Notes for

Data Analysis for Food Science

Contents

Preface

During the past decades the production of data in relation research, production, consumer behavior, social network etc. has increased dramatically. Today we are faced with data structures which were unimaginable just 50 years ago. Traditionally, a system under investigation were characterized by a few samples associated with say one to five descriptors and, carefully selected, responses. Today all aspects of the classical system interrogation has blown up, such that we have many more samples (e.g. production monitoring every minute), more descriptors (e.g. consumer characteristics), and by far more response variables (e.g. high throughput omic technologies). Tools developed for handling traditional scenarios still pertain the corner of how to approach todays data analytical challenges, however, by the development of computers, it is possible to carry out challenging mathematical procedures in no time and further produce visual graphics as resources for translating information into knowledge. Due to this fact, the traditional tools has gotten a makeover and new tools has been developed.

Food is, as such, an extremely inherent part of the human life, although one could argue that so is e.g. cardiovascular biology and governmental policy making, these subjects either work autonomously or does not demand everyday mental capacity. Everyday all humans need to eat- and drink in some social context, pay attention to the perception of the meal, and further deal with the possible health- and emotional implications of this process. When studying food science all these aspects are relevant.

Food science constitute a broad range of disciplines spanning controlled artificial model systems, over functional modification of real food matrices, production technology, to the relation between food- and meal composition, taste, perception and health. **All by means of data**. These notes are thought to cover data analysis within food science. That is to; provide a general understanding of the purpose of data analysis, found a theoretical- and practical basis for understanding various numeric and graphical tools and couple generic tools to concrete issues within related disciplines. To this end by theory, examples and exercises.

The Book material used in these notes are mostly from the notes for the course; Introduction to Statistics at DTU by P.B. Brockhoff and co workers. Additionally there are relevant chapters from other sources. All exercises are costum made and deal with real problems within food science.

Welcome to the course in Fødevaredataanalyse for second year bachelor students in food science - Hope you will enjoy learning about how to use data for getting insight on food systems.

August 2017
Morten Arendt Rasmussen

1. Week 1

This first week is going to introduce basic descriptive tools for getting a primary overview of data. These are divided into representative numerics, which we call descriptive statistics and various plots. Especially for the plotting part you will be needing a computer program. We strongly encourage you to get familiar with R, that, although not being as intuitive point-and-click as the widely used programs such as excel, is capable of conducting almost any type of sophisticated analysis you may wish, and further will strengthen you to become familiar with a scientific programing language - a generic competence useful whenever working with information.

Included in the week 1 notes are material related to working in R; Installing packages, importing data, working in scripts and debugging your code. This material you should try to cover briefly and then use it when you get stuck on a problem throughout the course (and after). However, there is a huge amount of videos, tutorials, etc. on the web which you can also use, and we encourage you to get familiar with these resources as well. Simply type your problem in google and check if others have experienced something similar.

1.1 Hand-in assignment

The exercise 1.7 *Analysis of Coffee Serving Temperature - PCA* is to be handed in (through Absalon or as hard-copy Wednesday night). You are welcome to put in R-code in the assignment, but it is your argumentation and interpretation that are the most important. If you have problems with R, then try to write what you would have done if you did not experience problems with the machinery.

1.2 Exercises

For Monday work through exercise 1.1, 1.2 and 1.4 and for Wednesday work through 1.5 and 1.8. You will most likely not be able to complete all exercises within the hours in the classroom, so we recommend that you use some time in advance to initiate the task.

1.3 Case I

The first, of a total of four cases, are described in the document "Case1.pdf". You should work on the case in groups of four, and hand in a slide-show with voice no later than Thursday evening next week. Be aware, that a lot of the technical stuff can be zacked from the exercises, so you might want to finalize those in advance of the Case.

1.4 Installing Packages

A package in R is a set of commands which are not a part of the base-set in R. Many of the R-commands which are used throughout this course requires a certain package to be installed on the computer/Mac. It is a good idea to get familiar with Installing packages and loading them onto your R-script mainly so you will not be missing them at the exercises, casework or examination.

In R there are two important commands concerning installation of packages.

- `install.packages()` installs the target package on your computer.
- `remove.packages()` uninstalls the package from your computer.

For example: `install.packages('readxl')` Installs the package `readxl` on the computer and `remove.packages('readxl')` uninstalls the package `readxl` from the computer.

1.4.1 Loading a package

When the packages are installed on the computer, you can load them onto your workspace/script at every occasion you initiate your analysis in R. To do this, you use the `library()` command. `library()` points at a package-library stored on your computer. Everytime you open a new session of R, you need to load the needed packages again.

For example, `library(readxl)` Loads the package "readxl" onto the workspace.

When you load a package, you might get warning messages like the following:

```
> library(readxl)
Advarselsbesked:
pakke 'readxl' blev bygget under R version 3.1.3
```

These are completely normal and won't affect the usage of the package.
After the package has been loaded, the associated commands can now be used.

1.4.2 Working directory

In R you are using something called a *working directory* or *wd* for short. This is the folder on your computer in which R saves and finds the projects that you are working on. This also makes it easier to load datasets. The working directory can be changed in R either manually or through code. `getwd()` and `setwd()` are the two important commands for changing the working directory.

```
> getwd() Shows the current working directory
[1] "/Users/madsbjorlie/Documents/Statistik/Exercises/Week 1"
> setwd("~/Documents/R-træning") Changes the working directory
> getwd()
[1] "/Users/madsbjorlie/Documents/R-træning"
```

1.5 Importing a dataset

Throughout this course you will need to import a lot of data into R. Getting familiar with the following packages and commands will help minimize your R-related frustration. Datasets can be imported into R in numerous ways. Like changing the working directory, it can be done both manually and through coding. I recommend doing it through coding since this makes it easier to maintain an overview.

Allmost all of the datasets that will be handed out in this course will be in the excel file-type `.xlsx`. R is also capable of importing text-files such as `.txt` or `.csv`.

`.xlsx`-files are Microsodt Excels standard projectfolder-filetype, whereas `.csv`-files are short for *comma separated values* and is a term for text-files where the values are separated by a comma (or in the Danish Excel, a semicolon).

You can either import datasets through R's inherent commands or use some data-import packages to import file-types such as `.xlsx` or `.xls`. Both methods work fine and which one you will use depends on your personal preference.

1.5.1 Importing a dataset through R's own commands

As a default, R can not import Excel-files such as `.xls` and `.xlsx`. To use R's `read.csv()` function, you need to save the Excel dataset as a `.csv` file. This is done by choosing *Save as* (in Excel) and then selecting the `.csv` file-type. This might seem a bit tedious, but it eliminates the demand for other packages.
`read.csv()` imports the dataset specified in the parenthesis. This can be done in two ways: by typing the path the file has on your computer or by using the command `file.choose()` which corresponds to opening a new file. If the dataset is in the working directory, you don not have to type the full path, but just the file-name.

For example: `Beer <- read.csv(file.choose(), header=TRUE, sep=";", dec=",")`

```
Beer <- read.csv("Beerdata.csv", header=TRUE, sep=";", dec=",")
Beer <- read.csv("~/Documents/R-træning/Øldata.csv", header=TRUE, sep=";", dec=",")
```

The different arguments: `header =`, `sep =` and `dec =` tells R how to import the data. `header=TRUE` tells R that the first row in the dataset is not a part of the data itself but carries the variablenames. `sep=";"` defines which separator the document uses. By using Danish Excel, this will always be semicolon. This can be checked by opening the dataset in NotePad on windows or TextEditor on Mac. `dec=","` defines which symbol is used for decimals. It is necessary to make sure that the dataset in R is separated by a full stop rather than a comma. This can be checked by using summary commands after the data has been imported.

1.5.2 Importing a dataset using packages

By using various packages, it is possible to import Excel-documents directly into R. This can be quite handy, but some of the packages will not run on Mac or on Windows due to other programs missing. The most used data-import packages are: `gdata`, verb—readxl—, `xlsx` and `rio`. `gdata` requires *Perl* which is default on Mac and Linux but not on Windows and therefor it will not run on Windows unless it is installed. The following is a couple of examples using the various packages.

```
Library(readxl)
Beer <- read_excel(file.choose())
Beer <- read_excel("~/Documents/R-træning/Beerdata.xls")
Beer <- read_excel("~/Documents/R-træning/Beerdata.xlsx")

Library(gdata)
Beer <- read.xls(file.choose())
Beer <- read.xls("~/Documents/R-træning/Beerdata.xls")

Libray(xlsx)
Beer <- read.xlsx(file.choose(), sheetIndex = 1)
Beer <- read.xlsx("~/Documents/R-træning/Beerdata.xls", sheetIndex = 1)
Beer <- read.xlsx("~/Documents/R-træning/Beerdata.xlsx", sheetIndex = 1)

Library(rio)
Beer <- import(file.choose())
Beer <- import("~/Documents/R-træning/Beerdata.xls")
Beer <- import("~/Documents/R-træning/Beerdata.xlsx")
```

1.5.3 Getting an overview of the dataset

When the dataset is imported into R, you can use different commands to check that it was imported correctly. The commands are `head()`, `str()` and `dim()`.

`head()` shows the first 6 rows in the dataset.
`str()` shows the types of the various columns, such as *numeric* and *factor*.
`dim()` shows the dimensions of the data-matrix.

```
> McDonalds <- read.csv(file.choose(), header=TRUE, sep=";", dec=",")
> head(McDonalds)
                       X Energy Protein Carbohydrate  Fat Saturated_Fat
1                Big Mac   9.51    0.12         0.20 0.11          0.04
2                McFeast   8.63    0.12         0.15 0.11          0.04
3   Quarter Pounder m/ost  10.46    0.16         0.18 0.13          0.06
4              McChicken   9.19    0.11         0.19 0.11          0.02
5 Grilled Chicken Caprese   8.09    0.13         0.15 0.09          0.00
6           Filet-O-Fish  11.79    0.11         0.26 0.15          0.03
```

Or

```
   > str(McDonalds)
'data.frame': 19 obs. of  6 variables:
$ X            : Factor w/ 19 levels "Apple-pie","Big Mac"...
$ Energy       : num  9.51 8.63 10.46 9.19 8.09 ...
$ Protein      : num  0.12 0.12 0.16 0.11 0.13 0.11 0.13 0.14 0.16 0.04 ...
$ Carbohydrate : num  0.2 0.15 0.18 0.19 0.15 0.26 0.29 0.26 0.18 0.19 ...
$ Fat          : num  0.11 0.11 0.13 0.11 0.09 0.15 0.08 0.1 0.13 0.03 ...
$ Saturated_Fat: num  0.04 0.04 0.06 0.02 0 0.03 0.03 0.04 0.06 0.02 ...

> dim(McDonalds)
[1] 19  6
```

1.6 Scripts

We highly recommend that you make your data analysis using a script. A script is simply a flat text file that is given the surname `.R` such that R can interpret the commands. Here you will have the commands needed to do the analysis from setting necessary functions, import of data, initial inspection, modeling and plots.

Each analysis task is slightly different, however, almost always there is a set of generic tasks which is always needed. That is: cleaning up the workspace, loading packages, setting work directory, loading data and checking the data structure. That typically fills up the first 5-10 lines of code in every script as follows:

```
rm(list = ls())
library(ggplot2)
library(rio)
setwd("~/MyComputer/Courses/FDA/Exercises/Week1")
X <- import("~/MyComputer/Courses/FDA/Data/SomeData.xlsx", format = 'Excel')
head(X)
qplot(data = X,x,y,color = treatment)
```

1.7 Descriptive statistics

The learning objectives for this theme is to:

- Understand what univariate, bi-variate and multivariate data is.
- Comprehend the concepts of centrality and dispersion.
- Be able to compute a range of metrics (numbers), that are informative with respect to centrality and dispersion.
- Know what correlated variables are, and be able to calculate a correlation coefficient.

Know how to use the functions `aggregate()` and `summary()` to create overview tables. And further what the very useful functions `str(), head(), tail()` and `view()` are doing.

1.7.1 Reading Material

Chapter 1 of *Introduction to Statistics* by Brockhoff https://02402.compute.dtu.dk/enotes/book-IntroStatistics especially 1.1 to 1.4.

1.7.2 Exercises

Exercise 1.1 Descriptive Statistics - by hand

Below is listed a vector of ranking (*Liking*) of coffee served at 56°C by 52 consumers. The data is sorted.

1	2	3	4	5	6	7	8	9	10
2	2	3	3	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5
5	6	6	6	6	6	6	6	6	6
6	6	7	7	7	7	7	7	7	7
7	7	7	7	7	7	7	8	8	8
8	9								

1. Calculate mean, variance, standard deviation, median and inner quartile range for this distribution of data.

Some useful numbers: $\sum X = 301$, $\sum (X_i - \bar{X})^2 = 122.7$

Exercise 1.2 Descriptive Statistics

Serving temperature of coffee seems of importance as to how this drink is perceived. However, it is not totally clear how this relation is. In order to understand this, studies on the same type of coffee served at different temperature is conducted. In this exercise we are going to use the data from a consumer panel of 52 consumers, evaluating coffee served at six different temperatures on a set of sensorical descriptors leading to a total of $52 \times 6 = 312$ samples.

In the dataset *Results Consumer Test.xlsx* the results are listed. Taking these data from A to Z involves descriptive analysis for understanding variation within judge, between judge and between different temperatures, further outlier detection, and finally determination of structure between sensorical descriptors. In this exercise we are only going through some of the initial descriptive steps.

In the table below a subset of the data is shown.

	Sample	Assessor	Liking	Intensity	Sour	Bitter	Sweet
1	31C	1	3	4	3	4	2
11	31C	11	3	1	9	1	7
21	31C	21	6	3	1	4	7
31	31C	31	3	3	3	8	2
41	31C	41	1	1	3	1	2
51	31C	51	3	7	7	3	2
61	37C	9	2	7	7	4	2
71	37C	19	6	3	4	4	1
81	37C	29	6	8	6	5	2
91	37C	39	5	2	3	5	1
101	37C	49	7	2	3	3	7
111	44C	7	3	7	7	8	2
121	44C	17	7	5	3	8	1
131	44C	27	5	3	3	9	3
141	44C	37	6	6	3	5	3
151	44C	47	7	6	1	6	1
161	50C	5	4	3	2	3	1
171	50C	15	2	8	8	6	1
181	50C	25	7	6	7	6	3
191	50C	35	4	2	1	2	2

1. Import the data (Be aware that the function `read.xls()` is not in the base library, so you need to add the specific library to your computer).
2. Subsample on one temperature (Below is listed two alternatives for doing this).

(Some code to zack from)

```
Coffee <- read.xls("Results Consumer Test.xlsx")
Coffee_t44_v1 <- Coffee[Coffee$Temperatur==44,]
```

3. Calculate the descriptive statistics for centrality (mean and median), dispersion (IQR, standard deviation and range) and extremes (min and max) for this distribution of datapoints for a single descriptor (e.g. *Liking*)
4. Now do it for all temperatures.

You should get something like the table below.

	Temp	N	Mean	Median	Std	Min	Max
1	31	52	3.58	3	1.65	1	7
2	37	52	4.75	5	1.78	1	7
3	44	52	5.83	6	1.61	2	9
4	50	52	5.96	6	1.60	2	8
5	56	52	5.79	6	1.55	2	9
6	62	52	6.17	6	1.37	2	8

This can be quite tedious, and result in a lot of coding. However, the function `summary()` and `aggreggate()` are very efficient in producing such results. Try to check out these functions and see if you can use those

to generate summary statistics. Below are shown some code which does exactly what you want without too many lines of code.

```
Coffee_t44_v2 <- subset(Coffee,Temperatur==44)
mean(Coffee_t44_v1$Liking)

# Include only responses.
CoffeeDT <- Coffee[,2:10]
# Run aggregate for each type of summary
tmpN <-aggregate(CoffeeDT,by=list(CoffeeDT$Temperatur),FUN
     = 'length')
tmpM<-aggregate(CoffeeDT,by=list(CoffeeDT$Temperatur),FUN
    = 'mean')
tmpM2<-aggregate(CoffeeDT,by=list(CoffeeDT$Temperatur),FUN
     = 'median')
tmpS<-aggregate(CoffeeDT,by=list(CoffeeDT$Temperatur),FUN
    = 'sd')
tmpMi<-aggregate(CoffeeDT,by=list(CoffeeDT$Temperatur),FUN
     = 'min')
tmpMx<-aggregate(CoffeeDT,by=list(CoffeeDT$Temperatur),FUN
     = 'max')

# merge these into a dataset
tmp <- cbind(tmpM$Temperatur,tmpN$Liking,tmpM$Liking,tmpM2
    $Liking,tmpS$Liking,tmpMi$Liking,tmpMx$Liking)
```

5. The above is done for *Liking*, try to do it for some of the other responses. HINT: This can be done by repeating the code and exchange `$Liking` with e.g. `$Bitter`. However, putting this in a for loop is another option.

6. What have you learned from analysing these data in terms of importance of serving temperature on the sensorical properties as percieved by consumers? HINT: You can run the code below to get a comprehensive overview. This is based on the mean aggreggate, but you might just as well check some of the other descriptive metrics. For instance, what does the standard deviation tells you about consumers in general, and does the type of sensorical attribute and serving temperature make a difference on the spread in scoring?

```
matplot(tmpM[,2],tmpM[,6:10],type='l',lwd=3)
text(cbind(60,t(tmpM[6,6:10])),colnames(tmpM[,6:10]))
```

(You might want to fix some of the labels in these figures. Check the documentation by typing `?matplot` and see how to add meaning full stuff to the plot)

1.8 Debugging - *Getting R to work*

When using any computer program you now and then encounter that it does not do as intended. However, it is so, that the program do exactly what it is told, which might not be in line with the task you anticipate conducted. R work by interpreting commands which are either written directly on the command line, or in the form of lines in a script which then is submitted to the R compiler. Sometimes instead of producing nice results and plots, R returns red stuff on the screen. Debugging is the process of figuring out why that is so, and change the code to do as anticipated.

We will throughout the course get used to debug code, why the learning objectives for this set of skills stretches over several weeks. In detail you should:

- Know that R distinguish small and capital letters.
- Know that some R functions comes in libraries.
- Know the top five most common, and trivial, reasons for R to produce errors.

1.8.1 Reading Material

The notes on debugging *Debugging in R.pdf* available through Absalon.

1.8.2 Exercises

Exercise 1.3 Debugging - Importing files and calling on objects and functions

For some of you, coding in Rstudio may seem simple. The aim with these debuggingstasks is to train you to analyze the errors Rstudio gives you, and to give you some tools to use to avoid issues when coding in Rstudio. Many of the debugging exercises throughout the course will be related to the datasets used in other exercises given in the same week, so you might find it sensible to do the debugging-exercises first, to get to know the datasets and their potential issues, before the struggle starts.

Importing a dataset, debugging function calls

Dataset: `Results Consumer Test.xlsx`

1. Why won't Rstudio read in the file in the following cases?

```
> Coffee <- read_excel('Results Consumer Test.xlsx')
Error: could not find function "read_excel"
>
> Coffee <- read_excel('Results Consumer Test')
Error: 'Results Consumer Test' does not exist in current working directory ('C:/Use
>
```

2. Why isn't the view-function working?

```
> Coffee <- read_excel('Results Consumer Test.xlsx')
> view(Coffee)
Error: could not find function "view"
>
```

3. What is missing when you call for the `aggregate` to do its calculations?

```
> CoffeeRP <- read_excel("Results Panel.xlsx")
> CoffeeAG <- aggregate (by=list(CoffeeRP$Assessor,CoffeeRP$Sample),FUN ="sd")
Error in is.ts(x) : argument "x" is missing, with no default
>
```

1.9 Plotting

The learning objectives for this theme is to be able to produce a range of plots reflecting the raw data distribution. That be; histogram, boxplot, jitterplot, scatterplot, lineplot, stem 'n' leaf plot, bar chart, pie chart and spline plot. Further, to be able to understand how desriptive statistics are connected to these plots. Utilizing the various functionality to infer information on the plot, by color, markersize, transparency etc. Optimally you will be using the ggplot2 package in R for this purpose.

1.9.1 Reading Material

Chapter 1 of *Introduction to Statistics* by Brockhoff `https://02402.compute.dtu.dk/enotes/book-IntroStatistics` especially 1.5 to 1.6.

Cheat Sheets: On https://www.rstudio.com/resources/cheatsheets/, there are a very nice cheat sheet on how to use the ggplot2 functions. Use some time to go through this to get hands on experience with these different calls.

Online lectures: There is a lot of how-to-plot on the web. Use it when you are getting stuck at a problem, or if you feel like being inspired. This one is pretty comprehensive http://www.mathtube.org/lecture/video/visualising-data-ggplot2

1.9.2 Exercises

Exercise 1.4 Plotting using R and ggplot2

The package `ggplot2` for R is a versatile tool for producing various plots with the option of modifying them in great detail. The following exercises should guide you through the basic functionality of the `qplot` function and provide the ability to produce plots satisfying 75% of your plotting needs ever.

Plotting distributions

Assume that you have a single response variable, for instance alcohol content of a series of various types of drinks, measurements of body weight from an nutritional experiment or content of antioxidants for a given product produced under different conditions. For all of these, the variable is continuous in form. As a starting point of every analysis an overview of the distribution of the variable of interest is crucial and plotting the distribution facilitate insight into character of distribution (bell shaped, skewness, bi modal, zero inflated etc.) and single point information for outlier identification.

Before you start...

Start by importing a dataset and specify some necessary packages (If you have not installed them on your local drive, you might need to do so). The following lines of code will do the job, if you specify the correct path.

```
Coffee <- read.xls('Results Consumer Test.xlsx')
```

There are several ways of importing data into R, depending on which format you have them in. From excel, the `read.xls()` (from the `gdata` library) or `read_excel()` (from the `readxl` library) are two ways of doing it. In either case, make sure that the data is correctly imported, by comparing the imported data with the original. Sometimes there are problems with decimals.

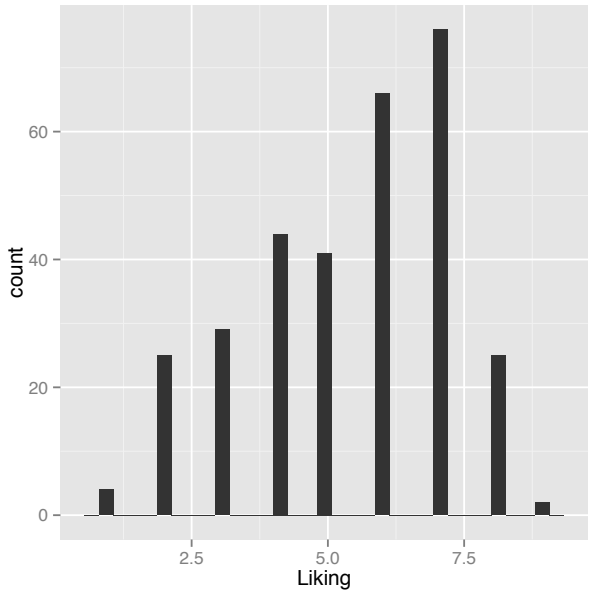
Histograms

The histogram is the most basic representation of continuous data. The very simple command:

```
qplot(data=Coffee,Liking)
```

produces the following plot.

Figure 1.1: Histogram

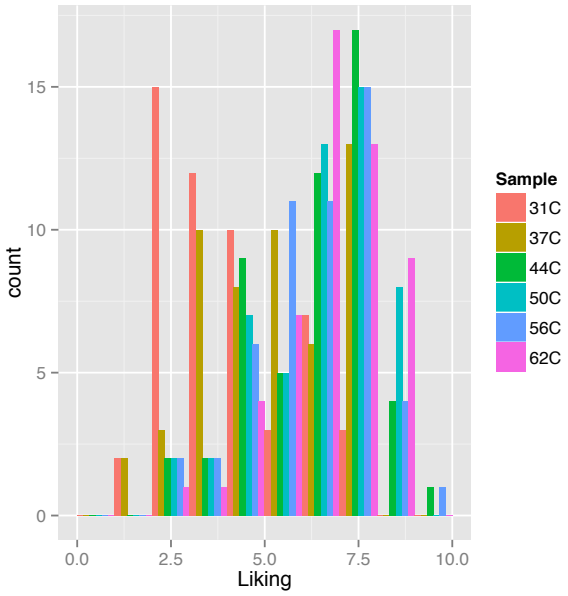


There are numerous additional arguments that can be used to modify this plot. Check out the documentation or google it to see how it is done. Try to modify the color, the bin width, the transparency and the main title on the plot.

The data here are "liking" of coffees at different temperatures, and so one might wish to infer this information on the histogram. This can be done by the following command.

```
ggplot(data=Coffee,aes(x = Liking,fill=Sample)) +
    geom_histogram(position = 'dodge',binwidth = 0.8)
```

Figure 1.2: Histogram with temperatur imposed



Interpreting this figure, how does the temperature affect the liking?

If you do not want to overlay the histograms, it is possible to plot them as individual panels. Try to run the following code and see what it does:

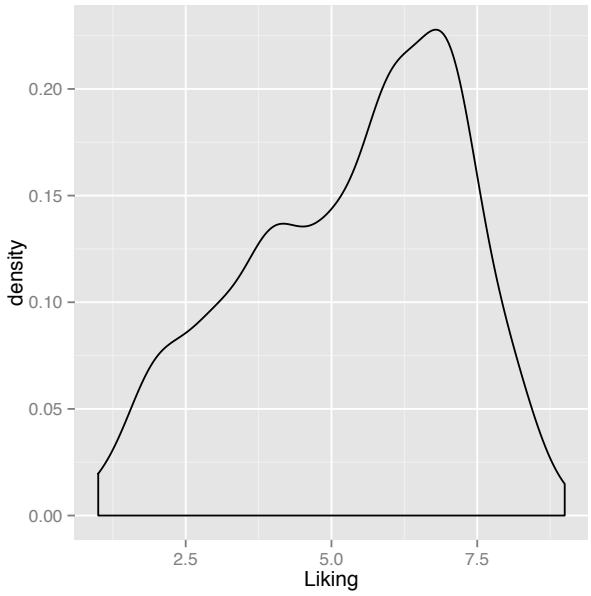
```
# as three different panels
```

Densitogramplot

A densitogram is a smoothed extension of the histogram, and as such represents the same type of information. The bin width in the histogram controls the resolution, whereas the counterpart in the densitogram is the degree of smoothing. By only adding a single option to the `qplot()` the plot is changed to a densitogram.

```
pdf('Densitogram1.pdf',width = w, height = h)
```

Figure 1.3: Densitogram



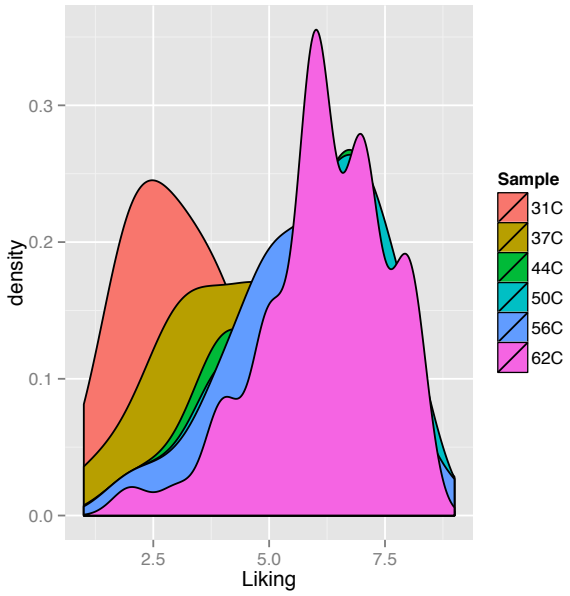
Try to modify the smoothing of the densitogram by changing the option `adjust = ..`

Try to make the smoothing very refined (e.g. `adjust = 0.3`), does this reflect the underlying distribution? What is a suitable smoothing option for these data?

Exactly as for the histogram, it is possible to infer additional information on the densitogram.

```
pdf('Densitogram2.pdf',width = w, height = h)
```

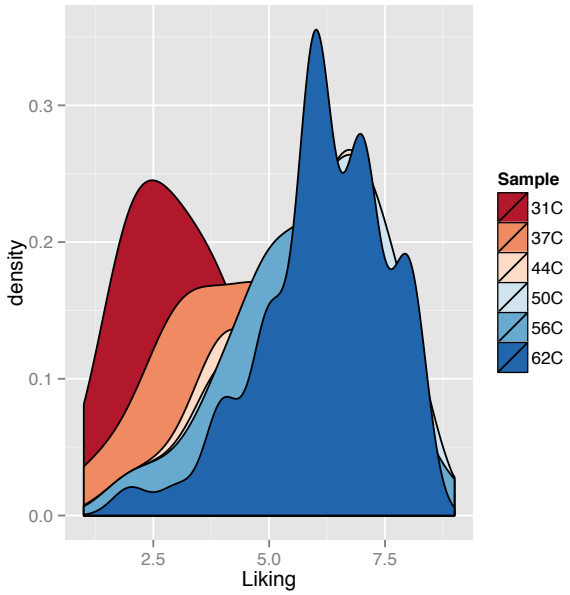
Figure 1.4: Densitogram with temperatur imposed



This plot is not optimal, as only the densitogram for temperature 62C is fully visible. Try to adjust the transparency by `alpha = I(0.5)` to get a better version of the same plot. The colors used in the plot could be more intuitive as they refer to temperature. There are several ways the color scheme. One is to add a layer to the plot specifying the either a predefined color scheme, a modification of a predefined color scheme or simply by specifying each of the colors used. Here we just use a predefined *Red to Blue* palette.

```
pdf('Densitogram3a.pdf',width = w, height = h)
```

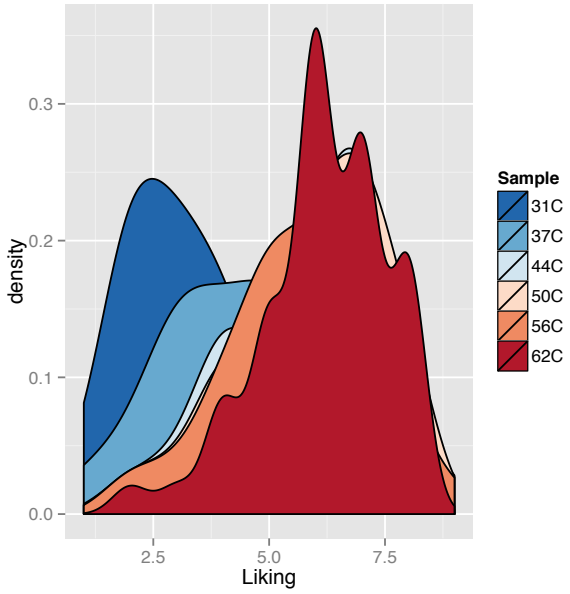
Figure 1.5: Densitogram with temperature imposed - Different color scheme



However, the colors are in a counter intuitive direction. If we instead manually specify the colors, by reverting the order of a predefined palette, then the coding goes as.

```
pdf('Densitogram3b.pdf',width = w, height = h)
library(RColorBrewer)
```

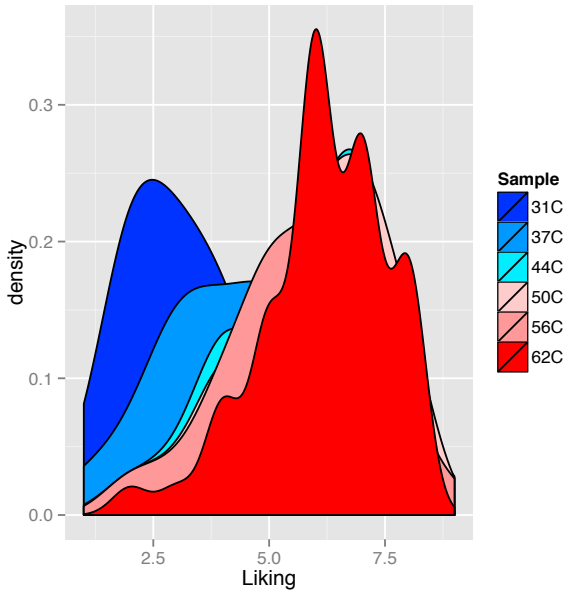
Figure 1.6: Densitogram with temperature imposed - reversed color scheme



If nothing seems to fit your ideal color-world, then you can simply specify the exact colors.

```
pdf('Densitogram3c.pdf',width = w, height = h)
colPalette <- c('#0033FF','#0099FF','#00EEFF','#FFCCCC','#
    FF9999','#FF0000')
```

Figure 1.7: Densitogram with temperature imposed - User defined color scheme



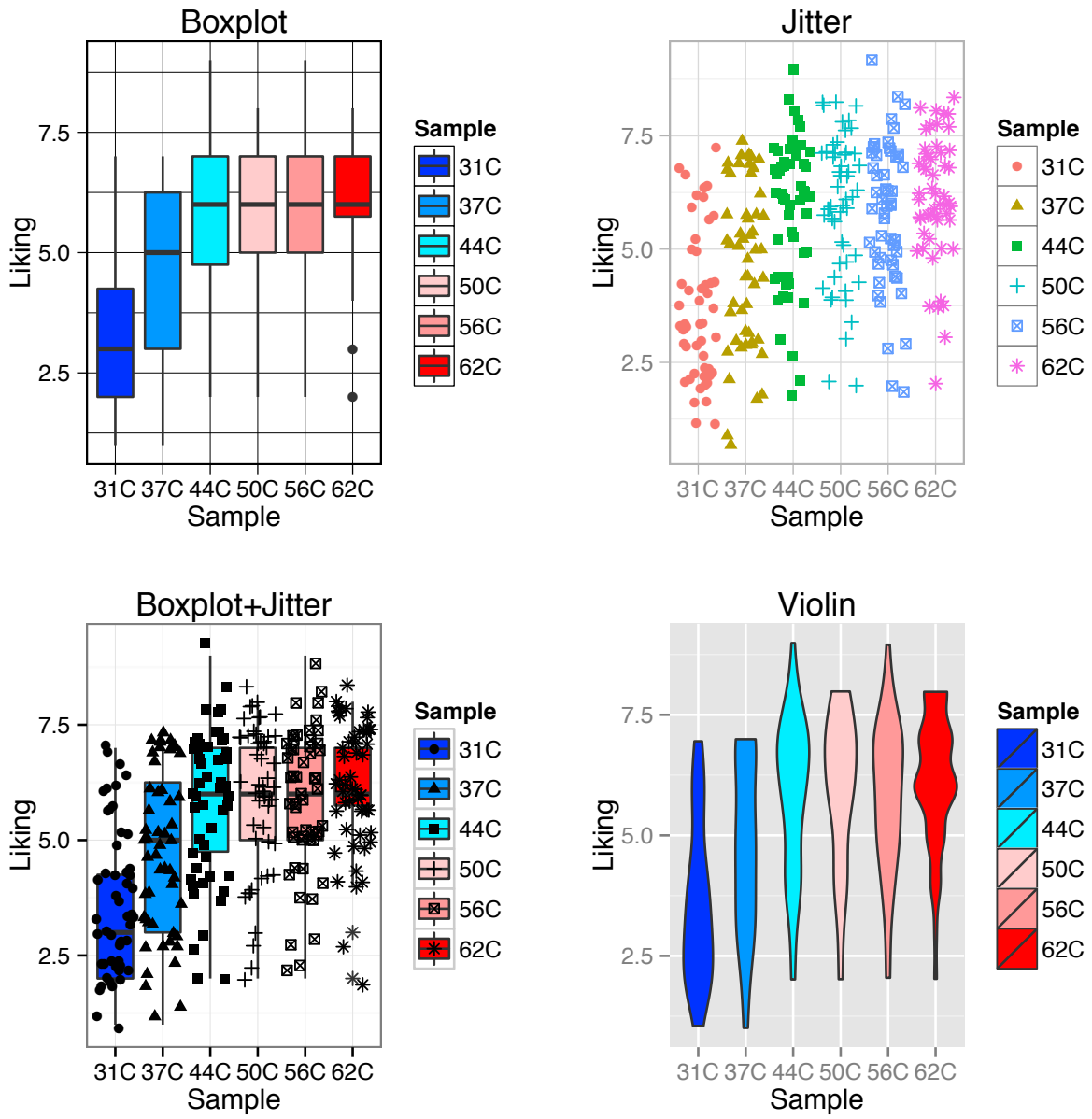
Try to reconstruct these plots, and try to use other predefined color schemes. Further all these plots suffers from lack of transparency, so include that as well.

Boxplot, Jitterplot and Violinplot

In the above, the different temperatures were infered on the plot by overlaying histograms. However, the x-axis can be used for keeping track of this information. This is especially useful when you are comparing more than four levels. The code below produces the four plots in Figure 1.8.

```
cll <- scale_fill_manual(values = colPalette)
p1 <- qplot(Sample,Liking,data=Coffee,geom='boxplot',fill=
    Sample) + theme_linedraw() + ggtitle('Boxplot') + cll
p2 <- qplot(Sample,Liking,data=Coffee,geom='jitter',color=
    Sample,shape=Sample) + theme_light() + ggtitle('Jitter
    ') + cll
p3 <- qplot(Sample,Liking,data=Coffee,geom=c('boxplot',"
    jitter"),fill=Sample, shape=Sample,main = 'Boxplot+
    Jitter') + theme_bw() + cll
```

Figure 1.8: Densitogram with temperature imposed - User defined color scheme



There are several things to notice from the code:

- The title on the individual plots can be inferred either using `qplot(..., main='The Title')` or by "adding" the title; |qplot() + ggtitle('The Title')—.
- The color scale is predefined and simply added to the individual plots.
- The background of the plots are different, and is inferred by adding `+ theme_XXX()` (The default is `theme_gray()`).
- In the *Jitter* plot the colors do not work. Try to use `scale_color_manual()` instead of `scale_fill_manual`, and see what happens.

The plots are written to a pdf file by the following command.

```
library(gridExtra)
pdf('Box_jitter_violin.pdf')
grid.arrange(p1,p2,p3,p4,ncol=2)
dev.off()
```

Try to reconstruct these plots.

Both the violin plot and the jitter plot are quite intuitive on what is actually being plotted. However, the boxplot have several features, such as a box with a line in the middle, some so-called whiskers and also maybe a few actual points. Check out what these refer to in the data, and calculate them directly on data to verify that the computer is not off.

Exercise 1.5 Analysis of Coffee Serving Temperature - Data inspection

Serving temperature of coffee seems of importance on how this drink is perceived. However, it is not totally clear how this relation is. In order to understand this, studies on the same type of coffee served at different temperature is conducted. In this exercise we are going to use the data from a trained Panel of eight judges, evaluating coffee served at six different temperatures on a set of sensorical descriptors. Each judge is presented with each temperature in a total of four replicates leading to a total of $6 \times 8 \times 4 = 192$ samples.

In the dataset *Results Panel.xlsx* the results are listed. Taking these data from A to Z involves descriptive analysis for understanding variation within judge, between judge and between different temperatures, further outlier detection, and finally determination of structure between sensorical descriptors. In this exercise we are only going to briefly explore the data with emphasis on uncertainty.

1. Import the data and check that it is matching the excel file using `head()` .

This is data from a trained panel, meaning each judge have been trained to be an objective instrument returning the *same* response when presented the *same* sample. However, there is always uncertainty in such responses, and especially when the instrument is a human being. We are interested in how big the deviation is between the four replicates, across judges and samples.

2. Use the `aggreggate()` function to extract certain descriptive measures (e.g. mean or standard deviation) from the data. As we are interested in the deviation, use `sd`.
3. Plot this descriptive measure for a single descriptor across temperature (x-axis) and join the points from the same judge.
4. What can you say about the individual judges? And is scoring more difficult for higher temperature than lower?

The code below does (some) of the job. Check out what is actually done by `rename.vars()`. Further, try to remove the `factor()` statement around Judge, and see what happens.

```
# remove replicate by averaging over this - that is; Keep
    data as AssessorXSample
CoffeeAG <- aggregate(Coffee,by=list(Coffee$Assessor,
    Coffee$Sample),FUN="mean")
# rename some variables
CoffeeAG <- rename.vars(CoffeeAG,c('Group.1','Group.2'),c(
    'Judge','Temp'))
# Make some initial plotting of the results
qplot(data=CoffeeAG,Temp,Intensity,group=Judge,color=
    factor(Judge)) + geom_line()
```

1.10 PCA

The learning objectives for this theme is to be able to compute a Principal Component Analysis (PCA) model and know that this model can be utilized for getting information on the **multivariate sample distribution** and **variable correlation structure**. Further, to comprehend *how* this is a generalization of the tools used for uni- and bivariate data such as; histograms/jitterplot/boxplot, scatterplots and correlation coefficient known for multivariate data. The mathematical formulation of the PCA is not a theme for this weeks learning.

1.10.1 PCA *In brief*

Principal Component Analysis is a method for understanding multivariate data. By multivariate data we mean a set of samples/observations (n) which are characterized on a number of different features (p). For example:

- Different beers ($n = 20$) analyzed for $p = 60$ chemical variables reflecting the aroma composition.
- Six coffee samples assessed on a set of sensorical descriptors ($p = 12$) by sensorical panel of eight judges ($n = 48$)
- A range of oil samples ($n = 40$) analyzed by near infrared spectroscopy ($p = 400$)

The data is often arranged in a data table (referred to as $\mathbf{X}$) with n rows (samples) and p columns (variables).

For almost all real life applications such multivariate data are correlated. That is; some of the variables carry the same type of information, and the interesting information in these data is captured by this so-called correlation structure. A nice visualization of the correlation is done via a scatter plot of two variables. For a few variables (say $p = 5$) it is possible to interpret all combinations of two variables. If $p = 5$ that amounts to $\frac{p(p-1)}{2} = 10$ plots. However, when p is high this becomes in-practical. PCA deals with this issue by compressing the data into a set of components:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \ldots + \mathbf{t}_k\mathbf{p}_k^T + \mathbf{E} \tag{1.1}$$

Notation wise, $\mathbf{X} \sim (n, p)$ is a matrix, $\mathbf{t}_i \sim (n, 1)$ and $\mathbf{p}_i \sim (p, 1)$ are vectors.

- $\mathbf{t}_1$ are scores for component 1 ($\mathbf{t}_2$ are scores for component 2 and so forth) and tells something about the multivariate sample distribution
- $\mathbf{p}_1$ are loadings for component 1 ($\mathbf{p}_2$ are loadings for component 2 and so forth) and tells something about the multivariate correlation structure between the variables.
- A set of $\mathbf{t}_1$ and $\mathbf{p}_1$ is referred to as a *component*.

The power of PCA is when the mathematical decomposition into scores and loadings is combined with visualization of these. That is:

- **Score plot** - scatter plots of combination of scores (often $\mathbf{t}_1$ vs. $\mathbf{t}_2$)
- **Loading plot** - scatter plots of combination of loadings (often $\mathbf{p}_1$ vs. $\mathbf{p}_2$)
- **Bi plot** - overlayed score- and loading plot

1.10.2 Example: Natural Phenolic Antioxidants for Meat Preservation - PCA

This data originates from a study investigating the effect of natural phenolic antioxidants against lipid and protein oxidation during sausage production and storage. Bologna-type sausages were prepared and treated with either green tea (GT) or rosemary extract (RE) as antioxidants, and a control batch was also included. The three types of sausages were evaluated by a sensory panel including 8 assessors, on 18 different descriptors within the categories *smell*, *color*, *taste* and *texture*. The sausages were evaluated immediately after production (*week*0) and after four weeks of storage (*week*4).

Data is from: Jongberg, Sisse, et al. "Effect of green tea or rosemary extract on protein oxidation in Bologna type sausages prepared from oxidatively stressed pork." *Meat Science* 93.3 (2013): 538-546.

Load libraries

Note that these need to be installed beforehand!

```
library(ggplot2)
library(ggbiplot)
```

Load the data

Remember to set your directory!

```
load('meat_data.rdata')
```

Calculate PCA model

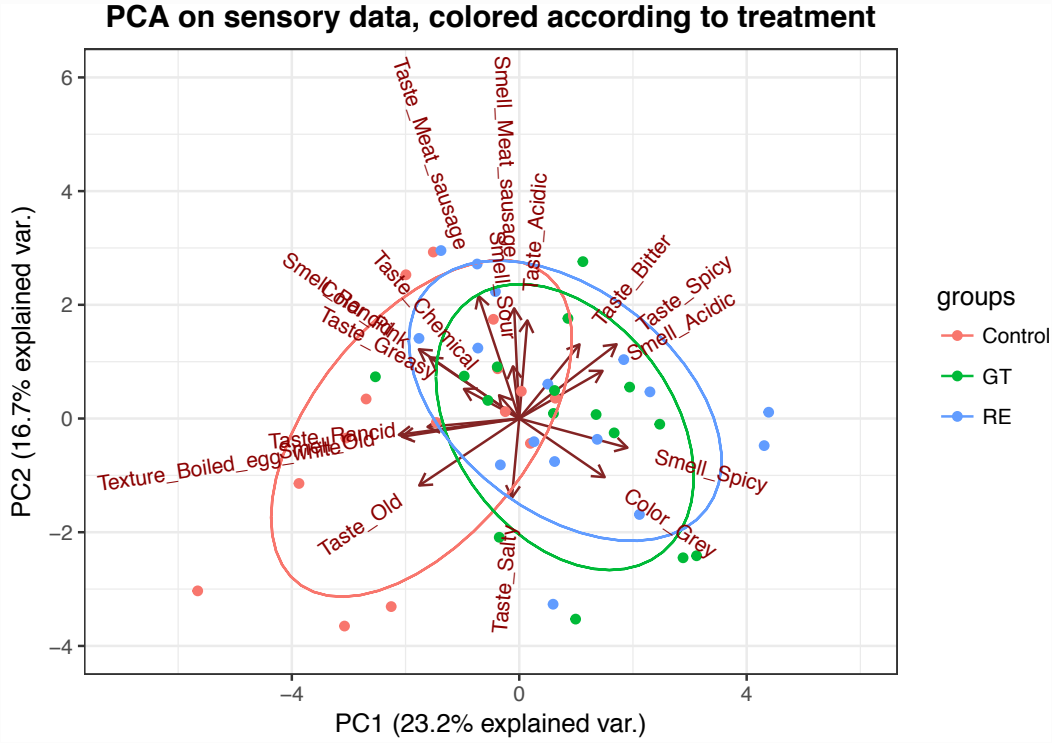
The data frame `X` consists of $p = 20$ columns, however only the last 18 are response variables, whereas the first two refers to the study design. In PCA only the response variables are used to calculate the model, whereas the design is used for e.g. coloring of the score plot.

```
PCA <- prcomp(X[,3:20],center=TRUE,scale.=TRUE)
```

Plot the model

The `ggbiplot()` function nicely plots a so-called biplot. Below a lot of features are added to the plot, but `ggbiplot(PCA)` will produce something which is similar.

```
g <- ggbiplot(PCA, obs.scale = 1, var.scale = 1,groups = X$Treatment,
              ellipse = T, circle = F) #create the biplot
g <- g + ggtitle("PCA on sensory data, colored according to treatment") # add title
g <- g+theme_bw() #black white background
g <- g+theme(plot.title=element_text(hjust=0.5,face="bold")) #center and bold title
g <- g+ylim(-4,6)+xlim(-7,6) #adjust axis limits
print(g)
```



Here you see a biplot of the PCA model on the sensory data on the sausages from both weeks. The points represent the scores for all samples, colored according to antioxidant treatment. The `ellipse=T` in the R-code draws ellipses representing the distribution of each treatment. The arrows indicate the loadings of the sensory variables. The scores can be used to evaluate the differences between samples, and the underlying reason behind the differences is interpreted through the directionality and magnitude of the loadings.

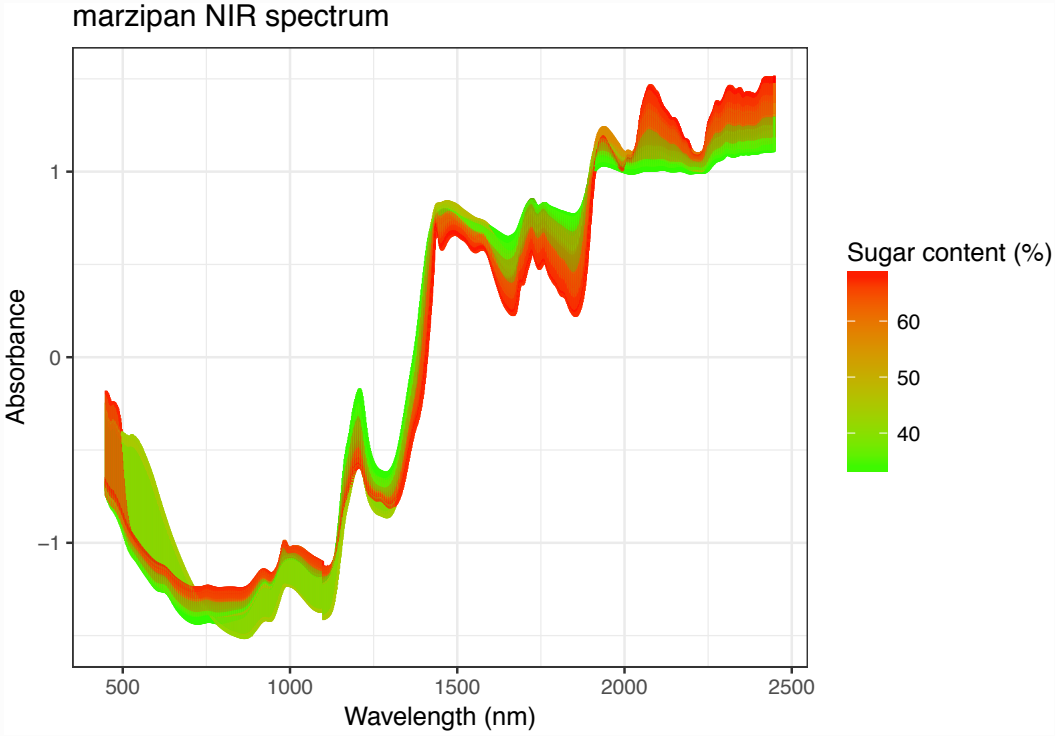
It is evident that there is a difference between the control and treated samples, since they are grouped differently, primarily in the $PC1$ direction. When compared to the loadings it seems that the control samples are characterised by *more old*, and *rancid* tastes and smells, which are sensorical attributes of lipid oxidation. Furhter the control samples have a boiled egg texture compared to the treated samples. On the other hand the treated samples are more spicy and bitter/acidic in taste and grey in color.
There are not registered any greater difference between the two green tea and rosemary extract samples.

1.10.3 Example: Near Infrared Spectroscopy of Marzipan - PCA

The following example illustrates how principal component analysis (PCA) can be used to explore your data. The dataset in this example consists of 32 measurements on marzipan bread (marzipanbrød) made from 9 different recipes. The measurements have been acquired using near infrared spectroscopy (NIR) where light is passed through a sample and the transmitted light analysed. The output measurement is a spectrum showing how much light the sample has absorbed at each wavelenght.

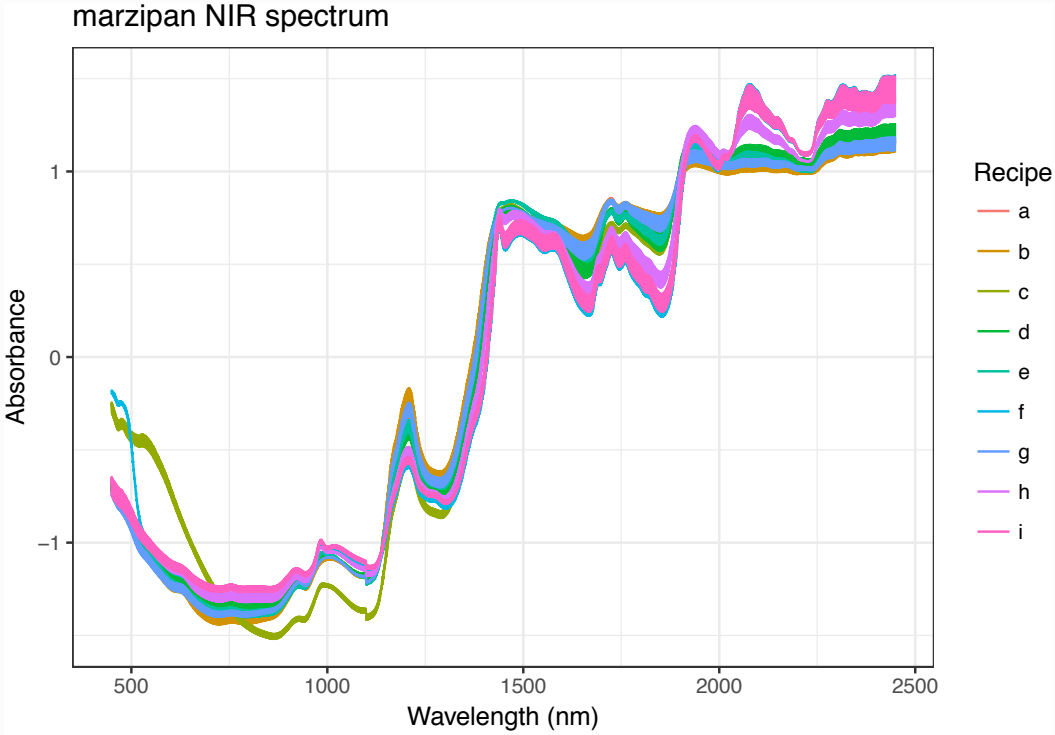
We now import and plot the data:

```r
# Loading data
load("Marzipan.Rdata")

# Loading libraries for plotting
library(ggplot2)
library(plotly)

# Plotting data according to sugar content
ggplot(data = Xm,aes(x = wavelength, y = value, colour=sugar))+
  geom_line()+theme_bw()+ylab("Absorbance")+xlab("Wavelength (nm)")+
  ggtitle("marzipan NIR spectrum") + scale_colour_gradient(low = "green", high="red")+
  labs(color='Sugar content (%)')
```



Looking at the raw spectral data we see that there is a concentration gradient in the spectra when we colour according to the sugar content. It seems that the main variation in the spectra has something to due with the sugar content.

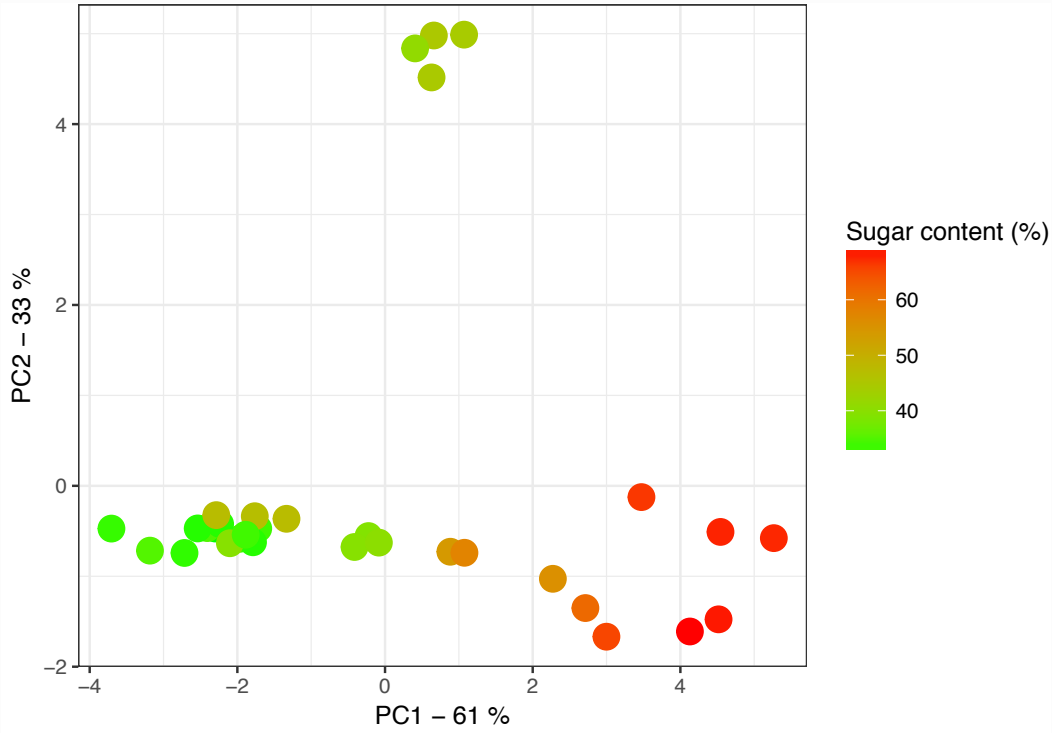
We can also plot the same data according to the recipe:



Here we see that we can distinguish some of the recipes from each other. This can be explained by the varying sugar content in the recipes. Also, if we look in the region below $1100nm$ and into the visible ($\tilde{}370 - 750nm$) we note that samples made with recipe **c** is different compared to the other samples.

We now make a PCA on the data and plot PC1 vs PC2 coloured according to sugar content:

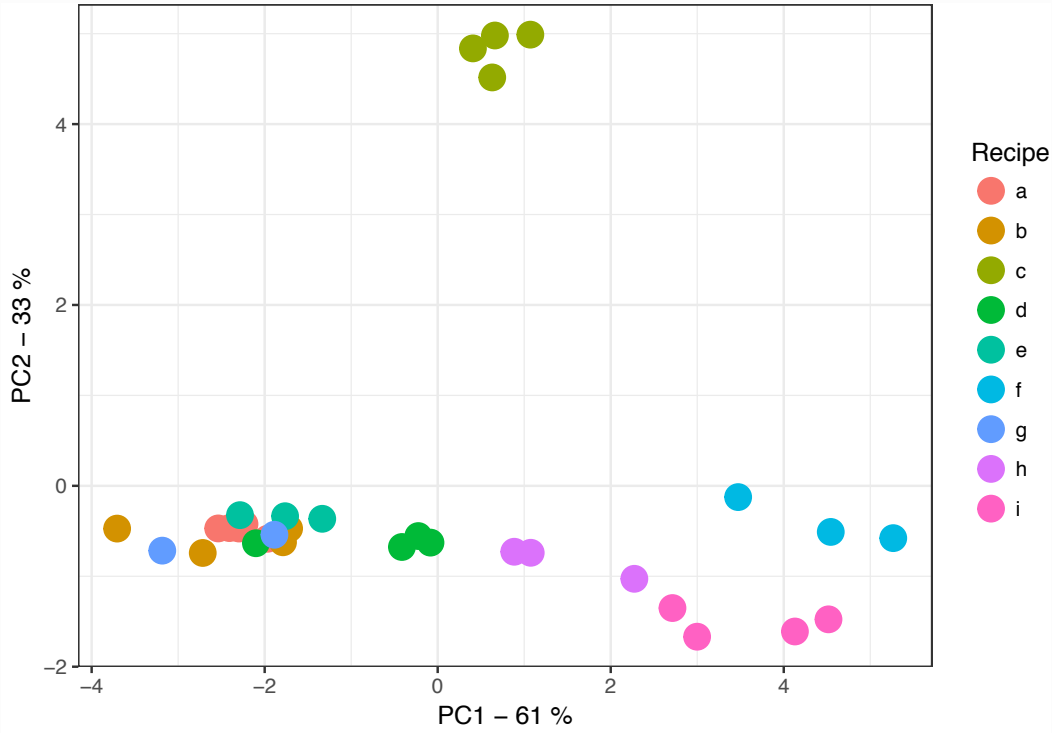
```
# Transposing the data and removing the wavelength column
Xt = t(X[,-1])

# Making PCA on mean centered Xt
marzipan = prcomp(Xt, center=TRUE, scale=FALSE)

# Extracting scores
scores = data.frame(marzipan$x, sugar = Y$sugar)

# Extracting % explained variance
varPC1 = round(summary(marzipan)$importance[2,1]*100)
varPC2 = round(summary(marzipan)$importance[2,2]*100)

# Plotting scores, PC1 vs PC2, coloured according to sugar content
ggplot(data=scores, aes(x=PC1,y=PC2))+
  geom_point(size=5, aes(colour=sugar))+
  scale_colour_gradient(low = "green",  high="red")+
  theme_bw() + labs(color='Sugar content (%)')+
  xlab(paste("PC1 -",varPC1, "%"))+ # Inserting % explained variance as label
  ylab(paste("PC2 -",varPC2,"%"))
```



We see that PC1 explains 61% of the variation in the data and that it seems to capture the variation in the sugar content. The samples are ordered from left to right in increasing concentration. Also, a group of samples are laying away from the rest when looking at PC2 which is explaining 33% of the variation in the data.

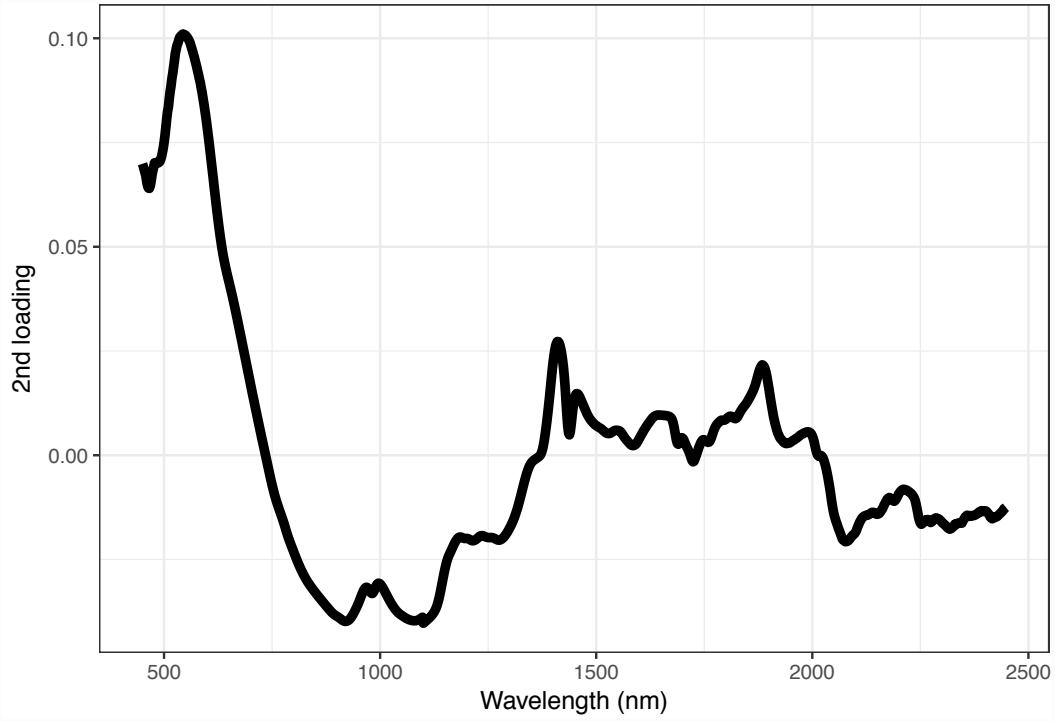
We now colour the scores according to recipe:

```
ggplot(data=scores, aes(x=PC1,y=PC2))+
  geom_point(size=5, aes(colour=substr(Y$sample,1,1)))+
  theme_bw() + labs(color='Recipe')+
  xlab(paste("PC1 -",varPC1, "%"))+
  ylab(paste("PC2 -",varPC2,"%"))
```



If we look at PC2 we see that it is the samples from recipe **c** that is laying away from the other samples. What is the reason for that? Let us look at the loadings. We start by looking at the second loading as it is dividing the samples from recipe **c** from the other samples:

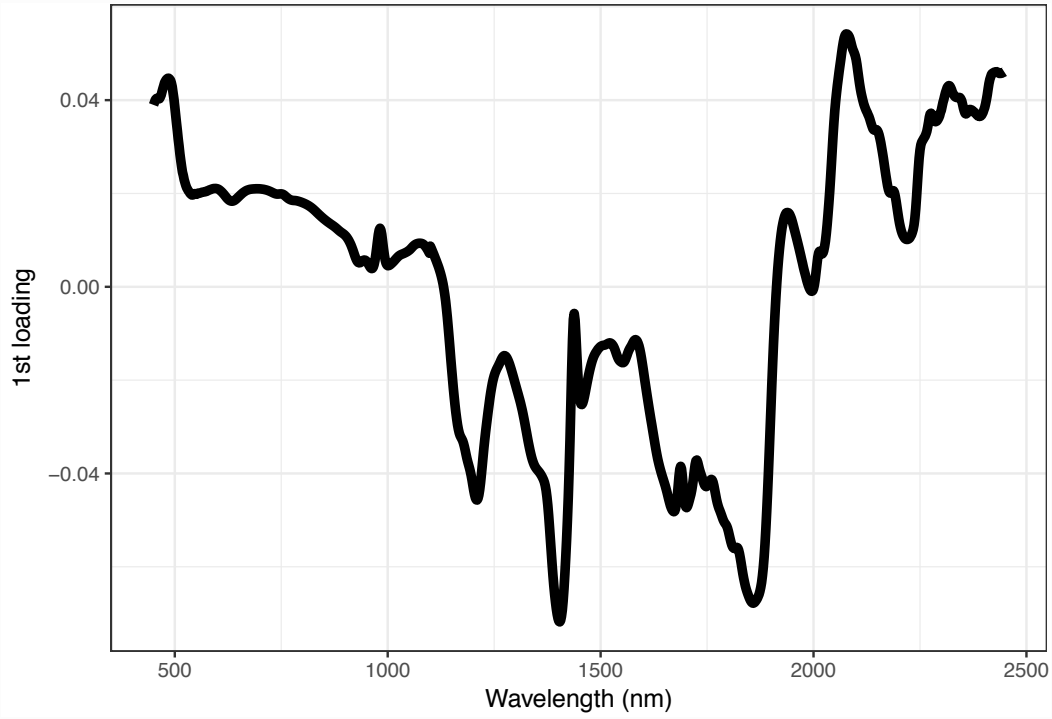
```
# Extracting loadings
loadings = as.data.frame(marzipan$rotation)
# Plotting the second loading
ggplot(data=loadings, aes(x=wl, y=PC2))+
  labs(x = "Wavelength (nm)", y ="2nd loading")+
  geom_line(linetype = "solid", size=2)+theme_bw()
```



The main contribution to PC2 is the peak around $550nm$. So the reason why the samples from recipe **c** is different from the other is related to colour. This actually makes sense as this recipe has cocoa powder added to the recipe which will influence the colour of the marzipan bread.

Lastly, we look at the first loading:

```
ggplot(data=loadings, aes(x=wl, y=PC1))+
  labs(x = "Wavelength (nm)", y ="1st loading")+
  geom_line(linetype = "solid", size=2)+theme_bw()
```



It is not straight forward to see which peaks is related to sugar. However, the peaks around 1200, 1400, 1875 and 2100nm has the highest magnitude and therefore the main reason for the sugar content gradient we see in PC1. Actually all 4 peaks is related to either the **C-H** (Carbon - Hydrogen) or **O-H** groups in sugar or **O-H** in water. You will learn more about assigning peaks to chemical information in other courses later on.

1.10.4 Reading Material

YouTube on PCA (PCA 1 - Introduction and PCA 2 - Introduction) by Rasmus Bro
(see `https://www.youtube.com/user/QualityAndTechnology`)

Chapter 2 in *Biological Data analysis and Chemometrics* by Brockhoff `http://27411.compute.dtu.dk/enote/`.

Chapter 4 (4.1 to 4.5) in *Chemometrics With R: Multivariate Data Analysis in the Natural Sciences and Life Sciences* by Ron Wehrens (2012). Springer, Heidelberg. Available in Absalon.

1.10.5 Exercises

Exercise 1.6 McDonalds data

The purpose of this exercise is to get familiar with PCA on a small intuitive dataset. The data - McDonaldsScaled.xlsx - constitutes of different fast food products and their nutritional content.

1. Read in the data using an appropriate package / function (e.g. `read.xls()` from the package `gdata` or `read.xlsx()` from the package `xlsx` works depending on the installation of dependencies), and set up the data with row names etc.

```
McD <- read.xls("McDonaldsScaled.xlsx")
rownames(McD)<-McD[,1]
McD <- McD[,2:6]
head(McD)
```

2. Make some initial descriptive plots for the five response variables, that indicate the distribution (center and spread).
3. Make some bi-variate scatter plots examining the relation between different variables, and comment on whether this relation is obvious and further, which types of samples are responsible for the relation. You can use the `ggplot2` tool `qplot()` and further add on tendency line and labels

```
library(ggplot2)
qplot(data=McD,x=Energy,y=Protein)

qplot(data=McD,x=Energy,y=Fat)
  + stat_smooth(method = 'lm',se=F)
  + geom_text(label=rownames(McD))
```

4. Now, make a PCA on the data. What does the two options `center = ...` and `scale. = ...` refer to?

```
McD.pca <- prcomp(McD,center = TRUE,scale. = TRUE)
```

5. If $X_1, X_2, .., X_{19}$ is the protein variable in the dataset (I.e. X_1 are the protein content of sample 1 and so forth), how do you calculate the centered and scaled representation of data, which could be called: $X_1^{auto}, X_2^{auto}, .., X_{19}^{auto}$)?
6. Plot the PCA results and comment on them. There are several ways of doing this. The first described here, is to zack out the parameters (scores and loadings) and then use the ggplot2 functionality to plot those. Try to comprehend what is actually produced from the list of functions listed below.

```
# Zack out the individual parameters (scores and
    loadings)
scores <- as.data.frame(McD.pca$x)
loads <- as.data.frame(McD.pca$rotation)

# Score plot
qplot(PC1,PC2,data=scores, label=rownames(McD),geom=c
    ('text','point'))
# Loading plot
qplot(PC1,PC2,data=loads, label = rownames(loads),
    geom=c('text','point'))
```

7. Alternatively you can utilize a package `ggbiplot` for making nice plots. However, this is a tool under development, so the installation is slightly different (see the code below)

```
library(devtools)
install_github("ggbiplot", "vqv")
library(ggbiplot)

g <- ggbiplot(McD.pca, obs.scale = 1, var.scale = 1)
g <- g + geom_text(label=rownames(McD))
g <- g + theme_bw()
print(g)
```

8. Make a vector that indicates the different fast food types (Burger, Drinks, etc.). You can either do this by extending the excel file with a column, or do it in R (see this below). Infer this class information on the plot as color (or maker shape or size). Be aware, that you simply just add this information onto the existing plot.

```
cl <- c(rep(c('Burger'),times=9),rep(c('Drinks'),
    times=3),rep(c('Icecream'),times=3),rep(c('Other'
    ),times=2),rep(c('Salad'),times=2))

g <- g +geom_text(aes(colour=factor(cl), label=
    rownames(McD)))
print(g)
```

9. Try to modify the PCA by removing scaling and/or centering. What happens to the plots of the results? What do you think is going on?

Exercise 1.7 Analysis of Coffee Serving Temperature - PCA

In the dataset *Results Panel.xlsx* the sensorical results from a panel of eight judges, evaluating coffee served at six different temperatures each four times are listed. In this exercise, we are going to first average over judge and temperature followed by PCA to evaluate sensorical descriptor similarity as well as the effect of serving temperature on the perception of coffee.

1. After import of data, and initial sanity check, calculate the average response across the four replicates. Use the `aggreggate()` function with `FUN="mean"` to make a dataset with the average response for the six different temperatures for each judge. HINT: The number of samples should be reduced by a factor of 4.

2. Use this data as input for construction of a PCA model. Which variables do you think should be included?

3. Make a biplot of this PCA model and interpret it.

   (a) Which descriptors go together and which are oppositely correlated?

   (b) Are there, from this analysis, a clear difference between the different serving temperatures?

   (c) What do you think blurs the picture?

The code below can be used for inspiration. Be aware, that we need to set a series of dependencies in order for this to work.

```
CoffeeAG <- aggregate(Coffee,by=list(Coffee$Assessor,
    Coffee$Sample),FUN="mean")
# rename some variables
CoffeeAG <- rename.vars(CoffeeAG,c('Group.1','Group.2'),c(
    'Judge','Temp'))
```

```
CoffeePCA <- prcomp(CoffeeAG[,yourlistofvariables],center
    = TRUE,scale. = TRUE)
# Make a meaningfull color palette
colPalette <- c('#0033FF','#0099FF','#00EEFF','#FFCCCC','#
    FF9999','#FF0000')
# Plot the model
g <- ggbiplot(CoffeePCA, obs.scale = 1, var.scale = 1,
    groups = CoffeeAG$Temp,ellipse = TRUE, circle = TRUE)
g <- g + theme(legend.direction = 'horizontal', legend.
    position = 'top')
g <- g + scale_color_manual(values= colPalette,name='
    Temperature')
print(g)
```

Check out the ggbiplot syntax (by `>?ggbiplot`). by adding stuff to the plot, it is modified to look exactly like you want it. Here we change legend appearance (for inferring temperature) and color scheme for the scores matching temperature.

In order to also remove variation due to differences between judges, the dataset is compressed such that the rows reflect the average for each temperature (across judges and replicates). Then this dataset is used for constructing a PCA model and visualized.

4. use the `aggreggate()` function to average across judges and replicates. HINT: modify the `by=list()` statement.
5. Make a PCA model on this dataset and visualise it.

The code below shows how to plot the model. Can you figure out why it is different from the previous biplot call?

```
g <- ggbiplot(CoffeePCA, obs.scale = 1, var.scale = 1,
    groups = CoffeeAG2$Temp,ellipse = FALSE, circle = TRUE
    )
g <- g + theme(legend.direction = 'horizontal', legend.
    position = 'top')
g <- g + scale_color_manual(values= colPalette,name='
    Temperature')
g <- g + theme_bw()
print(g)
```

There is something wrong with this plot. 1) Some of the labels are masked and 2) the points are way to small. Use the function `xlim(c(low,high))` and `geom_point(size=I(5),aes(color=CoffeeAG2$Temp))` and add them to the plot to fix these problems.

```
g <- g + geom_point(size=I(5),aes(color=CoffeeAG2$Temp))
g <- g + xlim(c(-4,4)) + ylim(c(-4,4))
```

Exercise 1.8 Wine Aromas

This exercise will take you through plotting, descriptive stats and PCA. Wine based on the same grape variety (Cabernet Sauvignon) from four different countries (Argentina, Australia, Chile and South Africa) were analyzed for aroma compound composition with GC-MS (gas chromatography coupled with mass spectrometry). The dataset can be found in the file "Wine.xlsx", and it will form the basis for working with basic descriptive statistics, plots and PCA.

Descriptive statistics

1. Start by importing the dataset "Wine.xlsx" to `R` and try to get an overview of it (Hint: use the `summary()` function in `R` and/or have a look on the raw data in the Excel file).

   (a) How many wines were analyzed from each country?

   (b) How many variables are there in the dataset, and how many constitutes the aroma profile?

For the descriptive statistics, only two of the aroma compounds are selected. Choose two on your own or make the calculations for the aroma compounds benzaldehyde (almond like aroma) and 3-Methylbutyl acetate (sweet fruit/banana like aroma).

2. Calculate mean, variance, standard deviation, median and inner quartile range for the selected aroma compounds from each of the four different countries. (Hint: it can be helpful to create a separate dataset for each country, which can be done with the R function subset.)

Plots

3. Make a boxplot, a jitterplot and a combination of the two with all 4 countries in one plot. Use the R-commands from the notes as inspiration.

4. What do you see? Discuss pros and cons of the different plots.

5. Adjust the layout of your favorite plots (e.g. color, background, title etc.). Think about how the data is presented in the best way. Actually, it can be rather beneficial to specify a generic theme including title and label font size, background color of the plot etc, which then can be added to each plot produced... Once and for all.

PCA

Working with a dataset with many variables, PCA provides a very nice tool to give an overview of the dataset. First we define the data we want to include in the analysis. With the functions scale, we specify which columns to include (i.e. we are only interested in analyzing the aroma compounds).

6. Use the PCA function from the package `ChemometricswithR` to calculate a PCA model on scaled Aroma data. What does the function `scale()` do to data?

7. Make score plot and loadingsplot.

```
library('ChemometricsWithR')
wine.pca <- PCA(scale(wine[,4:58])) #Only columns
    4-58 included#
par(mfrow = c(1,2))
```

```
scoreplot(wine.pca,col=wine$class,pch=wine$class)
loadingplot(wine.pca,show.names = T) #Score and
    loadings plot created. Use zoom to make plots
    bigger#
dev.off() #shuts down scores and lodings plot#
```

(Tip: Use the "zoom" function in the Plots window to make the plots bigger and easier to interpret.)

8. What do you see in the score and the loadings plot?
9. Can you see a grouping of the data? If so, how are the groups different?

2. Week 2

This week is going to focus on three very central subjects, namely correlation, the normal distribution and in relation to this the Central Limit Theorem.

2.1 Hand-in assignment

The exercise 2.2 *Covariance and Correlation - by hand* is to be handed in (through Absalon or as hard-copy Wednesday night). This exercise do not need any coding (except for maybe pocket calculations), so no coding (in the assignment) this week.

2.2 Exercises

For Monday work through exercise 2.1 and 2.3, and for Wednesday work through 2.5, 2.7 and 2.4.

2.3 Correlation

The learning objectives for this theme is to:

- Be able to compute the correlation between two variables.
- Know the relation between covariance and correlation.
- Comprehend that the correlation structure between variables is a key component in most pattern recognition techniques, and is a driver for the PCA analysis.
- Be able to visualize the correlation between two variables, and use this for judging whether the correlation coefficient is a valid representation for this relation.
- Visualize pairwise correlations for multivariate data.
- Know the relation between the PCA loadingplot and bi-variate scatterplot.

2.3.1 Correlation and Covariance - *in short*

A covariance or correlation is a scalar measure for the association between two (response-) variables.

Covariance bewteen two variables $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$ is defined as:

$$cov_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{2.1}$$

Covariance depends on the scale of data (X and Y), and as such is hard to interpret.

The correlation is however a scale in-variante version.

$$corr_{X,Y} = \frac{cov_{X,Y}}{s_X \cdot s_Y} \tag{2.2}$$

where s_X and s_Y are the standard deviation for X and Y respectively (see 2.4.1 for details).

Dividing the covariance by the individual standard deviations put the correlation coefficient in the range between -1 and 1:

$$-1 \leq corr_{X,Y} \leq 1 \tag{2.3}$$

A correlation (and covariance) close to zero indicates that there are no association between the two variables.

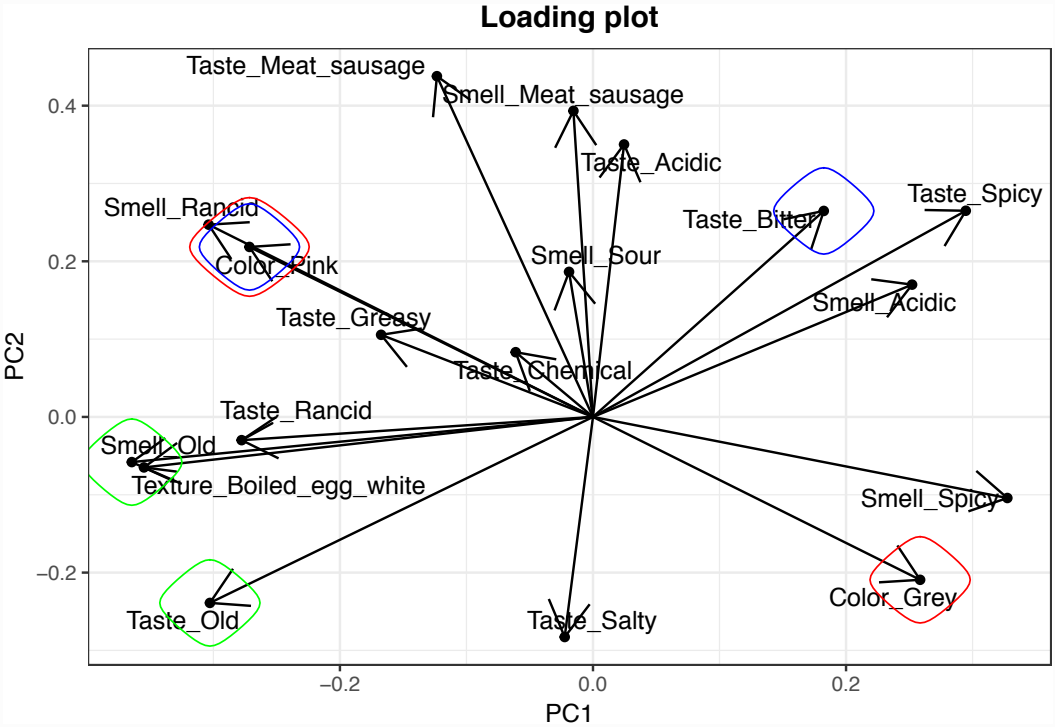
2.3.2 Example: Natural Phenolic Antioxidants for Meat Preservation - Correlation

This example continues where we left of in example 1.10.2 with the sensory data on the meat sausages treated with green tea (GT) and rosemary extract (RE) or control.

Load the data

Remember to set your directory

```
load('meat_data.rdata')
```



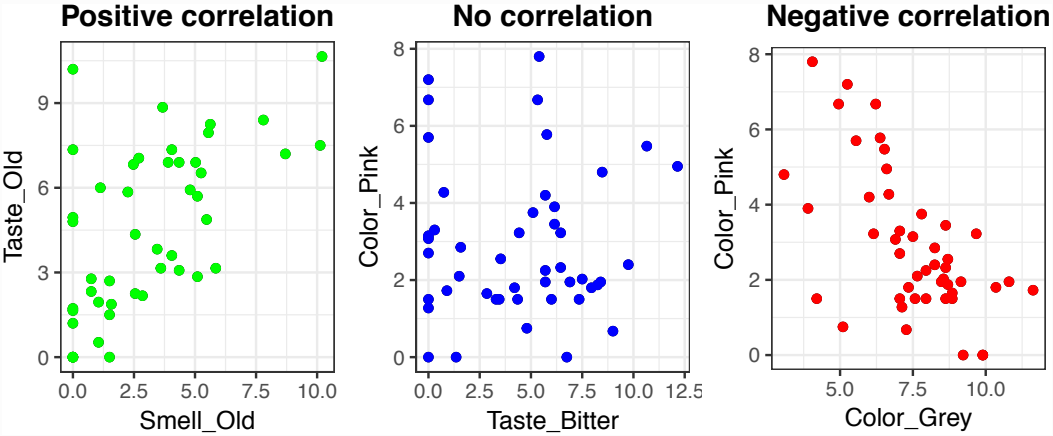
Above, the loading plot corresponding to the loadings found in the PCA model calculated in the previous example is shown. The colors encircling some of the variables indicate positive, negative and non-correlated loadings. In a PCA loading plot, loadings with *opposite directions* are *negatively correlated*, while loadings pointing in the *same direction* are *possitively correlated*. Loadings which are *orthogonal* (90 degree angle) with respect to each other, are *not correlated*.

To check if the correlation holds for the raw data, scatter plots are made for each of the encircled three examples.

(Note that the interpretation of these correlations are only valid with respect to the variation described in the model)

Create scatter plots

```
# define plot for example of positive correlation:
p<- qplot(data=X,x=Smell_Old,y=Taste_Old)
p<-p+geom_point(color="green") #color the points
p<-p+ggtitle("Positive correlation") # add title
p<-p+theme_bw()+
  theme(plot.title=element_text(hjust=0.5,face="bold"))
# Black white background and bold centered title

#no correlation:
q<- qplot(data=X,x=Taste_Bitter,y=Color_Pink)+
  geom_point(color="blue")+
  ggtitle("No correlation")+
  theme_bw()+
  theme(plot.title=element_text(hjust=0.5,face="bold"))

#negative correlation:
n<- qplot(data=X,x=Color_Grey,y=Color_Pink)+
  geom_point(color="red")+
  ggtitle("Negative correlation")+
  theme_bw()+
  theme(plot.title=element_text(hjust=0.5,face="bold"))

#All plots in one:
grid.arrange(p,q,n,ncol=3, heights=unit (7,"cm"))
```



The correlation values are calculated using the cor() command

```
#positive
cor(X$Smell_Old,X$Taste_Old)
```

```
## [1] 0.595475
```

```
# no correlation
cor(X$Taste_Bitter,X$Color_Pink)
```

```
## [1] 0.007446373
```

```
#negative
cor(X$Color_Grey,X$Color_Pink)
```

```
## [1] -0.604673
```

Conclusions

Here, it can be seen that the pink and grey color attributes, which are oppositely directed in the PCA loadings, display a moderate negative correlation in the raw data. The pink color and bitter taste, which are orthogonal to each other in the loadings, are not correlated at all. The old smell and old taste with similar directionality in the PCA are moderately postively correlated. With regards to food chemistry, do you think these conclusions make sense?

2.3.3 Reading Material

A youtube video briefly introducing the notion of correlations `https://www.youtube.com/watch?v=Ypgo4qUBt5o`

Chapter 1.4.3 and 2.5 of *Introduction to Statistics* by Brockhoff `https://02402.compute.dtu.dk/enotes/book-IntroStatistics`

Chapter 2 in *Biological Data analysis and Chemometrics* by Brockhoff `http://27411.compute.dtu.dk/enote/`.

2.3.4 Exercises

Exercise 2.1 Correlation between aroma compounds

Wine from four different countries (Argentina, Australia, Chile and South Africa) were analyzed for aroma compound composition with GC-MS (gas chromatography coupled with mass spectrometry). The dataset can be found in the file "Wine.xlsx", and it will form the basis for working with correlations and PCA.

Correlation

1. What does a correlation coefficient of -1, 0 or +1 tell you?
2. In the table below the amount of Nonanal and Ethyl.2.methyl.propanoate in the six Argentine wines are listed. Fill out the blank spaces in the table and calculate the correlation coefficient (r). (Help can be found in example 1.19 in Introduction to Statistics by Brockhoff http://introstat.compute.dtu.dk/enote/)

Wine number ($_i$)	1	2	3	4	5	6
Nonanal (X_i)	0.003	0.003	0.005	0.006	0.008	0.005
Ethyl.2.methyl.propanoate (Y_i)	0.106	0.165	0.150	0.155	0.149	0.141
$X_i - \bar{X}$						
$Y_i - \bar{Y}$						
$(X_i - \bar{X})(Y_i - \bar{Y})$						

$$s_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \tag{2.4}$$

$$r_{XY} = \frac{s_{XY}}{s_X \cdot s_Y} \tag{2.5}$$

3. Make a plot with Nonanal vs. Ethyl.2.methyl.propanoate and calculate the correlation coefficient in R (use the commando: `cor()` in R). Include only wines from Argentina.

4. Calculate the correlation coefficient for a, b, c and d, where you include wine data from all the countries

   (a) Diethyl.succinate and Ethyl.lactate

   (b) Ethyl.acetate and Ethanol

   (c) 2.Butanol and Ethyl.hexanoate

   (d) Benzaldehyde and Hexyl.acetate

PCA

5. Make a PCA including wines from all the countries (similar to the one from week 1 on the same dataset).
   Which variables are responsible for the grouping of the countries?

6. Compare the calculated correlation coefficients in question 4 with the loadings plot produced in question 5. How dose the position of the variables in the loading plot makes the correlation coefficients negative, positive, close to ± 1 or 0?

Exercise 2.2 Covariance and Correlation - by hand

Certain types of characteristics "go together", for instance will a variable that measures the *chocolate smell* of some food type be related to a variable measuring the *chocolate taste* of the same food type. We talk about these two types of information being correlated.

Below is 48 corresponding measures of the sensorical attributes *sour* and *roasted* for coffees at different serving temperatures and different judges (each based on the average of four measurements). Calculate the correlation between these two variables based on the metrics listed below.

	Sour	Roasted
1	1.70	1.06
2	1.93	1.10
3	3.59	3.87
4	2.30	1.86
5	1.32	1.72
6	2.67	1.24
7	2.21	1.50
8	2.70	1.65
9	0.54	0.96
10	2.02	0.82
11	3.78	3.22
12	2.61	1.59
13	1.81	1.56
14	2.55	1.80
15	2.06	0.87
16	1.92	1.66
17	1.92	1.70
18	2.21	1.42
19	2.04	1.56
20	1.31	2.03
21	1.57	1.18
22	1.86	2.06
23	2.05	1.75
24	3.40	2.32
25	0.26	0.75
26	1.34	0.26
27	3.81	1.55
28	1.72	3.18
29	1.40	0.46
30	1.49	1.14
31	0.77	1.43
32	1.46	0.54
33	0.85	0.44
34	1.69	0.37
35	2.97	1.83
36	1.44	1.68
37	1.41	1.31
38	1.88	2.32
39	2.62	0.69
40	1.24	1.85
41	0.37	0.23
42	0.77	0.68
43	4.12	3.40
44	1.39	1.51
45	1.30	1.89
46	0.26	1.93
47	1.12	1.05
48	3.84	5.41
$\sum$	91.59	76.42
$\hat{\sigma}^2$	0.90	0.94
$\sum XY$		172.73

1. Calculate mean and standard deviation for the two variables using the stats listed below data. That

is WITHOUT importing data into the computer!

2. Calculate $\sum(X-\bar{X})(Y-\bar{Y})$ from $\sum XY$, $\sum X$ and $\sum Y$ (where X is *Sour* and Y is *Roasted*). HINT: you need to multiply out the product of the two parenthesis and reduce the resulting part using the relation between $\bar{X} = \sum(X)/n$.
3. Calculate the covariance and the correlation between the two variables.
4. What happens with the covariance and correlation if we multiply the *Sour* ratings with 2 or -3.?
5. What happens with the covariance and correlation if we add 10 or -1234 to the *Roasted* ratings?
6. As an upcoming coffee expert, why do you think these two attributes are correlated?

Exercise 2.3 Correlation and PCA

In this exercise the sensorical data from ranking of coffee served at different temperatures are used. The aim is to see how correlations is the vehicle for PCA analysis. The data is named `Results Panel.xlsx`.

1. Import the data and remove the replicate effect by averaging over these.
2. Make a scatter plot of the attribute *Sour* versus *Roasted*. Comment on what you see in terms of relation between these two variables.
3. Now make a comprehensive figure where all the sensorical attributes (there are 8) are plotted against each other.
4. Calculate all the pairvise correlations between the variables. How does this correspond with the figure?
5. Make a PCA model on this dataset (Same as in exercise 1.5)
6. Comment on the (dis-)similarity between the correlation matrix, the multiple pairwise scatterplot and the PCA model.

Below are some code which might be useful for this purpose (You ned to install or add dependencies via `install.packages(GGally)` or `library(GGally)` in order for the functions to be recognized by R). In the pairwise scatterplot a straight line is added by `+ geom_abline()`, try also to add a smooth curve by `+ geom_smooth()`. What are the difference between those two representation of similarity between the variables?

```
qplot(Roasted,Sour,data=CoffeeAG,geom=c('point'),color=
    Temp) +geom_abline()+theme_bw()
g <- ggpairs(CoffeeAG,columns = 6:13,title='Matrix Plot')
print(g)
```

Exercise 2.4 Olive oil adulteration

Quick detection of adulteration of oils is of growing importance, since high quality oils, such as olive oil, are becoming increasingly popular and expensive, increasing the incentive for adulterating with cheaper oils. Spectroscopic techniques are the preferred measurement choice because they are quick, often non-destructive and in many cases highly selective for oil characterization.

The purpose of this exercise is to introduce you to how multivariate techniques, such as PCA, can be applied on spectral datasets. They are in fact, very useful on datasets such as these, because of the very high number of variables that can be included in the modelling.

The samples in this dataset are mixtures of olive oil and thistle oil. Olive oil and thistle oil are almost exclusively made up of triglycerides, consisting of a glycerol backbone with three fatty acid chains attached. An example of a triglyceride is shown in the top right of figure 2.1. Fatty acids are characterized by the amount of unsaturation. Olive oil consists mostly of monounsaturated fatty acids, while thistle oil is largely comprised of polyunsaturated fatty acids (ie: they have more double bonds).

Additionally a few of the samples were spiked with a free trans fatty acid. The structure of the added trans fatty acid is shown in the top left of figure 2.1.

The samples were measured with infrared (IR) spectroscopy. Some of the relevant peaks for oil characterization are shown in the raw spectra in figure 2.1. The dataset consists of 30 oil samples and 1794 variables. The first 1790 variables are the absorbance at 1790 wavelengths. The last 4 columns in the dataset describe the concentration of oliveoil, thistle oil and transfat, and lastly the sample ID.

Since it is a bit more difficult to work with spectral data in R, we have decided to be merciful and give you the data directly as an `.RData` file, which you simply get into R by `load('OliveOilAdult.RData')` (be sure to be in the correct directory, or add the path to the load command).

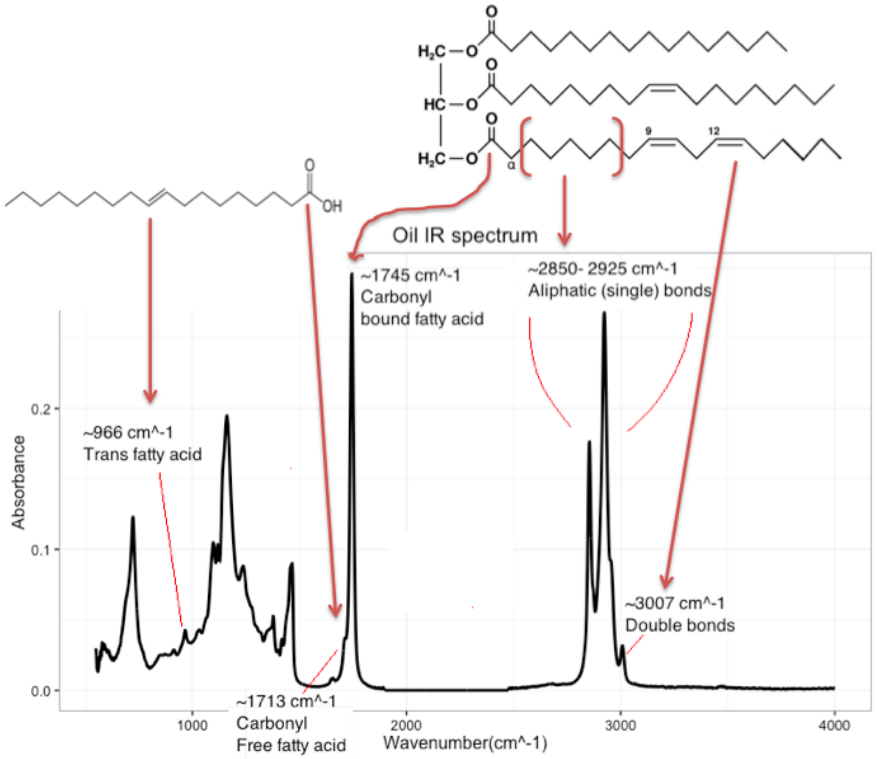


Figure 2.1: Top right) triglyceride. Left part: glycerol backbone. Right part from top to bottom: unsaturated, monounsaturated and polyunsaturated fatty acid. Top left) Trans fatty acid. Bottom) raw IR spectrum of mixed oil

1. Load the Oliveoil dataset in R. To be able to plot the raw spectral data, run the following R code. Use the `dim()` command to understand the difference between input and output to `melt()` on how the data is organised in the two formats.

```
# Before we can plot:
library(reshape)
ForRaw <- melt(Oliveoil,id.vars = c('sample_id','
    oliveoil',"thistleoil","transfat"))
```

```
# Define the correct x-axis with the wavelength
    vector:
ForRaw$wl <- sort(rep(wl,nrow(Oliveoil)))
```

2. Now make a plot of the raw data. Use the R-code below as a base, and customize it to your liking. Use the handy plotly package to plot an interactive plot that we can zoom in.

```
#make a simple plot:
library(ggplot2)
raw<-ggplot(data = ForRaw,aes(x = wl,y = value, group
     = sample_id))+geom_line()
#Plot "interactive plot" with ggplotly:
library(plotly)
ggplotly(raw)
```

3. Inspect the raw data plot.

   First off, how does the raw data look? Are there any outliers? in ggplotly you can identify specific samples by hovering the curser over them.

   Now, try to color according to the different types of oil, one at a time. Then, zoom in on some of the peaks highlighted in figure 2.1. Can you observe a correlation between these peaks and the oil concentration?

4. Now that you have a feel for the raw data, try to estimate the scores and loadings for a PCA model. Why do we only center, and not autoscale the spectral data?

```
#Define the PCA:
Oliveoil.pca <- prcomp(Oliveoil[1:1790],center = TRUE
    ,scale=FALSE)
scores <- as.data.frame(Oliveoil.pca$x)
loads <- as.data.frame(Oliveoil.pca$rotation)
```

5. Using `ggplot()` and `ggplotly()`, make a plot of the PC1 vs PC2 scores, and color according to *oliveoil* content.

6. Make two separate loading plots with first PC1 loads vs the wavelength (wl) vector and then PC2 loads vs. the wavelength. It is very important that the loadings are plotted separately, and not against each other when working with spectral data. Can you see why?

7. Inspect the scoreplot. Are there any outlying samples? If so, identify them. Did you notice this outlier in the raw spectra plot?
   Remove the outlier by using the code below and inserting the correct sample id number.

```
#remove the spectral outlier:
Oliveoil<-Oliveoil[-c("insert outlier ID number here"
    ),]
```

8. Now rerun your whole script and see how the removal of the outlier has changed the score and loading plots.

9. Using the score plot, try to colour them by different oil types and elucidate what information about the samples is being described by the first and second PCA component, respectively.

10. Inspect the loading plots and compare them to the information in figure 2.1. Can you re-find some of the important peaks in the loadings? And how does this relate to the findings in the scoreplot?.

    (HINT: remember that the dataset is mean centred, so the peaks will be positive or negative. Negative scores have a high absorbance in the negative loading "peaks", and positive scores have a high absorbance in the positive loading "peaks").

2.4 The Normal Distribution

The learning objectives for this theme is to introduce the predominant distribution within biological data; The normal distribution. In relation to this, be able to characterize the distribution by a set of parameters, and be able to calculate the precision of these parameters (give confidence intervals) based on data. Know which graphics that are useful for visualizing normally distributed data (histogram, boxplot, qq-plot), and be able to use this to judge whether normality is fulfilled. Last, for a given distribution to be able to calculate the probability of a given observation.

Formula: If X is normaly distributed with mean μ and variance σ^2, then the probability density function for a given draw from that distribution x is given by

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{2.6}$$

Further, the cumulative density function can be found by integration of this formula

$$P(X \leq x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\inf}^{x} e^{-(x-\mu)^2/2\sigma^2}\, dx \tag{2.7}$$

The short notation for this distribution is

$$X \sim \mathcal{N}(\mu, \sigma^2) \tag{2.8}$$

2.4.1 Estimation of μ and σ

More often than not, the distribution is unknown. That is the values of μ and σ are unknown and must then be estimated from data.

μ characterizes the center of the distribution, and is naturally estimated by the mean-value of the data-points ($\bar{X}$). σ reflects the spread around the mean, and is in a similar fashion estimated by the standard deviation ($\hat{\sigma}$ or s).

$$\hat{\mu} = \bar{X} = \sum_{i=1}^{n} X_i/n = (X_1 + X_2 + X_3 + \ldots + X_n)/n$$

$$\hat{\sigma} = s = \sum_{i=1}^{n} (X_i - \bar{X})/(n-1) = ((X_1 - \bar{(}X) + (X_2 - \bar{(}X) + (X_3 - \bar{(}X) + \ldots + (X_n - \bar{(}X))/(n-1)$$

2.4.2 Example: Effect of Caffeine on Activity

Caffeine is a central neurvous system stimulant, which can have several positive- and negative effects. In this study, the level of activity is up for examination, and for this purpose a model-system; the runing activity of mice in a wheel reflected as the number of rounds pr minutes recorded over 7 minutes. Here a total of 249 mice from two species; one breed to be high in running performance, and one control, where given either Water, Gatorade or Red Bull followed by measuring their voluentary wheel run activiy. Of interest is the average actvity within each mouse type and for each caffeine type and how large the spread is.

The data is listed in a table with 249 rows (here, the first five samples are shown):

```
kable(head(X),digits = 1)
```

MouseType	gender	Caffeine	RPM7
Control runner	Male	Red Bull	10.5
High runner	Female	Gatorade	24.3
High runner	Female	Gatorade	19.7
Control runner	Male	Red Bull	13.5
Control runner	Male	Red Bull	15.3
High runner	Male	Gatorade	30.8

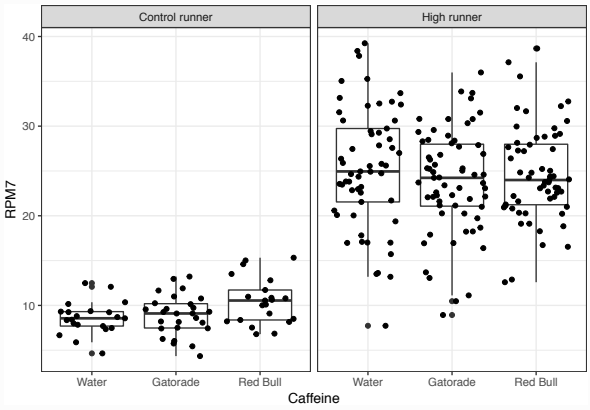
To get the experimental design, the `table()` function nicely summarize the numbers within each group:

```
kable(table(X$gender:X$MouseType,X$Caffeine))
```

	Water	Gatorade	Red Bull
Male:Control runner	10	14	12
Male:High runner	29	31	29
Female:Control runner	11	13	9
Female:High runner	28	32	31

Before doing anythingelse, the data is plotted to visualize the response as function of the design.

```
library(ggplot2)
qplot(data = X, Caffeine, RPM7,geom = c('boxplot','jitter')) +
  facet_wrap(~MouseType) +
  theme_bw()
```



As an example, the mean and standard deviation is calculated for females of the unselected breed given water.

First, the information is extracted from the dataset:

```
x <- X[X$MouseType=='Control runner' & X$gender=='Female' & X$Caffeine=='Water','RPM7']
x
```

```
##  [1]  8.4043 10.3545  9.3582  8.5620  9.3067  6.6723  4.6467  7.3518
##  [9]  7.8409 10.1589 12.4869
```

Using `x` as input to `mean()` and `sd()`, the mean and standard deviation is calculated:

```
c(mean(x),sd(x))
```

```
## [1] 8.649 2.078
```

So, an average of

$$\bar{X} = \sum_{i=1}^{n} X_i/n = 8.6$$

with a spread of 2.1 for this particular group:

$$\hat{\sigma} = \sum_{i=1}^{n} (X_i - \bar{X})/(n-1) = 2.1$$

An efficient way to do this for all combinations is to use the `aggregate()` function and combine in a table:

```
n <- aggregate(X$RPM7, by = list(X$MouseType,X$gender,X$Caffeine),length)
mn <- aggregate(X$RPM7, by = list(X$MouseType,X$gender,X$Caffeine),mean)
s <- aggregate(X$RPM7, by = list(X$MouseType,X$gender,X$Caffeine),sd)
descripStat <- cbind(n,mn[,'x'],s[,'x'])
colnames(descripStat) <- c('MouseType','Gender','Caffeine','N','mean','sd')
kable(descripStat,digits = 2)
```

MouseType	Gender	Caffeine	N	mean	sd
Control runner	Male	Water	10	8.55	1.59
High runner	Male	Water	29	26.03	6.35
Control runner	Female	Water	11	8.65	2.08
High runner	Female	Water	28	24.59	7.25
Control runner	Male	Gatorade	14	8.55	2.09
High runner	Male	Gatorade	31	25.04	5.52
Control runner	Female	Gatorade	13	9.31	2.42
High runner	Female	Gatorade	32	22.65	5.86
Control runner	Male	Red Bull	12	10.73	2.95
High runner	Male	Red Bull	29	24.81	3.65
Control runner	Female	Red Bull	9	10.18	2.14
High runner	Female	Red Bull	31	24.50	6.50

2.4.3 Example: Effect of Caffeine on Activity - Probability

Example 2.4.2 estimates the activity distribution of mice based on experimental data. Under the assumption that these metrics, the mean and the standard deviation, perfectly describe the distribution of the data, and further that this distribution is the normal distribution, we wish to calculate the probability of a given female mice, from normal breed, assigned water to perform **higher than 10 RPM7**.

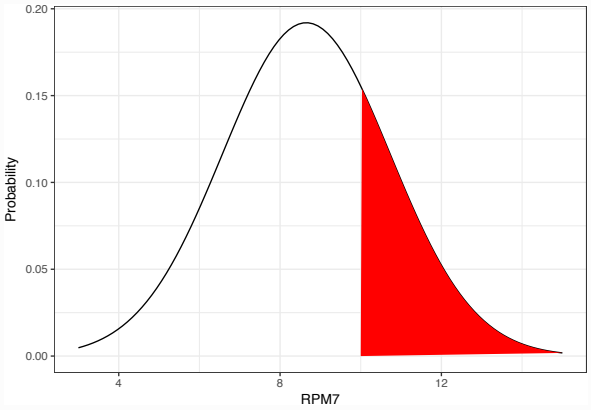
That is:

$$\begin{aligned} P(X \geq 10) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{10}^{\infty} e^{-\frac{(z-\mu)^2}{2\sigma}} dz \\ &= 1 - P(X < 10) \\ &= 1 - \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{10} e^{-\frac{(z-\mu)^2}{2\sigma}} dz \end{aligned}$$

where X is assummed normally distributed with mean μ and standard deviation σ.
That is: $X \sim \mathcal{N}(\mu, \sigma^2)$.

A small illustration below shows the area of interest

```
xx <- seq(3,15,length = 100)
yy <- dnorm(xx,mean(x),sd(x))
D <- data.frame(xx,yy)
ggplot(data = D,aes(xx,yy)) +
  geom_line() + theme_bw() +
  geom_polygon(data = rbind(D[D$xx>10,],c(10,0)),fill = 'red') +
  xlab('RPM7') + ylab('Probability')
```



This red area is simply calculated via the `pnorm()` function.

```
1 - pnorm(10,mean(x),sd(x))
```

```
## [1] 0.2578628
```

$$P(X \geq 10) = 0.26 \tag{2.9}$$

The probability of a single female mice, from normal breed, administered with water to perform higher than $10RPM7$ is hence 0.26.

2.4.4 Confidence interval for μ

The confidence interval for the center of a normal distribution (μ) is calculated as follows:

$$CI_{\mu,1-\alpha} : \hat{\mu} \pm t_{1-\alpha/2,df} \cdot \hat{\sigma}/\sqrt{(n)} \tag{2.10}$$

$$\bar{X} \pm t_{1-\alpha/2,df} \cdot s/\sqrt{(n)} \tag{2.11}$$

where $t_{1-\alpha/2,df}$ is a fractile, a number, which determines the coverage. Here, α is the left ot part. I.e. if a 90% confidence interval is wanted, the left out part is $\alpha = 0.10$. n is the number of samples on which the mean ($\bar{X}$) is estimated, and df is the degrees of freedom, which refers to how well the standard deviation is estimated.

For instance, if one needs a 95% confidence interval ($\alpha = 0.05$) based on a sample of 20 observations $t_{1-\alpha/2,df} = t_{0.975,20-1} = 2.093$.

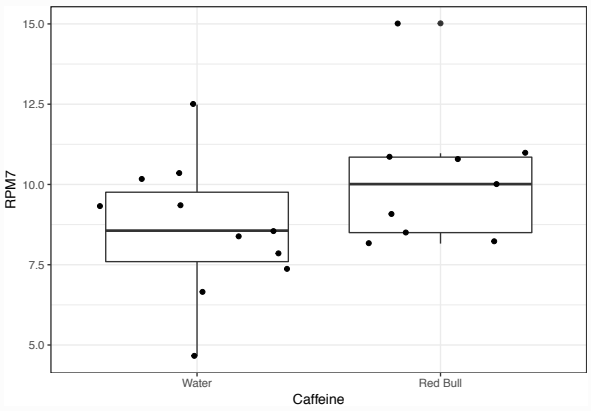
It is further noticed that the interval is symmetric around the mean, and that the more samples (n) the lower the spread as the standard deviation is divided by $\sqrt{n}$.

2.4.5 Example: Effect of Caffeine on Activity - Confidence Intervals

In example 2.4.2 concerning the relation between caffeine and voluntary activity of mice, the running activity of mice in a wheel is recorded. In this example we wish to estimate the mean activity of the two groups with either water- or Red bull as caffeine source in control runner female mice, and give a confidence interval for these means.

After subsetting the data to only include these two groups a plot of the raw data looks like the following, and consists of $n = n_1 + n_2 = 11 + 9 = 20$ samples.

```
library(ggplot2)
qplot(data = X2grps, Caffeine, RPM7,geom = c('boxplot','jitter')) +
  theme_bw()
```



The mean and standard deviation for the two groups can be calculated using the `aggregate()` function

```
n <- aggregate(X2grps$RPM7, by = list(X2grps$Caffeine),length)
mn <- aggregate(X2grps$RPM7, by = list(X2grps$Caffeine),mean)
s <- aggregate(X2grps$RPM7, by = list(X2grps$Caffeine),sd)
descripStat <- cbind(n,mn[,'x'],s[,'x'])
colnames(descripStat) <- c('Caffeine','N','mean','sd')
kable(descripStat,digits = 1)
```

Caffeine	N	mean	sd
Water	11	8.6	2.1
Red Bull	9	10.2	2.1

The confidence interval is calculated as follows:

$$CI_{\mu} : \hat{\mu} \pm t_{0.975,n-1} \cdot \hat{\sigma}/\sqrt{(n)} \tag{2.12}$$

$$\bar{X} \pm t_{0.975,n-1} \cdot s/\sqrt{(n)} \tag{2.13}$$

With a total of $df_1 = n_1 - 1 = 10$ degress of freedom the t-fractileat level 95% is $t_{0.975,10} = 2.23$ why the confidence interval for the water treated group is:

$$CI_{\mu_{water}} : [8.6 - 2.23 \cdot 2.1/\sqrt{11}; 8.6 + 2.23 \cdot 2.1/\sqrt{11}] = [7.3; 10.0] \tag{2.14}$$

```
n <- 11
tfrac <- qt(0.975,n-1)
s <- sd(X2grps[X2grps$Caffeine=='Water','RPM7'])
mn <- mean(X2grps[X2grps$Caffeine=='Water','RPM7'])
cilow <- mn - tfrac*s / sqrt(n)
cihigh <- mn + tfrac*s / sqrt(n)
c(cilow,cihigh)
```

```
## [1]  7.253336 10.045428
```

And similar for the Red bull treated group.

$$CI_{\mu_{RedBull}} : [8.5; 11.8] \tag{2.15}$$

In R this can be assesed via the `t.test()` functionallity:

```
t.test(X2grps[X2grps$Caffeine=='Water','RPM7'])
```

```
##
##  One Sample t-test
##
## data:  X2grps[X2grps$Caffeine == "Water", "RPM7"]
## t = 13.805, df = 10, p-value = 7.744e-08
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   7.253336 10.045428
## sample estimates:
## mean of x
##  8.649382
```

2.4.6 Reading Material

A youtube video giving a short condensed introduction to the Normal Distribution:
`https://www.youtube.com/watch?v=iYiOVISWXS4`

Chapter 2 and 3 of *Introduction to Statistics* by Brockhoff `https://02402.compute.dtu.dk/enotes/book-IntroStatistics` especially 2.1 to 2.4 (you can skip 2.2 on discrete distributions, which is the subject of week 4) and 3.1 - 3.1.6

2.4.7 Exercises

Exercise 2.5 Normal Distribution

Serum cholestorol level is a biomarker of health in relation to a range of life-style related diseases. A population of adults have mean values of $178mg/100mL$ and $207mg/100mL$, standard deviations of $31mg/100mL$ and $37mg/100mL$ in males and females respectively.

1. Draw curves and plug in the parameters for these distributions.
2. What percentage are below $150, 178$ and $200mg/100mL$ respectively (for males and females separately)?
3. What percentage are above $140mg/100mL$

A level above $240mg/100mL$ is considered high risk, whereas the range between 200 and $239mg/100mL$ is considered borderline high risk.

4. How big a proportion is considered at high risk (for males and females seperately)?
5. How big a proportion is considered borderline high risk (for males and females seperately)?
6. Now assume that the above metrics describing the distribution (mean and standard deviation) is estimated based on samples of 30 males and 30 females. What it the changes to the calculations above?

A population consists of 40% males and 60% females

7. What is the propability that a random person from that population have a serum cholesterol level higher than $240mg/100mL$.

Use that the mean and variance of a sum of two populations is the sum of the means and variance respectively. That is; Let X and Y be two indepedent random variables with mean μ_X and μ_y and variance σ_X^2 and σ_Y^2 respectively. And let:

$$Z = c_1 X + c_2 Y \tag{2.16}$$

Then the mean and variance of Z is:

$$\mu_Z = c_1 \mu_X + c_2 \mu_Y \tag{2.17}$$
$$\sigma_Z^2 = c_1 \sigma_X^2 + c_2 \sigma_Y^2 \tag{2.18}$$

8. What is the population mean and standard deviation of a population made up of 40% males and 60% females?
9. Use this mean and standard deviation to calculate the probability of a random person from that population having a serum cholesterol level higher than $240mg/100mL$. How is that different from question 7? and why so?

Exercise 2.6 Transformations and the Normal distribution

The normal distribution (also known as the Gaussian distribution) are central in biology as such, and in dataanalysis especially.

1. Why do we (sometimes) transform data before making statistical analysis?
2. Make a histogram plot and a qqplot (use qqnorm and qqline form the `base()` library) on the variable Ethyl.pyruvate from the `Wine data` (also used in exercise 1.8 and 2.1). Are the data normal distributed?
3. Try to Log transforms the variable Ethyl.pyruvate. Make a new histogram plot and a corresponding qqplot. Compare with question 1.
   (a) Use the command `par(mfrow = c(1,2))` (in front of your plotting commands) to get the plots side by side which makes it easier to compare the results.
   (b) Add a suitable title on both plots use the command `main='a very nice title'`.
4. Does it seem possible to get all samples to match the same distribution?
5. Do the same for the variable Benxyl.alcohol.
6. Does the log transformation work? What is the problem?
7. Try the square root transformation.

Exercise 2.7 WHO height and weight - standard normal distribution

Height and weight are important measures of growth during childhood. However, the largest factor impacting these measures are naturally the age of the child. In order to be able to compare children with slightly different ages such data are transformed using so-called growth curves. These are provided by the WHO and are constructed to represent the world wide distribution at specific ages for boys and girl. In this exercise you are supposed to take height and weight measures from African children living under different circumstances, transform them using WHO numbers for the world wide distribution, and further look for explanatory variables causing differences in growth.

1. Import the data (*GrowthData.xlsx*). Start by importing the dataset into R. Check that all the info from the excel file was imported correctly using the commands dim() and head(). You can 'chop' the dataframe in R to isolate the variables of importance (length, sex, age etc.).
2. WHO means and standard deviation. Using the following graphs, find the mean and standard deviation for boys and girls at age 15 months.

   Girls: `http://www.who.int/childgrowth/standards/cht_lfa_girls_z_0_2.pdf?ua=1`

   Boys: `http://www.who.int/childgrowth/standards/cht_lfa_boys_z_0_2.pdf?ua=1`
3. Based on the dataset imported, calculate mean and standard deviation for the girls and the boys, how does it relate to the data from WHO?

   You can use the command `X <- X[complete.cases(X),]` to remove the empty rows in the dataframe `X`, or google how to compute descriptive statistics in the case of missing values.
4. What is the Z-value? How can you use it to relate the dataset to the WHO data?
5. Calculate the Z-scores for all the boys and the girls.

   This can be done by creating a new vector inserting the length-vector from the previous dataframe to the equation for Z-score.

6. What does it mean that a child is above or below 0 in Z-score? How many children has a Z-score below -2 and -3?

7. How many children are longer/taller than the world average?

8. Investigate whether some of the other variables (source of water `primary_watersource` or treatment of water `water_treatment_type`) impact the length at 15 months. Use `qplot()` to do the plotting, and then use `qplot(...) + facet_wrap(~SEX)` to split the plot in two corresponding to the variable `SEX`. Are the trends similar for the two sexes? (OBS: The sign `~` works poorly copy-pasted, so type in the command)

2.5 Central Limit Theorem

The learning objectives for this theme is to understand what the central limit theorem (CLT) is, and how this is a useful theorem when dealing with uncertain data. Further, to know the difference between the distribution of a population and the distribution of the population parameters.

In short, the CLT says, that the central parameter (e.g. the mean) obtained from a sample (stikprøve) from ANY distribution is normal distributed with variance equal to the sample variance divided by the sample size. That is:

$$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n) \quad for\, n \to \infty \tag{2.19}$$

As we are dealing with samples of finite size, and further needs to estimate the parameters (mean and variance) based on this sample, the Normal distribution is exchanged by the T-distribution. The T-distribution take the uncertainty of having finite data into account. That is, if $X_1, X_2, .., X_n$ is randomly sampled from some distribution with mean μ and variance σ^2 (approximated by s_X^2), then:

$$\frac{\bar{X} - \mu}{s_X/\sqrt{n}} \sim \mathcal{T}(df) \quad for\, finite\, n \tag{2.20}$$

Where *df* (degrees of freedom) is equal to *how well* the variance is estimated (usually $df = n - 1$).

2.5.1 Reading Material

Chapter 8.3 in J.G. Kern *Introduction to Probability and Statistics Using R - IPSUR.pdf*.

A freaky lecture - `https://www.youtube.com/watch?v=zr-97MVZYb0` - But it is actually quite good... and short!

2.5.2 Exercises

Exercise 2.8 Quality Control and Central Limit Theorem

Ensuring the end product quality in food production is naturally of utmost interest. In doing so, the production is monitored at several critical points, in order to check that it is *on track*. This is referred to as statistical process control (SPC) and is under the umbrella of process analytical technology (PAT).

In short, a critical point in a production is monitored by a so-called control card, that register a value, e.g. pH, temperature, or other essential product/production parameters. When a point is above or below certain limits, there is an alarm. However, due to natural variation, now and then such alarms occur without there being any systematic faults. These are called false positives.

A cheese production monitors the $24hourspH$ for every production batch within a day. Under normal conditions the false alarms follow a distribution with mean $\mu = 3$ and variance $\sigma^2 = 3$ (the distribution is a Poisson distribution, but that is actually not so important here). The production runs very well, so the monitoring of these alarms are only used for retrospective follow up. Suddenly there are reclamations based on the last years production, and the boss want you to check up on whether it is the $24hourspH$ causing the trouble.

You check the control card and finds that during the last year there were on average 3.2 alarms. Does this indicate problems with the $24hourspH$?

Use the `rpois()` to generate data from 1000 years, and check whether your results match.

Now try to imagine that the aggregated measure were over a month ($30days$) or over a week ($7days$) or as low as average over $2days$. How does this affect the agreement between the analytical probability (via central limit theorem) and the simulated probability (based on the actual underlying distribution)?

You might want to use the following code as inspiration for how it can be done in R.

```
n<- # How many samples do we average over?
l<- # mean in population
s<- # std in population
x<- # the odd observation
1- pnorm(x,mean = ,sd = )
# simulation
k<-10000 # the number of e.g. years to simulate
X<- matrix(rpois(n*k,l),k,n)
mX<-apply(X, 1,mean)
hist(mX)
sum(mX>x)/length(mX)
```

3. Week 3

In this week we are introducing the concept of inferential statistics. That is to be able to answer a specific question based on observed data. Biological or scientific questions are, in statistical terms, formulated as a hypothesis, which can be tested using data. One of the most widely used tests, are the t-test for comparison of the mean from two distributions.

3.1 Hand-in assignment

The exercise 3.2 *Diet and Fat metabolism - T test - in R* is to be handed in (through absalon or as hard-copy Wednesday night). OBS: Complete 3.1 prior to the hand-in assignment. You are welcome to put in R-code in the assignment, but it is your argumentation and interpretation that are the most important.

3.2 Exercises

For Monday work through exercise 3.1, 3.3, 3.4 and 3.5, and for Wednesday work through 3.6, 3.7 to 3.8.

3.3 Case II

The second case is described in the document "Case II.pdf". You should work on the case in groups of four, and hand in a slide-show with voice no later than Thursday evening next week.

3.4 T-test

The learning objectives for this theme is to understand the basic idea behind this test. That amounts to:

- Understand the different ingredients in this test.
- Being able to calculate the test based on knowledge on sample means, standard deviations and number of observations.
- Be able to compute the t-test for two samples with equal and unequal variance.
- Understand the relation between confidence intervals for differences and the t-test for comparison of two means.
- Comprehend that the generic procedure of a null-hypothesis tests is based on some measure of distance to the null-assumption and some relevant distribution to look-up that distance.

In short, the t-test are working on two samples (to stikprøver), with the purpose of comparing the means. Here, a brief description of the so called two-sample t-test is given. For the paired and single sample t-test please consult the eNotes.

3.4.1 Models of two samples

First of all the data are characterized by two models, here normally distributed samples:

$$X_{11}, X_{12}, X_{13}, ..., X_{1n_1} \sim \mathcal{N}(\mu_1, \sigma_1^2) \tag{3.1}$$

$$X_{21}, X_{22}, X_{23}, ..., X_{2n_2} \sim \mathcal{N}(\mu_2, \sigma_2^2) \tag{3.2}$$

3.4.2 Hypothesis

The research question goes on the difference between the two samples. Such as question is formalized statistically using the parameters of the two models under investigation and called a hypothesis. Further, a question of *difference* is formalized as the opposite; similarity.

That is the null hypothesis of no difference between the two sample means is formalized as:

$$H0 : \mu_1 = \mu_2 \tag{3.3}$$

If this hypothesis turns out *not* to be true. That is; there is a difference between the two distribution means, then this is referred to as the alternative hypothesis:

$$HA : \mu_1 \neq \mu_2 \tag{3.4}$$

or directional:

$$H_A : \mu_1 > \mu_2 \quad or \quad H_A : \mu_1 < \mu_2 \tag{3.5}$$

$$H_A : \mu_1 < \mu_2 \tag{3.6}$$

3.4.3 Test statistics

From these data, a t-statistic t_{obs} can be calculated:

$$t_{obs} = \frac{\bar{X_1} - \bar{X_2}}{s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{3.7}$$

where s_{pooled} is the pooled standard deviation, which is simply a weighted mean of the variances.

$$s^2_{pooled} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{3.8}$$

$$\tag{3.9}$$

Alternatively, this formulation is equivalent:

$$t_{obs} = \frac{\bar{X_1} - \bar{X_2}}{\sqrt{\frac{s^2_{X_1}}{n_1} + \frac{s^2_{X_2}}{n_2}}} \tag{3.10}$$

3.4.4 Test probability

This t-statistics for the non-directional two-sided test can be translated into a probability under the null hypothesis (of no difference).

$$P = 2 \cdot P(T_{df} \geq |t_{obs}|) = 2 \cdot (1 - P(T_{df} \leq |t_{obs}|)) \tag{3.11}$$

Alternatively, one can calculate a confidence interval for the differences between the means, which leads to the same interpretation of the results:

$$\bar{X_1} - \bar{X_2} \pm t_{1-\alpha/2} s_{X_{pooled}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{3.12}$$

The standard deviation used here ($s_{X_{pooled}}$) is based on weighted average of the individual sample variances.

3.4.5 Example: Effect of Caffeine on Activity - Hypothesis test

In example 2.4.2 concerning the relation between caffeine and voluntary activity of mice, the running activity of mice in a wheel is recorded. In this example we wish to compare the mean activity of the two groups with either water- or Red bull as caffeine source in control runner female mice, and give a confidence interval for these means.

As such, from example 2.4.5 the activity seems higher in the Red Bull treated group. The question we wish to answer is: Is there a difference between the two population means. In statistical terms this question is formalized as a hypothesis.
First, models for the data is suggested

$$X_1 \sim \mathcal{N}(\mu_1, \sigma^2) \tag{3.13}$$
$$X_2 \sim \mathcal{N}(\mu_2, \sigma^2)$$

Note, that the spread (σ) is assumed to be similar in the two groups.

Hypothesis

If we are interested in a difference, then we formulate the opposite, that is; the two population means are equal.

$$H0 : \mu_1 = \mu_2 \tag{3.14}$$

If this turns out *not* to be true, then the alternative is suggested to be:

$$HA : \mu_1 \neq \mu_2 \tag{3.15}$$

A statistical test done via first constructing a measure of *how far* from the $H0$ the data is. This is referred to as the test-statistics and then secondly this measure is translated in to a probability.

Test statistics

$H0$ sets the two means to be equal, so naturally a measure of how well this fits with data is reflected as the *difference* between the observed averages:

$$\bar{X}_1 - \bar{X}_2 \tag{3.16}$$

where a large (positive or negative) value indicates descrepancy from $H0$.

This distance depends on the scale of the data, why some kind of normalization needs to be encountered.
Further, the number of samples used to calculate the means should also weight in.

In total this leads to the t-test statistics.

$$t_{obs} = \frac{\bar{X}_1 - \bar{X}_2}{s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{3.17}$$

We see that almost all the ingredients for calculating this is given via the descriptive statistics. The only thing needed is the pooled standard deviation (s_{pooled}), which is simply a weighted mean of the variances.

$$
\begin{aligned}
s^2_{pooled} &= \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \\
&= \frac{(11-1)\cdot 2.1^2 + (9-1)\cdot 2.1^2}{11+9-2} \\
&= 2.1^2
\end{aligned} \tag{3.18}
$$

This leads to:

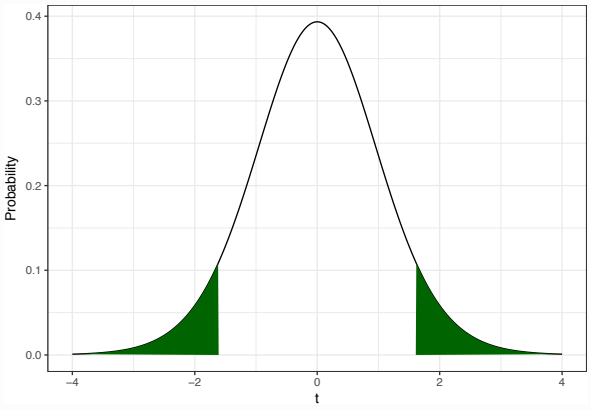
$$
\begin{aligned}
t_{obs} &= \frac{\bar{X}_1 - \bar{X}_2}{s_{pooled}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \\
t_{obs} &= \frac{8.6-10.2}{2.1\sqrt{\frac{1}{11}+\frac{1}{9}}} \\
&= 1.61
\end{aligned} \tag{3.19}
$$

P value

The test statistics is translated to a probability by the following:

$$
P = 2\cdot P(T_{df} \geq |t_{obs}|) = 2\cdot(1 - P(T_{df} \leq |t_{obs}|)) \tag{3.20}
$$

which corresponds to the colored areas shown below



This has no analytical solution, but can be assessed from tables (or via `pt()` in R)

```
2*(1-pt(1.610,20-2))
```

```
## [1] 0.1247948
```

In conclusion; Although a difference in activity is observed when mice are given Red Bull compared to water, it is Not unlikely ($p = 0.12$) that this could be due to chance. Actually this would happen one out of eight times.

The entire procedure can be done in R:

```
t.test(X2grps[X2grps$Caffeine=='Water','RPM7'],
       X2grps[X2grps$Caffeine=='Red Bull','RPM7'],
       var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  X2grps[X2grps$Caffeine == "Water", "RPM7"]
and X2grps[X2grps$Caffeine == "Red Bull", "RPM7"]
## t = -1.615, df = 18, p-value = 0.1237
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.5208153  0.4604011
## sample estimates:
## mean of x mean of y
##  8.649382 10.179589
```

3.4.6 Paired T-test

A special case, is where the two samples ($X_{11}, X_{12}, X_{13}, ..., X_{1n}$ and $X_{21}, X_{22}, X_{23}, ..., X_{2n}$) are paired. That is for example measurements before and after treatment on the same set of samples or when the same assessor in a sensorical panel scores two different product types.

The question of interest is still the same; namely is there a difference between the two populations, however, the statistics is slightly different.

3.4.7 Model

The model of the data is extended with a sequence of differences

$$\begin{aligned} & D_1, D_2, \ldots, D_n \\ =&(X_{12} - X_{11}), (X_{22} - X_{21}), \ldots, (X_{n2} - X_{n1}) \\ \sim&\mathcal{N}(\mu_D, \sigma_D^2) \end{aligned} \tag{3.21}$$

Where D_i is the difference between the responses for each of the paired observations i ($i = 1, 2, \ldots, n$), and s_D is the observed standard deviation calculated on $D_1, D_2, \ldots, D_n$.

3.4.8 Hypothesis

The hypothesis of similarity is then formalized on μ_D:

$$H0 : \mu_D = 0 \tag{3.22}$$

Often with the alternative of a difference:

$$HA : \mu_D \neq 0 \tag{3.23}$$

3.4.9 Test Statistics

$$t_{obs} = \frac{\bar{D}}{s_D/\sqrt{n}} \tag{3.24}$$

Which is translated to a p-value through a $\mathcal{T}_{df}$ with $df = n - 1$ degrees of freedom.

As such, the test statistics only sees the original data via paired differences, and hence become a specialty of the single-sample case.

3.4.10 Example: Natural Phenolic Antioxidants for Meat Preservation - Paired t-test

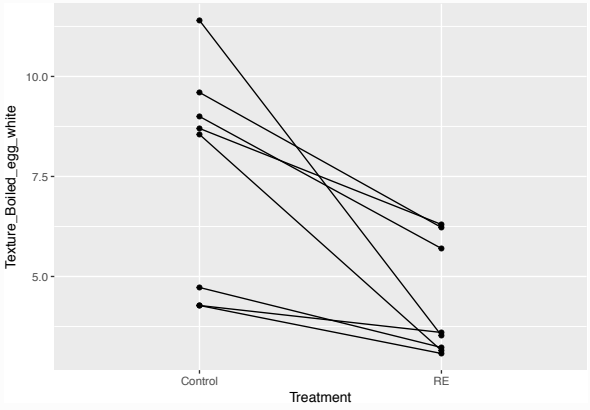
This example is based on the sensory data on the meat sausages treated with green tea (GT) and rosemary extract (RE) or control, used previously in a PCA example 1.10.2 and a Correlation example 2.3.2.

T-tests can be used to evaluate the effects (the response of a single variable) of two treatments against each other. Here, we use the *week*4 data, and compare the rosemary extract treatment to the control.

Plot data

First the data under investigation is plotted. Here the subset: Only week 4 and only Control and RE samples, is specificied directly in the `data = ...` command.

```
library(ggplot2)
qplot(data = X[X$Treatment %in% c('Control','RE') & X$week==4,],
      Treatment,Texture_Boiled_egg_white, group = Assessor) +
  geom_line()
```



The figure clearly shows that there is huge variation related to Assessor (connected by lines). Although the distributions of the two groups are overlapping, *ALL* assessors scores the *RE* sample lower than the *Control* sample.

Define subsets of data

```
# control treated samples, week 4:
Cw4<- X[X$Treatment=="Control"&X$week=="4",]
# Rosemary extract treated samples, week 4:
REw4<- X[X$Treatment=="RE"&X$week=="4",]
```

Be aware that the Assessors are ordered ensuring pairing in the two arrays `Cw4` and `REw4`. Take precaution, as this might not always be the case!.

Model of data

As it is the *same* assesors that are used for evaluation of both products the model of the data becomes:

$$X_{control} \sim \mathcal{N}(\mu_{control}, \sigma^2_{control}) \quad and \quad X_{rosemaryex} \sim \mathcal{N}(\mu_{rosemaryex}, \sigma^2_{rosemaryex}) \tag{3.25}$$

$$\begin{aligned} & D_1, D_2, \ldots, D_n \\ =&(X_{1,rosemaryex} - X_{1,control}), (X_{2,rosemaryex} - X_{2,control}), \ldots, (X_{n,rosemaryex} - X_{n,control}) \\ \sim&\mathcal{N}(\mu_D, \sigma_D^2) \end{aligned} \tag{3.26}$$

Where D_i is the difference between rosemary extract- and control treated sausages with respect to the sensorical score (texture - boiled egg) for assessor i ($i = 1, 2, \ldots, n$)

Hypothesis

If we are interested in a difference, then we formulate the opposite, that is; on average the difference between sensorical scores are 0.

$$H0 : \mu_D = 0 \tag{3.27}$$

If this turns out *not* to be true, then the alternative is suggested to be:

$$HA : \mu_D \neq 0 \tag{3.28}$$

T-test on the boiled egg texture variable

The T-test should be a paired T-test, since it is the same assosors tasting each of the treated sausages.

```
t.test(Cw4$Texture_Boiled_egg_white,REw4$Texture_Boiled_egg_white, paired=T)
```

```
##  Paired t-test
##
## data:  Cw4$Texture_Boiled_egg_white and REw4$Texture_Boiled_egg_white
## t = 3.7747, df = 7, p-value = 0.00694
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.20123 5.23002
## sample estimates:
## mean of the differences
##                3.215625
```

The p-value is $p = 0.007$ and the null-hypothosis must therefore be rejected $\rightarrow$ there is a very significant difference in boiled egg texture between the control and rosemary extracted treated samples.

T-test on the salty taste variable

In a similar fashion another sensorical variable is tested; namely *salty taste*

```
t.test(Cw4$Taste_Salty,REw4$Taste_Salty, paired=T)
```

```
##
##  Paired t-test
##
## data:  Cw4$Taste_Salty and REw4$Taste_Salty
## t = 0.98163, df = 7, p-value = 0.359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5811623  1.4061623
```

```
## sample estimates:
## mean of the differences
##                  0.4125
```

The p-value is $p = 0.36$ and the null-hypothosis is therefore accepted $\rightarrow$ there is no significant difference in salty taste between the control and rosemary extract treated samples.

Comparison with PCA results

When the results obtained from the T-tests are compared to a PCA on the *week*4 data, it can be seen that they are in agreement. The very significant effect of the rosemary extract treatment on the boiled egg texture corresponds well with the loading being strongly associated with the control samples, in a direction which seems to be explanatory for the difference between the roseary extract- and control treatment groups. The non-significant effect on the salty taste variable is congruent with its loading being located in a direction which does not seem to be of importance in the separation of the control and treated groups.

3.4.11 Reading Material

A video on the T-test https://www.youtube.com/watch?v=0Pd3dc1GcHc

Chapter 3 of *Introduction to Statistics* by Brockhoff https://02402.compute.dtu.dk/enotes/book-IntroStatistics

3.4.12 Exercises

Exercise 3.1 Diet and fat metabolism - T-test - *by hand*

The diet is a central factor involved in general health, and especially in relation to obesity, where a balance between intake of protein, fat and carbohydrates, as well as type of these nutrients seems important. Therefor various controlled studies are conducted to show the effect of different diets. A study examining the effect of protein from milk (casein or whey) and amount of fat on growth biomarkers of fat metabolism and type I diabetes was conducted in 40 mouse over an intervention period of 14 weeks.

For this exercise we are going to focus on cholesterol as a biomarker related to fat metabolism, and on a low fat diet (LF) and a high fat diet (HF). The cholesterol level at the end of the 14 week intervention is listed below.

											$\sum X$	$\sum X^2$
Cholesterol (LF)	3.97	3.69	2.61	4.03	2.98	3.51	3.62	2.81	3.62	3.53	34.37	120.20
Cholesterol (HF)	4.68	3.60	4.84	4.92	3.70	4.83	3.38	4.62	4.60	4.84	67.29	305.92
	4.84	4.54	5.27	4.26	4.37							

This exercise is supposed to be done *by hand* where the computer only is used as a pocket calculator.

1. Calculate descriptive statistics for these data.
2. Sketch these results in a graph (by pen and paper - no computer)
3. Give a frank guess on cholesterol differences between diets.
4. State a hypothesis, and calculate the test statistics and the corresponding p-value. What assumptions did you make concerning the variances in the two distributions? What alternatives do you have?

5. Give a confidence interval for the difference between diets.

6. How does the t-test result (p-value) corresponds to the confidence interval?

Exercise 3.2 Diet and fat metabolism - T-test - *in R*

This exercise repeats some of the elements from exercise 3.1 using R including an extension of the panel of relevant biomarkers and utilization of PCA to get a comprehensive overview.

The data can be found in the file `Mouse_diet_intervention.xlsx`. The code below imports the data, and subsets on two (of the three diets). The factor diet is called `diet_fat`, whereas cholesterol is called `cholesterol`.

1. Repeat the analysis of exercise 3.1 using R, including plotting and testing. (HINT: for testing you might want to predefine the response and factor and then use the function `t.test()` to run the analysis, which actually returns most of the relevant results)

2. How robust are the results towards transformation of the response or extreme samples?

In addition to cholesterol, other biomarkers have been measured, such as *insuline*, *triglycerides* etc.

3. Use one of these additional biomarkers as measure of health status, and repeat the analysis comparing dietary fat (including plotting).

4. Make a PCA on the biomarkers *insulin, cholesterol, triglycerides, NEFA, glucose* and *HbA1c*, and plot the results. (NEFA = nonesterified fatty acids).

5. Comment on the results. Do the T-test results fit with the PCA results? Are there other markers of dietary fat intake? Does the correlation structure make sense? (The later is hard to answer without physiological knowledge, but give it a shot)

```
Mouse <- read.xls('Mouse_diet_intervention.xlsx')
Mouse <- Mouse[Mouse$diet_protein=="casein",]
# Predefine response and factor
diet<- Mouse$diet_fat
y <- Mouse$cholesterol
```

Exercise 3.3 Fiber and Cholesterol

Fiber is suspected to have benificial physiological activity if being a part of a regular diet. In this study $n = 13$ healthy young men and women were enroled in a trial to unravel the effect of fiber supplement. At baseline (that is before dietary intervention), the persons were screned for a range of biomarkers in the blood including cholesterol fractions. Then, they were put on a diet with supplementation of dietary fibers for a period of 30 days. In the dataset *FiberData.xlsx*, the baseline levels (_t0), and end of trial levels (_t30) are listed for total-, hdl- and ldl cholesterol. The data is a part of a larger study conducted at Department of Nutrition Exercise and Sports.

Analyse if there is an effect of the fiber intervention on these biomarkers. That includes; inspection of raw data, possible transformations, visulization of effects, descriptive statistics, test of effect and presentation of relevant confidence intervals.

Exercise 3.4 Stability of oil under different conditions

Oil are primary made up of triglycerides, where some of the fatty acids are unsaturated. This causes such a product to be susceptible to oxidation both from chemical oxidative agents such as metal ions, or from exposure to light. Oxidation of the unsaturated fatty acids changes the sensorical and physical properties of the oil.

In the southern part of Africa grows a robust bean - the Marama bean. This bean has a favorable dietary composition, including dietary fibers, fats and proteins, as most similar types of nuts, therefor this crop could be utilized for making healthy products by the locals for the locals. One such product is Marama bean oil. A study has been conducted to investigate the oxidative stability of the oil under various conditions. In the dataset `MaramaBeanOilOx.xlsx` are listed the results from such an experiment (including data from both normal and roasted beans). The experimental factors are

- Storage time (`Month`)
- Product type (`Product`)
- Storage temperature (`Temperature`)
- Storage condition (`Light`)
- Packaging air (`Air`)

And the response variable reflecting oxidative stability is peroxide value and named `PV`.

1. Read in the data, and subset so that you only include data related to product type *Oil*.
2. Make descriptive plots of the response variable PV imposing storage- temperature, condition and time. (HINT: `factor(Temperature):Light` specifies all combinations of these two factors. `facet_grid(.~Month)` splits the plot into several plots according to Month).
3. What do you observe in terms of storage- time, temperature and condition from this plot?

Some of the differences is so pronounced that testing seems irrelevant. However, there are small differences between storage temperature at storage condition *dark*.

4. Zack out the data for these two groups at storage time $= 0.5mth$, and make a comparison.
5. Do the same at storage time $= 1mth$,

We would like to generate profiles for each condition over time, reflecting centrality and dispersion. In order to get there, we need to massage the data a little. The function `summarySE()` from the package `Rmisc` is a nice tool to construct a dataset with summary statistics. Try to run the code below - what have we learned about oil oxidation from this study?

```
library(Rmisc)
Marama.summarystat <- summarySE(Marama.oil, measurevar="PV
    ", groupvars=c("Light","Temperature","Month"))

# Repeat time=0 for all conditions
M1 <- Marama.summarystat[1,]
M1$Light <- 'light'
M1$Temperature <- 25

M2 <- Marama.summarystat[1,]
M2$Temperature <- 35

Mtot<-rbind(Marama.summarystat,M1,M2)
# Plot a nice oxidation graph with errorbars
```

```
ggplot(data=Mtot,aes(x=Month,y=PV,colour=factor(
    Temperature):Light)) + geom_point() + geom_line() +
    geom_errorbar(aes(ymin=PV-se, ymax=PV+se), width=.03)
    + theme_bw()
```

Exercise 3.5 Aroma in Milk and Cheese

Description of data In order to increase the biodiversity of hayfields it is of interest growing different kinds of herbs together with the hay. Hay is used for feeding of cows. Type of feed may affect the amount and type of flavor components in the milk as well as the flavor components in cheese made from the milk. Hence, increasing the biodiversity of hayfields used for cow feed may impact the taste of milk and cheese. An experiment was conducted in order to investigate this. In the experiment cows were fed three different diets.

- Feed 1: Ryegrass and white clover
- Feed 2: Feed 1 + red clover, chicory and ribwort plantain
- Feed 3: Feed 2 + lucern, birdsfoot trefoil, melilot, caraway, yarrow and salad burnet

Flavor components were quantified in the raw milk as well as in the cheese after 12 month of ripening. Three farmers (**G**, **H** and **P**) participated in the experiment. Furthermore, controls (**K**), consisting of bulk milk samples from the dairy, are included. The data are found in the file `Milk_Cheese_Aroma.xlsx`.

Data are originating from the EcoServe project and is kindly provided by Thomas Bæk Pedersen, Department of Food Science, University of Copenhagen.

1. Import the data set
2. Open the data and get familiar with the data set (that could include something like; how many samples, how many samples from each combination of feed and farmer, how many variables, and some descriptive plots of some of them)

In this exercise we are interested in making a pairwise comparison to investigate whether differences exists between the feedings. Hence, for now we will exclude the controls.

3. Make two subsets of the data; one subset for milk samples and one subset for cheese samples. Exclude controls in the subsets.

   (a) Investigate a given aroma compound (e.g. Limonene) in the milk subset. This can be done by e.g. making a boxplot with feed on the x-axis and filling color according to farmer. Do you see any systematic effect of feed?

```
# Investigate the following code and create the two
    subsets (One for cheese and one for milk). Exclude the
     controls. What is the &/|-statement doing? What is
    the !-statement doing?
Milk <- subset(Data, Sample == 'Milk' & Farmer == 'K')
Milk <- subset(Data, Sample == 'Milk' & !Farmer == 'K')
Milk <- subset(Data, Sample == 'Milk' | Farmer == 'K')
Milk <- subset(Data, Sample == 'Milk' | !Farmer == 'K')

# make a boxplot
M.box <- ggplot(your data,aes(x-axis,y-axis,fill=your
    factor))+geom_boxplot()
```

In a paired t-test we will investigate whether there is a difference between feed 1 and feed 3. In a paired t-test we are investigating if the mean of differences between pairs is significantly different from zero. The following steps will take you through the test.

4. Calculate the difference (for e.g. Limonene) between each pair. Note that these differences now represent your sample

```
F1 <- subset(Milk,Feed == 'I') # extract Feed I
F1 <- F1[with(F1,order(Farmer)),] # sort according to
    Farmer
F3 <- subset(Milk,Feed == 'III') # extract Feed III
F3 <- F3[with(F3,order(Farmer)),] # sort according to
    Farmer
D <- F1-F3 # Calculate the differences between Feed I and
    III
```

5. Make a boxplot of the differences.
   (a) Add to the boxplot a zero line and the mean of the differences
   (b) Discuss the plot
   (c) Why do we add a line with intercept y=0? Why do we add the mean?

```
D.box <- ggplot(D, aes(1,Limonene))+geom_boxplot() # box
    plot
D.hline <- geom_hline(aes(yintercept=0),color='red',
    linetype='dotted') # add dotted line with intercept y
    = 0
D.mean <- mean(D$Limonene) # calculate the mean of the
    differences
D.point <- geom_point(aes(x=1,y=D.mean),color='darkred',
    size=5) # add the mean as a point
D.plot <- D.box+D.hline+D.point # collect the plot
D.plot # plot data
```

We have now calculated the differences and we have calculated the mean of the differences. However, the mean is an estimated value and in order to find out whether there is a significant difference between feed 1 and feed 3, we need to calculate the confidence interval round the mean.

6. Calculate the mean and the standard deviation of the differences and use these calculations to find the confidence interval of the mean.
   For a 95% confidence interval you can find the critical t-value by calling `abs(qt(0.025,df))` for a two sided test. How would the R-call look if you were calculating the critical t-value for a one-sided test?

7. Is zero included in the confidence interval? What does this mean?

8. Calculate the t-statistics. Is there a significant difference between the two feedings?

9. Understand the following code and call it in R
   `t.test(F1$Limonene,F3$Limonene,mu=0,alt='two.sided',paired=T,conf.level=0.95)`.
   Understand the output and compare the confidence interval and the t-statistics with what you calculated in question 6 and question 7.

10. Now investigate the cheese-subset. Pick the same aroma compound you were working on during the milk-subset. Do you think there is a difference in the aroma compound between feed 1 and feed 3 when looking at the cheese-samples? Write the hypothesis and test it using a paired t-test. Use the R-call `t.test(.....)`.

If you were a farmer, would you be worried that increasing the biodiversity of your hayfields with herbs would result in altered taste in the milk/cheese?

Exercise 3.6 Power of Paired tests

In Exercise 3.3 *Fiber and Cholesterol* you have probably used either a paired t-test or a two-sample t-test. Try to conduct both of these tests and see the difference. What needs to be full filled in order to make a paired test? How many participants would have been enrolled in the study, in the case where a two sample t-test were used for inferential analysis of the data?

Exercise 3.7 Confidence intervals

The datafile *Wine.xlsx* lists aroma compounds from different wines from four countries. The compound *Acetic acid* have a skew distribution, and therefor needs an appropriate transformation. Calculate the center of the distribution for each country and give a confidence interval for this parameter, where you take into account the need for transformation, but still want to report the results in original units. The R function `t.test()` is capable of doing some of the calculations.

3.5 Hypothesis testing

The learning objectives for this theme is to usnderstand the statistical formalisation of posing- and answering questions, which we knows as null hypothesis testing, and that this is the key concept in inferential statistics. Further, to be able to suggest a relevant metric as a surrogate for testing in relation to the scientific question asked, a so called test statistics, and how this is related to a null-distribution. Understand the meaning of the p-value. Understand the concept of an alternative hypothesis.
Additionally the concepts of population, (random) sample, parameter and estimator-function are central repetitions for this theme.

3.5.1 Reading Material

Chapter 3 of *Introduction to Statistics* by Brockhoff https://02402.compute.dtu.dk/enotes/book-IntroStatistics especially 3.1 - 3.1.3 and 3.1.6.

3.5.2 Exercises

Exercise 3.8 Hypothesis Testing

This exercise is conceptual. The idea is, that you should think about a problem, and try to figure out what the null hypothesis is, and further what could be a relevant metric for measuring the distance to this null hypothesis.

1. You playing with a six-sided dice, and you are observing abnormally high number of fives. You set up an experiment to test whether the dice is skew, where you trow dice and register the outcome a number of time. What is the null hypothesis in this experiment?

2. Two response variables in a sample of size n seem to track, i.e. is positively correlated. What is the null hypothesis for testing this relation? and what measures the distance to this hypothesis?

3. In an experiment you have a treatment with three levels (placebo, treatment A and treatment B) and some relevant response variable. You are interested in whether there is a difference between the treatments. And in particular whether A is different from placebo and whether B is different from

placebo. What is the null-hypothesis for the former question? and what would be a relevant metric for measuring the distance to that hypothesis? What is the null hypothesis for the pairwise comparisons and what further relevant metrics for measuring this distance.

4. According to theory there is a proportional linear relation between y and x. You fit a line between the two ($y = a + bx$). What is the null hypothesis concerning proportionality? and what measures the distance to this hypothesis?

Exercise 3.9 Association or Causality?

This exercise is intended to show, that you need to be careful with drawing conclussions solely based on statistical numbers (confidence intervals, p-values,...), and that you need to be critical and think about the study design, biology, life, etc.

A study wants to investigate a certain biomarker in the discovery of cancer. From a population of cancer patients a sample of $n = 123$ patients is taken, and their blood is investigated for a specific biomarker (BMa). The mean and standard deviation of this sample is estimated to $\bar{x} = 3.4mg/L$ and $s_x = 1.5mg/L$ respectively.

1. Calculate the standard error for the mean of the distribution.
2. Make a confidence interval for the mean of the distribution.

The average in this population seems rather high from a biological point of view. However, the researchers want to verify that this is indeed the case, and therefore go out and recruites a population of healthy individuals of size $n = 130$. The discriptive statistics for this group is $\bar{x} = 2.9mg/L$ and $s_x = 1.3mg/L$.

3. Make a confidence interval for the mean in the healthy population.
4. Sketch the two population distributions. Are there an overlap?
5. Sketch the two confidence intervals and contemplate over similarity/differences between these two populations.

The researchers ask the question of whether the two distributions are similar.

6. Formulate the question as a null- and alternative hypothesis.
7. Test the hypothesis, and comment on the question raised.

The answer to the question seems to indicate differences between the two populations. Now the researchers take this one step further, and claims, that this must be due to cancer status.

8. What is problematic in drawing the conclusion, that differences BMa is caused by cancer status? HINT: Think about study-design, and other differences between the two populations such as age, lifestyle etc.
9. In order to be certain about cancer leading to increased levels of BMa, which circumstances must be fullfilled? Is possible to make such studies on humans? Mice?

4. Week 4

In this week we are going to discuss non continuous data, and look at binary and count data. This task involves probability distributions such as binomial distribution and the Poisson distribution.

4.1 Hand-in assignment

The exercise 4.3 *Fig bars* is to be handed in (through absalon or as hard-copy Wednesday night). You are welcome to put in R-code in the assignment, but it is your argumentation and interpretation that are the most important.

4.2 Exercises

For Monday work through exercise 4.1 and **??**. If you really think that this is easy stuff, then try to work through exercise 4.4. For Wednesday work through 4.5, 4.6 and 4.7.

4.3 Case II

The second case should be handed in as a slide-show with voice no later than Thursday evening.

4.4 Binomial data

The learning objectives for this theme is to understand the nature of binary data.

- Be able to calculate probabilities for a binary process.
- Understand what the meaning of independence.
- Understand the difference between point probability and cumulative probability, and be able to compute those.
- Based on data, be able to estimate parameters for the binomial distribution.
- Based on study design and data, be able to formulate null hypothesis, and test them.

In short, binomial data are of the type *either / or*, and is normally recorded as a list of $0's$ and $1's$. The binomial distribution is characterized by *how many* trials there is conducted (n) and the probability of a positive ("1") response (p). In notation that is:

$$X \sim \mathcal{B}(n, p) \tag{4.1}$$

The probability density function (pdf) evaluating the probability of x (out of n) positive responses are:

$$P(X = x) = \binom{n}{x} \cdot p^x (1-p)^{n-x} \tag{4.2}$$

Where

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!} \tag{4.3}$$

is the binomial coefficient, which calculates how many combinations of x in n there exists.

The cumulative density function (cdf) simply sums op the individual point probabilities.

$$P(X \leq x) = P(X = 0) + P(X = 1) + ... + P(X = x) \tag{4.4}$$

Based on data from n trials with x positive responses, the parameter p can be estimated as the frequency:

$$\hat{p} = \frac{x}{n} \tag{4.5}$$

With the following standard error:

$$S_p = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{4.6}$$

From the Central Limit Theorem follows that the point estimate for p is approximately normally distributed, why the confidence interval is as follows:

$$CI_p : \hat{p} \pm z_{1-\alpha/2} S_p \tag{4.7}$$

Be aware that the extreme cases where $x = 0$ or $x = n$ does not comply with the above mentioned method for calculating the confidence interval, as $S_p = 0$. A surpass in this situation is to find a value for p under which the observed data is still likely. For instance; find p such that:

$$P(X = 0) = (1-p)^n > \alpha \tag{4.8}$$

4.4.1 Example: Quality Control - Estimation

As a production engineer you are responsible for keeping the product quality high, which specifically means that the number of nonconforming products from the your production facility should be low.

In order to establish what the rate of nonconforming products is, you choose to randomly select 200 products and check their quality.

Of these 6 are nonconforming. So on average 3%.

However, the production facility produces a very large number of elements, and your boss wish to know whether the 3% holds for the entire production. In order to answer this you calculate a confidence interval:

```
n <- 200
x <- 6
phat <- x/n
sphat <- sqrt(phat*(1-phat)/n)
zfrac <- 1.96
CIlow <- phat - zfrac*sphat
CIhigh <- phat + zfrac*sphat
c(CIlow,CIhigh)
```

```
## [1] 0.006357817 0.053642183
```

You return these numbers to your boss, stating that there is probably 3% nonconforming products, and that it is very unlikely that there are more than 5.4% based on the 95% confidence bounds.

4.4.2 Reading Material

A introductory video to the binomial distribution.
https://www.youtube.com/watch?v=qIzC1-9PwQo

Chapter 2.1 and 2.2 of *Introduction to Statistics* by Brockhoff https://02402.compute.dtu.dk/enotes/book-IntroStatistics

4.4.3 Exercises

Exercise 4.1 Triangle Test

In sensorical science several types of experiments can be used for measuring (dis-)similarity between products. One of those is the Triangle Test.

1. Check out on the Internet how this test is conducted.
2. Assume that you run a Triangle test with n judges. What is the nature- and statistical model for the data obtained from such a trial?
3. State a null hypothesis, relating to the question of similarity of products in the study.
4. Assume that you include $n = 20$ judges in your experiment. How many correct answers do you need in order to statistically prove difference between products? (You need to define what *Statistically prove* means). HINT: The result is found by *trial and error* computation - see code below.

```
p <- (1/3)# set parameter under the H0 hypothesis
n <- 20 # set the number of trials
x <- 1:n # define all posible outcomes
pp <- 1 - pbinom(x-1,n,p) # calculate test probability
    under H0
pp
1 - pbinom(10,20,0.7)

# put the results into a data frame and plot it.
res <- data.frame(numberCorrect = x,nullprob = pp)
qplot(data=res,numberCorrect,nullprob) + geom_line() +
    geom_hline(y=(0.05))
```

5. Generalize this to $n = 1, 2, 3, ..., 100$ and plot the results (x-axis: number of trials, y-axis: frequency of correct answers). HINT you need to put this in a for loop over n and record the *least* number of samples needed for a significant results within each iteration. The code below can be used as inspiration.

```
p <- (1/3)# set parameter under the H0 hypothesis
a <- (0.05) # set significanse thresshold
X <- vector() # predefine vector for storing results.
# A for loop over the number of trials.
for (n in 1:100){
  x <- 1:n
  pp<-1 - pbinom(x-1,n,p)
  X[n]<-min(x[pp<a]) / n
}
```

6. In a population you think that 40% are capable of answering correct for discrimination of two products. The rest (60%) will simply just give a blind guess. How large a proportion of correct answers would you expect in such a population? HINT: Think of 100 persons.
7. Compare the results for the later two questions, and give a frank idea of the number of participants needed for an experiment with the purpose of discriminating the products.

Exercise 4.2 Uncertainty of the Binomial Distribution

This exercises examines the relation between the probability parameter p and the uncertainty on this S_p for the binomial distribution.

- Let X describe the number of successes out of n trial. Write op the model for X
- Let x be the observed number of successes from such a trial. Estimate the parameter of interest.
- Use the central limit theorem to approximate the standard error on this estimate, and write up the standard error for the parameter.
- Draw the relation between the parameter estimate and the standard error for the same estimate in a graph.
- At which point is the uncertainty lowest/highest? (HINT: You can either solve this analytically by differentiation with respect to $\hat{p}$) or use the graph.

Exercise 4.3 Fig bars

A company is producing fig and date bars and they are considering changing their date supplier. The company would like to produce fig and date bars with the same taste, smell and appearance as they have done for many years. It is therefore important that a new date supplier will not result in fig and date bars that differ from the original bars. The company asked the Department of Food Science at University of Copenhagen to perform a sensory test with five possible date suppliers. The Department of Food Science performed a triangle test with a trained sensory panel to detect if the new date suppliers would change the organoleptic (in this case smell, taste and sight) properties of the bars. The same 23 sensory judges performed the triangle test on smell, taste and appearance.

1. Open the file `dates.xlsx` in Excel or in R. The results of the triangle test were either 1 for a correct assignment of the deviating sample or 0 for an incorrect assignment. Get a first impression of the data by looking at the raw data (the $1's$ and $0's$). Is there a judge that is good at finding the deviating sample from the smell but bad a finding the deviating sample from the taste? Is there a judge that is good at finding the deviating sample from smell, taste and appearance?

2. Import the data into R by making a data matrix like this: (I.e. no need to read the excel file)

```
dates <-matrix(c(10, 11, 14, 19, 11, 9, 9, 16, 9, 9,
    18, 14, 22, 16, 11), ncol = 5, byrow = TRUE)
colnames(dates) <- c('A-R', 'B-R', 'C-R', 'D-R', 'E-R
    ')
rownames(dates) <- c('Smell', 'Taste', 'Appearance')
datespercent<-dates/23
summary(dates)
```

3. Make a descriptive plot to visualize the data (you can for example take inspiration from the plot in exercise 7.22 from Introduction to Statistics by Brockhoff http://introstat.compute.dtu.dk/enote/). By looking at the plotted data, does it seem likely that some of the new date producers could be used to produce fig and date bars that are similar to the reference/original bar?

4. State a model for X, where X is the number of correct answers in comparing a single product with the reference for a single sense, and state the chance probability. That is the probability of *by chance* guessing the correct deviating sample?

5. Now, state a null hypothesis and an alternative hypothesis. Is the alternative hypothesis directional ($>$ or $<$) or non-directional ($\neq$)? What do the three alternatives correspond to in terms of probability of correct answer compared to unqualified guessing?

6. Test this hypothesis for a triangle test result of 19 correct and 4 incorrect assignments (which is the result of the D-R smell triangle test).

NOTE: A number of tests could be applied to test the hypothesis but since the sample size is fairly small then an exact binomial test is a good choice. This is found in the R function `binom.test()` (Remember to specify the alternative hypothesis).

7. On the basis of this exercise, which date suppliers would you recommend to the company?

Exercise 4.4 Distribution of extreme values in the Normal distribution

This exercise combines the (standard) normal distribution with the binomial distribution in order to calculate the distribution of the maximum (or minimum) value from a sample of size n. OBS: This is a hard exercise, and not a part of the curriculum. However, the ideas used for solving the task is central in a number of applications.

1. Consider a single random draw X from the standard normal distribution. Calculate the probability of getting x or less. That is $P(X \leq x)$. If it helps, then set x to some specific number, say 1.2.
2. Now consider a draw of size n from the same distribution $(X_1, .., X_n)$. Write up the model for the number of data points less than x (from Q1).
3. Set n to a specific number, and calculate the probability of non of the data points being larger than x.
4. Generalize the above and write up the distribution for the maximum value in a finite sample from the standard normal distribution.
5. Simulate some data, and check that your analytical solution match.

4.5 Poisson data

The learning objectives for this theme is to understand the Poisson distribution.

- Understand the poisson data reflects a concentration (in time, in volume etc.)
- The relation between the poisson distribution and binomial distribution.
- Be able to calculate point probabilities, and cumulative probabilities in the poisson distribution.

In short, the Poisson distribution are characterized by observations that are non-negative integers. That is $X \in (0, 1, 2, ...)$, where each observation is a *concentration* in a fixed volume, time, space etc.. I.e. the number of bacteria in $1gram$ of sample or the number of events within one day. The Poisson distribution only have a single parameter λ, which is both the mean and the variance of the distribution. Formalized the distribution can be written as:

$$X \sim Po(\lambda) \tag{4.9}$$

For calculation of a point probability, the probability density function (pdf) is as follows:

$$P\left(X = x\right) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{4.10}$$

The cumulative probability function (cdf) is simply the sum over the individual point probabilities (just as in the binomial distribution).

In the case where we have data $(X_1, X_2, ..., X_n)$ following the Poisson distribution, we can use these to estimate the parameter λ. This is simply done by using the mean of X.

$$\hat{\lambda} = \bar{X} \tag{4.11}$$

A confidence interval for this parameter is found by using the Central Limit Theorem, which states that the distribution of the mean approximately follow a normal distribution. That is:

$$CI_\lambda : \hat{\lambda} \pm z_{1-\alpha/2}\sqrt{\hat{\lambda}/n} \tag{4.12}$$

Where $z_{1-\alpha/2}$ is a fractile from the standard normal distribution. If $\alpha = 0.05$, then $z_{1-\alpha/2} = 1.96$.

4.5.1 Reading Material

A video with an introduction to the Poisson distribution
https://www.youtube.com/watch?v=jmqZG6roVqU

Chapter 2.2 of *Introduction to Statistics* by Brockhoff https://02402.compute.dtu.dk/enotes/book-IntroStatistics especially 2.2.4

4.5.2 Exercises

Sometimes, when solving exercises you end up by doing what you are asked, without understanding why. The next two exercises tries to surpass this, by presenting the task in a short form, from which you should specify the questions which will give answers to the task. It is up to you, but maybe try to use the short version, and see how far that will get you.

Exercise 4.5 Quality Assurance

The short version A production of a food material is checked batch wise for contamination of unwanted bacteria. The procedure is to take out n samples of a given size, innoculate each sample in a relevant media at a relevant temperature for a relevant number of hours, and then check each sample for growth. Derive the underlying distribution of bacteria in the production batch. Which assumptions do you impose? and how would you ensure the validity of these in the sampling procedure?.

A batch were sampled following the procedure with $n = 5$ and there were not observed any growth. What is the upper confidence bound for the bacteria concentration in the sample under this observation?

You probably assume that in order to observe growth in a single sample there need to be at least 1 viral bacteria cell present. However, the way growth is measured is via checking the optical density a method which is not extremly sensitive, why there need to be at least 500 viral bacteria cells present in original sample in order to detect growth. Recalculate the upper confidence bound for the bacteria concentration under this opservation.

The long version A production of a food material is checked batch wise for contamination of unwanted bacteria. The procedure is to take out n samples of a given size, inoculate each sample in a relevant media at a relevant temperature for a relevant number of hours, and then check each sample for growth.

1. Which distribution do the number of bacteria cells in each sample follow?
2. Calculate the probability of observing growth. I.e. that a sample contain at least a single bacteria cell.
3. Which distribution does the observation of growth/no growth in the n samples follow? and what are the parameters for this distribution?
4. What assumptions naturally follow? and how would you ensure the validity of these in the sampling procedure?

A batch were sampled following the procedure with $n = 5$ and there were not observed any growth in any of the 5 samples.

5. Estimate the binomial parameter and calculate a confidence interval for this. HINT: For this extreme case (0 positive) you can not use the approximation by the normal distribution. Instead look for a value of p that would produce the observed result with some certainty (you have to specify). You can do this analytically or trial and error using the `dbinom()` in R.
6. What is the upper confidence bound for the bacteria concentration in the sample under this observation?

You probably assume that in order to observe growth in a single sample there need to be at least 1 viral bacteria cell present. However, the way growth is measured is via checking the optical density a method which is not extremely sensitive, why there need to be at least 500 viral bacteria cells present in original sample in order to detect growth.

7. Recalculate the upper confidence bound for the bacteria concentration under this observation. HINT: This can not be done analytically, so you need to try with different values of λ and see which one that matches

Exercise 4.6 Quality Assurance - Poisson and Binomial distribution

The short version A quality assurance program uses the test described in exercise 4.5 with n samples. Derive the sensitivity of the test, that is; the probability of detecting contamination, given that is indeed there, as a function of the number of samples and the true concentration in the batch under investigation. Draw some curves visualizing this relation.

The long version A quality assurance program uses the test described in exercise 4.5 with n samples.

1. Set $\lambda = 1$, calculate the binomial parameter for detecting growth in a single sample.
2. Set $n = 5$, calculate the probability of observing no growth in any of the samples.
3. What is the sensitivity of the test, that is; the probability of detecting contamination, given that it is indeed there.
4. What happens if the sample size is changed for instance by a factor of 2 or 1/2?
5. Generalize this for other values of λ and n, and draw curves for varying $n$with sensitivity on the y-axis and λ on the x-axis.

Exercise 4.7 Quality Assurance - Chance of False Rejections

For some types of microorganisms (bacteria or yeast) a product is damaged for very low concentrations, due to the possibility of growth during storage. However, a number of organisms are only damaging the product when present in high amounts - if present in low amounts, they do not affect the quality.

For a specific type of bacteria a concentration above a value C (measures in CFU/g) leads to a damaged product, that should not be put on market, whereas a concentration below this value results in a good product, at least within the labeled shelf life.

You have the responsibility for quality assurance at your company. In order to be able to export to Japan and USA (those countries are the most pernickety with respect to product safety) you need to document a thorough eigen-control system. So you set up a procedure.

This procedure is set up to measure the concentration in a product sample. That is, a predefined number of samples (n) is sampled from the batch, diluted, spread on plates, incubated and the number of colony forming units are counted. That leads to the observations: $X_1, X_2, \ldots, X_n$.

1. Based on these observations give an estimate of the concentration in the entire batch.
2. Additionally, give a confidence interval for this concentration, where you approximate the distribution of the mean concentration with a normal distribution (using the central limit theorem).
3. Which of the three numbers (central estimate, lower and upper confidence bound) do you think is essential for determination of whether the batch is ok or not?
4. Which rule would you suggest for making this decision?
5. What is the change of rejecting an ok product under this rule?
6. Simulate the scenario for varying concentration parameters, varying sample size (n) and determine the rate of rejection of ok batches.

For construction of random Poisson data use the function `rpois()` of fictious samples.

5. Week 5

In this week we are going to discuss a last subject for non continuous data, that is, multinomial data and the multinomial distribution. The Chi squared test is a central test for discrete data types and can be used for inferential testing. Further, we are going to deal with the notion of power, that is; if there truly is a difference, how likely is it that the statistical tests will find it?

5.1 Hand-in assignment

The exercise 5.3 *Power calculation - Triangle test* is to be handed in (through absalon or as hard-copy Wednesday night). You are welcome to put in R-code in the assignment, but it is your argumentation and interpretation that are the most important.

5.2 Exercises

For Monday work through exercise 5.1 and 5.2, and for Wednesday work through 5.4 and 5.5. Further, this week might allow you to recap on some of the exercises you did not make during the last weeks.

5.3 Case III

The third case should be handed in as a slide-show with voice no later than Thursday evening next week.

5.4 Multinomial data

The learning objectives for this theme is to comprehend tools for analyzing multinomial data (categorical data).

- Identify the distribution from the study design and/or data.
- Be able to formalize the probability model, from which the data were generated.
- Know the basic principle behind goodness of fit test.
- Be able to perform a goodness of fit test for comparison of categorical data from several groups.

Multinomial data are categorical data with *more* than two groups. If there is only two groups, the data follow the binomial distribution. These data are naturally organized in a table (see for example Exercise 5.1). In general terms such a table with n rows and k columns can be shown as follows:

	Cat I	Cat II	$\cdots$	Cat k
Sample I	N_{11}	N_{12}	$\cdots$	N_{1k}
Sample II	N_{21}	N_{22}	$\cdots$	N_{2k}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Sample n	N_{n1}	N_{n2}	$\cdots$	N_{nk}

Such data can be collected in two ways.

- N samples, that are put into nk categories. For example, 256 people are selected and categorized according to gender and color of hair.
- Several N_i samples, that are put into k categories. In this case, we predefine the number of samples within each row and distribute those over the categories. For example, 100 men and 123 women, distributed on color of their hair.

The models for those two cases are:

- $N_{11}, N_{12}, ..., N_{nk} \sim Multinomial(N, p_{11}, p_{12}, ..., p_{nk})$
  where
  $N = N_{11} + N_{12} + ... + N_{nk}$ and $p_{11} + p_{12} + ... + p_{nk} = 1$. I.e. One distribution.

- $N_{i1}, N_{i2}, ..., N_{ik} \sim Multinomial(N_i, p_{i1}, p_{i2}, ..., p_{ik})$
  where
  $N_i = N_{i1} + N_{i2} + ... + N_{ik}$ and $p_{i1} + p_{i2} + ... + p_{ik} = 1$ for $i = 1, ..., n$. I.e. several (n) distributions.

The natural question for both types of data is whether there is independence between the columns (or rows). For the example that is; Is the distribution similar regardless of gender. However, the null hypothesis is stated differently depending on the model.

- $H0 : p_{ij} = p_{i\cdot}p_{\cdot j}$ Where $p_{i\cdot}$ refers to the row probability (I.e. the probability of having gender i), and $p_{\cdot j}$ refers to the column probability (I.e. the probability of having hair color j).

- $H0 : p_{1j} = p_{2j} = ... = p_{nj} = p_{\cdot j}$ I.e. equal probability of hair color across gender.

In both cases the test is the same, and based on calculating the expected number of observations under the null hypothesis, and comparing those with the observed number of observations. If this number is large, then there is large differences between the observed and the expected, why the null hypothesis is rejected.

	Cat I	Cat II	$\cdots$	Cat k	
Sample I	N_{11}	N_{12}	$\cdots$	N_{1k}	$N_{1\cdot} = \sum N_{1j}$
Sample II	N_{21}	N_{22}	$\cdots$	N_{2k}	$N_{2\cdot} = \sum N_{2j}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Sample n	N_{n1}	N_{n2}	$\cdots$	N_{nk}	$N_{n\cdot} = \sum N_{nj}$
	$N_{\cdot 1} = \sum N_{i1}$	$N_{\cdot 2} = \sum N_{i2}$	$\cdots$	$N_{\cdot k} = \sum N_{ik}$	$N_{\cdot\cdot} = \sum N_{ij}$

The expected value E for each cell in the table is calculated as:

$$E_{ij} = \frac{N_{i\cdot}N_{\cdot j}}{N_{\cdot\cdot}} \tag{5.1}$$

I.e. the row sum multiplied with the column sum and divided by the total sum.
The test statistic X^2_{obs} is calculated as:

$$X^2_{obs} = \sum_{ij} \frac{(E_{ij} - N_{ij})^2}{E_{ij}} \tag{5.2}$$

I.e. the (squared) discrepancy between the expected (E_{ij}) and the observed (N_{ij}) divided by the expected value. Summed across all cells.

Under the null hypothesis X^2_{obs} follow a so-called chi-squared distribution (χ^2) with $df = (n-1)(k-1)$ degrees of freedom. The P-value is one-sided:

$$P = P(\chi^2_{df} > X^2_{obs}) \tag{5.3}$$

OBS: Be aware that too low expected values (E_{ij}) makes this test unstable. A rule of thumb is that 80% of the cells should be above 5 and ALL should be above 1. In the case where this is violated, cells can be merged by summing both the expected and observed values.

5.4.1 Reading Material

A video going through Goodness of fit test. Be aware that it does not concern a two dimensional table, but the principle is general.
https://www.youtube.com/watch?v=OGctoeNVn5A

Chapter 7 of *Introduction to Statistics* by Brockhoff https://02402.compute.dtu.dk/enotes/book-IntroStatistics

5.4.2 Exercises

Exercise 5.1 Comparison of Senses

A study wants to compare two types of trout samples, being meat stored under different conditions. The instrument used is a sensorical panel of 23*judges* using either their visual sense, smelling sense or tasting sense. At each trial, each judges is presented with three pieces of meat - two similar and one odd. The task for the judge is to identify the odd sample using one of the senses. Data from such an experiment is presented below.

	Correct	Not correct
Smell	14	9
Taste	16	7
Visual Apperance	22	1

1. For now, stick to the sense *Taste*. State a statistical model for the outcome of each trial.
2. Formulate a null hypothesis based on the model, and test how different the observed results are compared to this hypothesis.

Based on the previous result, it seems like the different meat pieces is identifiable based on tasting. Now the questions is whether the two other senses performs similar in identification of the odd sample? (HINT: In this exercise you should use the method sketched in *Method 7.19* and *Method 7.21* in the eNotes)

3. Give a frank ranking of the senses based on the observed data.
4. Formulate a model for each of the three senses.
5. State a null hypothesis in relation to the question of similarity between senses.
6. Compute *by hand* the expected values under this null hypothesis, X^2_{obs} and the degrees of freedom.
7. Use the `pchisq()` to test the null hypothesis.
8. Test the hypothesis using a function in R (try to figure out which one that does the job in a single line) - compare the results with your own calculation.
9. Report the results in such a way, that differences between the three senses are communicated. HINT: This can either be done by pairwise contrasts or confidence intervals for the central parameters.

5.5 Power calculation

The learning objectives for this theme is to understand the idea behind statistical power. That is:

- Know that a true difference might not be statistical detected due to size of data.
- Understand that statistical power depends on effect size, uncertainty and number of samples.
- Be able to make appropriate assumptions and calculate the power for a given study design.
- Be able to calculate power for studies evaluated by the t-test and by the binomial distribution.

5.5.1 Example: Quality Control - Power Calculation

This example extends 4.4.1

The 3% nonconforming products is unsatisfactory, and so you use a lot of money and (hopefully) make a lot of improvements. After a year you and your colleagues feel that the end quality has improved, and you wish to test that this improvement is indeed also statistically provable.

You believe that the number of nonconforming products is reduced by a factor of 2, that is down to 1.5% nonconforming products.

The question is: In a trial, how many samples (n) should you select in order to statistically prove this change?.

The null hypothesis is $H0 : p = p_0 = 0.03$

With the *one-sided* alternative: $HA : p < p_0 = 0.03$

From a study on say $n = 200$ ($X_{HO} \sim \mathcal{B}(n, p = 0.03)$) with x nonconforming products the probability for accepting $H0$ is:

$$P(X_{HO} \leq x) \geq \alpha = 0.05 \tag{5.4}$$

```
n <- 200
x <- 0:3
pbinom(x,n,0.03)
```

```
## [1] 0.002261241 0.016248299 0.059290946 0.147151194
```

So, in order to *reject* $H0$ (at level $\alpha = 0.05$) you should among 200 samples find 0 or 1 nonconforming products (at $x = 2$, the p-value is $p = 0.059$).

If you believe that the true probability is $p = 0.015$ ($X_{HA} \sim \mathcal{B}(n, p = 0.015)$) then the probability for getting this number of positive samples (or what is even more extreme compared to $H0$) can be calculated.

$$P(X_{HA} \leq 1) = P(X_{HA} = 0) + P(X_{HA} = 1) \tag{5.5}$$

```
x <- 1
pbinom(x,n,0.015)
```

```
## [1] 0.1968966
```

This is the power of the trial under these assumptions.
In detail that means, that only one out of five times you would be able to prove that the investment were worth the effort.

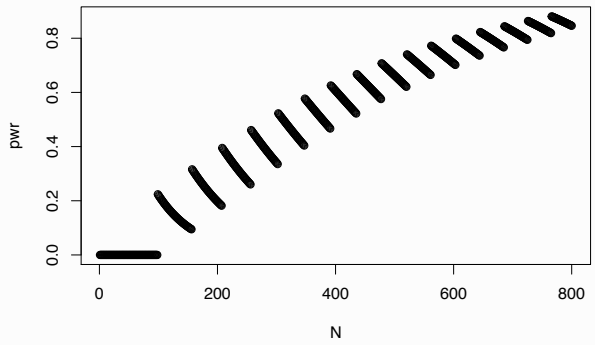
If you which to have a higher power for the study, then the number of samples should be increased e.g. to $n = 400$:

```
alpha <- 0.05
n <- 400
x <- 0:100
pH0 <- pbinom(x,n,0.03)
xmax <- max(x[pH0<alpha])
pbinom(xmax,n,0.015)
```

```
## [1] 0.6063058
```

Or assessed for a sequence of n values.

```
alpha <- 0.05
pwr <- vector()
N <- 1:800
for (n in N){
x <- 0:n
pH0 <- pbinom(x,n,0.03)
xmax <- max(x[pH0<alpha])
pwr[n] <- pbinom(xmax,n,0.015)
}
plot(N,pwr)
```



It is seen that indeed a very high number of samples needs to be collected in order to be fairly certain that the trial will statistically confirm that the number of nonconforming products has dropped.

5.5.2 Example: Effect of Caffeine on Activity - Power

The test in Example 3.4.5 reveal a p-value of $p = 0.12$ (see test below), which is not to be considered as strong support/evidence of a difference between the two groups. However, we do believe that there should be a difference between whether a mice is given water or red bull on their level of activity, and hence speculates that the study design is too small to be able to show a statistical significant difference between the two groups.

```
t.test(X2grps[X2grps$Caffeine=='Water','RPM7'],
       X2grps[X2grps$Caffeine=='Red Bull','RPM7'],
       var.equal = T)
```

```
##  Two Sample t-test
##
## data:  X2grps[X2grps\$Caffeine == "Water", "RPM7"] ...
```

```
##      and X2grps[X2grps\$Caffeine == "Red Bull", "RPM7"]
## t = -1.615, df = 18, p-value = 0.1237
...
```

The null hypothesis assummes equal means between the two groups. However, if we believe that this is *not* true. I.e. that there is a difference of say 2 points, then given a sample size of n $(= n1 + n2)$ what is the chance that the study turns out to give data that accepts (or rejects) the null hypothesis of no difference.

We try to similate a bunch of trial and check how many times $H0$ is accepted:

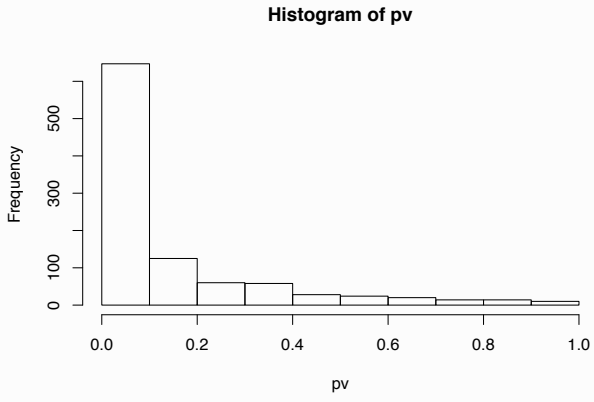
```
m1 <- 8 # mean in group1
m2 <- 10 # mean in group2
n1 <- n2 <- 10 # number in each group
sp <- 2.1 # standard deviation in both groups

set.seed(2000)
x1 <- rnorm(n1,m1,sp) # simulated data for group1
x2 <- rnorm(n2,m2,sp) # simulated data for group2

pv <- t.test(x1,x2,var.equal = T)$p.value
pv
```

```
## [1] 0.08542594
```

In this *first trial* $p = 0.085$ and $H0$ is hence accepted.

Repeat 1000 times

```
pv <- vector()
for (i in 1:1000){
  x1 <- rnorm(n1,m1,sp) # simulated data for group1
  x2 <- rnorm(n2,m2,sp) # simulated data for group2
  pv[i] <- t.test(x1,x2,var.equal = T)$p.value

}
hist(pv)
```




```
sum(pv<0.05) / length(pv)
```

```
## [1] 0.501
```

For the 1000 repeated trials *half of them* (501 out of 1000) accepts $H0$ even thoug WE KNOW that the two groups are different (as they were simulated to be).

As such this is unacceptable.

There are several ways to increase the power.

- The two groups can be selected to be even more different (increase the **effect size**).
- The spread could be lowered
- The number of observations ($n1$ and $n2$) could be increased.

The function in R `power.t.test()` calculates the power as a function of these different settings.

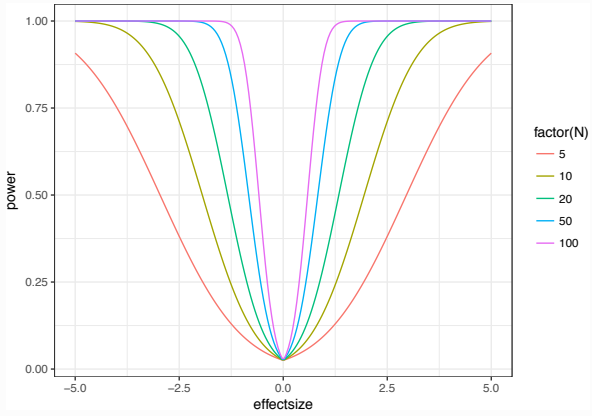
```
power.t.test(delta = m2-m1,sd = sp,power = 0.8)
```

```
##
##      Two-sample t test power calculation
##
##               n = 18.31862
##           delta = 2
##              sd = 2.1
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

So, in order to achieve a power of 0.8 with the population settings (mean and std) at least 18.3 samles ($n1 = n2 = 18.3$) should be included in each arm of the study.

A more generic overview can also be achieved, where the power is calculated based on effect size (x-axis) and number of samples.

```
library(ggplot2)
effS <- seq(-5,5,length.out = 200)
N <- c(5,10,20, 50, 100)
PWR <- data.frame()
for (n in N){
  pwr <- power.t.test(delta = effS,sd = sp,n = n)$power
  PWR <- rbind(PWR, data.frame(effectsize = effS,power = pwr,N = n))
}
ggplot(data = PWR, aes(x = effectsize,y = power, color = factor(N))) +
  geom_line() + theme_bw()
```



5.5.3 Reading Material

A video on power calculation for normal distributed data
https://www.youtube.com/watch?v=STO1NtVR2HI

A video on how to do power calculation in R
https://www.youtube.com/watch?v=7xghHcmQC50

Chapter 3.2.4 and 3.1.9 of *Introduction to Statistics* by Brockhoff https://02402.compute.dtu.dk/enotes/book-IntroStatistics

5.5.4 Exercises

Exercise 5.2 Our Sensorical trial

In September 2015 the students in food science conducted two sensorical trials. A triangle test and a duo trio test. Both on apple juice for discrimination between normal juice and juice with added citric acid. Below is listed the results from the 69 judges in the study.

	N	Correct
Duo-Trio	29	12
Triangle	40	20

1. What is the chance probability for answering correct in the two tests?
2. Initially, inspect data to see how the frequencies of correct answers compare with the chance probability. Can you draw conclusions without doing any stats?
3. State two models for the data obtained from the Duo-trio test and from the triangle test.
4. State null- and alternative hypothesis for the relevant data.
5. Calculate the probability of observing the results (and what is more extreme compared to the null hypothesis) under the null-assumption.
6. In case of rejection of the null hypothesis, give a estimate (and confidence interval) for the parameter.
7. How big is the proportion of people who are actually able to taste differences based on this result? HINT: you need to take into account the amount of people who gives correct answer by guessing.

Exercise 5.3 Power calculation - Triangle test

A chocolate company have a product, which they want to improve the quality of. In order to do so, they change some of the raw material to a more expensive alternative. In order to test whether this change actually "pays of" they set up an experiment - A triangle test. Among a set of consumers (n), each consumer is presented to three pieces of chocolate, two of type A and one of type B, with the task of finding type B. The recorded data from such an experiment may look like:

$$0, 0, 1, 0, 1, 1, \ldots, 0$$

1. Write up the statistical model underlying these data.

2. If it is not possible to identify type B, what is the value of the parameter?

A trial based on responses from 20 consumers results in half ($x = 10$) correctly identified.

3. What is the probability of observing such a results under the null hypothesis of no differences between type A and type B?

The product manager is not satisfied with the results, after all half of the consumers were able to distinguish he says.

You think 50% - *Aaahhh that is not actually what the data says - some of them were random!*, but you leave it there, as your boss is from CBS and basically without any technical insight!, so instead you assumme that this is indeed the case for the entire population (that the probability of selecting type B is $p = \frac{1}{2}$).

4. How many customers should you include in your trial in order to achieve a significant result? HINT: calculate for $n = 20, 21, \ldots$ the least number of correct findings in order for the results to be significant under the null hypothesis. Calculate the probability of observing this (and more extreme) given the population probability of $p = \frac{1}{2}$.

5. Plot the combination of number of customers (n) and the probability of getting significant results, and device a new trial.

Exercise 5.4 Triangel or Duo-Trio?

There exist a bunch of sensorical discrimination test, which all have the same aim, but are different in setup. In this exercise we are going to deal with the power of Triangle test and Duo-trio test.

1. State the statistical model for the two test types.

2. State the chance probabilities, and formulate null- and alternative hypothesis, for both tests.

3. Now assume you conduct both trials $n = 20$ times. What is the least number of correct answers needed to get significant discrimination between the products.

4. Assume that 50% of the people in the population are actually able to discriminate the samples. What is the expected frequency of correct answers under this assumption? HINT: think of 100 persons.

5. Calculate the power for a Duo-trio and a Triangle test with $n = 20$ assuming these populations probabilities.

6. Comment on why somebody still uses the Duo-trio test.

Exercise 5.5 Power calculation in T-test

The short version You have been able to isolate a fiber from a novel source. With the increasing lifestyle related health problems, you think that your new fiber might actually *save the world*. You just need to prove it. In order to do so, you set up an experiment similar to the one described in *Fiber and Cholesterol*. I.e. a paired setup with baseline measurements, followed by an intervention period and end-of-trial measurements. You will have proven your case if the fiber supplemented to the diet is able to lower the cholesterol level. The question is: How many patients do you need in order to run such a trial?

The long version You have been able to isolate a fiber from a novel source. With the increasing lifestyle related health problems, you think that your new fiber might actually *save the world*. You just need to prove it. In order to do so, you set up an experiment similar to the one described in *Fiber and Cholesterol*. I.e. a paired setup with baseline measurements, followed by an intervention period and end-of-trial measurements. You will have proven your case if the fiber supplemented to the diet is able to lower the cholesterol level. The question is: How many patients do you need in order to run such a trial?

1. Based on your knowledge on fiber and health give some numbers for effect size, and standard deviation. HINT: You might use the results from a previous similar study to get reliable numbers.
2. Write up the test statistics for this experiment.
3. Calculate the test statistics for varying number of included persons, with the parameters (effect size and standard deviation) fixed.
4. Use the function `power.t.test()` to calculate how many persons are needed for a trial with power of $\beta = 0.80$ and a significant threshold of $\alpha = 0.05$.

6. Week 6

In this week we are going to expand on inferential statistics by introdcuing the very general notion of statistical model formulation. In short, that is to set up an equation or function that specifies how the observed response is made up. In addition we are going to work with the analysis of variance (ANOVA) and the associated F statistics and distribution.

6.1 Hand-in assignment

The exercise 6.2 *Diet and fat metabolism - ANOVA - by hand* Q1 to Q6 is to be handed in (through absalon or as hard-copy Friday night). Q7 to Q11 verifies the questions, so please use those for checking your results - however, we DO NOT correct these if you include them in the assignment. You are welcome to put in R-code in the assignment, but it is your argumentation and interpretation that are the most important.

6.2 Exercises

For Monday work through exercise 6.1 and 6.3. For Wednesday work through 6.4, 6.5 and 6.6 (Exercise 6.6 includes more than two factors and further interaction between those - this is not part of the curriculum, but being able to compute the numbers in an ANOVA table is)

6.3 Case III

The third case should be handed in as a slide-show with voice no later than Friday evening.

6.4 Model formulation

The learning objectives for this theme is to comprehend the idea of model specification. That amounts to:

- Comprehend that a model describes the underlying process by which the data were generated, and that this ideally should cover the entire population which data is sampled from.
- Understand that a model have a systematic part and a random part.
- Be able to relate the systematic part of the model to central statistics.
- Know what the systematic parameters of a model is.
- Be able to formulate hypothesis based on the parameters of a model.
- Know that the random part of a model is the vehicle for testing the systematic part.

6.4.1 Reading Material

Chapter 3.1.10 and 8 of *Introduction to Statistics* by Brockhoff https://02402.compute.dtu.dk/enotes/book-IntroStatistics

6.5 Oneway- and Twoway analysis of variance (ANOVA)

The learning objectives for this theme is to understand the statistical comparison of means from e.g. several different treatments. That is:

- ANOVA is based on separation of the total variance in the data.

- The inferential test is relating the systematic variance with the random variance.
- To grasp, that the ratio of variances can be tested in a F distribution.
- Be able to compare responses to several "treatments", by model specification, hypothesis formulation and testing.
- Be able to validate model assumptions.
- To understand the relation between t-test and oneway ANOVA, and use a oneway ANOVA model to do pairwise comparisons between treatment 1 and treatment 2, treatment 1 and treatment 3, and so forth.
- To be able to compute anova with one or two factors.

In short, ANOVA can be seen as an extention of the t-test. Oneway ANOVA is for two sample t-test with more than two samples. And Two way anova is for paired t-test where there are more than just a pair.

6.5.1 Model formulation

Depending on the notes, there are several ways to write up the data for an ANOVA model. The following three versions are exactly similar for a setup with k groups, and possibly not equal number of observations within each group.

Version 1:

$$\begin{aligned} X_{11}, X_{12}, X_{13}, ..., X_{1n_1} &\sim \mathcal{N}(\mu_1, \sigma^2) \\ X_{21}, X_{22}, X_{23}, ..., X_{2n_2} &\sim \mathcal{N}(\mu_2, \sigma^2) \\ &\vdots \\ X_{k1}, X_{k2}, X_{k3}, ..., X_{kn_k} &\sim \mathcal{N}(\mu_k, \sigma^2) \end{aligned} \tag{6.1}$$

Version 2:

$$\begin{aligned} X_{ij} &\sim \mathcal{N}(\mu_i, \sigma^2) \\ for j = 1, ..., n_i \text{ and } i = 1, .., k \end{aligned} \tag{6.2}$$

Version 3:

$$\begin{aligned} X_{ij} &= \mu_i + e_{ij} \\ where\ e_{ij} &\sim \mathcal{N}(0, \sigma^2) \text{ and independent} \\ for\ j = 1, ..., n_i &\text{ and } i = 1, .., k \end{aligned} \tag{6.3}$$

For all three versions, the parameters describing the groups (μ_i for $i = 1, .., k$) and the dispersion within the groups (σ) appear, and are exactly similar. Naturally, all three versions are correct, but in order to harmonize with notation for ANOVA with several factors and regression we stick with version 3.

6.5.2 Distributional assumptions

For these models it is assumed that the residuals are coming from the *same* normal distribution. That is: $e_{ij} \sim \mathcal{N}(0, \sigma^2)$. By construction, this distribution has mean equal to zero. Further it is assumed that these residuals are independent, that is: Changing one will not affect the others. If these two assumptions (same distribution and independent) are not fulfilled, the model is not valid, and we cannot trust the tests. Therefor this is checked by visual inspection of the residuals for normality (qq-plot, histogram, residuals vs. predicted value etc.) and independence (residuals vs. predicted value, line-plotting). As this is a very inherent part of model validation, it is made easy in R, where `plot()` on the object created by `aov()` or `lm()` produces graphics which can be used directly; fast'n'easy.

6.5.3 Hypothesis

$\mu_1, .., \mu_k$ describes the center of each of the $1, .., k$ samples, and is naturally what is interesting to compare. Usually we are looking for differences between these means, why the null hypothesis is:

$$H0 : \mu_1 = \mu_2 = ... = \mu_k \tag{6.4}$$

with the alternative, that at least one group *sticks out*.

6.5.4 ANOVA table and test

If we have two groups, the natural choice would be to use the differences between the observed means. However, for k groups there are $k(k-1)/2$ combinations, why this approach would not lead to a single test, but merely $k(k-1)/2$ comparisons.

As an alternative for the differences between the observed means $\bar{X}_1, ..., \bar{X}_k$, the variance across these numbers are used (only totally true for groups of equal size), that is: $MS_{bewteen} = var(\bar{X}_1, ..., \bar{X}_k)$. If there is large differences between the group means, then this measure $MS_{bewteen}$ is big, and the null hypothesis should be discarded. Formalized, this needs to be compared to the variance within the groups, that is the average spread around the points.

It turns out, that this process can be seen as a partitioning (splitting) of the total variance in the entire dataset into i) how much can be ascribed to the fact that samples are from different groups, and ii) how much which is caused by natural variation within the groups. A natural way to organize these variances is in an ANOVA table. (The table here corresponds to a oneway ANOVA - a single factor. In models with several factors, the table is extended with rows similar to the row: *Factor(A)*).

Source	$SumSq$	Df	$MeanSq$	F_{obs}	$Pr(F_{df_A,df_e} > F_{obs})$
Factor(A)	SS_A	$df_A = k-1$	$MS_A = SS_A/df_A$	$F_A = MS_A/MS_e$	p_A
Residuals	SS_e	$df_e = n-k$	$MS_e = SS_e/df_e$		
Total	SS_{tot}	$df_{tot} = n-1$			

Here the total variance is described as the sums of squares across all samples: $SS_{tot} = \sum (X_{ij} - \bar{\bar{X}})^2$ (for all $j = 1, ..., n_i,$ *and* $i = 1, .., k$). Where $\bar{\bar{X}}$ is the overall average, and X_{ij} is the individual observations. OBS: In the columns: $SumSq$ and Df, *Total* is the sum of the elements.

In the two sample t-test we use the pooled variance $s^2_{X_{pooled}}$, In ANOVA this number is reflected by the residual variation MS_e, which is also the estimate of σ^2 (I.e. $\hat{\sigma}^2 = MS_e$).

6.5.5 ANOVA with several factor

Experiments often consist of several factors. For instance Coffee served at different *Temperatures* evaluated by different *Judges*, beer produced from different *Hops* and different *Yeast cultures* or oil stored at different *Temperature*, over *Time*, at different *Oxygen levels* and different *Light conditions*.

MODEL Imagine an experiment with two factors: A and B, where A has two levels, and B has three levels, then the model is simply an extension of the oneway model:

$$\begin{gathered} X_i = \alpha(A_i) + \beta(B_i) + e_i \\ where\ e_i \sim \mathcal{N}(0, \sigma^2)\ and\ independent \\ for\ i = 1, ..., n \end{gathered} \tag{6.5}$$

In this equation, $\alpha()$ describes the level with respect to A, and has two levels: $\alpha(A=1)$ and $\alpha(A=2)$, and $\beta()$ describes the level with respect to B, and has three levels: $\beta(B=1)$, $\beta(B=2)$ and $\beta(B=3)$.

HYPOTHESES The associated null hypothesis are:
Factor A: $H0 : \alpha(A=1) = \alpha(A=2) = ... = \alpha(A=k_A)$
Factor B: $H0 : \beta(B=1) = \beta(B=2) = ... = \beta(B=k_B)$

ANOVA TABLE In the test of these hypothesis, the ANOVA table is simply extend with rows in relation to these effects:

Source	$SumSq$	Df	$MeanSq$	F_{obs}	$Pr(F_{df_{Source},df_e} > F_{obs})$
Factor(A)	SS_A	$df_A = k_A - 1$	$MS_A = SS_A/df_A$	$F_A = MS_A/MS_e$	p_A
Factor(B)	SS_B	$df_B = k_B - 1$	$MS_B = SS_B/df_B$	$F_B = MS_B/MS_e$	p_B
Residuals	SS_e	$df_e = n - df_A - df_B - 1$	$MS_e = SS_e/df_e$		
Total	SS_{tot}	$df_{tot} = n - 1$			

INTERACTION For some experiments the effect of one factor might be dependent on the other factor, that is referred to as *interaction*. The interaction model for a two factor experiment is written as:

$$\begin{gathered} X_i = \alpha(A_i) + \beta(B_i) + \gamma(A_i, B_i) + e_i \\ where\ e_i \sim \mathcal{N}(0, \sigma^2)\ and\ independent \\ for\ j = 1, ..., n \end{gathered} \tag{6.6}$$

where $\gamma()$ describes the interaction between the two factor. In a model with e.g. 2 and 3 levels for the factors, this interaction term naturally has $2 \cdot 3 = 6$ levels. However, in the model above, the main effects (α and β) are included, so that consumes some of the levels. Actually, this factor has $(k_A - 1)(k_B - 1)$ levels.

ANOVA TABLE - INTERACTION Testing the interaction leads to an extension of the anova table:

Source	$SumSq$	Df	$MeanSq$	F_{obs}	$Pr(F_{df_{Source},df_e} > F_{obs})$
Factor(A)	SS_A	$df_A = k_A - 1$	$MS_A = SS_A/df_A$	$F_A = MS_A/MS_e$	p_A
Factor(B)	SS_B	$df_B = k_B - 1$	$MS_B = SS_B/df_B$	$F_B = MS_B/MS_e$	p_B
Factor(A*B)	SS_AB	$df_{AB} = (k_A - 1)(k_B - 1)$	$MS_{AB} = SS_{AB}/df_{AB}$	$F_{AB} = MS_{AB}/MS_e$	p_{AB}
Residuals	SS_e	$df_e = n - df_A - df_B - df_{AB} - 1 = n - (k_A)(k_B)$	$MS_e = SS_e/df_e$		
Total	SS_{tot}	$df_{tot} = n - 1$			

We see that the only thing that changes is the sums of squares and the degrees of freedom for the residuals.

6.5.6 Example: Natural Phenolic Antioxidants for Meat Preservation - ANOVA

This example is based on the sensory data on the meat sausages treated with green tea (GT) and rosemary extract (RE) or control, used previously in example 1.10.2, 2.3.2 and 3.4.10.

A two-way anova can be used to evaluate the effect of several factors, and possible interaction effects

on a single response variable. Here, we define a model with systematic effects of Treatment (3 levels), Assessor (8 levels) and Week (2 levels). Since it may be that the effect of the storage time differs for each of the antioxidant treatments, an interaciton effect between Treatment and Week is also included.
The starting model is therefore formulated as:

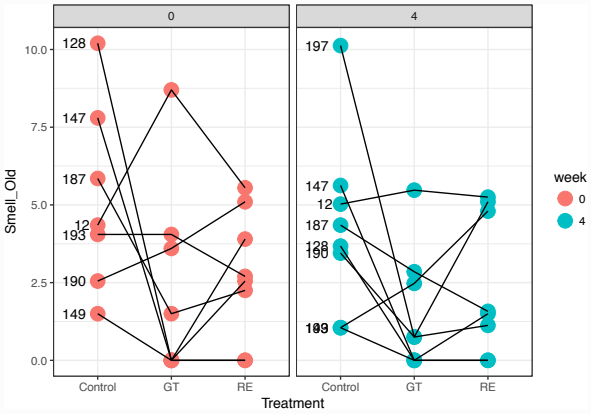
$$Y_i = \alpha(Treat_i) + \beta(Assessor_i) + \gamma(Week_i) + \theta(Treat_i \times Week_i) + e_i \tag{6.7}$$

where $e_i \sim \mathcal{N}(0, \sigma^2)$ and independent for $i = 1, \ldots, n$

Plot the data

We wish to visualize three factors on a single response variable. Here we use the x-axis for Week and Treatment, whereas the points are connected within Assessor.

```
library(ggplot2)
ggplot(data = X, aes(x = Treatment,y = Smell_Old,
                     group = Assessor, color = week)) +
  geom_point(size = 5) +
  geom_line(color = 'black')  +
  facet_wrap(~week) +
  geom_text(data = X[X$Treatment=='Control',],
            aes(label = Assessor), color = 'black',hjust = 1.5) +
  theme_bw()
```



From the figure there are a few observations: First, some assessors (e.g. 128 and 197) use the entire range whereas some only scores in a narrow range (e.g. 12, 193 and 187). Generally it seems as the control treatment (at both timepoint) obtain higher scores compared to the treated samples. There are no apparent indications for the two treatments being different.

Calculate two-way anova

```
# Here we choose to evaluate theeffects on the old smell response variable.
m<- aov(data=X,Smell_Old~factor(Assessor)+factor(Treatment)*factor(week)) # defines starting model
anova(m) # shows anova results
```

```
## Analysis of Variance Table
##
## Response: Smell_Old
```

```
##                                Df  Sum Sq Mean Sq F value   Pr(>F)
## factor(Assessor)                7  77.833  11.119  1.9861 0.086116 .
## factor(Treatment)               2  70.153  35.077  6.2653 0.004826 **
## factor(week)                    1   5.629   5.629  1.0054 0.323084
## factor(Treatment):factor(week)  2   1.062   0.531  0.0948 0.909763
## Residuals                      34 190.352   5.599
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here it can be seen that the only effect which is found significant ($p < 0.05$) for the old smell response is the treatment factor. No significance was found for the Week or interaction factors. The Assessor effect is not strictly signifcant, but on the borderline ($p = 0.09$). If we plot the raw data (see below), it seems each of the assesors have different base levels of scoring the old smell descriptor. Therefore, it is chosen to keep the systematic effect of Assessor in the model. Furthermore, it can be seen that there is some variation in the Assessor evaluation of the treatment effect (non-parallel), indicating that the sensory panel is not very precise, and may be poorly trained.

Final model

The final model is therefore the model only including the effects of the treatment and Assessor factors:

$$Y_i = \alpha(Treat_i) + \beta(Assessor_i) + e_i \tag{6.8}$$

where $e_i \sim \mathcal{N}(0, \sigma^2)$ and independent for $i = 1, \ldots, n$.
The estimates for the factors $(\alpha(Control), \alpha(GT), \alpha(RE), \beta(Ass12), \ldots, \beta(193))$ as well as the standard deviation $(\hat{\sigma})$ can be obtained by first calculating a new model including only significant terms followed by the `anova()` and `summary()` commands.

6.5.7 Contrasts

If it turns out that there is a significant effect of the factor, a natural next step is to compare the individual levels by estimation of confidence intervals and/or testing differences. This is quite similar to comparing two means by t-test. The only difference is, that $\hat{\sigma}^2 = MS_e$ is used as the measure of random uncertainty (instead of $s^2_{X_{pooled}}$), and the degrees of freedom is NOT $n_1 + n_2 - 2$, but the degrees of freedom associated with the uncertainty estimate: df_e. Apart from that, the method is similar.

Assume that Factor(A) has five levels, then the comparison of e.g. level 2 and 5 is done as follows:

Confidence intervals

$$\bar{X}_5 - \bar{X}_2 \pm t_{1-\alpha/2, df_e} \sqrt{MS_e} \sqrt{\frac{1}{n_5} + \frac{1}{n_2}} \tag{6.9}$$

Where $t_{1-\alpha/2, df_e}$ is the t-fractile at a user-specified level α with df_e degrees of freedom.

Test of contrast

The associated hypothesis are stated as:

$$H_0 : \mu_5 = \mu_2 \tag{6.10}$$

With the alternative:

$$H_A : \mu_5 \neq \mu_2 \tag{6.11}$$

Then the test statistics is calculated as:

$$t_{obs} = \frac{\bar{X_5} - \bar{X_2}}{\sqrt{MS_e}\sqrt{\frac{1}{n_5} + \frac{1}{n_2}}} \tag{6.12}$$

With the associated (2-sided) p-value: $p = 2 \cdot P(\mathcal{T}_{df=df_e} > t_{obs})$.

6.5.8 Example: Natural Phenolic Antioxidants for Meat Preservation - Contrasts

In this example, we wish to compare pairs of levels within a factor from an ANOVA model. In example 6.5.6 the *Treatment* effect from a two way model as shown below is significant, however, that does not imply that all levels are different, but only that at least a single strikes out.

```
library(knitr)
m<- aov(data=X,Smell_Old~factor(Assessor)+factor(Treatment)) # defines starting model
kable(anova(m),digits = 3) # shows anova results
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(Assessor)	7	77.833	11.119	2.088	0.069
factor(Treatment)	2	70.153	35.077	6.587	0.004
Residuals	37	197.043	5.325	NA	NA
And print out the me	an va	lues for e	ach Treatm	ent	

```
mTB <- model.tables(m,type = 'means')
TAB <- data.frame(mean = as.matrix(mTB[[1]][[3]]),
                  n = as.matrix(mTB[[2]][[2]]))
kable(TAB,digits = 2)
```

	mean	n
Control	4.75	15
GT	1.86	16
RE	2.57	16

Contrasts

We wish to compare the treatments RE and GT. This a basically a two sample t-test, just where the entire dataset and model is used to calculate the standard deviation.

The hypothesis are:

$$H_0 : \mu_{GT} = \mu_{RE} \quad and \quad H_A : \mu_{GT} \neq \mu_{RE} \tag{6.13}$$

Naturally, the driver of this test is the observed difference between the two means. In total the test-statistics amount to:

$$t_{obs} = \frac{\bar{X_{GT}} - \bar{X_{RE}}}{\sqrt{MS_e}\sqrt{\frac{1}{n_{GT}} + \frac{1}{n_{RE}}}} \tag{6.14}$$

With the associated (2-sided) p-value: $p = 2 \cdot P(\mathcal{T}_{df=df_e} > t_{obs})$.

```r
m1 <- 1.86 # GT mean
m2 <- 2.57 # RE mean
n1 <- 16 # GT n
n2 <- 16 # RE n
s <- sqrt(5.325) # standard deviation from ANOVA tab
df <- 37 # df on residuals

tobs <- (m1 - m2) / (s*sqrt(1/n1 + 1/n2)) # test statistics

2*(1 - pt(abs(tobs),df)) # pvalue for GT vs RE
```

```
## [1] 0.3897755
```

In line with the raw data, the observed difference between the two treatments, is not statistically singificant, and we conclude that the two treatments has similar properties in terms of *Old Smell*.

Confidence interval on difference between RE and Control

As an alternative, the confidence interval on the difference between two means can be calculated. As example this is done for RE vs Control.

```r
m3 <- 4.75 # Control mean
n3 <- 15 # Control n

CIlow <- m3- m2 - qt(0.975,df)*s*sqrt(1/n2 + 1/n3) # lower bound
CIhigh <- m3- m2 + qt(0.975,df)*s*sqrt(1/n2 + 1/n3) # upper bound
c(CIlow,CIhigh)
```

```
## [1] 0.4995882 3.8604118
```

We see that the confidence interval on the difference between the control- and RE treated samples are positive and do not overlap 0, that is, in general the control samples has a higher score with respect to *Old Smell*, and furhter that this is a significant difference on 95% level.

6.5.9 Reading Material

A brief video going through the concepts of both oneway ANOVA (first 3 : 40 minutes) and twoway/ two factor ANOVA (last minute)https://www.youtube.com/watch?v=ITf4vHhyGpc

A video going though how to do ANOVA in R. https://www.youtube.com/watch?v=Dwd3ha0P8uw

Chapter 8 of *Introduction to Statistics* by Brockhoff https://02402.compute.dtu.dk/enotes/book-IntroStatistics

6.5.10 Exercises

Exercise 6.1 Wine and One-way ANOVA

Wines from four different countries (Argentina, Australia, Chile and South Africa) were analyzed for aroma compounds with GC-MS (gas chromatography coupled with mass spectrometry). The dataset can be found in the file “Wine.xlsx”.
The wine data should already be familiar to you otherwise look at the exercises from week 1 and 2.

1. When working with ANOVA which assumption is then made about the data, and how does the model look?
2. We would like to investigate if the wines from the four different countries are different. State the null hypothesis and the alternative hypothesis.
3. Make a combined jitter and boxplot on the variable X3.Hexenol. Are there any suspicious samples?
4. Make an ANOVA on X3.Hexenol, look at the `summary()` and draw conclusions.
5. Compare the result with the boxplot from Q3. Was the result of the ANOVA expected? If yes; why?
6. We would like to state which countries are significantly different from the others. For this you need to install the R-package `multcomp`. Use the command `TukeyHSD()` to investigate the differences/contrasts between the countries. Use the confidence intervals for judging differences.
7. Which countries are significantly different and which are not? Compare with the boxplot produced in Q3.
8. Make the boxplot and the ANOVA for the following variables too; Diethyl.succinate, X1.Hexanol, Ethyl.hexanoate and Ethyl.propanoate. Are any of these variables different between countries?
9. Check for normality of the model residuals by making a qq-plot (Hint: You need to extract these from the model, for instance by the function `resid()`. Use `qqnorm()` and `qqline()` to make the plot). Can we trust the assumption about normality? After doing this manually, try the function `plot()` on the `aov()` object, which produces four plots for assumption checking.
10. For some of the results, the jitter plot indicates, a few extreme samples. What happens if these are removed? Does the conclusions change?

Exercise 6.2 Diet and fat metabolism - ANOVA - by hand

The diet is a central factor involved in general health, and especially in relation to obesity, where a balance between intake of protein, fat and carbohydrates, as well as type of these nutrients seems important. Therefor various controlled studies are conducted to show the effect of different diets. A study examining the effect of protein from milk (casein or whey) and amount of fat on growth biomarkers of fat metabolism and type I diabetes was conducted in 40 mouse over an intervention period of 14weeks.
The data for this exercise is the same as for Exercise 3.2
For this exercise we are going to focus on cholesterol as a biomarker related to fat metabolism, and on three types of diet:

1. High fat with the milk protein casein *HF casein* ($n = 15$)
2. High fat with whey protein from milk *HF whey* ($n = 15$)
3. Low fat with the milk protein casein *LF casein* ($n = 10$)

The cholesterol level at the end of the 14 week intervention is listed below including some descriptive stats.

	1	2	...	9	10	11	...	15	$\sum X$	$\sum (X - \bar{X})^2$	$\bar{X}$	s_X
Cholesterol (HF casein)	4.68	3.60	...	4.60	4.84	4.84	...	4.37	67.29	4.06		
Cholesterol (HF whey)	3.79	2.82	...	3.77	4.63	3.44	...	3.24	54.86	4.69		
Cholesterol (LF casein)	3.97	3.69	...	3.62	3.53	...	...	...	34.37	2.07		

The total sums of squares is: $SS_{tot} = \sum (X_{ij} - \bar{\bar{X}})^2 = 18.99$.

The first six questions are supposed to be done *by hand* where the computer only is used as a pocket calculator.

1. Calculate descriptive statistics for the three groups.
2. Sketch these results in a graph (by pen and paper - no computer)
3. State a model for these data, and a hypothesis of similarity between the three dietary treatments (wrt cholestorol level).
4. Construct an ANOVA table and calculate and fill in the numbers including test statistics (F_{obs}) and the corresponding p-value (for the translation of F_{obs} to p, use the function `|pf()|`in R).
5. This p-value should be significant. Guided by the initial descriptive stats, do you believe that *all* three diets are different? Or is there two groups which are close in estimate?
6. Formulate a hypothesis of similarity between the two most similar groups. Test this contrast.

The data can be found in the file `Mouse_diet_intervention.xlsx`.

7. Import the data, and make a plot of cholesterol (`$cholesterol`) and dietary intervention (`$Fat_Protein`)
8. Repeat the ANOVA analysis using R with the function `aov()` for construction of model and the function `anova()` for analysis of this model.
9. Check that the model assumptions are ok. (`plot(model)` where `model` is the object created by `aov()`).
10. For the individual contrasts you can use the function `TukeyHSD()` on the model. Be aware that the p-values in this analysis is adjusted for multiple testing, and therefor are bigger than the ones done one by one.
11. Based on these results, which dietary component do you think causes differences in cholesterol level?
12. There are some indications of differences between *HF whey* and *LF casein*, however, not significant. How many samples would have been needed in order to achieve significance with an appropriate power level?

Exercise 6.3 Diet and fat metabolism - ANOVA - Multivariate

In this exercise, we are going to extend the analysis from Exercise 6.2 to include several biomarker.

1. Start out by construction of a PCA on the biomarkers: *insulin,cholesterol,triglycerides,NEFA,glucose* and *HbA1c* (NEFA = nonesterified fatty acids), including all 40 samples, and comment on the results with respect to differences between diets. This is computationally identical to the task in exercise 3.2.
2. Repeat the ANOVA analysis of for several biomarkers including plotting.
3. Zack out the components in the PCA model (you can find inspiration on how to do this in exercis 1.6), and glue them together with the original data (use: `cbind()`).
4. Use the components from the PCA model as response variables, and repeat the ANOVA (including plotting).
5. Comment on the similarity / dis-similarity between the univariate results, and the ones based on the PCA.
6. Compare the results (plot and ANOVA) for $PC1$ and $PC6$. Why do you think the dietary signal is more pronounced in the first component? HINT: Think of how much, and which type of variation that is captured in the individual components.

Exercise 6.4 Analysis of Coffee Serving Temperature

Serving temperature of coffee seems of importance of how this drink is perceived. However, it is not totally clear how this relation is. In order to understand this, studies on the same type of coffee served at different temperature is conducted. In this exercise we are going to use the data from a trained Panel of eight judges, evaluating coffee served at six different temperatures on a set of sensorical descriptors. Each judge is presented with each temperature in a total of four replicates leading to a total of $6 \times 8 \times 4 = 192$ samples.

In the dataset *Results Panel.xlsx* the results are listed. In this exercise we are going to analyse the differences between the individual temperatures while utilizing the design of repeated scoring by the same judges.
The exercise is an extension of Exercise 1.7 from week one.

1. We want to summarize data, such that each judge have *one* score for each coffee sample, each attribute (instead of four). Use the `aggreggate()` function to average over replicates, which is to retain *Assessor* and *Sample*.

2. Plot this descriptive measure for *Intensity* across temperature (x-axis) and join the points from the same judge.

3. Make an ANOVA model with the response *Intensity* including both factors (Temperature and Judge).

```
m <- aov(data=CoffeeAG,Intensity ~ Temp + factor(
    Judge))
anova(m)
```

4. Check out what the function `factor()` is doing? What happens if it is removed (check the degrees of freedoms consumed by the two options).

5. Check that the model assumptions are ok. (`plot(model)` where `model` is the object created by `aov()`).

6. Comment on these results (in relation to Temperature and Judge).

7. Use this data as input for construction of a PCA model, and make a bi-plot (This task is computationally identical to Exercise 1.7).

8. What do you see with respect to differences in temperature and judges?

9. Zack out the first couple of component and glue them together with the original data.

10. Make ANOVA model for component 1 to 5 and find the component with the strongest *Temperature* and *Judge* signature respectively.

11. Plot the PCA model with these two components (This is specified by the option `choice = ...`). Color this plot according to *Temperature* and *Judge*.

12. Which sensorical markers seems mostly related to differences between *Temperature* and *Judge* respectively?

Exercise 6.5 Carcass suspension

Toughness is found to be the most important quality characteristic in beef. Variation in toughness is primarily related to the muscle fibers and the connective tissue. Connective tissue is thought to be responsible for the relative fixed background toughness, largely affected by animal age. The toughness of the muscle fibers depends on a couple of factors and is more likely to be manipulated by e.g. carcass suspension to

prevent post-mortem contraction of the muscle fibers.
An experiment was conducted in order to test the effects of Animal Age, Carcass Suspension and the interaction between Animal Age and Carcass Suspension. The muscle of interest was longissimus dorsi. The hypothesis for the experiment was that pelvic suspension would prevent post-mortem contraction of the muscle fibers and thereby result in a more tender longissimus dorsi compared with suspension in the Achilles tendon. Furthermore, the toughness was believed to increase with increasing age. This increase in toughness was assigned to connective tissue.
In the experiment, animals were slaughtered and the carcasses were spilt. One side was suspended in in the pelvic bone (Figure 6.1 A) and the other side was suspended in the Achilles tendon (Fig. 6.1 B). The experiment was balanced and 15 carcass sides were assigned to each group.

Fig. 6.1 Carcass Suspension: (A) Pelvic Suspension and (B) Achilles Suspension

In the table below are shown the average results from each cell in the design.

	Age 2y	Age 3y	Age 4y
Pelvic(A)	58.5	65.0	66.0
Achilles(B)	72.0	82.5	93.0

Table 6.1: Mean values for the design ($n = 15$ in each group)

1. Sketch these results with the response on the y-axis, one factor on the x-axis, and join the points based on the other factor. What is your first impression on the effect of *age* of the animal and of *suspenssion* method. Further, does the lines seem to be deviating from being parallel? (This indicates interaction between the factors).

2. Write up a model including interaction term.

3. Write up the $H0$ (and the alternative) hypotheses for the main effects in this experiment.

	SS	df	MS	F	p-value
Age	3083.7				
Suspension	8410.0				
Age*Suspension					
Error			375.0	-	-
Total	43715.0	89	-	-	-

Table 6.2: ANOVA table for response *shear force*

4. Fill the missing values in the table above.
5. What factors are having a significant effect (significant level of 0.05)
6. Modify the two-way ANOVA model. Should the interaction be included in the final model for the present study? Why/Why not?
7. Toughness is explained by muscle fibers and connective tissue. The contribution from muscle fibers is impacted by suspension method. The contribution from connective tissue is impacted by age. Would you expect an effect of the interaction term between suspension method and age? Why/Why not? How does your answer correspond with the answers in question 6?
8. Traditionally, carcasses suspension is done by Achilles, what would you recommend based on this study? What is the limitation/assumptions for such a recommendation? (HINT: Have we checked all muscles?, look at the picture - do you think the conclusions extrapolates to other parts of the animal?)

Exercise 6.6 Stability of oil under different conditions

This Exercise is a direct extention of Exercise 3.4.
Oil are primary made up of triglycerides, where some of the fatty acids are unsaturated. This causes such a product to be susceptible to oxidation both from chemical oxidative agents such as metal ions, or from exposure to light. Oxidation of the unsaturated fatty acids changes the sensorical and physical properties of the oil.

In the southern part of Africa grows a robust bean - the Marama bean. This bean has a favorable dietary composition, including dietary fibers, fats and proteins, as most similar types of nuts, therefor this crop could be utilized for making healthy products by the locals for the locals. One such product is Marama bean oil. A study has been conducted to investigate the oxidative stability of the oil under various conditions. In the dataset `MaramaBeanOilOx.xlsx` are listed the results from such an experiment (including data from both normal and roasted beans). The experimental factors are

- Storage time (`Month`)
- Product type (`Product`)
- Storage temperature (`Temperature`)
- Storage condition (`Light`)
- Packaging air (`Air`)

And the response variable reflecting oxidative stability is peroxide value and named `PV`.

The first three questions are identical to Exercise 3.4.

1. Read in the data, and subset so that you only include data related to product type *Oil*.

2. Make descriptive plots of the response variable PV imposing storage- temperature, condition and time. (HINT: `factor(Temperature):Light` specifies all combinations of these two factors. `facet_grid(.~Month)` splits the plot into several plots according to Month).

3. What do you observe in terms of storage- time, temperature and condition from this plot?

4. Based on the plot, suggest a model including the three factors *Month, Temperature* and *Light*. Which factors do you think is additive (i.e. have the same effect regardless of the other factors), and which do you think interacts.

5. Build a model including all combinations of the three factors. That is:
   `m <- aov(data=Marama.oil, PV~factor(Temperature)*Light*factor(Month))`.

6. Check the assumption concerning the distribution of the residuals. If nessesary, make an appropriate transformation of the response variable, and check that it is doing better in terms of distributional assumptions for the model residuals.

7. It seems like we need all combinations of everything. Construct a single variable, that has the seven levels indicating the design, and use this in an ANOVA model.
   (HINT: `Allcomp <- factor(Marama.oil$Temperature):Marama.oil$Light:factor(Marama.oil$Month`)

8. Use `TukeyHSD()` on that model to make all pairwise comparisons of the seven individual levels. And comment on what you obtain. Does the initial plot correspond to the pairwise observations?

7. Week 7

This week is going to be on regression analysis. Regression might be the most influential and most widely used technique for relating response with outcomes. Althought regression is a general framework for different types of responses (continous, binomial, poisson, categorical,...) we are only going to deal with continous outcomes in this course. However, knowing the basis for continous data, makes it easy to extend into other types of data.

Using regression as an example, this week will introduce the mathematical concept behind analysis of continous normally distributed data. Namely Least Squares estimation. In the estimation of parameters for a model, ANOVA, linear regression, correlation, PCA and also non linear models, there is often stated a so-called objective in the form of a minimization problem. I.e. *give me the parameters that results in the minimum sum of squared errors*, that is the least squares parameter estimates. The most simple situation is the center of a distribution; Using the mean/average as the center results in a minimum *overall* distance to the center.

7.1 Hand-in assignment

The exercise 7.6 is to be handed in (through absalon or as hard-copy Wednesday night). You are welcome to put in R-code in the assignment, but it is your argumentation and interpretation that are the most important.

7.2 Exercises

For Monday work through exercise 7.1 and for Wednesday work through 7.2, 7.4 and 7.5 Further, this week might allow you to recap on some of the exercises you did not make during the last weeks.

7.3 Case IV

The fourth case is open!

7.4 Regression

The learning objectives for this theme is to:

- Graphically show regression problems.
- State a statistical model for linear regression with a single predictor.
- Compute a regression model.
- Know that a (oneway) ANOVA model can be formulated as a regression problem.
- Be able to formulate hypothesis for a given question in relation to regression.

7.4.1 In short

Regression models are simply extensions of the ANOVA framework allowing for descriptive variables being continuous.
The simplest form is univariate regression with a single response variable (Y) described by a single predictor (X).

The model is :

$$\begin{gathered} Y_i = \alpha + \beta \cdot X_i + e_i \\ where\ e_i \sim \mathcal{N}(0, \sigma^2)\ and\ independent \\ for\ i = 1, ..., n \end{gathered} \tag{7.1}$$

Where α is the level of (Y) at $x = 0$ and β is the slope of the line.

Each of these two parameters consumes one degree of freedom.

The natural hypothesis is to check whether (X) has an effect on (Y), that is: $H0 : \beta = 0$ with the alternative $HA : \beta \neq 0$. This can be tested with either an F-test or a T-test (yielding the same results). For some applications the intercept (α) might be related to a question, just as other values for β could be relevant to test, but the most common is testing the relation between X and Y.

Estimation of the model is done by least squares fit, which is explained in detail elsewhere in these notes.

Model assumptions Similarly to other types of ANOVA models, regression models assumes independent residuals with the same variance, why regression model should be exposed to the same procedures for model validation.

Transformations The model assumes a linear relation between X and Y. However, inspection of data or mechanistic insight might reveal, that this relation is suboptimal, why the response and/or the descriptor variable might be subjected to a relevant transformation. This is investigated and checked by visual plotting of the raw and transformed data.

7.4.2 Reading Material

Chapter 5 of *Introduction to Statistics* by Brockhoff https://02402.compute.dtu.dk/enotes/book-IntroStatistics

7.4.3 Exercises

Exercise 7.1 Diet and fat metabolism - Regression by hand

The diet is a central factor involved in general health, and especially in relation to obesity, where a balance between intake of protein, fat and carbohydrates, as well as type of these nutrients seems important. Therefor various controlled studies are conducted to show the effect of different diets. A study examining the effect of protein from milk (casein or whey) and amount of fat on growth biomarkers of fat metabolism and type I diabetes was conducted in 40 mouse over an intervention period of 14weeks.
The data for this exercise is the same as for Exercise 3.2
For this exercise we are going to focus on predicting different biomarkers from weight data.

1. Set the independent variable (X) to be weight at 14 weeks (measured in grams), and let the dependent variable (Y) be the cholesterol level. Write up the model for the relation between X and Y.

2. Make a small drawing, and infer all the model parameters on this figure.

3. Estimate the parameters based on the discriptive numbers in the table below.

$\sum X$	$\sum (X - \bar{X})^2$	$\sum Y$	$\sum (Y - \bar{Y})^2$	$\sum (X - \bar{X})(Y - \bar{Y})$	$\sum (Y - \hat{Y})^2$
1426.2	1099.3	156.5	18.99	116.44	6.65

4. Construct confidence intervals for the intercept and the slope for this model, and judge whether there seem to be a relation between weight and cholesterol level. (HINT: Chapter 5.4 in the enotes have formulas for the variance of the parameters)

5. Calculate a prediction interval for a *new* mouse with a weight of $40g$ (HINT: Box 5.17 in the enotes have the relevant formulas).

6. Load data, plot it, and verify the results using `aov()` for construction of a regression model, and `predict()` for the prediction of new samples.

Exercise 7.2 Diet and fat metabolism - Regression and PCA

This exercise is an extension of Exercise 7.1.

1. Make regression models between weight at week 14 and all 7 biomarkers. That includes; Scatter plot, regression modeling and check of model assumptions.

2. For the response variable insuline, there are 3 outlying points. Is the results robust towards removal of those?

3. Make a PCA on the biomarkers and plot it using `ggbiplot()`.

4. Infer the external information `$bw_w14$` on the plot using `geom_point(aes(color = ...))`. If you are not satisfied with the colors, then use `scale_color_gradient(low=-..., high=...)` to modify it.

5. Extract the first couple of components and use those as response variables in regression models versus weight.

6. Compile the results, and give an answer to which pattern of biomarkers that are related to weight.

Exercise 7.3 Standard Addition

This exercise shows how regression modeling is used for calculating the concentration of a chemical substance in a sample. The setup is. We have a sample with a molecule of unknown concentration (C). We have an indirect technique which gives a response (y) proportional to the concentration of this molecule. That is $y \propto C$ or $y = \alpha C$. However, we do not know the coefficient α. In order to determine this, a series of five measurements with added volume of the molecule is conduceted. Schematically that results in measurements like these:

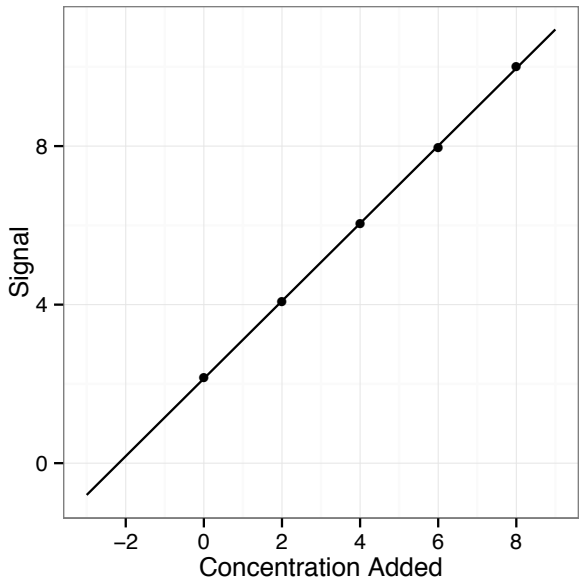


Figure 7.1: Principle in standard addition. Sample concentration is read off in the point where the line crosses the x-axis.

Due to proportionality then the concentration in the sample is simply the value where the line crosses the x-axis, that is: $-x_{y=0}$.

In the table below are listed five measurements from a standard addition experiment.

	1	2	3	4	5
ConcentrationAdded (x)	0.0	0.8	1.5	2.2	3.0
Signal (y)	2.2	3.3	4.0	4.8	6.0

1. Chuck the data into R.
2. Make a scatter plot, and add a straight line indicating the best fit. (If you use `ggplot2` then adding `+ geom_smooth(method = 'lm',fullrange=T)`) to the plot will do it.
3. State a statistical model between y and x, and fit it using either `aov()` or `lm()`.
4. Based on the model parameters (slope and intercept) derive an expression for $-x_{y=0}$.
5. Based on the estimated model, estimate the concentration in the sample.

It should be quit easy to give a central estimate for the concentration. However, giving a confidence interval for this measure is not so easy. In order to do so, confidence limits for the estimated line is used to asses the bounds of the confidence interval.

6. Construct confidence intervals for a broad range of x-values. This is done by constructing a new data frame (`nDT`) and predicting responses for this using the estimated model (here: `m`).
```
>nDT <-data.frame(x = seq(-3,-1,length.out = 1000))
>p <-predict(m,newdata = nDT,interval ="confidence")
```
7. Use the upper, central and lower limits of this series of x-values to estimate the limits for $-x_{y=0}$.

Exercise 7.4 Standard curve – Calcium in milk– *Hand* calculations using R

Milk coagulation is needed during cheese making. A number of factors influence the coagulation properties of milk. Obviously κ-casein is important. However, also the amount of calcium (Ca^{2+}) is essential during milk coagulation.
Calcium in a milk sample may be quantified by atom absorption spectroscopy. This technique relies on Beer's law:

$$Abs = \epsilon * L * C \tag{7.2}$$

Where Abs is the absorbance, ϵ is the molar absorptivity (constant), L is the path length of the sample (constant) and C is the concentration. Hence, Beer's law reveals proportionality between absorbance and concentration.

In this exercise we determine the concentration of Ca^{2+} from a standard curve. Five standard solutions with known Ca^{2+} concentrations were prepared and the absorbance for each sample was measured (See Table below).

ppm Ca^{2+}	Absorbance, $422.7nm$
0.0	0.000
2.0	0.063
5.0	0.141
8.0	0.218
10.0	0.265

Table 7.1: Table – Measured absorbance of standard solutions

Use the following code to get the data into Rstudio and plot the data. We want to find the relationship between concentration of Ca^{2+} and Absorbance. This is a linear model: $y = b + ax + e$
In this case we control the concentration of Ca^{2+} and we measure the absorbance. Hence, in the linear model y is the absorbance and x is the concentration of Ca^{2+}. Here, e is the part of the variation not captured by the model.

```
  Ca <- matrix(c(0.00, 2.00, 5.00, 8.00, 10.00, 0.000, 0.063, 0.141, 0.218, 0.265),ncol = 2)
colnames(Ca) <- c('ppmCa2','Absorbance')
Ca <- data.frame(Ca)
g <- ggplot(data = Ca, mapping = aes(x=ppmCa2, y = Absorbance)) + geom_point()
```

1. Calculate slope and offset of the straight line (the standard curve) that best describes the relationship between Ca^{2+} and absorbance. The slope, a is estimated by

$$\hat{a} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{7.3}$$

   And the offset, b is estimated by

$$\hat{b} = \bar{y} - \hat{a}\bar{x} \tag{7.4}$$

   Here is some R-code to help you:

```
x <- Ca$ppmCa2
y <- Ca$Absorbance
mx <- mean(x)
my <- mean(y)
Sxx <- sum((x-mx)^2)
```

Add the best line to the plot. This can be done by:

```
g <- g + geom_abline(intercept = b, slope = a)
```

2. Test for proportionality, i.e. no offset ($b = 0$). First calculate a confidence interval around your offset. Then state the null hypothesis and the alternative hypothesis. Test the null hypothesis. Does the t-test result correspond with the confidence interval?

   Here is some R code to help you:

```
n <- 5 # number of standard solutions
e <- y-(a*x+b) # model error
```

3. Three new milk samples with unknown concentration of Ca^{2+} were measured with atom absorption spectroscopy. Use the standard curve to quantify the concentrations of Ca^{2+} in the new samples. The absorbance values for the three new samples are found in the table below.

Sample	Absorbance, 422.7nm	Ca^{2+} concentration, ppm
U1	0.101	
U2	0.147	
U3	0.243	

Table 7.2: Table – Absorbance of unknown samples

   Use the following code to get the unknown samples into R and add the samples to the standard curve.

```
Abs.U <- matrix(c(0.101,0.147,0.243)) # measured absorbance
c <-  # Here you estimate the concentrations of unknown samples
U <-cbind(Abs.U,c)
colnames(U) <- c('Abs', 'Conc')
U <- data.frame(U)
# Add the new samples to the standard curve plot
g <- g + geom_point(data = U, mapping = aes(x = Conc, y = Abs), color='red')
```

   In order to get the confidence interval for these estimates, we need to calculate the confidence limits for the estimated regression line.

4. Calculate confidence bounds for a sequence of x-values (concentrations). Add the confidence bounds to the plot.

   Here is some R-Code to help you:

```
xx <- seq(0,10,length.out = 1000) # sequence of x-values (Concentrations)
line.upper <- a*xx+b+ ...
line.lower <- a*xx+b-...
CL <- cbind(xx,line.lower,line.upper)
colnames(CL) <- c('Conc', 'Lower.Limit', 'Upper.Limit')
CL <- data.frame(CL)
# Add confidence interval to the plot
g <- g+geom_line(data = CL, mapping = aes(x=Conc, y=Upper.Limit),color='blue')+
```

```
   geom_line(data = CL, mapping = aes(x=Conc, y=Lower.Limit),color='blue')
```

Look for the point in each confidence band (lower and upper) where y equals the measured absorbance for each sample.

Here is some R-code to help you

```
Upper <- vector()
Lower <- vector()
for (i in 1:3) {
 Upper[i] <- xx[which.min(abs(line.upper-Abs.U[i]))]
 Lower[i] <- xx[which.min(abs(line.lower-Abs.U[i]))]
}
U <- cbind(U,data.frame(cbind(Lower,Upper)))
g+geom_hline(yintercept = Abs.U, linetype = 'dotdash',col = 'red')+
 geom_vline(xintercept = c(Lower,Upper), linetype = 'dotdash',col = 'red')
```

5. Have a look at the widths of the confidence intervals (Lower to Upper) for each estimated concentration of Ca^{2+} (the three new samples). Are the confidence intervals having the same widths? Why/why not?

Exercise 7.5 Standard curve – Quantification of phenol content in spice extracts

Oxidation of food stuff is one of the main reasons leading to decreased quality. During lipid oxidation, lipid radicals are formed, which will increase the speed of oxidation. However, a number of spices contain antioxidants. These antioxidants are often phenol containing molecules, which are able to donate a hydrogen atom to the lipid radicals and thereby decrease the speed of oxidation. By using a standard curve, we will in this exercise estimate the phenol content (and thereby the antioxidative effect) of extracts originating from oregano and tarragon (estragon).

Six standard solutions have been prepared and measured with a spectrophotometer (absorbance at $765nm$). Find the data in the table below.

In the previous exercise on standard curves we did *hand* calculations. In this exercise we will solve the problems using the built-in function in R.

Concentration (mg/L)	Absorbance
2.5	0.2747
5.0	0.5541
7.5	0.8363
6.5	1.2375
10.0	1.0931
12.5	1.3234

Table 7.3: Table – standard solution

1. State the statistical model for the linear relationship between the phenol concentration and the absorbance at $765nm$.

Type the data into Rstudio and plot it using the following code. Here the shaded area corresponds to the confidence interval for the regression line.

```
Conc <- c(2.5, 5.0, 7.5, 6.5, 10.0, 12.5)
Abs <- c(0.2747, 0.5541, 0.8363, 0.7375, 1.0931, 0.7234)
X <- data.frame(cbind(Conc,Abs))
qplot(data = X, Conc, Abs) + geom_smooth(method = 'lm',fullrange = T)
```

2. Fit the model using the function `lm(y~x)`, where `y` is your absorbance values and `x` is the concentrations.

3. State the null hypothesis and the alternative hypothesis for both the slope and the intercept. Check your coefficients with the function `coef(mod)`, where `mod` is the estimated model object. Anything that seems suspicious? Look into the confidence interval for each coefficient. You can get the confidence interval using the function `confint(mod)`. Anything here that seems suspicious? In R, try to call; `summary(mod)$coefficients`. This will return your estimates, std. errors, t-values and the probabilities of your null hypotheses being true. Is the slope significantly different from 0?

4. Inspect the plot of the standard curve and decide whether some points should be excluded from the analyses. If you remove some data points, plot the reduced data, recalculate your model (reduced data) and investigate the coefficients (like question 3).

5. Make a Q-Q plot to investigate whether the model residuals are normally distributed. You can extract your model residuals by `mod_new$residuals`, where `mod_new` is your model object. Are the residuals approximately normally distributed?

6. Use the standard curve to estimate the antioxidative effect (phenol concentration) of the oregano and tarragon extracts. The two extracts were measured with a spectrophotometer (absorbance at $765nm$). Find the absorbance values in the table below.

Sample	Absorbance ($765nm$)	Concentration (mg/L)
Oregano	0.2911	
Tarragon	0.9413	

Table 7.4: Table – Spices

7. Estimate the confidence bounds around the phenol concentration of both Oregano and Tarragon extracts. Again we need to calculate confidence bounds for a sequence of x-values (concentrations) and then look for the point in each confidence band (lower and upper) where y equals the measured absorbance for each sample.

   Here is some code to help you:

```
xx <- seq(0,12,length.out = 1000) # sequence of x-values (Concentrations)
yy <- predict(your_model_object, newdata = data.frame(Conc = xx),interval = 'confidence')

Upper <- vector()
Lower <- vector()
for (i in 1:2) {
 Upper[i] <- xx[which.min(abs(yy[,3]-UK[i,1]))]
 Lower[i] <- xx[which.min(abs(yy[,2]-UK[i,1]))]
}
```

   Use these confidence bounds to decide if the concentration of phenol is different in the two extracts. What extract is having the best antioxidative effect?

7.5 Least squares

The learning objectives for this theme is to understand the mathematical formulation of least squares problems. That is:

- Be able to formulate a objective given a model.
- Understand how the first derivative of the objective is interesting.
- For the most simple examples, with one and two parameters, be able to state the analytical solution that minimizes the objective.
- Based on data, give least squares central estimates for parameters.
- Know that this a very generic framework, which is conducted "under the shelf" by computer algorithms.
- Relate least squares to PCA, ANOVA, linear regression and correlation analysis.

A model (ANOVA, linear regression, PCA,...) can be seen as a representation of the observed data, such that:

$$Observed = Systematic + Residuals \tag{7.5}$$

Here the aim is to choose some parameters for the *Systematic* part, which makes the *Residuals small*. Specifically *small* often refers to a small sum of squares. The enotes describes the case for linear regression, so here we will mention ANOVA problems and PCA.

7.5.1 ANOVA - Least Squares

The formula for an additive ANOVA model with two factors is listed below:

$$\begin{gathered} X_i = \alpha(A_i) + \beta(B_i) + e_i \\ where\ e_i \sim \mathcal{N}(0, \sigma^2)\ and\ independent \\ for\ i = 1, ..., n \end{gathered} \tag{7.6}$$

In this case the aim is to estimate some numbers for the parameters ($\alpha(1), .., \alpha(k_A)$, and $\beta(1), .., \beta(k_B)$) - $k_A + k_B$ in total - such that $\sum e_i^2$ is as small as possible. For what is called balances studies, it turns out that using the group means within each factor, as estimates for $\alpha()$ and $\beta()$, gives the least $\sum e_i^2$ (sum of squared errors). The proof for this is found by calculating $\sum e_i^2$:

$$L = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (X_i - \alpha(A_i) + \beta(B_i))^2 \tag{7.7}$$

Differentiation of this and setting it to zero:

$$\frac{\delta L}{\delta \alpha, \delta \beta} = ... = 0 \tag{7.8}$$

Followed by isolation of the parameters, it is possible to derive the estimates. These are called the *Least Squares* estimates.

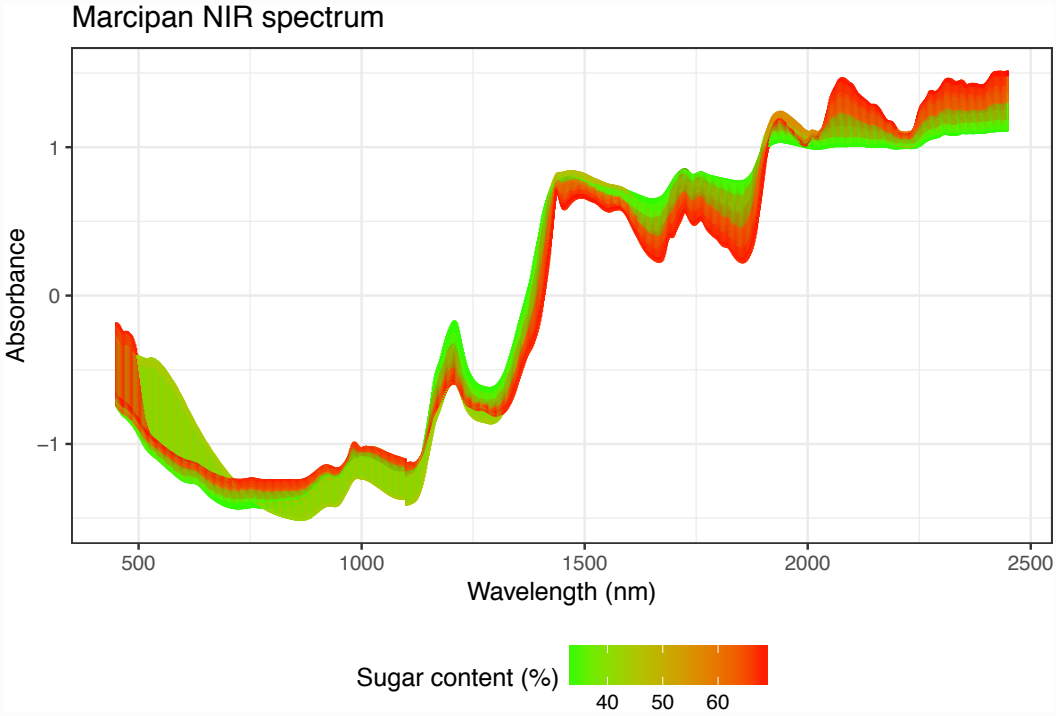
The math is very similar to regression. It is however, beyond the curriculum to be able to do it for ANOVA and PCA problems.

7.5.2 Example: Near Infrared Spectroscopy of Marzipan - Least Squares

The following example illustrates how we can use scores from a principal component analysis to predict the sugar content in marzipan bread (marcipanbrød). The examples is continued from example 1.10.3.

First we import and plot the data:

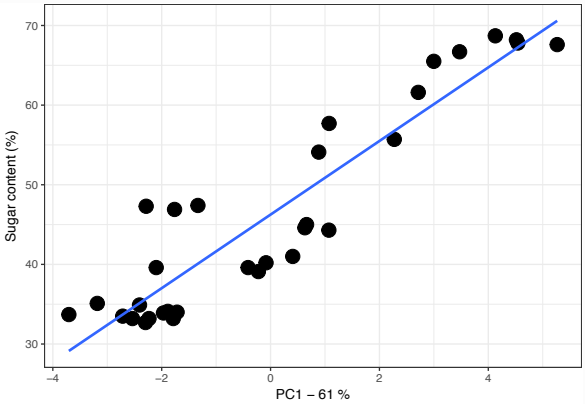
```
# Loading data
load("Marzipan.Rdata")

# Loading libraries for plotting
library(ggplot2)
library(plotly)

# Plotting data according to sugar content
ggplot(data = Xm,aes(x = wavelength, y = value, colour=sugar))+
  geom_line()+theme_bw()+ylab("Absorbance") +
  xlab("Wavelength (nm)") +
  ggtitle("Marcipan NIR spectrum") +
  scale_colour_gradient(low = "green", high="red") +
  theme(legend.position = 'bottom') + # place colorbar below figure
  labs(color='Sugar content (%)')
```



In this example we want to make a model which can predict the sugar content from a spectrum.

We now make a PCA on the data and plot PC1 vs sugar content:

```
# Transposing the data, removing the wavelength column
Xt = t(X[,-1])

# Making PCA on mean centered Xt
marcipan = prcomp(Xt, center=TRUE, scale=FALSE)

# Extracting scores
scores = data.frame(marcipan$x, sugar = Y$sugar)

# Extracting % explained variance
varPC1 = round(summary(marcipan)$importance[2,1]*100)

# Plotting scores, PC1 vs sugar content, coloured according to sugar content
ggplot(data=scores, aes(x=PC1,y=sugar))+
  geom_point(size=5)+
  geom_smooth(method = "lm", se=FALSE) + # Showing regression line
  ylab("Sugar content (%)") +
  xlab(paste("PC1 -",varPC1, "%"))+ # Inserting % explained variance as label
  theme_bw()
```



There is indeed a linear relation between the scores on PC1 and the sugar content in the marzipan breads. Let us make a linear regression model using the least squares approach with sugar content as dependent variable and the scores from PC1 as predictors:

```
linreg = lm(data = scores, sugar ~PC1)
```

From which we can extract the intercept, slope and R^2:

```
summary(linreg)$r.squared
```

```
## [1] 0.8531228
```

```
linreg$coefficients
```

```
## (Intercept)         PC1
##   46.253124    4.620843
```

Our model of sugar content (Y) can be written as:

$$Y = 4.6 \cdot PC1_{scores} + 46.3 \tag{7.9}$$

This model is explaining 85% of the variance of the response variable (sugar content).

7.5.3 Principal Component Analysis - Least Squares

In PCA the multivariate dataset ($\mathbf{X}$) is parameterized by scores ($\mathbf{T}$) and loadings ($\mathbf{P}$):

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{7.10}$$

Contrary to the univariate cases mentioned above, the data and parameters are here matrices, however, the aim is still to find some scores and loadings that minimizes $\mathbf{E}$ in a least squares sense: $\sum \mathbf{E}_{ij}^2$, where i refers to the sample i and j refers to variable j. I.e. $\mathbf{E_{3,4}}$ is the residual for sample 3 variable 4.

7.5.4 Reading Material

A video on LS in linear regression: `https://www.youtube.com/watch?v=0T0z8d0_aY4`
Chapter 5 (5.1 and 5.2) of *Introduction to Statistics* by Brockhoff https://02402.compute.dtu.dk/enotes/book-IntroStatistics

7.5.5 Exercises

Exercise 7.6 Least Squares Estimation

This exercise has the purpose of showing least squares estimation of the center of a distribution.

Let $X_1, X_2, ..., X_n$ be some observed values. The least squares problem for the center of the distribution μ is defined as:

$$L(\mu) = \sum (X_i - \mu)^2 \tag{7.11}$$

1. Show that the solution that minimizes L is $\mu = \bar{X}$ (the mean of X). That is done by differentiation of L with respect to μ and setting this to zero:

$$\frac{\delta L(\mu)}{\delta \mu} = ... = 0 \tag{7.12}$$

2. Simulate some numbers in R using the function `rnorm()` and calculate the mean and median.

3. Calculate $e_{mean,i} = X_i - \bar{X}$ and $e_{median,i} = X_i - X_{median}$. (I.e. subtracting the mean and median from each value).

4. Plot the residuals e_{mean} and e_{median}, and add a horizontal line at 0:
```
>par(mfrow=c(1,2))
> plot(x - mean(x)); lines(c(0,23),c(0,0),col='red')
> plot(x - median(x)); lines(c(0,23),c(0,0),col='red')
```

5. Ad two positive outliers to the data by e.g. `x <- c(rnorm(30),20,23)`, and redo the plotting.

6. Comment on what you see. In case of outliers, which method produces meaningful estimates and residuals?

8. Week 8

This week is going to extend regression to several predictors. Further, least squares for more complicated problems will be pursued.

8.1 Hand-in assignment

The exercise 8.1 *Diet and fat metabolism - Regression with several variables* is to be handed in (through absalon or as hard-copy Wednesday night). You are welcome to put in R-code in the assignment, but it is your argumentation and interpretation that are the most important.

8.2 Exercises

For Monday work through exercise 8.2 to 8.3, and for Wednesday work through 8.4 and 8.5. Further, this week might allow you to recap on some of the exercises you did not make during the last weeks.

8.3 Case IV

The fourth case should be handed in as a slide-show with voice no later than the Friday evening this week.

8.4 Multiple Linear Regression

The learning objectives for this theme is to:

- Graphically show regression problems with several predictors.
- State a statistical model for linear regression with several predictors.
- Compute a regression model.
- Understand the difference between marginal and crude parameter estiamtes.
- Be able to formulate hypothesis for a given question in relation to regression.

8.4.1 In short

Regression models with several predictors are simply extensions of the simple linear regression with a single predictor.

A model with two predictors $X1$ and $X2$ and a single response variable (Y) can be formalized as:

$$\begin{gathered} Y_i = \alpha + \beta_1 \cdot X1_i + \beta_2 \cdot X2_i + e_i \\ where\ e_i \sim \mathcal{N}(0, \sigma^2)\ and\ independent \\ for\ i = 1, ..., n \end{gathered} \tag{8.1}$$

Where α refers to the level of Y at $X1 = X2 = 0$ and β_1 and β_2 are slopes in relation to $X1$ and $X2$.

Each of these three parameters consumes one degree of freedom.

The natural hypothesis is to check whether $X1$ in the presence of $X2$ has an effect on (Y), that is: $H0 : \beta_1 = 0$ with the alternative $HA : \beta_1 \neq 0$ (or the other way around). This can be tested with either an F-test or a T-test (yielding the same results).

Marginal and Crude estimates

In a setup as described above, the investigation of the effect of $X1$ on Y can be done by two approaches:

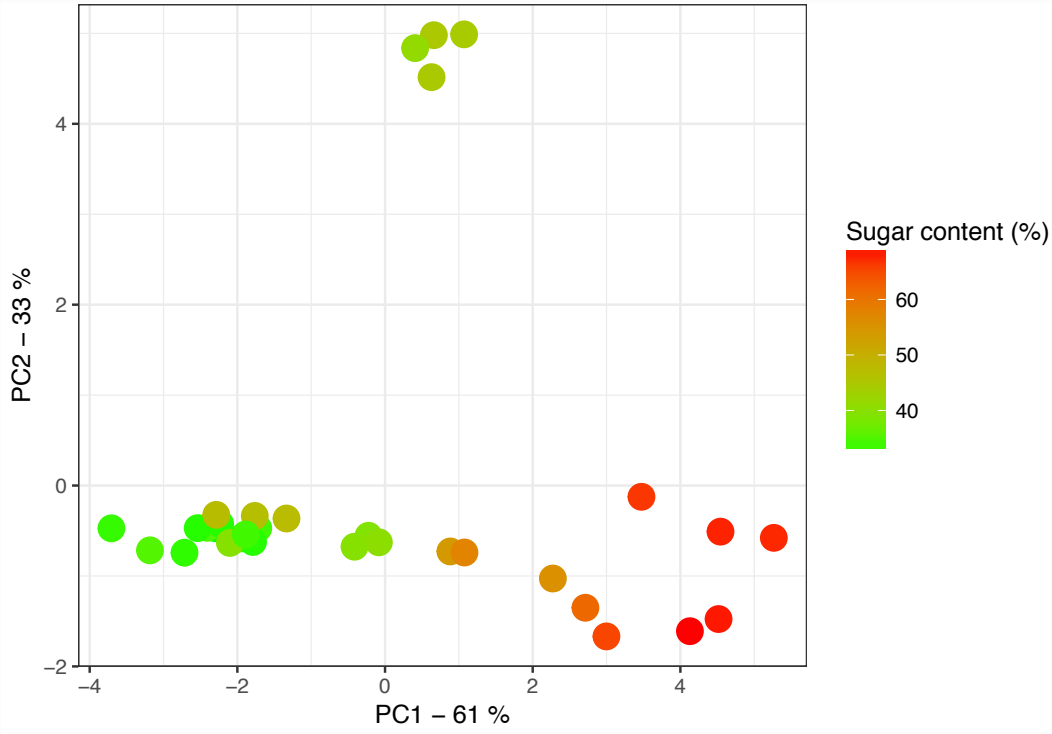
- A simple regression model: $Y_i = \alpha + \beta X1_i + e_i$
- A multiple regression model: $Y_i = \alpha + \beta_1 X1_i + \beta_2 X2_i + e_i$

Naturally these two approaches yields different results due to the absence/presence of the variable $X2$. From the simple regression model β shows the crude effect of $X1$ on Y, whereas in the multiple regression model β_1 shows the marginal effect of $X1$ on Y. Sometimes comparison of these models are referred to as *adjusting* for $X2$ or *controlling* for $X2$. These models often elucidate the direct and indirect relation between predictors and responses as is seen in exercise 8.1

8.4.2 Example: Near Infrared Spectroscopy of Marzipan - Regression

In the following example we want to investigate if we can improve a prediction model by using scores from more than one principal component to predict the sugar content in marzipan bread (marcipanbrød).

The PCA model is exactly the same as in example 1.10.3 and 7.5.2.



We now make two models:

- A linear regression model on the sugar content versus the scores from PC1 (see example 7.5.2).
- A linear regression model on the sugar content versus the scores form PC1 and PC2

```
linreg1 = lm(data = scores, sugar ~PC1)
linreg2 = lm(data = scores, sugar ~PC1+PC2)
```

Let us look at the summary of the two models:

```
summary(linreg1)
```

```
##
## Call:
## lm(formula = sugar ~ PC1, data = scores)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1351 -4.3117 -0.6425  3.6072 11.6073
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  46.2531     0.8872   52.13  < 2e-16 ***
## PC1           4.6208     0.3501   13.20 4.97e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.019 on 30 degrees of freedom
## Multiple R-squared:  0.8531, Adjusted R-squared:  0.8482
## F-statistic: 174.3 on 1 and 30 DF,  p-value: 4.969e-14
```

```
summary(linreg2)
```

```
##
## Call:
## lm(formula = sugar ~ PC1 + PC2, data = scores)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7699 -3.5350 -0.7164  2.7634 11.2274
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  46.2531     0.8063  57.367  < 2e-16 ***
## PC1           4.6208     0.3181  14.526 7.67e-15 ***
## PC2          -1.1724     0.4331  -2.707   0.0113 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.561 on 29 degrees of freedom
## Multiple R-squared:  0.8827, Adjusted R-squared:  0.8747
## F-statistic: 109.2 on 2 and 29 DF,  p-value: 3.178e-14
```

From the summary of linreg2 we see that the PC2 estimate is slightly significantly different from 0 ($p - value = 0.01$) and does contribute to improving the model. If we think back on example 1.10.3, we learned that the second principal component had something to do with the colour of the marzipan bread. In that regard it does not make a lot of sense that adding a colour term (captured by the scores on PC2) into your model improve the prediction of the sugar content. However, when comparigng the R^2 values for the model including only one component ($R^2 = 0.85$) and a model with two components ($R^2 = 0.88$) it is seen that the improvement does not practically make a difference. One should always be carefull not to build models on meaningless data. PCA is a powerfull tool for exploring data, but you need to choose meaningfull principal components for your analysis. This is

also relevant when selecting variables for the analysis.

Maybe the wavelengths covering the visible part of the spectrum should have been left out of the analysis from the beginning as they do not contain information about the sugar content?

8.4.3 Reading Material

A video on multiple linear regression https://www.youtube.com/watch?v=G4ZlC9zKfII
Chapter 6.1 of *Introduction to Statistics* by Brockhoff https://02402.compute.dtu.dk/enotes/book-IntroStatistics

8.4.4 Exercises

Exercise 8.1 Diet and fat metabolism - Regression with several variables

This exercise examines the relation between a biomarker, dietary intervention and weight in order to disentangle the effects causing elevated levels of cholesterol.
The data for this exercise is the same as for Exercise 3.2
In this exercise we are going to focus on predicting cholesterol from both weight and dietary intervention.

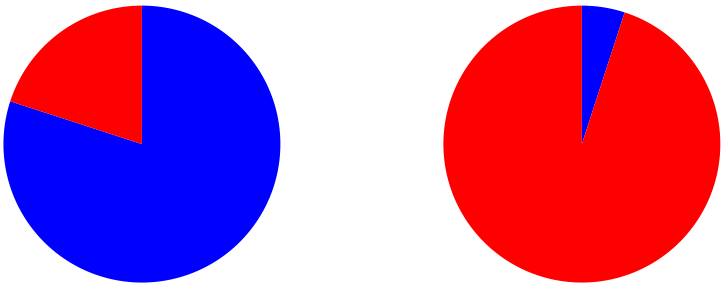
1. Plot, formulate and build univariate models predicting cholesterol level from:
   (a) Dietary intervention (`$Fat_Protein`) - A oneway ANOVA model.
   (b) Body weight at 14 weeks (`$bw_w14`) - A regression model.
2. State and test relevant hypothesis for these two models.
3. Comment on the relations.
4. Make a scatter plot of cholesterol versus weight colored or shaped according to dietary intervention.
5. Make a regression model for cholesterol with two predictors: i) Weight at week 14 and ii) Dietary intervention, and test the factors.

   OBS: In contrast to anova with one dependent variable, the order of the dependent variables makes a differences in the `anova()`. In order to surpass this, use the `drop1(aov(),test='F')` function to get results independent of order.
6. What happens to the significance of the individual factors?
7. Do you think that the dietary intervention *directly* affects the cholesterol level, or that is is mediated through body weight?

8.5 Explained Variance

The learning objectives for this theme is to understand the general notion of Explained Variance and its relation to least squares.

Explained variance refers to how well the estimated model describes the observed data. That is, how much of the variation is systematic and how much is residual. In the figure below are shown two examples.



Model which describes data fairly well (High R^2)

Model which describes data poorly (Low R^2)

Figure 8.1: Blue - Systematic variation, Red - Residual variation

The explained variance is summarized in the so-called $R^2 \in [0, 1]$ metric. Where a value close to 1 indicates high degree of explained variance, and a value close to 0 the contrary.

R^2 is based on the sums of squares of the different model terms. Given the model:

$$X = S + E \tag{8.2}$$

Where X refers to the observed data, S the systematic parameterized part of the model and E the residuals, then:

$$SS(X) = SS(S) + SS(E) \tag{8.3}$$

I.e. the sums of squares are additive. From this the R^2 value is defined as:

$$R^2 = \frac{SS(S)}{SS(X)} = \frac{SS(X) - SS(E)}{SS(X)} = 1 - \frac{SS(E)}{SS(X)} \tag{8.4}$$

In both ANOVA and linear regression these $SS()$ values are readily available from the analysis table, where $SS(X) = SS_{tot}$ and $SS(E) = SS_e$. Further, it is possible to calculate the R^2 for the entire model, but also for individual factors in twoway ANOVA and regression with multiple descriptors, simply by using the individual $SS()$ contributions.

Model estimates ($\hat{y}$) and R^2

In linear regression the model is stated as:

$$\begin{gathered} y_i = \alpha + \beta \cdot x_i + e_i \\ where\ e_i \sim \mathcal{N}(0, \sigma^2)\ and\ independent \\ for\ i = 1, ..., n \end{gathered} \tag{8.5}$$

The predicted values for y (referred to as $\hat{y}$) can be expressed as:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} \cdot x_i \tag{8.6}$$

where the estimates of the slope and the intercept is used for calculating y based on x.
The R^2 value for such a model can also be seen as:

$$R^2 = r^2_{\hat{y},y} \tag{8.7}$$

That is, taking the squared value of the correlation coefficient between the observed response and the predicted response yields the explained variance by the model.

This concept is valid for for a range of models including ANOVA, PCA and linear regression.

Correlation coefficient and R^2

In bi-variate correlation analysis, the correlation coefficient (see e.g. equation 2.5) is direcly related to R^2, as: I

The implication of this, is that the *direction* is lost ($R^2 = 0.3^2 = (-0.3)^2$), but the *strength* of the relation is not.

R^2 for PCA

In a PCA model, it is possible to talk about R^2 values for the entire model or for the individual components. It is so, by definition, that the contribution from the individual components is decreasing, that is:

$$R^2_{PC1} \geq R^2_{PC2} \geq \ldots \geq R^2_{PCk} \tag{8.8}$$

Further, from a k component model follows that total explained variance is:

$$R^2_{PC1-PCk} = R^2_{PC1} + R^2_{PC2} + \ldots + R^2_{PCk} = \sum_{i=1}^{k} R^2_{PCi} \tag{8.9}$$

OBS OBS OBS In PCA, by increasing the number of components you will eventually reach explained variance of 1. This is due to the parameters of this model (scores and loadings - see equation 7.10) is solely based on the response variables, so there is no set of independent variables, as there is in ANOVA and linear regression. This has the implication, that comparing R^2 values between e.g. an ANOVA model and a PCA model is not at all straight forward.

8.5.1 Reading Material

A video on R^2 `https://www.youtube.com/watch?v=IMjrEeeDB-Y`

8.5.2 Exercises

Exercise 8.2 Explained Variance (R^2) - Regression n' Correlation

This exercise deals with explained variance and its relation to the residual deviation. This exercise uses the dietary intervention on mice and the effect on biomarkers related to fat metabolism.

1. First, we are going to deal with insulin and cholesterol. Make three analyses:
   (a) Regress cholesterol on insulin.
   (b) Regress insulin on cholesterol.
   (c) Make a correlation analysis between the two.
2. Make a drawing (not exact, just some dots) of cholesterol versus insulin and indicate a regression line, and the the residuals for the first two models in Q1.
3. For each model, calculate R^2. (HINT: You can use the SS measures of a `aov()` model, or use `summary()` of a `lm()` model directly).
4. Comment on what you observe.

Exercise 8.3 R^2 and outliers

This exercise should show how extreme outliers can influence the R^2 measure by making it unrealistically high.

1. Simulate two sets of vector x and y, each of $n = 15$ points which are not related. (use `rnorm()` in R for doing so).
2. Scatter plot x and y and add the best regression line. (code hint: `qplot(x,y) + stat_smooth(method='lm')`)
3. Calcualte the correlation coefficient (r) and the R^2 value.
4. Do the stats (r and R^2) and the plot tells the same story?
5. Now add an extreme point to both x and y (`x <- c(x,extremenumber)`).
6. Repeat plotting and stat calculation (r and R^2).
7. What happended?

TAKE HOME: R^2 alone without visual inspection can be misleading.

Exercise 8.4 R^2 and transformations

This exercise should show how transformation influences the R^2 measure. The exercise is an extension of exercise 2.6 *Transformations and the Normal distribution*, and is using the wine dataset.

1. Import data, and make a plot of the response variable `..$Ethyl.pyruvate` infering country membership.
2. Make a oneway anova model for this response variable, and check (or calculate) the R^2 value.
3. Make a transformation of the response variable, and repeat plotting, modelling and R^2 calculation.
4. Compare the R^2 for the raw response and transformed response, and figure out (based on the plots) which samples that are causing the difference.
5. Make a check of model assumptions for both models, and try to fix the isssues by outlier removal. HINT: a simple way to remove samples and update models is:
   ```
   > ic <- wine$Ethyl.pyruvate < ...
   > m <- lm(data=wine[ic,],...)
   ```

Exercise 8.5 Explained Variance and PCA

This exercise deals with explained variance in relation to PCA. This exercise uses the dietary intervention on mice and the effect on biomarkers related to fat metabolism.

1. Initially extract and scale the biomarkers in the dataset (use `X <- scale()` for this purpose).
2. Calculate the total sums of squares for these data? (`sum(X^2)`)
3. Calculate a PCA model on these preprocessed data, and plot it using `ggbiplot()`.
4. Calculate the residuals after the first component. This is done by calculating the predicted values and subtracting those from the dataa:
   `>Xhat <- mm$x[,1]%o%mm$rotation[,1]`
   `>E <- X - Xhat` (where `mm` refers to the PCA model build by `prcomp()`)
5. Calculate the residual sums of squares after removing the first component.

6. Calculate the R^2 for the first component.
7. Try to calculate the R^2 value from the correlation between $\mathbf{X}$ and $\hat{\mathbf{X}}$:
   `>cor(as.vector(X),as.vector(Xhat))`
8. Try to modify the code in Q4 to be able to calculate for component 2, 3,...

Eventually you can match your results with the ones produced by `ggbiplot()`.