange")
plot(anscombe$x3, anscombe$y3, ylab = "Y3", xlab = "X3", col="darkseagreen4", pch=
abline(reg3, col = "orange")
plot(anscombe$x4, anscombe$y4, ylab= "Y4", xlab= "X4", col="darkseagreen4", pch=19
```

```
abline(reg4, col = "orange")
```



b) For all four models the estimates of the axis intercept β_0 and the slope β_1 are almost identical:

	model 1	model 2	model 3	model 4
Axis intercept (a)	3.000	3.001	3.002	3.002
Slope (β_1)	0.500	0.500	0.500	0.500

Conclusion: It is *not* enough to look at a and b. In all models these estimates are almost the same, but the records look very different. A (graphical) check of the model assumptions is therefore inevitable.

c)
```
cor(anscombe$x1, anscombe$y1)

## [1] 0.8164205

cor(anscombe$x2, anscombe$y2)
```

```
## [1] 0.8162365

cor(anscombe$x3, anscombe$y3)

## [1] 0.8162867

cor(anscombe$x4, anscombe$y4)

## [1] 0.8165214
```

Although the scatter plots look very different, the correlation coefficients are the same except for the 3rd digit after the decimal point.

Again: Don't just look exclusively at the correlation coefficient, but at the corresponding scatter plot.

Solution 8.3

a) Table:

```
library(ISLR)
head(Auto)

##   mpg cylinders displacement horsepower weight acceleration year
## 1  18         8          307        130   3504         12.0   70
## 2  15         8          350        165   3693         11.5   70
## 3  18         8          318        150   3436         11.0   70
## 4  16         8          304        150   3433         12.0   70
## 5  17         8          302        140   3449         10.5   70
## 6  15         8          429        198   4341         10.0   70
##   origin                       name
## 1      1 chevrolet chevelle malibu
## 2      1         buick skylark 320
## 3      1        plymouth satellite
## 4      1             amc rebel sst
## 5      1               ford torino
## 6      1          ford galaxie 500
```

```
help(Auto)
```

Auto {ISLR}

Auto Data Set

Description

Gas mileage, horsepower, and other information for 392 vehicles.

Usage

```
Auto
```

Format

A data frame with 392 observations on the following 9 variables.

`mpg`
: miles per gallon

`cylinders`
: Number of cylinders between 4 and 8

`displacement`
: Engine displacement (cu. inches)

`horsepower`
: Engine horsepower

`weight`
: Vehicle weight (lbs.)

`acceleration`
: Time to accelerate from 0 to 60 mph (sec.)

`year`
: Model year (modulo 100)

`origin`
: Origin of car (1. American, 2. European, 3. Japanese)

`name`
: Vehicle name

The orginal data contained 408 observations but 16 observations with missing values were removed.

Source

This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. The dataset was used in the 1983 American Statistical Association Exposition.

References

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) *An Introduction to Statistical Learning with applications in R*, www.StatLearning.com, Springer-Verlag, New York

b) Linear regression:

$$\texttt{mpg} = \beta_0 + \beta_1 \cdot \texttt{horsepower}$$

c) Output:

```
fit <- lm(mpg ~ horsepower, data = Auto)
# Or: fit <- lm(Auto$mpg ~ Auto$horsepower)

summary(fit)

##
## Call:
## lm(formula = mpg ~ horsepower, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059,Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

i) The p value for **horsepower** is almost 0 and so the null hypothesis ($\beta_1 = 0$) is rejected. The fuel consumption depends on the horsepower.

ii) The value 39.93 for the **intercept** indicates the fuel consumption (miles per gallon) at 0 hp. Of course this value has no practical meaning here.

More interesting is the value -0.15 for **horsepower**. This means that per hp the car gets 0.15 miles less for one gallon ($\approx 3.8\,\text{l}$) of petrol.

So the correlation is negative: the more horsepower the less distance is travelled per gallon.

iii) confidence interval:

```
confint(fit)

##                 2.5 %     97.5 %
## (Intercept) 38.525212 41.3465103
## horsepower  -0.170517 -0.1451725
```

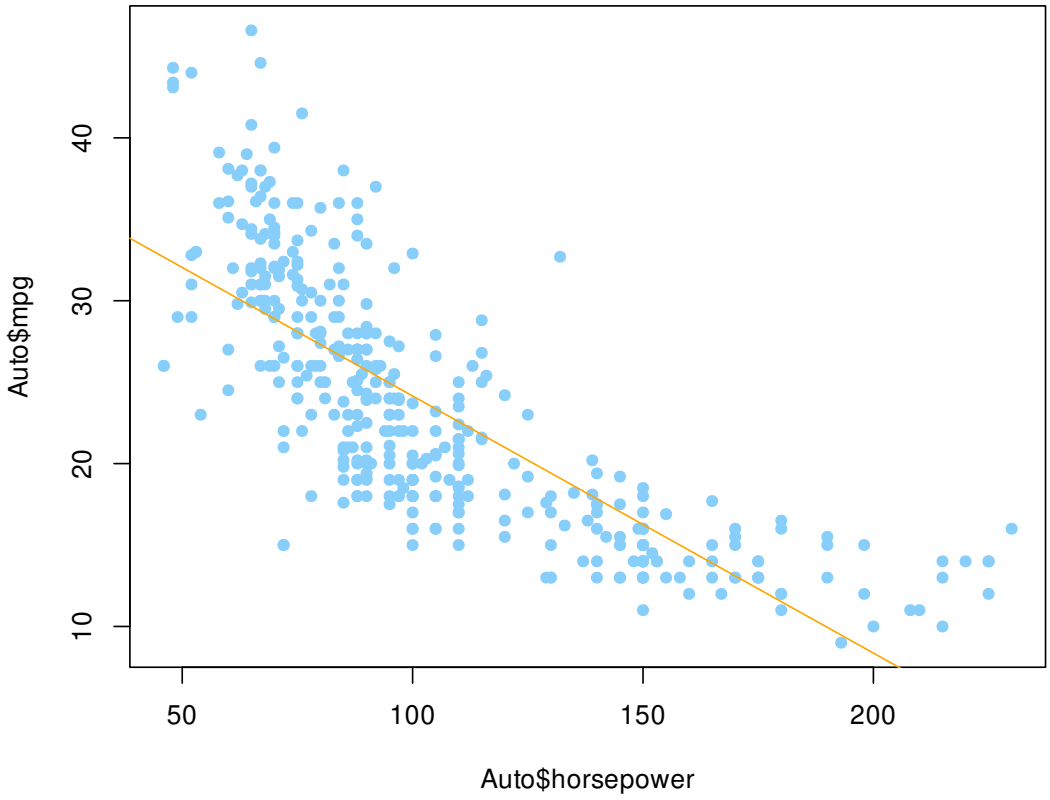
The true values for **intercept** and **horsepower** lie at 95 % in the corresponding intervals. The intervals are quite narrow, so that the significance of these intervals is quite large.

iv) The R^2 value is 0.606. This indicates that the variability to 60 % is through the model.

This is ok, but not very good, because other predictors also have an influence on the fuel consumption.

d) Plot:

```
plot(Auto$horsepower, Auto$mpg, pch=16, col="lightskyblue")
abline(lm(Auto$mpg~Auto$horsepower), col="orange")
```



The downward trend is clearly visible, hence the low p value. However, the point cloud does not fall linearly (weak R^2 value).

Solution 8.4

a)

b) Column names

```
library(MASS)
colnames(Boston)

##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"
##  [7] "age"     "dis"     "rad"     "tax"     "ptratio" "black"
## [13] "lstat"   "medv"
```

c) Attach

```
attach(Boston)
```

d) i) The model is defined as follows:

$$\texttt{medv} = \beta_0 + \beta_1 \cdot \texttt{lstat}$$

ii) Output

```
lm.fit <- lm(medv ~ lstat)
summary(lm.fit)

##
## Call:
## lm(formula = medv ~ lstat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.168  -3.990  -1.318   2.034  24.500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.55384    0.56263   61.41   <2e-16 ***
## lstat       -0.95005    0.03873  -24.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.216 on 504 degrees of freedom
## Multiple R-squared:  0.5441,Adjusted R-squared:  0.5432
## F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

e)

```
names(lm.fit)

##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

f)

```
coef(lm.fit)

## (Intercept)       lstat
##  34.5538409  -0.9500494
```

Substitute these values in the simple linear regression model above

$$\texttt{medv} = 34.554 - 0.95 \cdot \texttt{lstat}$$

The values 34.55 is the intercept, which is the value for `lstat = 0` (zero percent of lower status of the population). The median house value is \$34 554 in neighborhoods with 0 percent population of lower status.

The value -0.95 is the slope of the regression line. We can interpret this value as follows: for each additional percent in population of lower status, the median house value drops by \$950.

g)
```
confint(lm.fit)

##                 2.5 %     97.5 %
## (Intercept) 33.448457 35.6592247
## lstat       -1.026148 -0.8739505
```

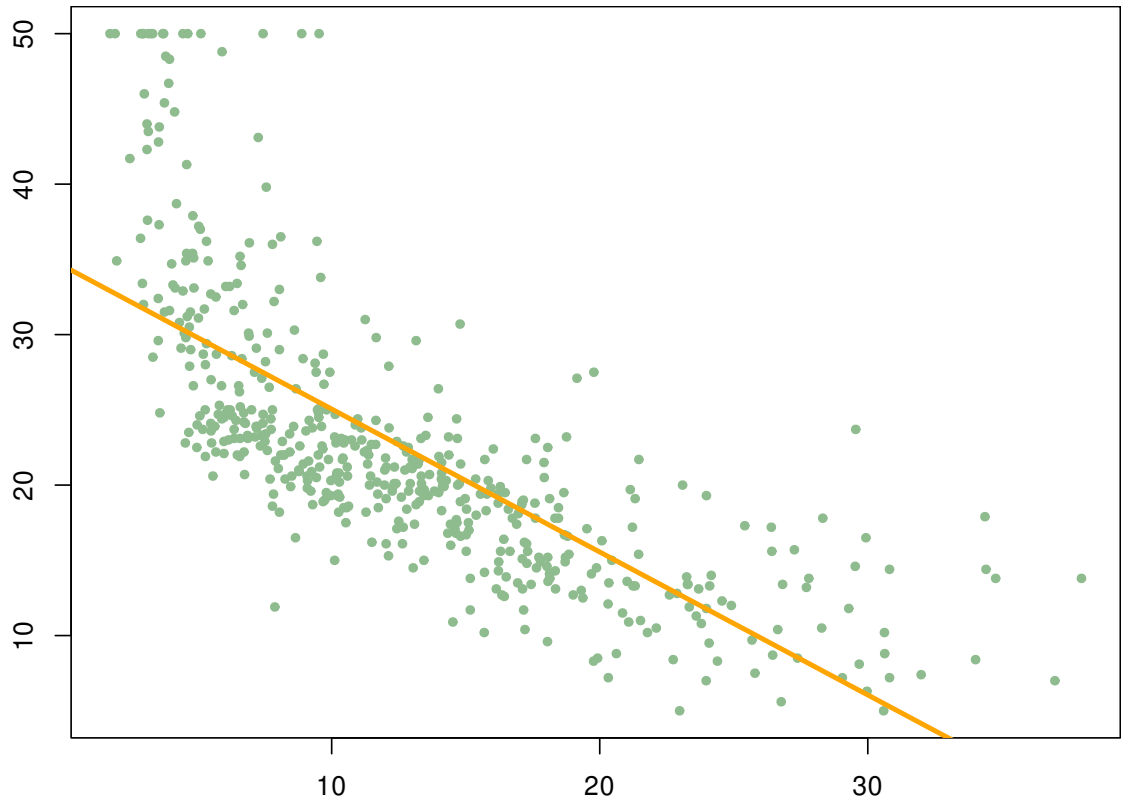
The true value of the intercept is with 95 % probability in the interval

$$[33.45,\ 35,66]$$

The true value of the slope is with 95 % probability in the interval

$$[-1.02,\ -0.87]$$

h) Plot:
```
plot(lstat, medv, col = "darkseagreen", pch = 20)
abline(lm.fit, col= "orange", lwd = 3)
```



Classical and Bayesian Statistics

Problems 9

Problem 9.1

We investigate further the data set `Boston`.

In order to fit a multiple linear regression model using least squares, we again use the lm() function. The syntax `lm(y ~ x1 + x2 + x3)` is used to fit a model with three predictors, `x1` , `x2` , and `x3` . The `summary()` function now outputs the regression coefficients for all the predictors.

(∗∗) a) Fit a multiple linear regression model with response variable `medv` and predictors `lstat` and `age`.

Define the model and interpret all values in the `summary()` output which we discussed in class (coefficients, its P values, R^2 value, P value of the F-statistics).

(∗∗) b) The `Boston` data set contains 13 variables, and so it would be cumbersome to have to type all of these in order to perform a regression using all of the predictors. Instead, we can use the following short-hand `lm(medv ~., data = Boston)`.

In the `summary()` output interpret the coefficient of `age` and the corresponding P value compare this with the output in a) and explain the difference.

(∗∗) c) The R^2 value is bigger than the one calculated in a). Explain.

(∗∗) d) It is easy to include interaction terms in a linear model using the `lm()` function. The syntax `lstat:black` tells `R` to include an interaction term between `lstat` and `black`.

The syntax `lstat * age` simultaneously includes `lstat`, `age`, and the interaction term `lstat × age` as predictors; it is a shorthand for `lstat + age + lstat:age`.

Again, discuss all the values in the `summary()` of `lstat*age` as in a).

Problem 9.2

We want to perform a multiple linear regression for `Auto`.

(**) a) Produce with `pairs` scatterplot containing all variables of the data set.

(**) b) Calculate the correlation matrix between the variables with `cor()`. To do this, we must first remove the variable `name`, as it is qualitative.

```
library(ISLR)

head(Auto)

##   mpg cylinders displacement horsepower weight acceleration year
## 1  18         8          307        130   3504         12.0   70
## 2  15         8          350        165   3693         11.5   70
## 3  18         8          318        150   3436         11.0   70
## 4  16         8          304        150   3433         12.0   70
## 5  17         8          302        140   3449         10.5   70
## 6  15         8          429        198   4341         10.0   70
##   origin                       name
## 1      1 chevrolet chevelle malibu
## 2      1         buick skylark 320
## 3      1        plymouth satellite
## 4      1             amc rebel sst
## 5      1               ford torino
## 6      1          ford galaxie 500

Auto.1 <- within(Auto, rm(name))

head(Auto.1)

##   mpg cylinders displacement horsepower weight acceleration year
## 1  18         8          307        130   3504         12.0   70
## 2  15         8          350        165   3693         11.5   70
## 3  18         8          318        150   3436         11.0   70
## 4  16         8          304        150   3433         12.0   70
## 5  17         8          302        140   3449         10.5   70
## 6  15         8          429        198   4341         10.0   70
##   origin
## 1      1
## 2      1
## 3      1
## 4      1
## 5      1
## 6      1
```

Interpret the values for `horsepower` and `displacement` using the scatter plot above.

c) We use `lm()` to perform a multiple regression with the target variable `mpg` and all other variables (except `name`) as predictors. Use again to interpret the output of the `summary()` command.

i) Is there a relationship between the predictors and the response variable? Justify this with the p value to the F value.

(∗∗) ii) Which predictors seem to have a statistically significant influence on the target variable?

(∗∗) iii) What does the coefficient for `year` indicate?

(∗∗) d) Examine the model from c) still for interaction effects.

Problem 9.3

The library `ISLR` contains the data set `Carseats`. We want `Sales` (number of child car seats) based on different predictors in 400 different locations.

The data set contains qualitative predictors, such as `ShelveLoc` as an indicator of the location in the rack, i.e. the space in a shop where the car seat is displayed. The predictor assumes the three values `Bath`, `Medium` and `Good`. For qualitative variables `R` generates dummy variables automatically.

(∗) a) Examine the data set with `head(Carseat)` and `?Carseat`.

(∗∗) b) Find a multiple regression model with `lm()` to predict `Sales` from `Price`, `Urban` and `US`.

(∗∗) c) Interpret the coefficients in this model. Be aware that some variables are qualitative.

(∗) d) Write the model as an equation. Make sure that you treat the qualitative variables correctly.

(∗) e) For which predictors can the null hypothesis $H_0 : \beta_j = 0$ be rejected?

(∗∗) f) Based on the previous question, find a smaller model that only uses predictors for which there is evidence of a relationship with the response variable.

(∗∗) g) How exactly do the models in b) and f) fit the data?

Classical and Bayesian Statistic

Sample solution for Problems 9

Solution 9.1

a) Model:

$$\texttt{medv} = \beta_0 + \beta_1 \cdot \texttt{lstat} + \beta_2 \cdot \texttt{age}$$

```
library(MASS)
fit <- lm(medv ~ lstat + age, data = Boston)

summary(fit)

##
## Call:
## lm(formula = medv ~ lstat + age, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.981  -3.978  -1.283   1.968  23.158
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.22276    0.73085  45.458  < 2e-16 ***
## lstat       -1.03207    0.04819 -21.416  < 2e-16 ***
## age          0.03454    0.01223   2.826  0.00491 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.173 on 503 degrees of freedom
## Multiple R-squared:  0.5513,Adjusted R-squared:  0.5495
## F-statistic:   309 on 2 and 503 DF,  p-value: < 2.2e-16
```

The estimates are

$$\widehat{\beta}_0 = 33.22; \qquad \widehat{\beta}_1 = -1.03; \qquad \widehat{\beta}_2 = 0 - 03$$

We get for the model

$$\texttt{medv} = 33.22 - 1.03 \cdot \texttt{lstat} + 0.03 \cdot \texttt{age}$$

Interpretation of the estimates:

- $\widehat{\beta}_0 = 33.22$

In neighborhoods where there is no population of lower status and no units build before 1940, the medium value of houses is $ 33 220.

- $\widehat{\beta}_1 = -1.03$

  For each additional percent of population of lower status, the medium value decreases by $ 1030.

- $\widehat{\beta}_2 = 0.03$

  For each additional percent of units build before 1949, the medium value increases by $ 30.

- All p-values are significant (below the significance level of 5 %), so all estimates individually contribute significantly to the model.
- The R^2 value is 0.5513, therefore about 55 % of the variation is explained by the model.
- The p-value of the F value is below the significance level and therefore significant. The null hypothesis H_0

$$\beta_1 = \beta_2 = 0$$

  is rejected. One of β's is signigficantly different from 0. At least one variables contributes significantly to the model.

b)
```
fit <- lm(medv ~ ., data = Boston)

summary(fit)$coefficients[,"Pr(>|t|)"][[2]]

## [1] 0.00108681
```

The p-value is almost 1, so not significant at all. But in a), the p-value is 0.005, which is significant. That means that the variable `age` must correlate strongly with other variables (see d)).

c) The more variables you have the bigger the R^2 value. That means that the R^2 is not a good indicator to compare different models.

d) Model:

$$\texttt{medv} = \beta_0 + \beta_1 \cdot \texttt{lstat} + \beta_2 \cdot \texttt{age} + \beta_{12} \cdot \texttt{lstat*age}$$

Remark: $*$ in `lstat*age` does *not* signify multiplication, it just means interaction.

```
fit <- lm(medv ~ lstat * age, data = Boston)

summary(fit)
```

```
##
## Call:
## lm(formula = medv ~ lstat * age, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.806  -4.045  -1.333   2.085  27.552
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.0885359  1.4698355  24.553  < 2e-16 ***
## lstat       -1.3921168  0.1674555  -8.313 8.78e-16 ***
## age         -0.0007209  0.0198792  -0.036   0.9711
## lstat:age    0.0041560  0.0018518   2.244   0.0252 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.149 on 502 degrees of freedom
## Multiple R-squared:  0.5557,Adjusted R-squared:  0.5531
## F-statistic: 209.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

The estimates are

$$\widehat{\beta}_0 = 36.10; \qquad \widehat{\beta}_1 = -1.39; \qquad \widehat{\beta}_2 = -0.0007; \qquad \widehat{\beta}_{12} = 0.004$$

We get for the model

$$\texttt{medv} = 36.10 - 1.39 \cdot \texttt{lstat} - 0.00072 \cdot \texttt{age} + 0.0041 \cdot \texttt{lstat*age}$$

Interpretation of the estimates:

- $\widehat{\beta}_0 = 36.10$

  In neighborhoods where there is no population of lower status and no units build before 1940, the medium value of houses is $ 36 100.

- $\widehat{\beta}_1 = -1.39$

  For each additional percent of population of lower status, the medium value decreases by $ 1930.

- $\widehat{\beta}_2 = -0.00072$

  For each additional percent of units build before 1949, the medium value decreases by $ 0.27.

  As you can imagine, this value is not significant, as you can see from the output.

- $\widehat{\beta}_{12} = 0.004$

  This coefficient is somewhat difficult to interpret and we did't do it in class.

- Not all p-values are significant (below the significance level of 5 %) anymore.

  The p value for `age` ist 0.97, so this not significance anymore, whereas without interaction it was. What is the reason for this?

  The p-value of the interaction term is 0.0252 which is below the significance level of 5 %. The null hypothesis H_0, that there is no interaction, is rejected. There is statistically significant interaction.

  Now, let's take a look at the correlation coefficient of the two explanatory variables `lstat` and `age`.

```
cor(Boston["lstat"],Boston["age"])

##             age
## lstat 0.6023385
```

  This value is quite high. An explanation *could* be that in the poorer neighborhoods, people didn't have the money to build new houses, so there are more houses built before 1940.

- The R^2 value is 0.56, therefore about 56 % of the variation is explained by the model.

- The p-value of the F value is below the significance level and therefore significant. The null hypothesis H_0

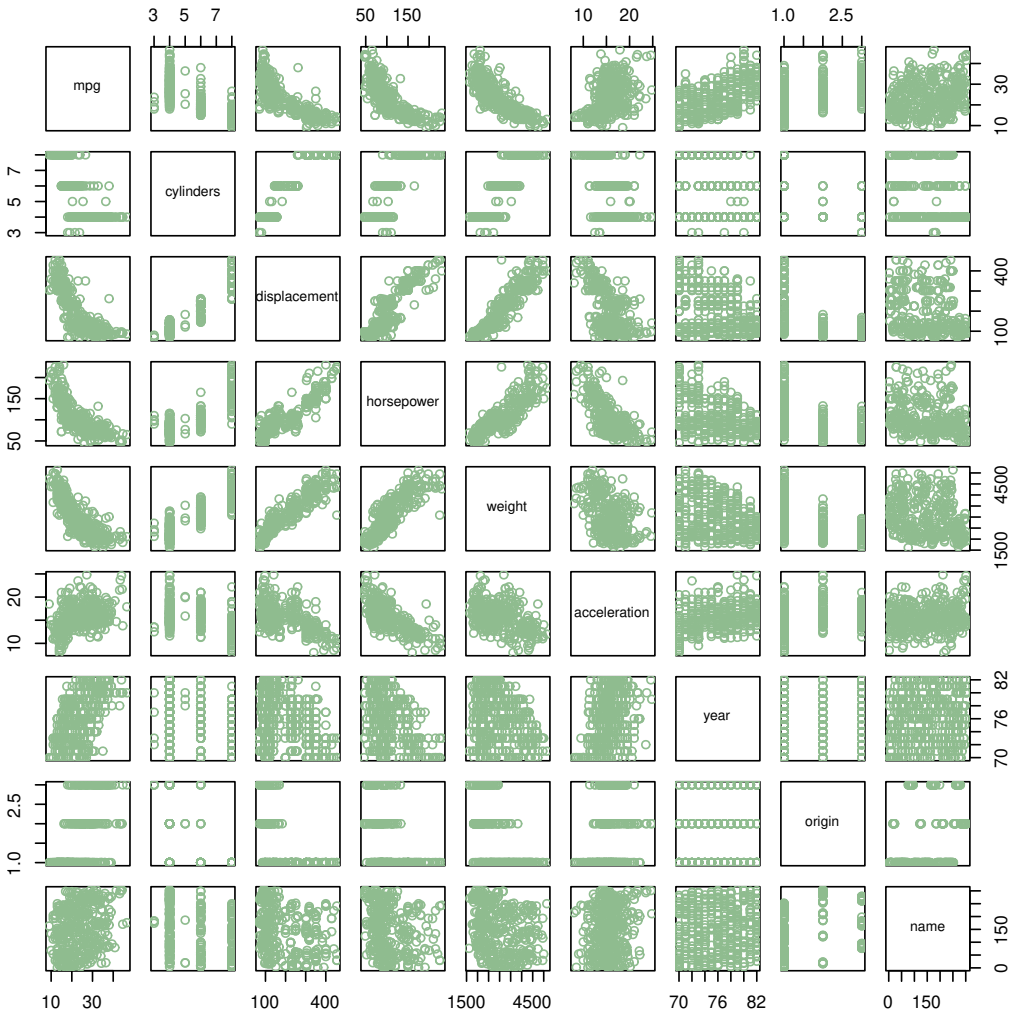
$$\beta_1 = \beta_2 = \beta_{12} = 0$$

  is rejected. One of β's is signigficantly different from 0. At least one variables contributes significantly to the model.

Solution 9.2

a) Scatter diagram:

```
pairs(Auto, col="darkseagreen")
```



b) Correlation matrix

```
cor(Auto.1)

##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
## origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
##               acceleration       year     origin
## mpg              0.4233285  0.5805410  0.5652088
## cylinders       -0.5046834 -0.3456474 -0.5689316
## displacement    -0.5438005 -0.3698552 -0.6145351
```

```
## horsepower     -0.6891955 -0.4163615 -0.4551715
## weight         -0.4168392 -0.3091199 -0.5850054
## acceleration    1.0000000  0.2903161  0.2127458
## year            0.2903161  1.0000000  0.1815277
## origin          0.2127458  0.1815277  1.0000000
```

c) Output

```
fit <- lm(mpg ~ . , data=Auto.1)
summary(fit)

##
## Call:
## lm(formula = mpg ~ ., data = Auto.1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215,Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

i) The p value to the corresponding F value is practically 0 and thus there is a statistically significant relationship between the response variable and the predictors.

ii) These are the coefficients with $**$ or $***$ (`displacement`, `weight`, `year` and `origin`)

iii) The coefficient for `year` is positive. This means that with younger cars you can get further per gallon of petrol. The newer cars are more fuel efficient in general.

d) Output:

```
fit <- lm(mpg ~ weight * year, data=Auto)
summary(fit)

##
## Call:
## lm(formula = mpg ~ weight * year, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0397 -1.9956 -0.0983  1.6525 12.9896
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.105e+02  1.295e+01  -8.531 3.30e-16 ***
## weight       2.755e-02  4.413e-03   6.242 1.14e-09 ***
## year         2.040e+00  1.718e-01  11.876  < 2e-16 ***
## weight:year -4.579e-04  5.907e-05  -7.752 8.02e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.193 on 388 degrees of freedom
## Multiple R-squared:  0.8339,Adjusted R-squared:  0.8326
## F-statistic: 649.3 on 3 and 388 DF,  p-value: < 2.2e-16
```

The p value of the interaction term is of the order of 8^{-14}, i.e. very close to 0, so the null hypothesis that there is no interaction is rejected.

This can be explained by the fact that the weight has become smaller and smaller with younger cars.

Solution 9.3

a) Data set:

```
library(ISLR)
head(Carseats)

##   Sales CompPrice Income Advertising Population Price ShelveLoc Age
## 1  9.50       138     73          11        276   120       Bad  42
## 2 11.22       111     48          16        260    83      Good  65
## 3 10.06       113     35          10        269    80    Medium  59
## 4  7.40       117    100           4        466    97    Medium  55
## 5  4.15       141     64           3        340   128       Bad  38
## 6 10.81       124    113          13        501    72       Bad  78
##   Education Urban  US
## 1        17   Yes Yes
## 2        10   Yes Yes
## 3        12   Yes Yes
## 4        14   Yes Yes
## 5        13   Yes  No
## 6        16    No Yes
```

b) Output:

```
fit <- lm(Sales~Price+Urban+US, data=Carseats)
summary(fit)

##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393,Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

c) Interpretation of the coefficients:

- The coefficient 13.04 is a bit difficult to interpret. According to the model under d), this is the average sales figures in shops reached in rural areas outside the USA, with the price of child seats still being \$0 (not very realistic).
- The coefficient -0.05 indicates that for an increase of one dollar, an average of 0.05 units of child seats are sold less.
- The coefficient -0.021 means that on average 0.021 less units are sold in urban areas compared to rural areas. However, the p value is very high, so this is more of a random variation.
- The 1.2 coefficient means that 1.2 more units are sold within the US compared to shops outside the USA. Perhaps child seats are compulsory in the USA.

d) Model: For `Urban` we choose the dummy variable:

$$x_{2i} = \begin{cases} 1 & \text{if } i\text{th person lives in urban area} \\ 0 & \text{if } i\text{-th person lives in rural area} \end{cases}$$

For `US` we choose the dummy variable

$$x_{3i} = \begin{cases} 1 & \text{if } i\text{th person lives in the USA} \\ 0 & \text{if } i\text{-th person does not live in the USA} \end{cases}$$

The model is then

$$\begin{aligned} y_i &= \beta_0 + \beta_1 \cdot \texttt{Price} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \\ &= \beta_0 + \beta_1 \cdot \texttt{Price} + \begin{cases} \beta_2 + \beta_3 + \varepsilon_i & \text{if } i\text{-th person lives in urban are in the USA} \\ \beta_2 + \varepsilon_i & \text{if } i\text{th person lives in urban area outside the USA} \\ \beta_3 + \varepsilon_i & \text{if } i\text{th person lives in rural area in the USA} \\ \varepsilon_i & \text{if } i\text{th person lives in rural area outside the USA} \end{cases} \end{aligned}$$

e) For all except `Urban`

f) Output:

```
fit <- lm(Sales~Price+US, data=Carseats)
summary(fit)

##
## Call:
## lm(formula = Sales ~ Price + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393,Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

Model: For `US` we choose the dummy variable

$$x_{2i} = \begin{cases} 1 & \text{if } i\text{th person lives in the USA} \\ 0 & \text{if } i\text{-th person does not live in the USA} \end{cases}$$

The model is then

$$\begin{aligned} y_i &= \beta_0 + \beta_1 \cdot \texttt{Price} + \beta_2 x_{2i} + \varepsilon_i \\ &= \beta_0 + \beta_1 \cdot \texttt{Price} + \begin{cases} \beta_2 + \varepsilon_i & \text{if } i\text{th person lives in the USA} \\ \varepsilon_i & \text{if } i\text{-th person does not live in the USA} \end{cases} \\ &= 13.03 - 0.055 \cdot \texttt{Price} + \begin{cases} 1.2 + \varepsilon_i & \text{if } i\text{-th person lives in the USA} \\ \varepsilon_i & \text{if } i\text{-th person does not live in the USA} \end{cases} \end{aligned}$$

g) In both models the correlation is proven (p-value for F-value practically 0), but if we look at the R^2-values, the one with 0.2393 is relatively bad. That means that although the correlation is verified the fit is bad, because only 23 % of the variability of the `Sales` can be explained by the model.

Classical and Bayesian Statistics

Problems 10

Problem 10.1

In a third world country, 1 % of the people suffer from a certain infectious disease. A test correctly indicates the disease in those actually ill with a probability of 98 %. Unfortunately, the test also indicates that 3 % of healthy people are sick.

Let S denote a sick person and T a person who tested positive.

$(*)$ a) Interpret (do not calculate!) the probabilities

$$P(S), \qquad P(\overline{T}), \qquad P(S \mid T), \qquad P(T \mid S), \qquad P(\overline{T} \mid \overline{S})$$

$(*)$ b) Denote the given probabilities in the problem definition using the notation as in a).

$(*)$ c) Calculate $P(\overline{S})$.

$(*)$ d) What is the probability that the test show a positive result for a randomly selected person?

Hint: Use the law of total probability.

$(*)$ e) What is the probability that a person who tested positive is actually sick? Interpret the result.

Hint: Use the Bayes' Theorem.

$(**)$ f) What is the probability that a person tested negative is actually healthy? Interpret the result.

Hint: Use the Bayes' Theorem.

Problem 10.2

A doping test is carried out at a sporting event. If an athlete has doped, the test is positive with a probability of 99 %.

However, if an athlete has not doped, the test will still show a positive result with a probability of 5 %. From experience we know that 20 % of the athletes are doped.

(∗∗) a) What is the probability that a doping test is positive.

(∗∗) b) What is the probability that the test is negative even though the athlete has doped?

(∗∗) c) What is the probability that an athlete has doped, if the doping test is negative.

Problem 10.3

An polygraph (lie detector) test is routinely performed on employees who work in sensitive positions. Let $+$ denote the event that the test is positive, i.e., that the polygraph indicates that the employee has lied. With W we denote the event that the employee told the truth and with L that the employee lied.

From former investigations of polygraphs test, we know that

$$P(+ \mid L) = 0.88 \quad \text{and} \quad P(- \mid W) = 0.86$$

Furthermore, we know that

$$P(W) = 0.99$$

(∗) a) Interpret the probabilities $P(W)$ and $P(+ \mid L)$.

(∗∗) b) For a person, the detector indicates that a lie has been told. What is the effective probability that this person has lied?

(∗∗) c) Interpret the result from b) in 2-3 sentences. How significant do you hold polygraph test?

Problem 10.4

(∗∗) The serum test examines pregnant women for babies with Down syndrome. The test is a very good but not perfect test. About 1 % of the babies have Down syndrome. If the baby has Down syndrome, there is a 90 % probability that the result will be positive. If the baby is not affected, there is still a 1 % probability that the result will be positive. A pregnant woman has been tested and the result is positive. What is the probability that your baby actually has Down syndrome?

Problem 10.5

The smoke sensors in a factory report a fire with a probability of 0.95. On a day without fire, they will give a false alarm with a probability of 0.01. One fire is expected per year.

(∗∗∗) a) The alarm system reports a fire. What is the probability that there is in fact a fire? Interpret the result.

(∗∗∗) b) In one night it is quiet (no alarm). What is the probability that there is really a fire? Interpret the result.

Problem 10.6

An insurance company believes that you can divide people into two classes - unlucky and others. Their statistics show that an unlucky person will have an accident within one year with a probability of 0.4, while for all others the probability is only 0.2. We assume that 30 percent of the population is unlucky.

(∗∗) a) What is the probability that a new customer will have an accident within one year after signing the contract?

(∗∗) b) A new customer has an accident within one year. What is the probability that he is an unlucky person?

Classical and Bayesian Statistic

Sample solution for Problems 10

Solution 10.1

a)
- $P(S)$: Probability that a randomly selected person in this country is really sick.
- $P(\overline{T})$: Probability that a person is tested negative (illness is not displayed).
- $P(S \mid T)$: Probability that a person who tested positive is actually sick.

  Or: A person is tested positive. Probability that she is really ill.
- $P(T \mid S)$: Probability that a sick person is tested positive.

  Or: A person is sick. Probability that she is tested positive.
- $P(\overline{T} \mid \overline{S})$: Probability that a healthy person is tested negative.

  Or: A person is healthy. Probability that she is also tested negative.

b)
- In a third world country 1 % of the people suffer from a certain infectious disease:
$$P(S) = 0.01$$
- A test correctly indicates the disease in those actually ill with a probability of 98 %:
$$P(T \mid S) = 0.98$$
- Unfortunately, the test also indicates that 3 % of healthy individuals are ill:
$$P(T \mid \overline{S}) = 0.03$$

c) It follows
$$P(\overline{S}) = 1 - P(S) = 1 - 0.01 = 0.99$$

d) Sought: $P(T)$
$$\begin{aligned} P(T) &= P(T \mid S) \cdot P(S) + P(T \mid \overline{S}) \cdot P(\overline{S}) \\ &= P(T \mid S) \cdot P(S) + P(T \mid \overline{S}) \cdot (1 - P(S)) \\ &= 0.98 \cdot 0.01 + 0.03 \cdot 0.99 \\ &= 0.0395 \end{aligned}$$

This means that 3.95 % of the tested persons are tested positive.

e) Wanted: $P(S \mid T)$

$$\begin{aligned} P(S \mid T) &= \frac{P(T \mid S) \cdot P(S)}{P(T)} \\ &= \frac{0.98 \cdot 0.01}{0.0395} \\ &= 0.2481 \end{aligned}$$

This means that only about 25 % of all those who test positive are effectively also ill.

f) Wanted: $P(\overline{S} \mid \overline{T})$:

$$\begin{aligned} P(\overline{S} \mid \overline{T}) &= \frac{P(\overline{T} \mid \overline{S}) \cdot P(\overline{S})}{P(\overline{T})} \\ &= \frac{(1 - P(T \mid \overline{S})) \cdot P(\overline{S})}{P(\overline{T})} \\ &= \frac{(1 - 0.03) \cdot 0.99}{1 - 0.0395} \\ &= 0.999792 \end{aligned}$$

This means that if the test is negative, we are very sure that we are healthy.

While a positive test does not say very much about having the disease, a negative test says tells very much.

Solution 10.2

Notation:

- D: Doped
- T: Tested positive

The following applies from the task definition

$$P(D) = 0.2, \qquad P(T \mid D) = 0.99, \qquad P(T \mid \overline{D}) = 0.05$$

a) Sought: $P(T)$.

We use the law of total probability:

$$\begin{aligned} P(T) &= P(T \mid D) \cdot P(D) + P(T \mid \overline{D}) \cdot P(\overline{D}) \\ &= P(T \mid D) \cdot P(D) + P(T \mid \overline{D}) \cdot (1 - P(D)) \\ &= 0.99 \cdot 0.2 + 0.05 \cdot 0.8 \\ &= 0.238 \end{aligned}$$

This means that 23.8 % of the tested will test positive.

b) Wanted: $P(\overline{T} \mid D)$

$$P(\overline{T} \mid D) = 1 - P(T \mid D) = 1 - 0.99 = 0.01$$

c) Sought: $P(D \mid \overline{T})$

We use the theorem of Bayes:

$$\begin{aligned} P(D \mid \overline{T}) &= \frac{P(\overline{T} \mid D) \cdot P(D)}{P(\overline{T})} \\ &= \frac{0.01 \cdot 0.2}{1 - 0.238} \\ &= 0.00262 \end{aligned}$$

Only 0.262 % of all negatively tested athletes are also doped. So a negative test is quite significant.

Solution 10.3

a) $P(W)$: Probability that a randomly selected employee is telling the truth.
$P(+ \mid L)$: Probability that for a person who lied, the polygraph indicates so.

b) Sought: $P(L \mid +)$

It follows

$$\begin{aligned} P(L \mid +) &= \frac{P(+ \mid L)P(L)}{P(+ \mid L)P(L) + P(+ \mid W)P(W)} \\ &= \frac{P(+ \mid L)(1 - P(W))}{P(+ \mid L)(1 - P(W)) + (1 - P(- \mid W))P(W)} \\ &= \frac{0.88 \cdot (1 - 0.99)}{0.88 \cdot (1 - 0.99) + (1 - 0.86) \cdot 0.99} \\ &= 0.0597 \end{aligned}$$

c) So, if the test indicates that the employee lied, then he is only a real liar with a probability of 6 %. Or in 94 % of positive tests, the employees, who allegedly lied, are telling the truth.

Thus, the test shows a very high percentage of false results and is therefore worthless.

This is also the reason why lie detector tests are not allowed in court.

Solution 10.4

Let D denote that the baby has Down syndrome and $+$ that the test is positive. Given are the following probabilities:

$$P(D) = 0.01. \qquad P(+ \mid D) = 0.9. \qquad P(+ \mid \overline{D}) = 0.01$$

Applying Bayes' theorem and the law of total probability, it follows

$$\begin{aligned} P(D \mid +) &= \frac{P(+ \mid D) \cdot P(D)}{P(+ \mid D) \cdot P(D) + P(+ \mid \overline{D}) \cdot P(\overline{D})} \\ &= \frac{0.9 \cdot 0.01}{0.9 \cdot 0.01 + 0.01 \cdot 0.99} \\ &= 0.4761905 \end{aligned}$$

If the test result is positive, there is a 48 % probability that the baby has Down syndrome.

Solution 10.5

Notation:

F : Event that fire breaks out
A : Event that the alarm goes off

The probability that a fire will break out is

$$P(F) = \frac{1}{365}$$

The probability that the alarm will go off, given a fire breaks out, is

$$P(A \mid F) = 0.95$$

The probability that there is an alarm, given that no fire has broken out, is

$$P(A \mid \overline{F}) = 0.01$$

a) The probability that a fire has broken out, given there was an alarm, is

$$P(F \mid A) = \frac{P(A \mid F) \cdot P(F)}{P(A)}$$

The probability of an alarm is expressed by the law of total probability:

$$P(A) = P(A \mid F) \cdot P(F) + P(A \mid \overline{F}) \cdot P(\overline{F})$$

So:

$$\begin{aligned} P(F \mid A) &= \frac{P(A \mid F) \cdot P(F)}{P(A \mid F) \cdot P(F) + P(A \mid \overline{F}) \cdot P(\overline{F})} \\ &= \frac{0.95 \cdot \frac{1}{365}}{0.95 \cdot \frac{1}{365} + 0.01 \cdot (1 - \frac{1}{365})} \\ &= 0.207 \end{aligned}$$

In only 1 in 5 cases of alarm there is actually a fire.

b) The probability that no fire has broken out, if there was no alarm, is

$$P(\overline{F} \mid \overline{A})$$

We now apply the Bayes Theorem:

$$P(\overline{F} \mid \overline{A}) = \frac{P(\overline{A} \mid \overline{F}) \cdot P(\overline{F})}{P(\overline{A})}$$

The three probabilities on the right hand side of the equation are unknown, but we can calculate these from the known ones:

- For $P(\overline{A} \mid \overline{F})$ it follows

$$\begin{aligned} P(\overline{A} \mid \overline{F}) &= 1 - P(A \mid \overline{F}) \\ &= 1 - 0.01 \\ &= 0.99 \end{aligned}$$

- For $P(\overline{F})$ it follows

$$\begin{aligned} P(\overline{F}) &= 1 - P(F) \\ &= 1 - \frac{1}{365} \\ &= \frac{364}{365} \end{aligned}$$

- For $P(\overline{A})$ it follows

$$P(\overline{A}) = 1 - P(A)$$

We can calculate the probability $P(A)$ with the law of total probability:

$$\begin{aligned} P(A) &= P(A \mid F)P(F) + P(A \mid \overline{F})P(\overline{F}) \\ &= 0.95 \cdot \tfrac{1}{365} + 0.01 \cdot \tfrac{364}{365} \\ &= 0.01257534 \end{aligned}$$

So we find

$$\begin{aligned} P(\overline{F} \mid \overline{A}) &= \frac{P(\overline{A} \mid \overline{F}) \cdot P(\overline{F})}{P(\overline{A})} \\ &= \frac{\big(1 - P(A \mid \overline{F})\big) \cdot P(\overline{F})}{1 - P(A)} \\ &= \frac{(1 - 0.01) \cdot \frac{364}{365}}{1 - (0.95 \cdot \frac{1}{365} + 0.01 \cdot (1 - \frac{1}{365}))} \\ &= 0.999 \end{aligned}$$

That means, if there is no alarm, we are pretty sure that there is no fire.

Finally

$$P(F \mid \overline{A}) = 1 - P(\overline{F} \mid \overline{A}) = 1 - 0.999 = 0.001$$

Solution 10.6

a) Event A: The new customer is an unlucky one;

Event B: The new customer has an accident within one year

Known: $P(B \mid A) = 0.4$, $P(B \mid \overline{A}) = 0.2$, $P(A) = 0.3$, $P(\overline{A}) = 0.7$

$$P(B) = P(B \mid A) \cdot P(A) + P(B \mid \overline{A}) \cdot P(\overline{A}) = 0.4 \cdot 0.3 + 0.2 \cdot 0.7 = 0.26$$

b) We are looking for $P(A \mid B)$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A) \cdot P(A)}{P(B)} = \frac{0.3 \cdot 0.4}{0.26} = 0.4615$$

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page this extra page has been added to receive it.

If you rerun the document (without altering it) this surplus page will go away, because LaTeX now knows how many pages to expect for this document.

Classical and Bayesian Statistics

Problems 11

Problem 11.1

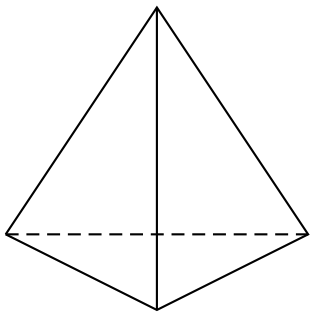
The result of a public-opinion poll for a presidential election in three provinces (A, B and C) are as follow: In province A the percentage of voters supporting the Liberal candidate is 50 %. In province B the percentage of voters supporting the Liberal candidate is 60 %. In province C the percentage of voters not supporting the Liberal candidate is 65 %.

The population of the three provinces is distributed as follows: A has 40 % of the total population in A, B and C, 25 % of the total population live in B and the remaining 35 % live in C.

Let us randomly choose a supporter of the Liberal candidate in province A or B or C. Determine the probability that such a voter was chosen from province B.

Problem 11.2

On a tetrahedral cube, each side is an equilateral triangle.



When you throw the cube, it lands face down and the other three faces are visible as a three-sided pyramid. The faces are labeled 1–4 points, and the value of the bottom face is labeled x.

a) Consider the following three mathematical descriptions of the probabilities of x:

- model A : $P(x) = 1/4$

- model $B : P(x) = x/10$
- model $C : P(x) = 12/(25x)$

For each model, determine the value of $P(x)$ for each value of x. Describe in words what kind of „unfairness“ is expressed by each model.

b) Suppose we have a) presented tetrahedral cube, along with the three candidate models for the probabilities of the cube.

Suppose that initially we are not sure what to make of the cube. On the one hand, the die could be fair, with each side landing with the same probability.

On the other hand, the cube could be skewed, so that the faces with more points are more likely to land on the ground (because the points are created by embedding heavy jewels in the cube, so the sides with more points are more likely to land on the ground).

Or, that more dots on a side make it less likely to land at the bottom (because the dots may be made of springy rubber or protrude from the surface).

Initially, then, our beliefs about the three models as

$$p(A) = p(B) = p(C) = \frac{1}{3}$$

can be described.

Now we roll the dice 100 times and find these results:

$$\text{no1} = 25, \quad \text{no2} = 25, \quad \text{no3} = 25, \quad \text{no4} = 25$$

D̈o these data change our beliefs about the models? Which model now seems most likely?

Suppose that when we rolled the dice 100 times, we found these results:

$$\text{no. 1} = 48, \quad \text{no. 2} = 24, \quad \text{no. } = 16, \quad \text{No. 4} = 12.$$

Which model now seems most likely?

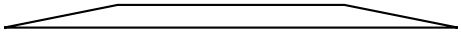
Problem 11.3

In the malaria example, we examined a positive test twice. Calculate the probability that we test positive first, then negative.

What does it look like if we reverse the order?

Problem 11.4

Consider a coin that has asymmetric cross section:



We do not know whether H or T is up. What are the prior probabilities?

Problem 11.5

School children were asked about their favorite foods. Of the total sample, 20% were first graders, 20% were sixth graders, and 60% were eleventh graders. The following table shows for each grade the percentage of respondents who selected each of the three foods as their favorite:

	Ice cream	Fruit	French fries
1st graders	0.3	0.6	0.1
6th graders	0.6	0.3	0.1
11th graders	0.3	0.1	0.6

From this information, construct a table of joint probabilities of grade and favorite food. Also tell whether or not the grade and favorite food are independent and how you determined the answer.

Note: You will get $P(\text{class})$ and $P(\text{food} \mid \text{class})$. You need to determine $P(\text{food} \cap \text{class})$.

Problem 11.6

This task refers to lecture notes where one coin was tossed once and H occured. We determined the posterior distribution.

a) Now you toss the coin again and H occurs. What is the new posterior distribution?

b) Instead of an H a T occurs. What is the new posterior distribution? Compare it to the original prior distribution. What can you tell?

Classical and Bayesian Statistic

Sample solution for Problems 11

Solution 11.1

Designations:

- L: Voter supports liberal candidate
- $\overline{L}$: Voter does not support Liberal candidate
- A: Voter is from province A
- B: Voter is from province B
- C: Voter comes from province C

Known (from task):

$$P(L \mid A) = 0.5, \quad P(L \mid B) = 0.6, \quad P(\overline{L} \mid C) = 0.65$$

and

$$P(A) = 0.4, \quad P(B) = 0.25, \quad P(C) = 0.35$$

We are looking for $P(B \mid L)$. According to Bayes' theorem and the marginal probability:

$$\begin{aligned} P(B \mid L) &= \frac{P(L \mid B)P(B)}{P(L)} \\ &= \frac{P(L \mid B)}{P(L \mid A)P(A) + P(L \mid B)P(B) + P(L \mid C)P(C)} \\ &= \frac{0.6 \cdot 0.25}{0.5 \cdot 0.4 + 0.6 \cdot 0.25 + (1 - 0.65) \cdot 0.35} \\ &= 0.3175 \end{aligned}$$

About 32 % of the voters who support the Liberal candidate are from province B.

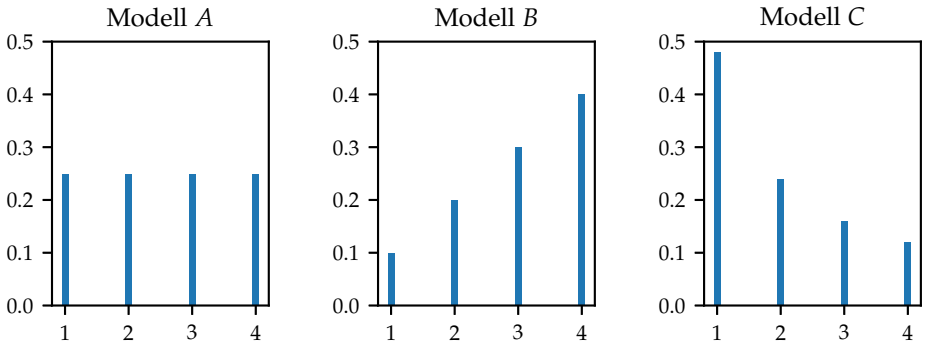
Solution 11.2

a) We can list the probabilities as a table:

	x	1	2	3	4
Model A :	$P(X=x)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
Model B :	$P(X=x)$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{4}{10}$
Model C :	$P(X=x)$	$\frac{12}{25}$	$\frac{12}{50}$	$\frac{12}{75}$	$\frac{12}{100}$
		$\frac{48}{100}$	$\frac{24}{100}$	$\frac{16}{100}$	$\frac{12}{100}$

Note that all the probabilities in the rows add up to 1.

We still sketch these probabilities:



For model A, all sides have equal probability of being at the bottom.

For model B, side 1 has the smallest probability and this increases linearly with number.

For model C, side 1 has the largest probability and it decreases nonlinearly (hyperbolically) with the number. So here the probability $P(1)$ is weighted much more than the other probabilities.

b) Before we roll the dice with the tetrahedron we have no idea about the throw probabilities of the individual sides. In a) we considered three possible models that we can consider as prior probabilities for the sides.

However, we do not know which model best „fits“ and before we make any experiments, we make the assumption that all models are equally likely, i.e.

$$p(A) = p(B) = p(C) = \frac{1}{3}$$

which we can consider as the prior probability for the models.

Note that we have two prior probabilities:

- One for the *throw*probability.
- One for the *model* probability.

Now we make a trial of 100 throws and get

$$\text{no. } 1 = 25, \quad \text{no. } 2 = 25, \quad \text{no. } 3 = 25, \quad \text{no. } 4 = 25$$

that is, all sides exit with equal frequency on *this* trial. So we will change our assumption about the probability of the models and assign a larger probability to the model A. However, this probability is not 1, since on another trial the litter numbers will be different.

In the same way, we change the assumption

$$\text{no. } 1 = 48, \quad \text{no. } 2 = 24, \quad \text{no. } = 16, \quad \text{No. } 4 = 12.$$

and assign a greater probability to model C and, most importantly, less to model B, since this model assigns the smallest probability to the number 1.

Solution 11.3

As in theory, we first use the formula for a positive test

$$P(M \mid +) = \frac{P(+ \mid M) \cdot P(M)}{P(+)} = \frac{P(+ \mid M) \cdot P(M)}{P(+ \mid M) \cdot P(M) + P(+ \mid \overline{M}) \cdot P(\overline{M})}$$

and then for the negative test

$$P(M \mid -) = \frac{P(- \mid M) \cdot P(M)}{P(-)} = \frac{P(- \mid M) \cdot P(M)}{P(- \mid M) \cdot P(M) + P(- \mid \overline{M}) \cdot P(\overline{M})}$$

Where $P(M)$ equals $P(M \mid +)$ from the first trial.

Since only the prior probability changes, we do this with Python:

```
# sensitivity P(+|M)
sens = 0.917

# specificity P(-|M across)
spec = 0.935

# formula for positive test
post_pos <- function(prior) {
    post = (sens * prior)/(sens * prior + (1 - spec) * (1 - prior))
    return(post)
```

```
}

# formula for negative test
post_neg <- function(prior) {
    post = (1 - sens) * prior/((1 - sens) * prior + spec * (1 - prior))
    return(post)
}
```

If the test is first positive then negative, we get:

```
post_neg(post_pos(0.03))

## [1] 0.03728794
```

In the reverse case:

```
post_pos(post_neg(0.03))

## [1] 0.03728794
```

So the order does not matter here.

What happens if there are two negative tests?

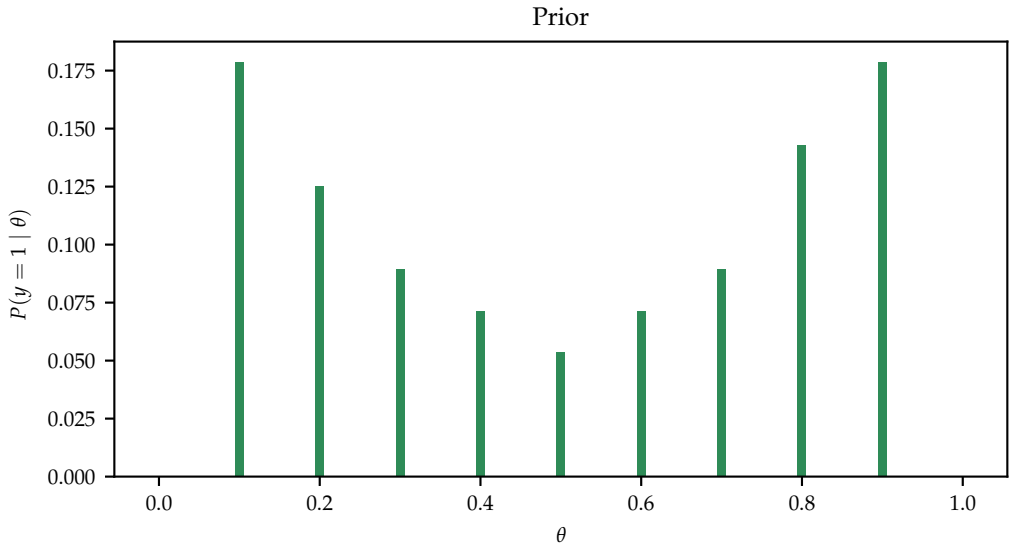
```
post_neg(post_neg(0.03))

## [1] 0.0002436557
```

Then we are already pretty sure that we don't have malaria.

Solution 11.4

We need to consider the situation where T or H get a relatively large probability, since we don't know which side H and T are on.

Thus, an a priori probability might look like the following:



Solution 11.5

First, a note: the table gives *not* $P(\text{food} \cap \text{class})$. The reason is that the sum of the probabilities in the rows adds up to 1, and in the columns it generally does not.

Thus, the first value in the upper left is the probability that ice cream is the favorite food for a first grader:

$$P(G \mid 1) = 0.3$$

According to the definition of the certain probability

$$P(G \cap 1) = P(G \mid 1) \cdot P(1) = 0.3 \cdot 0.2 = 0.06$$

This is the value in the upper left corner of the following table. The other values are calculated analogously. You have to multiply the 1st line with 0.2, the 2nd with 0.2 and the third with 0.6

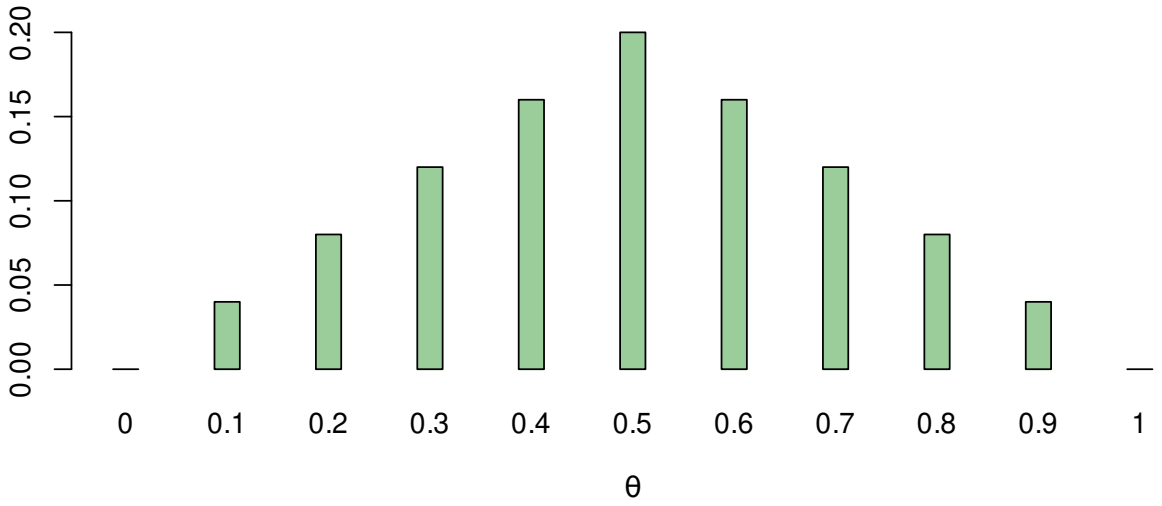
	Ice Cream	Fruit	French Fries	
1st graders	0.06	0.12	0.02	0.2
6th graders	0.12	0.06	0.02	0.2
11th graders	0.18	0.06	0.36	0.6
	0.36	0.24	0.4	1

Solution 11.6

We used the prior distribution

```
y <- c(0, 1, 2, 3, 4, 5, 4, 3, 2, 1, 0)
prior <- y/sum(y)
```

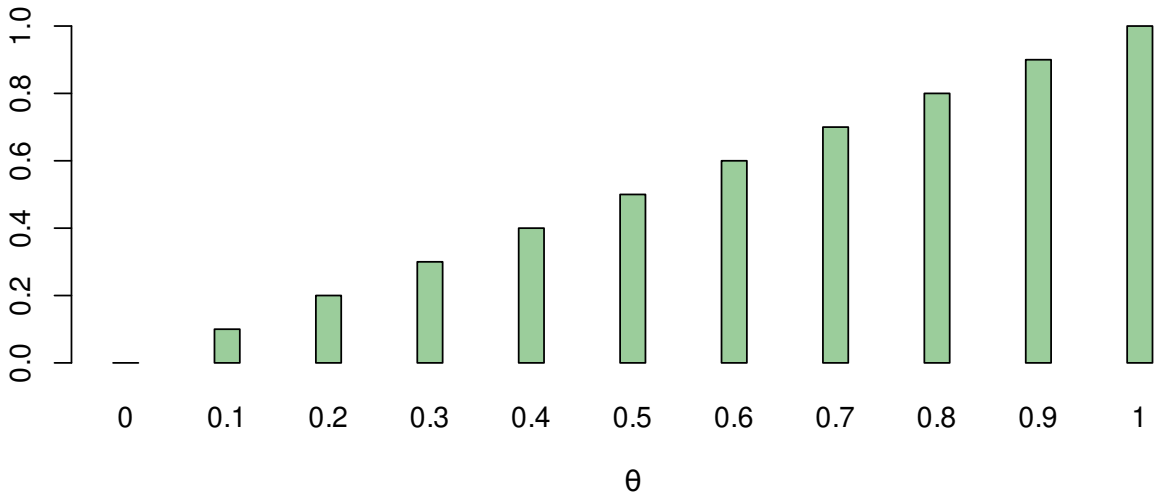
```
library(latex2exp)
steps = seq(0, 1, 0.1)
barplot(prior, space = 3, names = steps, col = "darkseagreen3", xlab = TeX("$\\theta$"))
```



and the likelihood function

```
like <- seq(0, 1, 0.1)
like


##  [1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
```

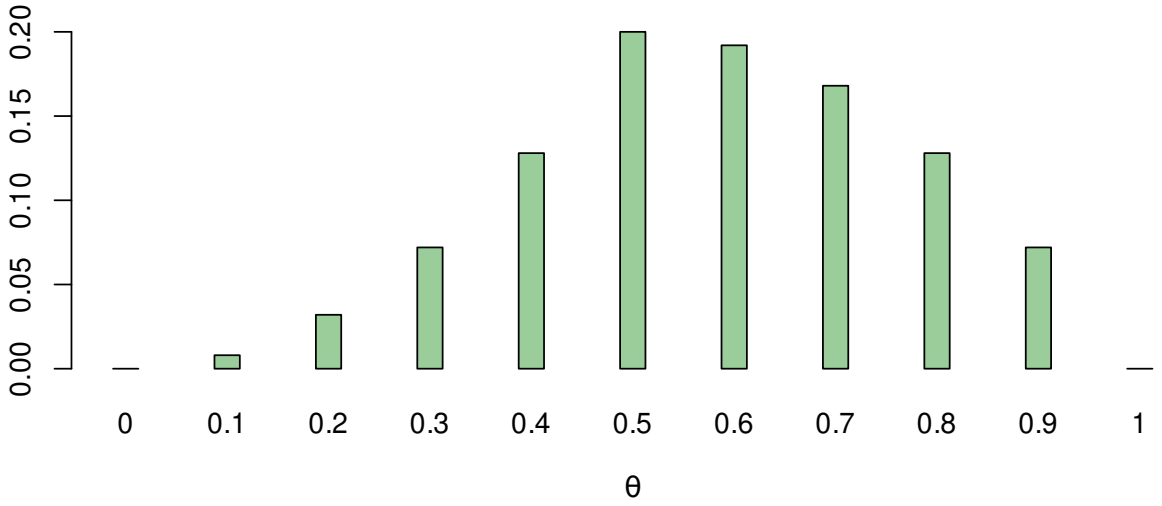
```
barplot(like, space = 3, names = steps, col = "darkseagreen3", xlab = TeX("$\\theta$"))
```



and obtained the posterior distribution

```
margin = sum(prior * like)
post = prior * like/margin
post


##  [1] 0.000 0.008 0.032 0.072 0.128 0.200 0.192 0.168 0.128 0.072 0.000
```

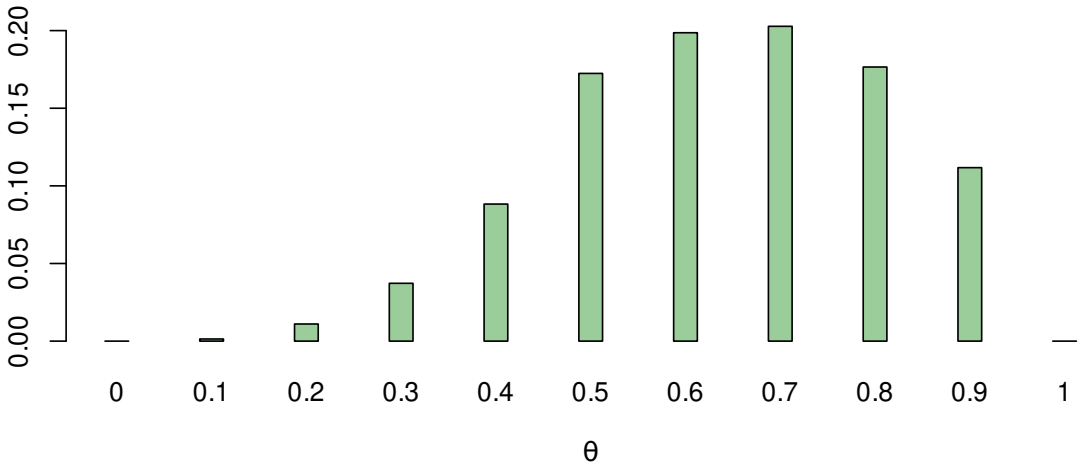
```
barplot(post, space = 3, names = steps, col = "darkseagreen3", xlab = TeX("$\\theta$"))
```



a) The likelihood function stays the same but we use the old posterior as new prior distribution.

```
prior_H = post
margin = sum(prior_H * like)
post_H = prior_H * like/margin
post_H

##  [1] 0.00000000 0.00137931 0.01103448 0.03724138 0.08827586 0.17241379
##  [7] 0.19862069 0.20275862 0.17655172 0.11172414 0.00000000
```

```
barplot(post_H, space = 3, names = steps, col = "darkseagreen3", xlab = TeX("$\\th
```



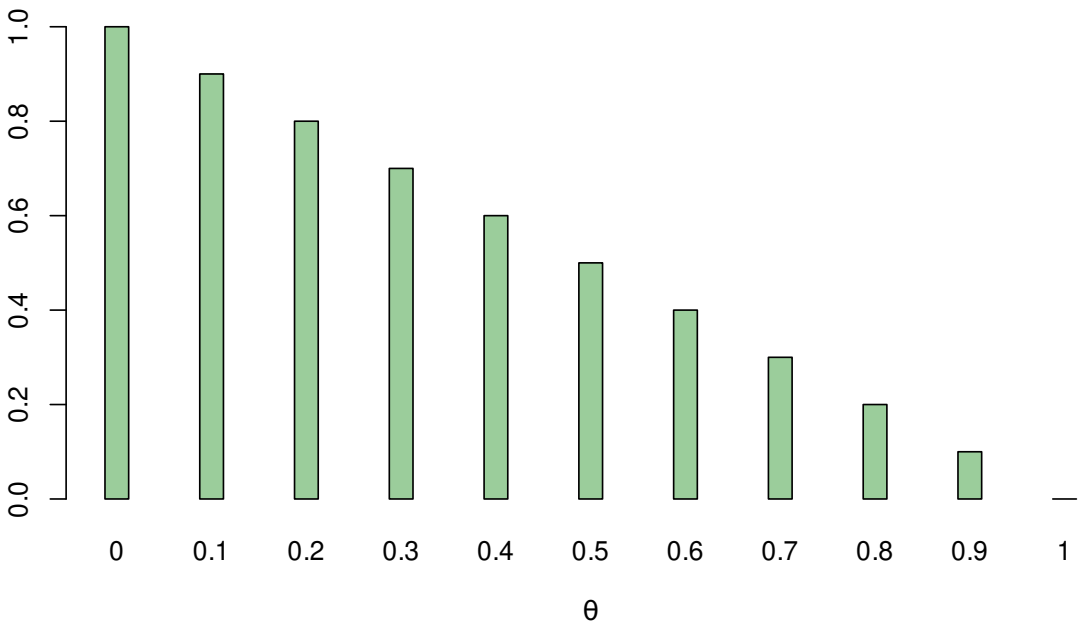
As expected the distribution moves further to the right because with an additional H, we can assume that the probability θ is greater than 0.5. If $\theta = 0.5$ we would expect that the H and T are evenly distributed.

However, it is entirely possible that we toss two H in a row even if $\theta = 0.5$.

b) If we toss a T instead, we have to change the likelihood function to `1-like`

```
like <- seq(0, 1, 0.1)
like

##  [1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
```

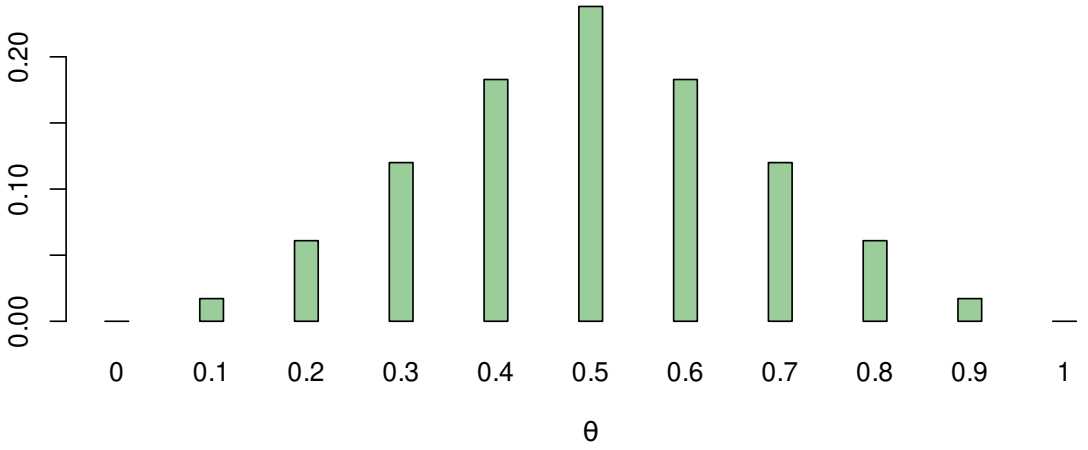
```
like_T = 1 - like
barplot(like_T, space = 3, names = steps, col = "darkseagreen3", xlab = TeX("$\\th
```



For the new prior distribution we use the old posterior distribution

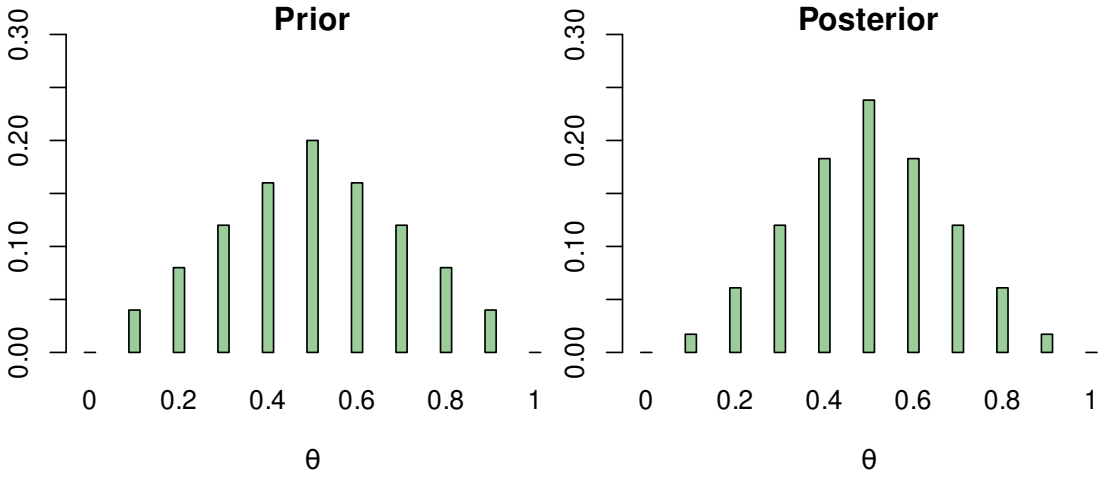
```
prior_T = post
margin = sum(prior_T * like_T)
post_T = prior_T * like_T/margin
post_T

##  [1] 0.00000000 0.01714286 0.06095238 0.12000000 0.18285714 0.23809524
##  [7] 0.18285714 0.12000000 0.06095238 0.01714286 0.00000000
```

```
barplot(post_T, space = 3, names = steps, col = "darkseagreen3", xlab = TeX("$\\th
```



It *appears* that we are back to square one, namely the original prior distribution. However, this is *not* the case. If we plot the original prior distribution and the posterior distribution we recognize differences.

```
par(mfrow = c(1, 2))
barplot(prior, space = 3, names = steps, col = "darkseagreen3", xlab = TeX("$\\the
    ylim = c(0, 0.3), main = "Prior")
barplot(post_T, space = 3, names = steps, col = "darkseagreen3", xlab = TeX("$\\th
    ylim = c(0, 0.3), main = "Posterior")
```



Both distributions are symmetrical, but the probabilities have reallocated to middle. Why is this the case? If we toss a H and then a T, this is more evidence that the coin is fair, namely $\theta = 0.5$. Hence the θ's in the middle get more weight compared to the prior distribution.

Classical and Bayesian Statistics

Problems 12

Problem 12.1

a) Start with a prior distribution expressing some uncertainty that a coin is fair: $\text{Beta}(\theta \mid 4, 4)$.

   Flip the coin once. Suppose you get heads. What is the posterior distribution?

b) Use the posterior of the previous toss as the prior for the next toss. Suppose, You throw again and get heads. What is the new posterior now?

c) Using this posterior as the prior for the next throw, throw a third time and you get tails. What is the new posterior now?

d) Make the same three updates, but in the order H, T, T instead of T, T, H.

   Is the final posterior distribution the same for both orders of the toss results?

Problem 12.2

Suppose an election is coming up and you want to know whether the general population prefers candidate A or candidate B. There is a poll just published in the newspaper which says that out of 100 people randomly surveyed, 58 prefer candidate A and the rest prefer candidate B.

Note: To determine the HDI, use the command.

```
hdi <- function(a,b, prob=0.95){
  k <- 0
  x <- seq(0, 1, 0.001)
  y <- dbeta(x, a, b)
  while(TRUE){
     k <- k + 0.001
     if (sum(y[y > k])/length(x) < prob){
       break
     }
     }
  return(c(x[(y > k)][1], x[(y > k)][length(x[(y > k)])]))
}
```

You then only need to specify a and b:

```
hdi(5,5)


## [1] 0.212 0.788
```

a) Assume that before the newspaper survey, your prior assumption was a uniform distribution. What is the 95 %-HDI for your beliefs after you learn about the newspaper survey results?

b) You would like to conduct a follow-up survey to narrow down your estimate of the narrow down your estimate of the population's preference.

   In your follow-up survey, you take a random sample of 100 more people and find that 57 people prefer candidate A and the rest prefer candidate B.

   Assuming that people's opinions have not changed between surveys, what is the 95 %-HDI for the posterior distribution?

Problem 12.3

Suppose you are training people in a simple learning experiment as follows: When people see the two words „radio" and „ocean" on the computer screen, they are to press the F key on the computer keyboard.

The subjects see several repetitions and learn the response well. Then introduce another relationship for them to learn: Whenever the words „radio" and „mountain" appear, have them press the J key on the computer keyboard.

You train the subjects until they know both relationships well. Now you check what they have learned by asking them to do two new test items. For the first test, show them the word „radio" as such and instruct them to give the best answer (F or J) based on what they have learned before.

For the second test, show them the two words „ocean" and „mountain" and ask them to give the best answer. You carry out this procedure with 50 people.

Your data show that for „radio" alone, 40 people chose F and 10 people chose J.

For the word combination „ocean" and „mountain", 15 people chose F and 35 people chose J.

Are people biased towards F or towards J for either sample type? To answer this question, assume a uniform prior and use a 95 %-HDI to decide which variance can be classified as credible.

Problem 12.4

Suppose we have a coin that we know comes from a trick toy shop.

We therefore believe that the coin shows either heads or tails, but we do not know which. Express this belief as a beta prior.

Now we toss the coin 5 times and heads comes out in 4 of the 5 tosses. What is the posterior distribution?

Problem 12.5

a) Suppose you have a coin that you know has been minted by the government and has has not been tampered with. Therefore, you have a strong belief that the coin is fair. You flip the coin 10 times and get 9 heads.

   What is your predicted probability of heads for the 11th toss? Explain your answer carefully; justify your choice of prior.

b) Now you have another coin, which is made of a strange material and is with the inscription (in small print) "Patent Pending, International Magic, Inc.". You flip the coin 10 times and get 9 heads. What is your predicted probability of heads on the 11th toss.

   Explain your answer carefully; justify your choice of prior. Hint: Use the prior from the from the task before.

Problem 12.6

We again assume a fair coin, i.e. $a = b$. We now want to determine a and b such that the HDI of the prior distribution has a certain width.

Again, use the `hdi` command from Task 2 and

```
a <- 2
(hdi(a,a)[2] - hdi(a,a))[1]


## [1] 0.812
```

a) We are not very convinced that the coin is fair and assume that 95 % of the most credible parameters are in the range of 0.2 and 0.8. How should we choose the parameter $a = b$?

   Try values for a until you get approximately to the width of the interval.

b) The coin looks fair and we assume that 95 % of the most credible parameters are in the range 0.4 and 0.6. How should we choose the parameter $a = b$?

c) We have examined the coin closely and are very optimistic that the coin is fair and assume that 95 % of the most credible parameters are in the range 0.48 and 0.52. How should we choose the parameter $a = b$?

Classical and Bayesian Statistic

Sample solution for Problems 12

Solution 12.1

a) We denote by N the number of tosses and by z the number of tosses with heads.

We have $N = 1, z = 1, a = b = 4$ and use the formula

$$p(\theta \mid z, N) = \text{Beta}(\theta \mid z + a, N - z + b)$$

If we put in all the values, we get

$$p(\theta \mid 1, 1) = \text{Beta}(\theta \mid 5, 4)$$

b) Now $\text{Beta}(\theta \mid 5, 4)$ is the new prior distribution: again we have $N = 1, z = 1$, but $a = 5$ and $b = 4$:

$$p(\theta \mid 1, 1) = \text{Beta}(\theta \mid 6, 4)$$

c) Now $\text{Beta}(\theta \mid 6, 4)$ is the new prior distribution: we have $N = 1, z = 0$, but $a = 6$ and $b = 4$:

$$p(\theta \mid 1, 1) = \text{Beta}(\theta \mid 6, 5)$$

d) The result is the same. This has to do with the fact that the throws are stochastically independent and thus the permutation does not matter.

Solution 12.2

a) We have $N = 100, z = 58$ and $a = b = 1$. The posterior distribution is:

$$p(\theta \mid 58, 100) = \text{Beta}(\theta \mid z + a, N - z + b) = \text{Beta}(\theta \mid 59, 43)$$

The HDI is

```
hdi(59,43)

## [1] 0.483 0.673
```

The HDI is $(0.483, 0.673)$. Has a width of approximately 0.2. The HDI describes the range where 95 % of the most likely values are. So we have got a rough sense of what the election result will be. However, there is still a probability that candidate B will win.

b) We take as a new prior distribution $\text{Beta}(\theta \mid 59, 43)$ and get $N = 100$ and $z = 57$ with $a = 59$ and $b = 43$

$$p(\theta \mid 57, 100) = \text{Beta}(\theta \mid z + a, N - z + b) = \text{Beta}(\theta \mid 116, 86)$$

The HDI is

```
hdi(116,86)

## [1] 0.507 0.642
```

The HDI has shrunk, the width is now only about 0.14. Most importantly, the interval has shifted further to the right, so the chances of candidate B winning have worsened. These values are no longer in the HDI at all.

Solution 12.3

We again take as prior distribution a beta distribution with $a = 1$ and $b = 1$ with $N = 50$ and $z = 10$ in the 1st case. We obtain

$$p(\theta \mid 10, 50) = \text{Beta}(\theta \mid z + a, N - z + b) = \text{Beta}(\theta \mid 11, 41)$$

With the HDI

```
hdi(11,41)


## [1] 0.107 0.323
```

If the combinations are equally probable, then 0.5 "must" be in the HDI. This is not the case, so we can assume that one of the samples has a higher probability of being chosen.

We take as prior distribution again a beta distribution with $a = 1$ and $b = 1$ with $N = 50$ and $z = 15$ in the 2nd case. We get

$$p(\theta \mid 15, 50) = \text{Beta}(\theta \mid z + a, N - z + b) = \text{Beta}(\theta \mid 36, 16)$$

With the HDI

```
hdi(36,16)


## [1] 0.567 0.813
```

If the combinations are equally probable, then 0.5 "must" be in the HDI. This is not the case, so we cannot assume that the deviation is rather random.

Solution 12.4

For the prior distribution, we choose a beta distribution that is very low in the middle at $\theta = 0$ and very high at both ends $\theta \approx 0$ and $\theta \approx 1$. Such a beta distribution has about $a = b = 0.1$.

We have $N = 5$ and $z = 4$ and therefore

$$p(\theta \mid 4, 5) = \text{Beta}(\theta \mid z + a, N - z + b) = \text{Beta}(\theta \mid 4.1, 1.1)$$

Solution 12.5

a) We are very sure that the coin is fair, so we take a beta distribution with large a and b but equal, so for example $a = b = 100$. Now we have $N = 10$ and $z = 9$, which actually contradicts a fair coin. We calculate the posterior distribution.

$$p(\theta \mid 9, 10) = \text{Beta}(\theta \mid z + a, N - z + b) = \text{Beta}(\theta \mid 109, 101)$$

The distribution has shifted slightly, but by how much?

One possibility is that we calculate the mode. This is calculated by

$$\omega = \frac{a - 1}{a + b - 2} = \frac{109 - 1}{109 + 101 - 2} = \frac{108}{208} \approx 0.519$$

Note that we use the new a and b of the posterior distribution to calculate the mode.

The new probability that we still throw heads is now about 0.52, which is still very close to 0.5.

b) We choose the prior distribution from task 4 with $a = b = 0.1$ and $N = 10$ and $z = 9$

$$p(\theta \mid 9, 10) = \text{Beta}(\theta \mid z + a, N - z + b) = \text{Beta}(\theta \mid 9.1, 1.1)$$

We calculate again the mode

$$\omega = \frac{a - 1}{a + b - 2} = \frac{9.1 - 1}{9.1 + 1.1 - 2} = \frac{8.1}{8.2} \approx 0.988$$

Here we get a clear shift towards the data.

Solution 12.6

a) The width of the interval is 0.6 and with a little trial and error we get $a \approx 4$ or 5.

b) The width of the interval is 0.2 and with a little trial and error you get $a \approx 50$.

c) The width of the interval is 0.04 and with a little trial and error we get $a \approx 1000$.

Classical and Bayesian Statistics

Problems 13

Problem 13.1

Use the file `stan_beta_compare.R` to check the approximation of the posterior distribution by `stan`. The top figure is the approximated distribution and the bottom figure is the exact posterior distribution we know in this case.

Play with the options `iter=...` (number of steps after tune-in) and `warmup=...` (The first steps that are ignored) around. What can you say about the HDI and mode compared to the exact distribution?

Problem 13.2

The pharmaceutical company Life Co. has developed a new drug to combat ADHD. To determine its effectiveness, the drug was tested with $n = 10$ patients. The current standard method shows an effect in 30 % of the treated patients.

Treatment with the new drug was successful in 4 patients. Carry out a hypothesis test to see if the new drug is more effective than the standard method.

a) Investigate the data using the Bayesian method from the lecture. Make explicit statements about the prior distribution and the ROPE. Justify your answer.

b) around with the data so that the zero value is discarded or accepted.

Use the file `stan_beta_rope.R`.

Hint: Run

```
library(rstan)


## Loading required package: StanHeaders

## Loading required package: ggplot2

## rstan (Version 2.21.7, GitRev: 2e1f913d3ca3)

## For execution on a local, multicore CPU with excess RAM we recommend calling
## options(mc.cores = parallel::detectCores()).
## To avoid recompilation of unchanged Stan programs, we recommend calling
## rstan_options(auto_write = TRUE)
```

```
library(bayestestR)
library(latex2exp)
rstan_options(auto_write = TRUE)


plot_posterior <- function(params, Rope = c(0.4, 0.5)) {
    dens <- density(params$theta)
    max_dens <- max(dens$y)
    hdi_l <- as.numeric(hdi(params$theta)[2])
    hdi_r <- as.numeric(hdi(params$theta)[3])

    plot(dens, main = "Posterior", col = "darkseagreen", TeX("$\\theta$"))

    lines(c(hdi_l, hdi_r), c(0.12 * max_dens, 0.12 * max_dens), col = "orange")
    text(hdi_l, 0.15 * max_dens, round(hdi_l, 3), cex = 0.6)
    text(hdi_r, 0.15 * max_dens, round(hdi_r, 3), cex = 0.6)
    text((hdi_l + hdi_r)/2, 0.15 * max_dens, "95% HDI", cex = 0.6, col = "orange")

    lines(c(Rope[1], Rope[2]), c(0.08 * max_dens, 0.08 * max_dens), col = "blue")
    text(Rope[1], 0.05 * max_dens, Rope[1], cex = 0.6)
    text(Rope[2], 0.05 * max_dens, Rope[2], cex = 0.6)
    text((Rope[1] + Rope[2])/2, 0.05 * max_dens, "ROPE", cex = 0.6, col = "blue")

}
```

only once after that run only

```
modelString = "
data{
int<lower=0> N ;
int y[N] ; // y is a length-N vector of integers
}
parameters {
real<lower=0,upper=1> theta ;
}
model {
theta ~ beta(2,4) ;
y ~ bernoulli(theta) ;
}
"
# close quote for modelString

model <- stan_model(model_code = modelString)

N = 50
z = 10
y = c(rep(1, z), rep(0, N - z))

stanFit = sampling(object = model, data = list(y = y, N = N), iter = 2000,
```

```
    warmup = 200)

params = rstan::extract(stanFit)

plot_posterior(params, Rope = c(0.6, 0.8))
```

Problem 13.3

In a pilot medical study, 5 out of 16 patients responded to a *new* treatment. The probability of response to the standard treatment is given as 15 %. Is the new treatment superior to the standard treatment?

a) Investigate the specifications using the Bayesian method from the lecture. Make explicit statements about the prior distribution and the ROPE. Justify your answer.

b) Play around with the data so that the zero value is discarded or accepted.

Use the file **stan_beta_rope.R**.

Problem 13.4

("Quality control of screws") A manufacturer of screws guarantees his customers that the proportion of inferior screws is significantly less than 10 %. For quality assurance purposes, he takes a random sample of fifty screws from a large delivery. It turns out that 3 of these fifty screws are substandard. The problem for the manufacturer is: can he really assume with confidence that the proportion of substandard bolts in the whole delivery is really significantly less than 10 % (at a proportion of 10 % of substandard bolts, the quality standards are no longer met).

a) Investigate the data using the Bayesian method from the lecture. Make explicit statements about the prior distribution and the ROPE. Justify your answer.

b) Play around with the data so that the zero value is discarded or accepted.

Use the file **stan_beta_rope.R**.

Classical and Bayesian Statistic

Sample solution for Problems 13

Solution 13.2

In order to solve the problem with Bayes inference, we need a prior distribution. We choose the beta distribution as prior distribution, where we have to define the two parameter values a and b. In the prior distribution expresses our prior knowledge or our estimates of the efficacy of the drug.

We may well assume that the efficacy of the new drug is probably in the range of the standard method, i.e. around 30 %. To do this, we need to know or estimate how certain our assumption of 30 % is. This information is not included in the task, so we have to make appropriate assumptions here. To do this, we consider the value $\kappa = a + b$: If κ is large, then our certainty is large that the value of 30 % is correct. The beta distribution function then becomes very narrow. If our uncertainty is large, we choose a small value of κ so that the beta distribution function becomes broad.

We specify κ and the mode ω and, based on these two specifications, calculate the parameters a and b for the beta distribution as follows (see script):

$$a = \omega(\kappa - 2) + 1$$

and

$$b = (1 - \omega)(\kappa - 2) + 1$$

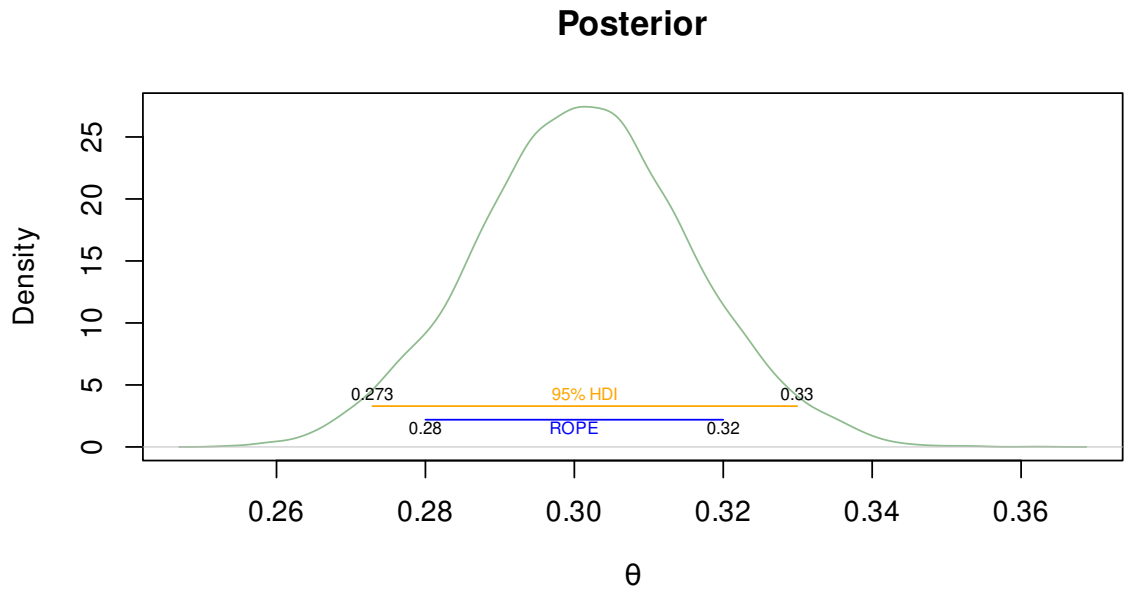
The mode of the prior distribution ω is consequently chosen as 0.3. Now for the choice of κ, which is much more difficult. If we are pretty sure that the standard method is already very good and difficult to improve, we are pretty sure that the efficacy of the new drug should also be around 30 %. We choose in this case $\kappa = 1000$. This gives $a = 300.4$ and $b = 699.6$.

The ROPE indicates which values are equivalent to the zero value 0.3. This information is provided by experts. We choose the ROPE as 0.3 ± 0.02.

We have 10 trials and 4 successes. Consequently, the **rstan** output looks like this:

```
modelString = "
data{
int<lower=0> N ;
int y[N] ; // y is a length-N vector of integers
}
parameters {
real<lower=0,upper=1> theta ;
```

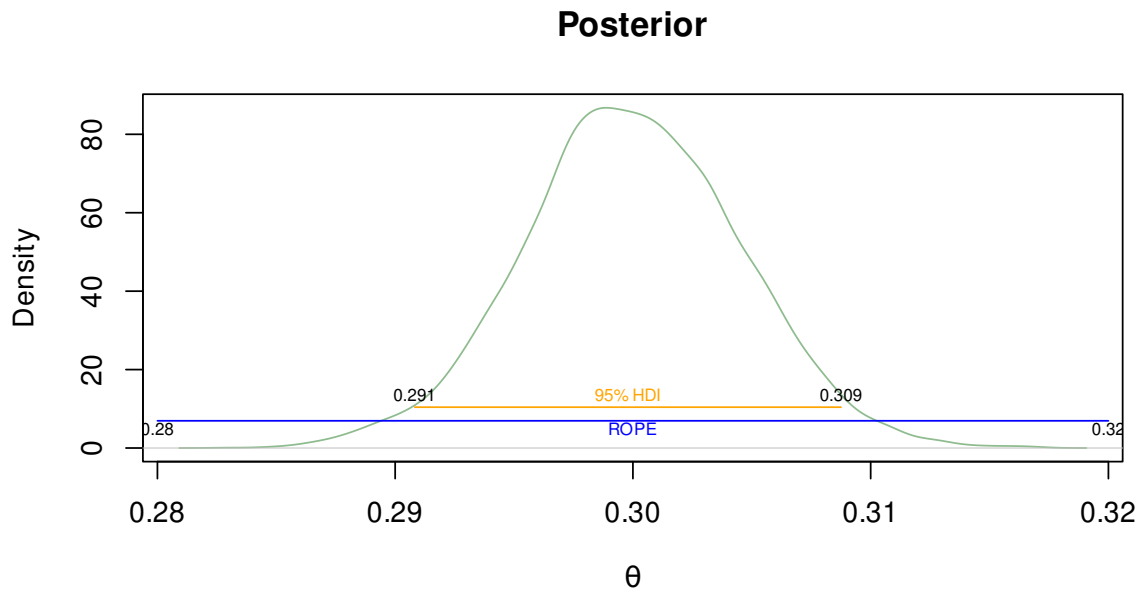
```
}
model {
theta ~ beta(300.4,699.6) ;
y ~ bernoulli(theta) ;
}
"
# close quote for modelString

model <- stan_model(model_code = modelString)

N = 10
z = 4
y = c(rep(1, z), rep(0, N - z))

stanFit = sampling(object = model, data = list(y = y, N = N), iter = 2000,
    warmup = 200, refresh = 0)

params = rstan::extract(stanFit)

plot_posterior(params, Rope = c(0.28, 0.32))
```



Since we have very little data, it does not have much influence on the prior distribution. The mode has not shifted. The ROPE is fully contained in the 95 %-HDI, i.e. some of the 95 % most likely values are outside the ROPE. In this case, we refuse to decide whether the new treatment method is more successful or equivalent to the standard method.

Let us now try $\kappa = 10000$, i.e. $a = 3000.4$ and $b = 6999.6$.

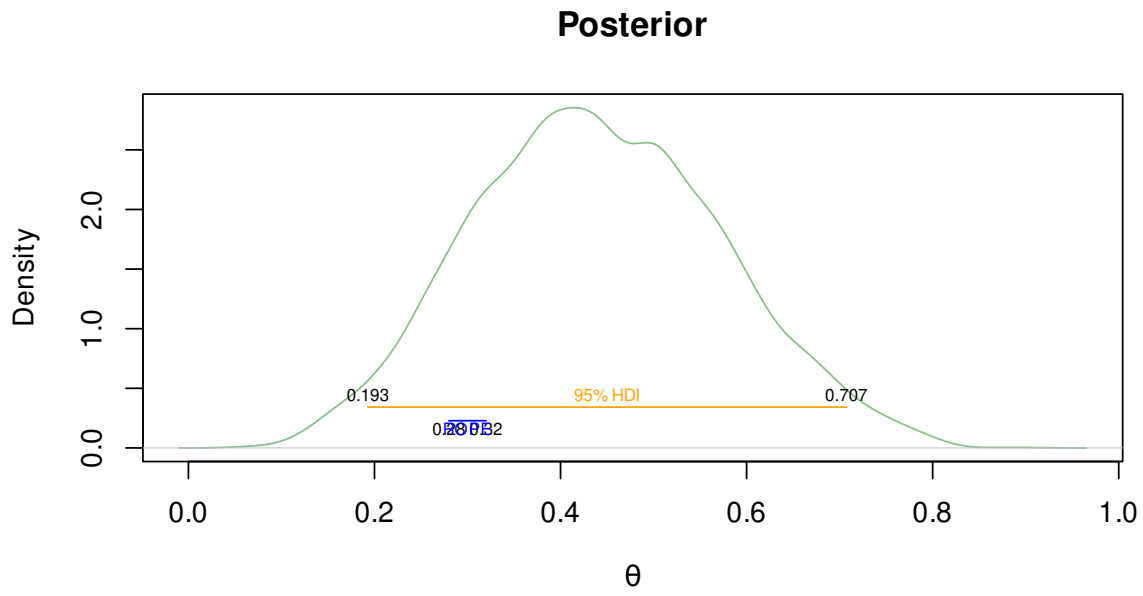
```
modelString = "
data{
int<lower=0> N ;
int y[N] ; // y is a length-N vector of integers
}
parameters {
real<lower=0,upper=1> theta ;
}
model {
theta ~ beta(3000.4,6999.6) ;
y ~ bernoulli(theta) ;
}
"
# close quote for modelString

model <- stan_model(model_code = modelString)

N = 10
z = 4
y = c(rep(1, z), rep(0, N - z))

stanFit = sampling(object = model, data = list(y = y, N = N), iter = 2000,
    warmup = 200, refresh = 0)

params = rstan::extract(stanFit)

plot_posterior(params, Rope = c(0.28, 0.32))
```



The ROPE now fully contains the 95 % most likely values, and thus we accept the zero value 0.3. Consequently, the efficacy of the new drug is equivalent to the standard method.

Let us now try a small κ, say $\kappa = 3$, i.e. $b = 1.7$ and $a = 1.3$. In this case, we are very unsure whether the 30 % are correct. Perhaps because the studies on the standard method are unreliable/dubious.

```
modelString = "
data{
int<lower=0> N ;
int y[N] ; // y is a length-N vector of integers
}
parameters {
real<lower=0, upper=1> theta ;
}
model {
theta ~ beta(1.7, 1.3) ;
y ~ bernoulli(theta) ;
}
"
# close quote for modelString

model <- stan_model(model_code = modelString)

N = 10
z = 4
y = c(rep(1, z), rep(0, N - z))

stanFit = sampling(object = model, data = list(y = y, N = N), iter = 2000,
    warmup = 200, refresh = 0)

params = rstan::extract(stanFit)

plot_posterior(params, Rope = c(0.28, 0.32))
```



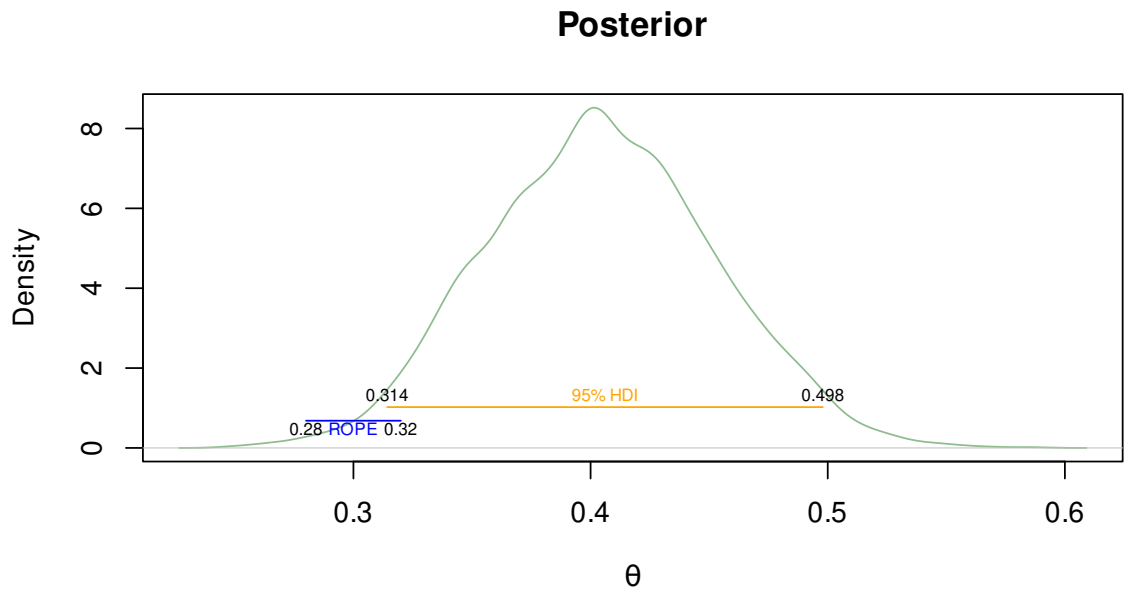
The mode has shifted from 0.3 to 0.39. The ROPE now sweeps only a very small range of the 95 % HDI, and thus very few of the most likely values are in the ROPE. In this case, we refuse to decide whether the new treatment method is more successful or equivalent to the standard method.

The reason why we cannot make a clear decision except for very large κ is because of the small number of trials. If we increase the number of trials, the posterior distribution becomes narrower.

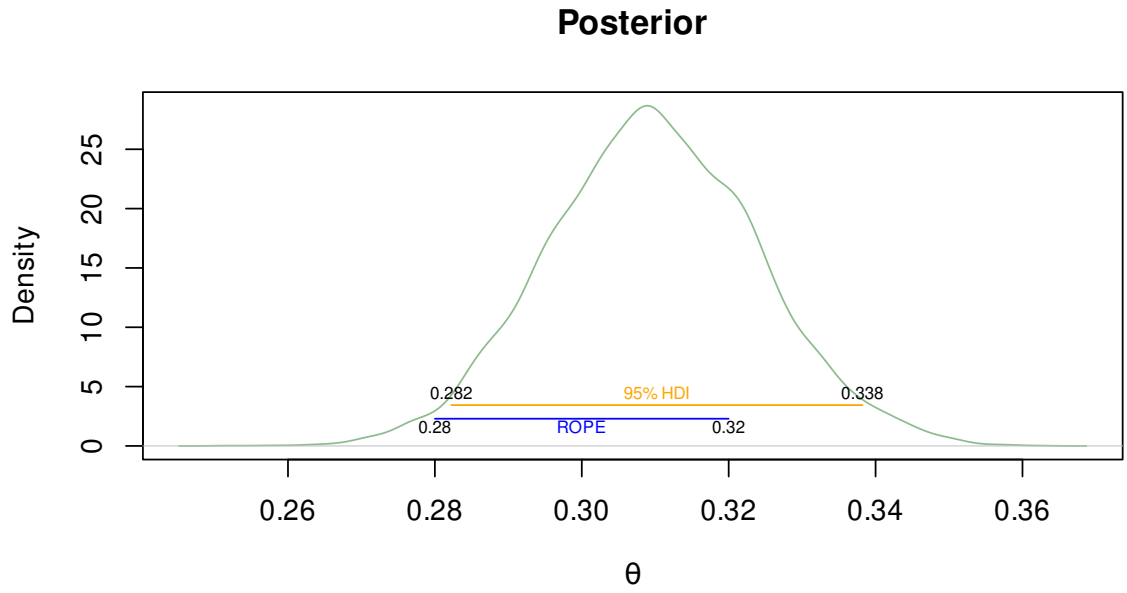
The posterior is always a trade-off between data and the prior:

- Large uncertainty and lots of data: Posterior looks more like the likelihood function (data).
- Large certainty and few data: Posterior looks more like the prior.
- In all other cases, a mixture between likelihood function and prior.

In the following, we consider 100 trials and 40 successes. Thus the ratio of successes to trials remains the same. For $\kappa = 3$ (large uncertainty), i.e. $b = 1.7$ and $a = 1.3$, the posterior is similar to the likelihood function. The ROPE is almost completely outside the HDI, i.e. there is some evidence that the new drug has a higher efficacy than the standard method. However, even in this case we refuse to decide whether the new treatment method is more successful or equivalent to the standard method.

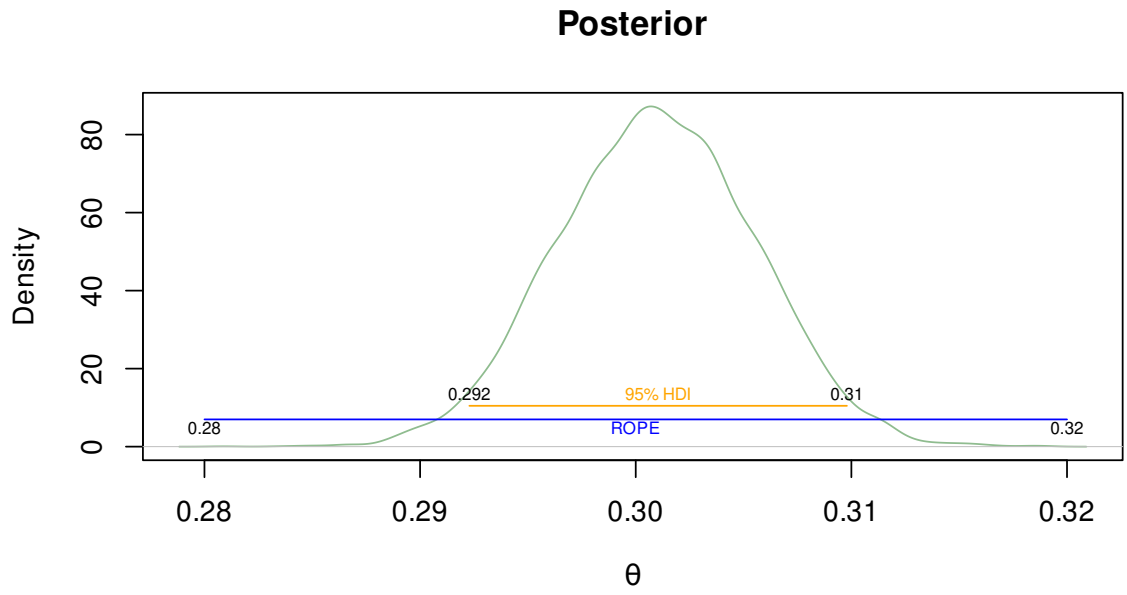


For $\kappa = 1000$ ($a = 300.4$ and $b = 699.6$) the data no longer has the överhand. Since there is a significant overlap between ROPE and HDI, we cannot make a decision on whether the new treatment is also effectively better.



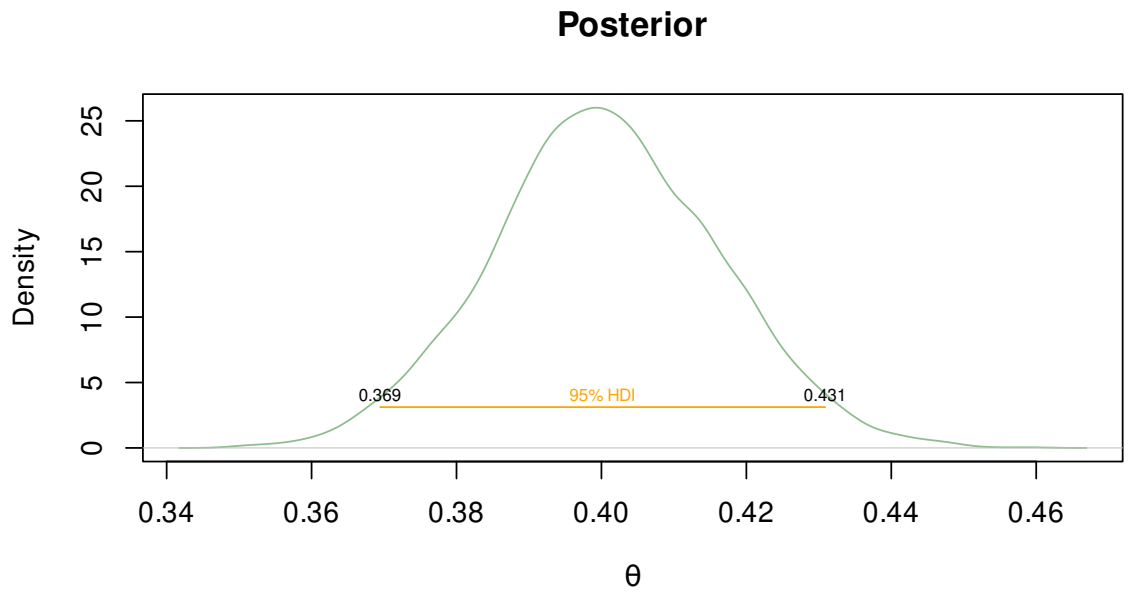
For $\kappa = 10000$ ($a = 3000.4$ and $b = 6999.6$), it does not have enough data to dissuade us from our 0.3 conviction, the ROPE fully contains the HDI. That is, the new drug

is not better than the standard method, but equivalent to it. We accept the zero value 0.3 in this case.

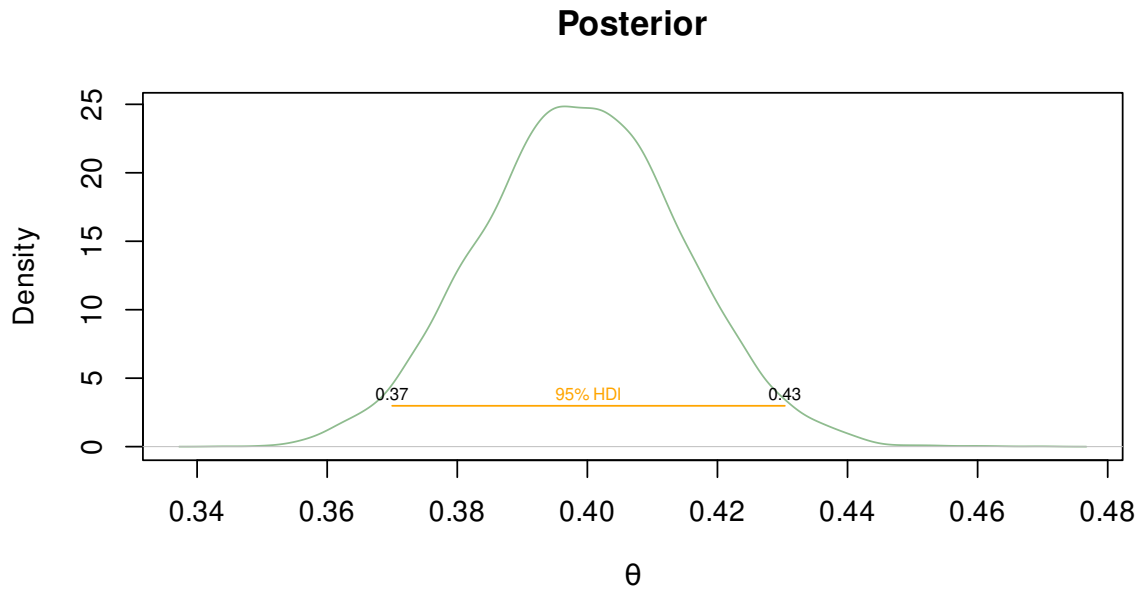


Below we consider 1000 trials and 400 successes and the prior with $\kappa = 3$ ($b = 1.7$ and $a = 1.3$). The data now convince us to abandon our conviction since the ROPE and the HDI do not overlap. We can now assume that the new treatment method is better.

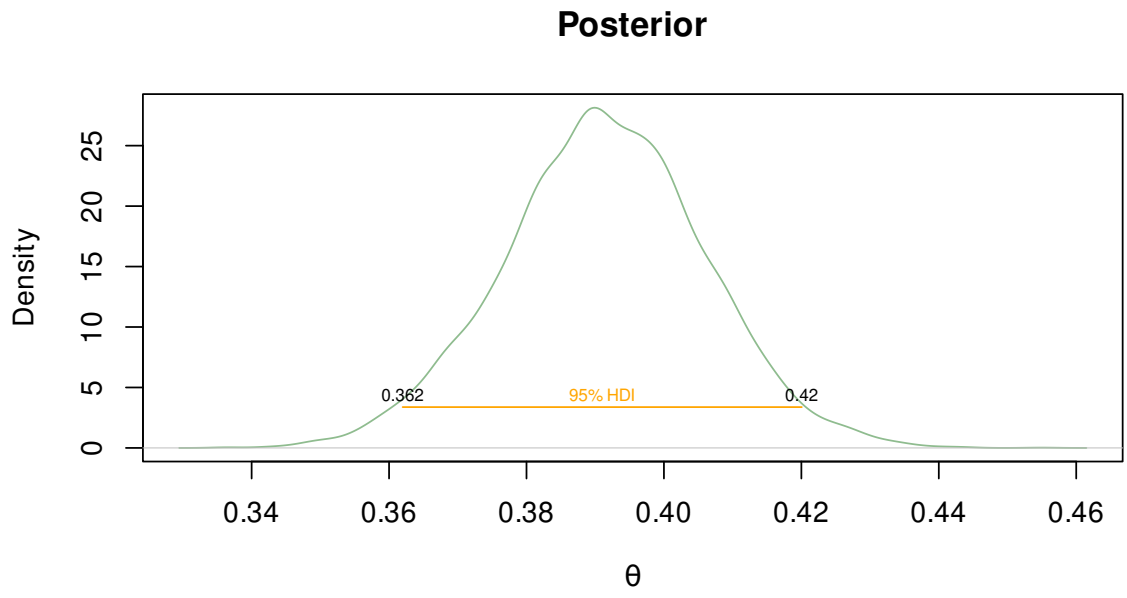
```
## recompiling to avoid crashing R session
```



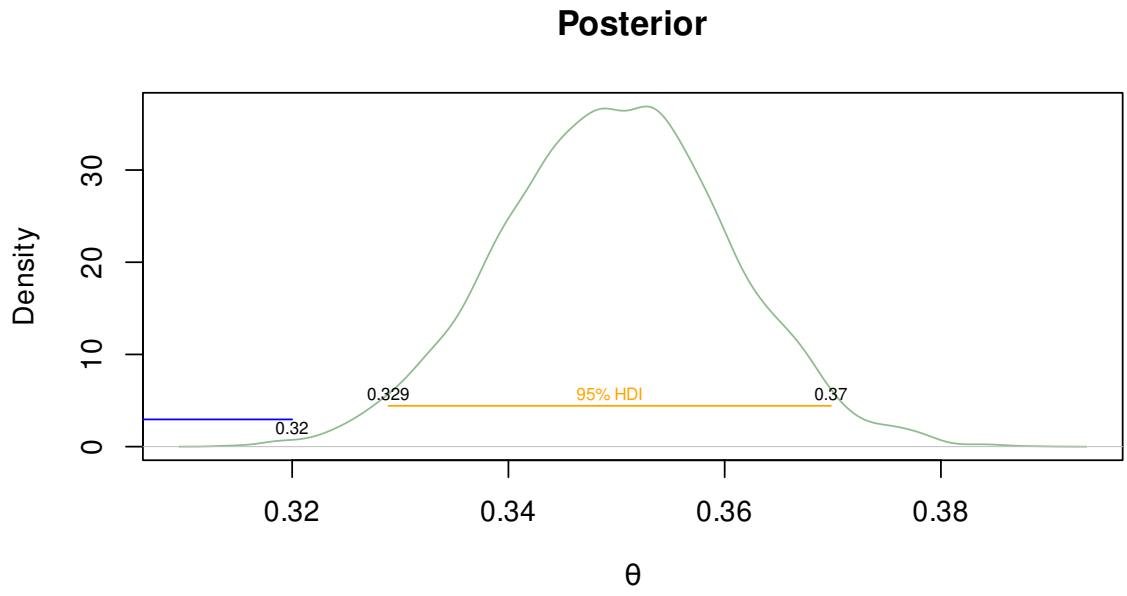
We reach the same conclusion with $\kappa = 10$.



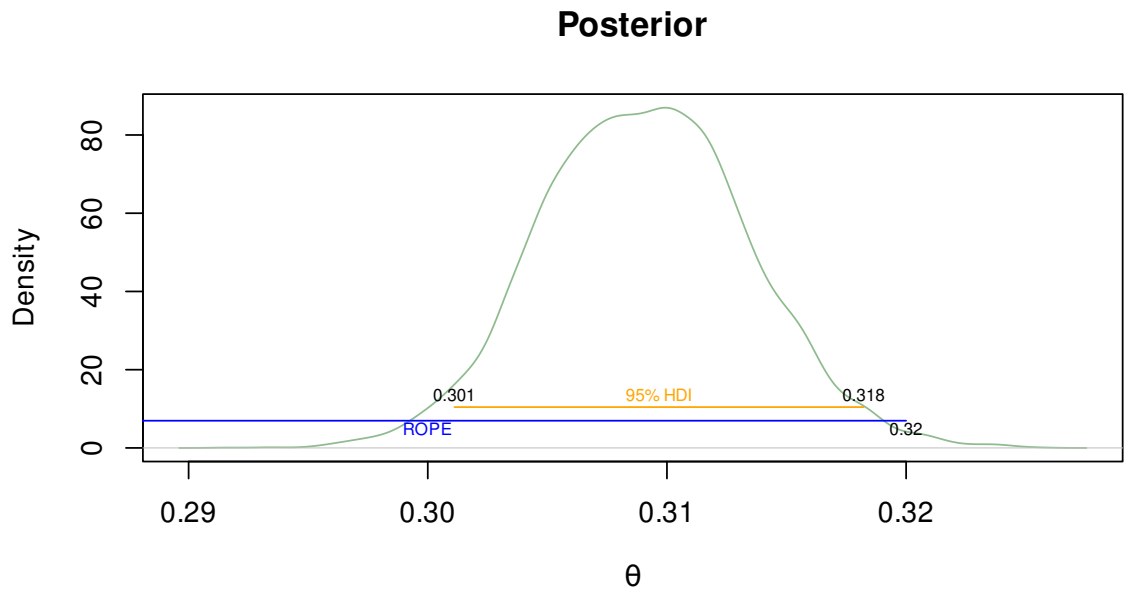
We reach the same conclusion with $\kappa = 100$.



We come to the same conclusion with $\kappa = 1000$



For $\kappa = 10000$ it looks different again. It does not have enough data to dissuade us from our 0.3 conviction, the ROPE fully contains the HDI. That is, the new drug is not better than the standard method, but equivalent to it. We accept the zero value 0.3 in this case.



Solution 13.3

To solve the problem with Bayesian inference, we need a prior distribution. We choose the beta distribution as prior distribution, where we have to define the two parameter values a and b. In the prior distribution expresses our prior knowledge or our estimates of the efficacy of the drug.

We may well assume that the efficacy of the new drug is probably in the range of the standard method, i.e. by 15 %. To do this, we need to know or estimate how certain our assumption of 15 % is. This information is not included in the task, so we have to make appropriate assumptions here. To do this, we consider the value $\kappa = a + b$: If κ is large, then our certainty is large that the value of 15 % is correct. The beta distribution function then becomes very narrow. If our uncertainty is large, we choose a small value of κ so that the beta distribution function becomes broad.

We specify κ and the mode ω and, based on these two specifications, calculate the parameters a and b for the beta distribution as follows (see script):

$$a = \omega(\kappa - 2) + 1$$

and

$$b = (1 - \omega)(\kappa - 2) + 1$$

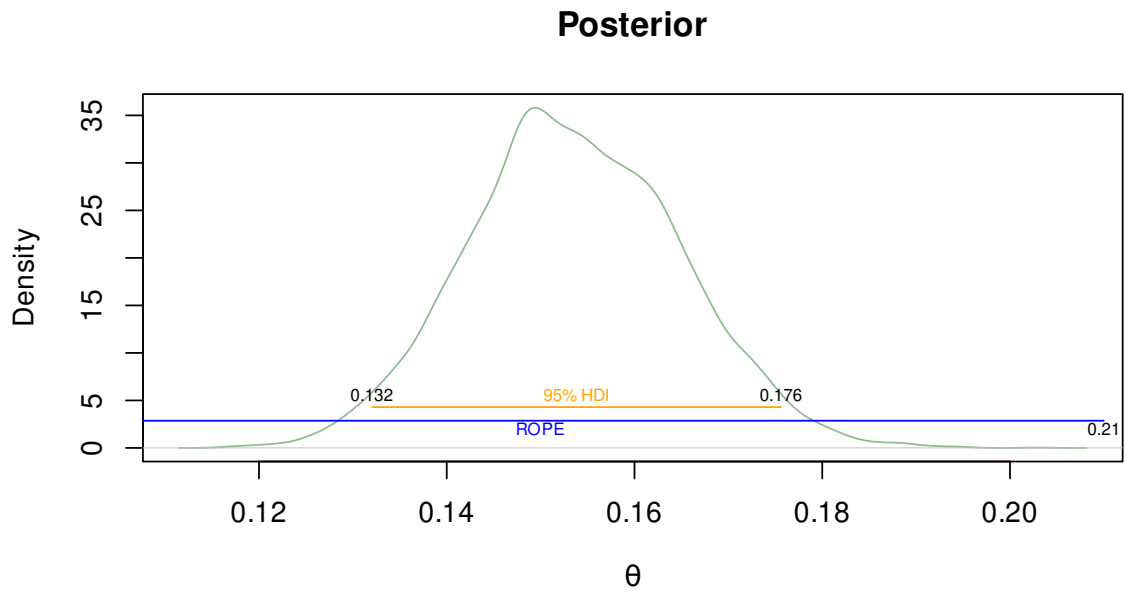
The mode of the prior distribution ω is consequently chosen as 0.15. Now for the choice of κ, which is much more difficult. If we are pretty sure that the standard method is already very good and difficult to improve, we are pretty sure that the

efficacy of the new drug should also be around 15 %. We choose in this case $\kappa = 1000$. This gives $a = 150.7$ and $b = 849.3$.

The ROPE indicates which values are equivalent to the zero value 0.15. This information must be provided by experts. We choose the ROPE as 0.15 ± 0.06.

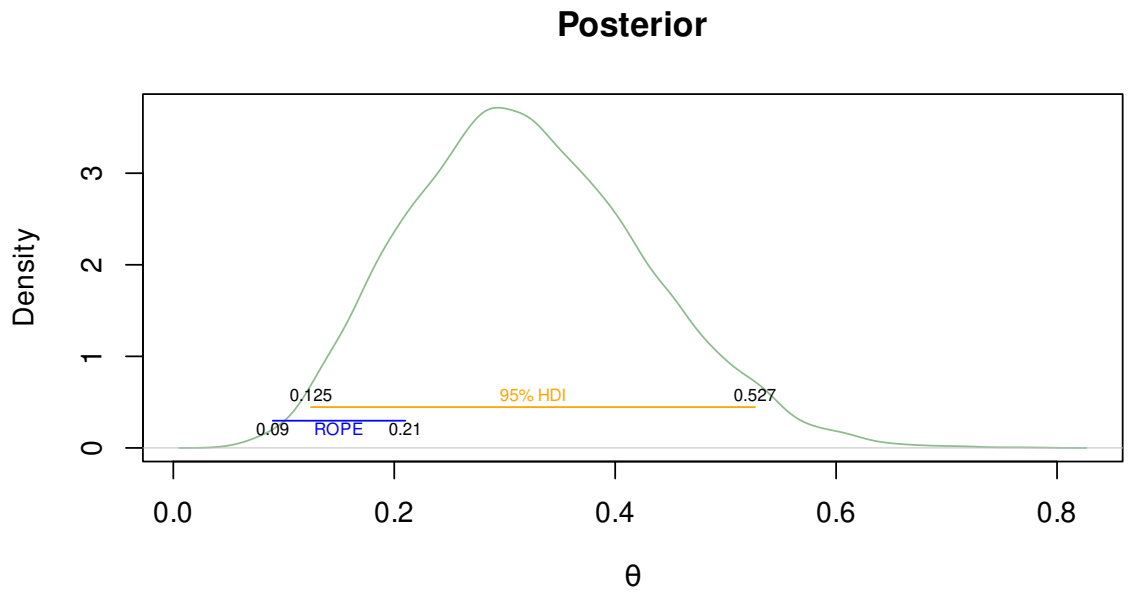
We have 16 trials and 5 successes. Consequently, the **rstan** input looks like this:

```
modelString = "
data{
int<lower=0> N ;
int y[N] ; // y is a length-N vector of integers
}
parameters {
real<lower=0, upper=1> theta ;
}
model {
theta ~ beta(150.7, 849.3) ;
y ~ bernoulli(theta) ;
}
"
# close quote for modelString

model <- stan_model(model_code = modelString)

N = 16
z = 5
y = c(rep(1, z), rep(0, N - z))

stanFit = sampling(object = model, data = list(y = y, N = N), iter = 2000,
    warmup = 200, refresh = 0)

params = rstan::extract(stanFit)

plot_posterior(params, Rope = c(0.09, 0.21))
```



Since we have very little data, it does not have much impact on the prior distribution. The mode has not shifted. The 95 %-HDI is fully contained in the ROPE, i.e. all of the 95 % most likely values are within the ROPE. In this case, we accept the zero value and conclude that the new treatment type is equivalent to the standard method.

Let us now try a small κ, say $\kappa = 3$, i.e. $b = 1.85$ and $a = 1.15$, and increase the number of trials: 160 trials and 50 successes. In this case, we are very unsure whether the 30 % are correct. Perhaps because the studies on the standard method are unreliable/dubious. In addition, we now have much more data.



The mode has shifted from 0.15 to 0.31. The ROPE intersects with the 95 % HDI. In this case, based on the data, we cannot decide whether we accept the zero value or not.

Solution 13.4

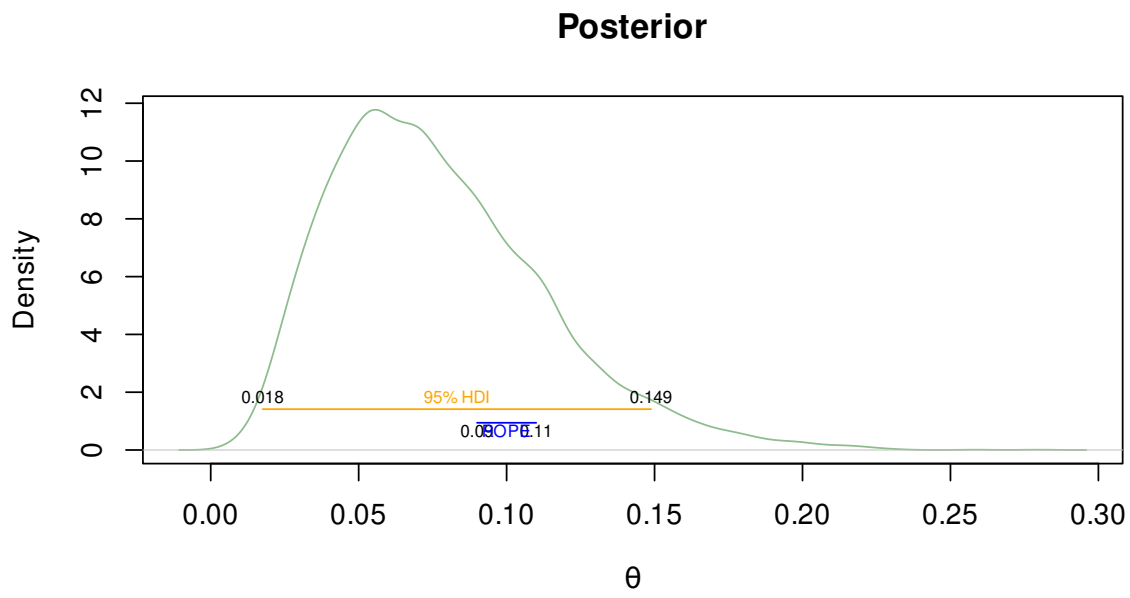
To solve the problem with Bayesian inference, we need a prior distribution. We choose the beta distribution as prior distribution, where we have to define the two parameter values a and b. In the prior distribution expresses our prior knowledge or our estimates of the efficacy of the drug. In this task, we assume complete ignorance of the proportion of inferior bolts. We consequently choose the Uniform Distribution, i.e. $a = 1$ and $b = 1$.

The ROPE indicates which values are equivalent to the zero value 0.1. This information must be provided by material testing experts. We choose the ROPE from 0.09 to 0.11. If the percentage of substandard bolts falls within this range, then we consider the quality standards not met.

We have 50 samples and 3 substandard bolts. Consequently, the **rstan** output looks like this:

```
modelString = "
data{
int<lower=0> N ;
int y[N] ; // y is a length-N vector of integers
}
parameters {
```

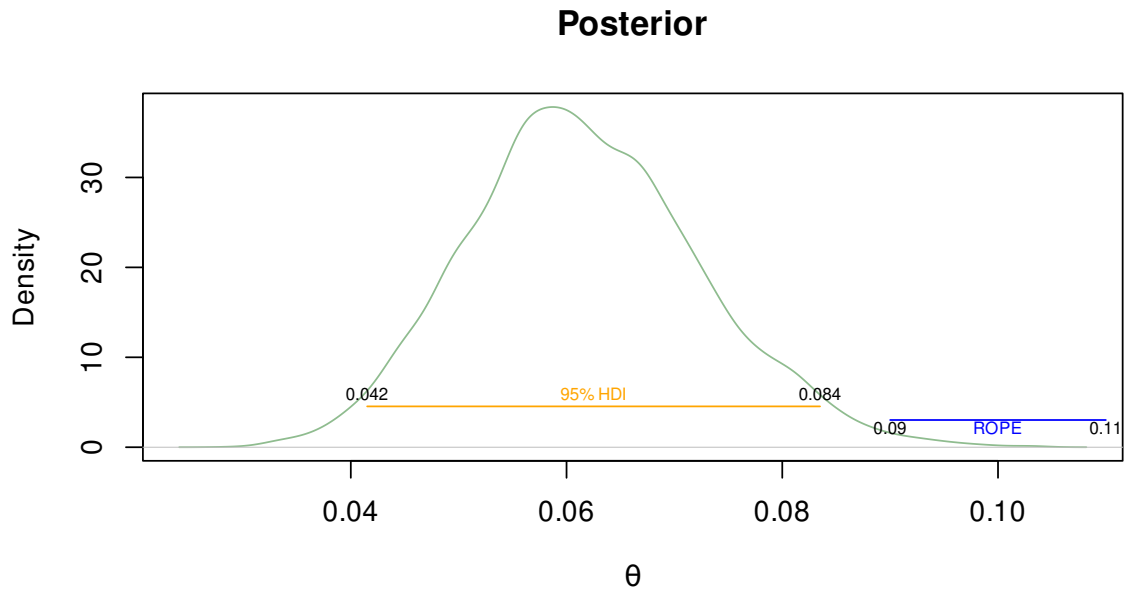
```
real<lower=0,upper=1> theta ;
}
model {
theta ~ beta(1,1) ;
y ~ bernoulli(theta) ;
}
"
# close quote for modelString

model <- stan_model(model_code = modelString)

N = 50
z = 3
y = c(rep(1, z), rep(0, N - z))

stanFit = sampling(object = model, data = list(y = y, N = N), iter = 2000,
    warmup = 200, refresh = 0)

params = rstan::extract(stanFit)

plot_posterior(params, Rope = c(0.09, 0.11))
```



Since we have chosen a uniform distribution for the prior distribution, the data has a very large influence on the posterior distribution. The mode of the posterior distribution is now at the percentage of inferior bolts in the sample. The ROPE overlaps with the 95 %-HDI, i.e. some of the 95 % most likely values lie outside the ROPE. In this case, we refuse to decide whether the bolts meet the quality standards.

Suppose we drew a sample of 500 screws and 30 inferior screws.

```
modelString = "
data{
int<lower=0> N ;
int y[N] ; // y is a length-N vector of integers
}
parameters {
real<lower=0,upper=1> theta ;
}
model {
theta ~ beta(1,1) ;
y ~ bernoulli(theta) ;
}
"
# close quote for modelString

model <- stan_model(model_code = modelString)

N = 500
z = 30
y = c(rep(1, z), rep(0, N - z))

stanFit = sampling(object = model, data = list(y = y, N = N), iter = 2000,
    warmup = 200, refresh = 0)

params = rstan::extract(stanFit)

plot_posterior(params, Rope = c(0.09, 0.11))
```



The ROPE and the HDI are separate in this case. We discard the zero value 0.1 here and conclude that the screws meet the quality standards.