Chapter 1

Introduction, descriptive statistics, R and data visualization

Chapter 1

Contents

This is the first chapter in the eight-chapter DTU Introduction to Statistics book. It consists of eight chapters:

1. Introduction, descriptive statistics, R and data visualization
2. Probability and simulation
3. Statistical analysis of one and two sample data
4. Statistics by simulation
5. Simple linear regression
6. Multiple linear regression
7. Analysis of categorical data
8. Analysis of variance (analysis of multi-group data)

In this first chapter the idea of statistics is introduced together with some of the basic summary statistics and data visualization methods. The software used throughout the book for working with statistics, probability and data analysis is the open source environment R. An introduction to R is included in this chapter.

1.1 What is Statistics - a primer

To catch your attention we will start out trying to give an impression of the importance of statistics in modern science and engineering.

In the well respected *New England Journal of medicine* a millennium editorial on the development of medical research in a thousand years was written:

EDITORIAL: Looking Back on the Millennium in Medicine, *N Engl J Med*, 342:42-49, January 6, 2000, NEJM200001063420108.

They came up with a list of 11 points summarizing the most important developments for the health of mankind in a millennium:

- Elucidation of human anatomy and physiology
- Discovery of cells and their substructures
- Elucidation of the chemistry of life
- Application of statistics to medicine
- Development of anaesthesia

- Discovery of the relation of microbes to disease
- Elucidation of inheritance and genetics
- Knowledge of the immune system
- Development of body imaging
- Discovery of antimicrobial agents
- Development of molecular pharmacotherapy

The reason for showing the list here is pretty obvious: one of the points is *Application of Statistics to Medicine*! Considering the other points on the list, and what the state of medical knowledge was around 1000 years ago, it is obviously a very impressive list of developments. The reasons for statistics to be on this list are several and we mention two very important historical landmarks here. Quoting the paper:

"One of the earliest clinical trials took place in 1747, when James Lind treated 12 scorbutic ship passengers with cider, an elixir of vitriol, vinegar, sea water, oranges and lemons, or an electuary recommended by the ship's surgeon. The success of the citrus-containing treatment eventually led the British Admiralty to mandate the provision of lime juice to all sailors, thereby eliminating scurvy from the navy." (See also James_Lind).

Still today, clinical trials, including the statistical analysis of the outcomes, are taking place in massive numbers. The medical industry needs to do this in order to find out if their new developed drugs are working and to provide documentation to have them accepted for the World markets. The medical industry is probably the sector recruiting the highest number of statisticians among all sectors. Another quote from the paper:

"The origin of modern epidemiology is often traced to 1854, when John Snow demonstrated the transmission of cholera from contaminated water by analyzing disease rates among citizens served by the Broad Street Pump in London's Golden Square. He arrested the further spread of the disease by removing the pump handle from the polluted well." (See also John_Snow_(physician)).

Still today, epidemiology, both human and veterinarian, maintains to be an extremely important field of research (and still using a lot of statistics). An important topic, for instance, is the spread of diseases in populations, e.g. virus spreads like Ebola and others.

Actually, today more numbers/data than ever are being collected and the amounts are still increasing exponentially. One example is Internet data, that internet companies like Google, Facebook, IBM and others are using extensively. A quote from New York Times, 5. August 2009, from the article titled “For Today's Graduate, Just One Word: Statistics" is:

"I keep saying that the sexy job in the next 10 years will be statisticians," said Hal Varian, chief economist at Google. 'and I'm not kidding.' "

The article ends with the following quote:

"The key is to let computers do what they are good at, which is trawling these massive data sets for something that is mathematically odd," said Daniel Gruhl, an I.B.M. researcher whose recent work includes mining medical data to improve treatment. "And that makes it easier for humans to do what they are good at - explain those anomalies."

1.2 Statistics at DTU Compute

At DTU Compute at the Technical University of Denmark statistics is used, taught and researched mainly within four research sections:

- Statistics and Data Analysis
- Dynamical Systems
- Image Analysis & Computer Graphics
- Cognitive Systems

Each of these sections have their own focus area within statistics, modelling and data analysis. On the master level it is an important option within DTU Compute studies to specialize in statistics of some kind on the joint master programme in Mathematical Modelling and Computation (MMC). And a *Statistician* is a well-known profession in industry, research and public sector institutions.

The high relevance of the topic of statistics and data analysis today is also illustrated by the extensive list of ongoing research projects involving many and diverse industrial partners within these four sections. Neither society nor industry can cope with all the available data without using highly specialized people in statistical techniques, nor can they cope and be internationally competitive without continuously further developing these methodologies in research projects. Statistics is and will continue to be a relevant, viable and dynamic field. And the amount of experts in the field continues to be small compared to the demand for experts, hence obtaining skills in statistics is for sure a wise career choice for an engineer. Still for any engineer not specialising in statistics, a basic level of statistics understanding and data handling ability is crucial for the ability to navigate in modern society and business, which will be heavily influenced by data of many kinds in the future.

1.3 Statistics - why, what, how?

Often in society and media, the word *statistics* is used simply as the name for a summary of some numbers, also called data, by means of a summary table and/or plot. We also embrace this basic notion of statistics, but will call such basic data summaries *descriptive statistics* or *explorative statistics*. The meaning of *statistics* goes beyond this and will rather mean *"how to learn from data in an insightful way and how to use data for clever decision making"*, in short we call this *inferential statistics* . This could be on the national/societal level, and could be related to any kind of topic, such as e.g. health, economy or environment, where data is collected and used for learning and decision making. For example:

- Cancer registries
- Health registries in general
- Nutritional databases
- Climate data
- Macro economic data (Unemployment rates, GNP etc.)
- etc.

The latter is the type of data that historically gave name to the word *statistics*. It originates from the Latin 'statisticum collegium' (state advisor) and the Italian word 'statista' (statesman/politician). The word was brought to Denmark by the Gottfried Achenwall from Germany in 1749 and originally described the processing of data for the state, see also History_of_statistics.

Or it could be for industrial and business applications:

- Is machine A more effective than machine B?
- How many products are we selling on different markets?
- Predicting wind and solar power for optimizing energy systems
- Do we produce at the specified quality level?
- Experiments and surveys for innovative product development
- Drug development at all levels at e.g. Novo Nordisk A/S or other pharmaceutical companies
- Learning from "Big Data"
- etc.

In general, it can be said say that we learn from data by analysing the data with statistical methods. Therefore *statistics* will in practice involve *mathematical* *modelling*, i.e. using some linear or non-linear function to model the particular phenomenon. Similarly, the use of *probability theory* as the concept to describe randomness is extremely important and at the heart of being able to "be clever" in our use of the data. Randomness express that the data just as well could have come up differently due to the inherent random nature of the data collection and the phenomenon we are investigating.

Probability theory is in its own right an important topic in engineering relevant applied mathematics. Probability based modelling is used for e.g. queuing systems (queuing for e.g. servers, websites, call centers etc.), for reliability modelling, and for risk analysis in general. Risk analysis encompasses a vast diversity of engineering fields: food safety risk (toxicological and/or allergenic), environmental risk, civil engineering risks, e.g. risk analysis of large building constructions, transport risk, etc. The present material focuses on the statistical issues, and treats probability theory at a minimum level, focusing solely on the purpose of being able to do proper *statistical inference* and leaving more elaborate probability theory and modelling to other texts.

There is a conceptual frame for doing *statistical inference*: in *Statistical inference* the observed data is a *sample*, that is (has been) taken from a *population*. Based on the sample, we try to generalize to (infer about) the population. Formal definitions of what the sample and the population is are given by:

|||| Definition 1.1 Sample and population

- An *observational unit* is the single entity about which information is sought (e.g. a person)
- An *observational variable* is a property which can be measured on the observational unit (e.g. the height of a person)
- The *statistical population* consists of the value of the observational variable for all observational units (e.g. the heights of all people in Denmark)
- The *sample* is a subset of the statistical population, which has been chosen to represent the population (e.g. the heights of 20 persons in Denmark).

See also the illustration in Figure 1.1.

This is all a bit abstract at this point. And likely adding to the potential confusion about this is the fact that the words *population* and *sample* will have a "less

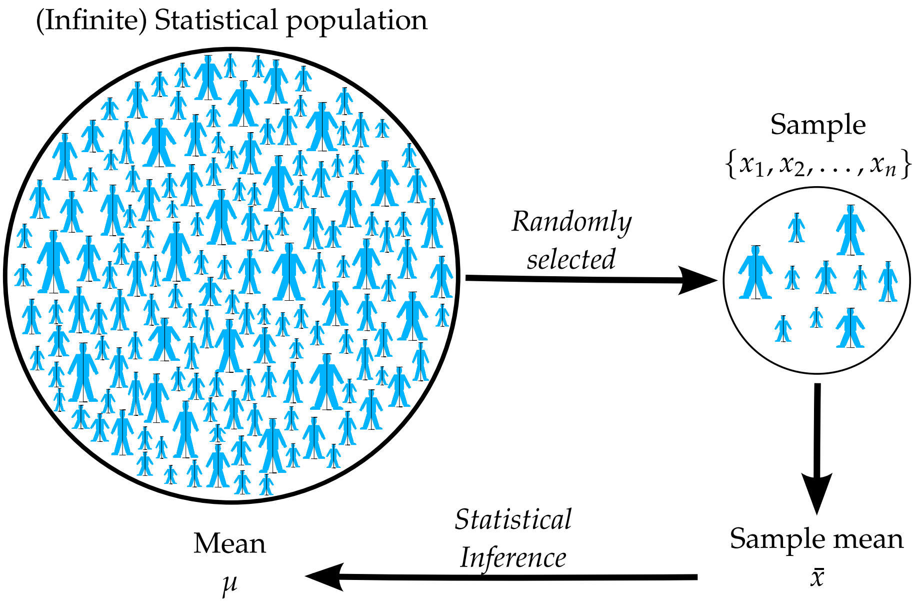


Figure 1.1: Illustration of statistical population and sample, and statistical inference. Note that the bar on each person indicates that the it is the height (the observational variable) and not the person (the observational unit), which are the elements in the statistical population and the sample. Notice, that in all analysis methods presented in this text the statistical population is assumed to be very large (or infinite) compared to the sample size.

precise" meaning when used in everyday language. When they are used in a statistical context the meaning is very specific, as given by the definition above. Let us consider a simple example:

|||| Example 1.2

The following study is carried out (actual data collection): the height of 20 persons in Denmark is measured. This will give us 20 values $x_1, \ldots, x_{20}$ in cm. The *sample* is then simply these 20 values. The statistical *population* is the height values of all people in Denmark. The *observational unit* is a person.

The meaning of *sample* in statistics is clearly different from how a chemist or medical doctor would use the word, where a sample would be the actual substance in e.g. the petri dish. Within this book, when using the word sample, then it is always in the statistical meaning i.e. a set of values taken from a statistical population.

With regards to the meaning of *population* within statistics the difference to the everyday meaning is less obvious: but note that the *statistical population* in the example is defined to be the height values of people, not actually the people. Had we measured the weights instead the statistical population would be quite different. Also later we will realize that statistical populations in engineering contexts can refer to many other things than populations as in a group of organisms, hence stretching the use of the word beyond the everyday meaning. From this point: *population* will be used instead of *statistical population* in order to simplify the text.

The population in a given situation will be linked with the actual study and/or experiment carried out - the data collection procedure sometimes also denoted the *data generating process*. For the sample to represent relevant information about the population it should be *representative* for that population. In the example, had we only measured male heights, the population we can say anything about would be the male height population only, not the entire height population.

A way to achieve a representative sample is that each observation (i.e. each value) selected from the population, is randomly and independently selected of each other, and then the sample is called a *random sample*.

1.4 Summary statistics

The descriptive part of studying data maintains to be an important part of statistics. This implies that it is recommended to study the given data, the sample, by means of *descriptive statistics* as a first step, even though the purpose of a full statistical analysis is to eventually perform some of the new inferential tools taught in this book, that will go beyond the pure descriptive part. The aims of the initial descriptive part are several, and when moving to more complex data settings later in the book, it will be even more clear how the initial descriptive part serves as a way to prepare for and guide yourself in the subsequent more formal inferential statistical analysis.

The initial part is also called an *explorative* analysis of the data. We use a number of summary statistics to summarize and describe a sample consisting of one or two variables:

- Measures of centrality:
  - Mean
  - Median
  - Quantiles

- Measures of "spread":
  - Variance
  - Standard deviation
  - Coefficient of variation
  - Inter Quartile Range (IQR)
- Measures of relation (between two variables):
  - Covariance
  - Correlation

One important point to notice is that these statistics can only be calculated for the sample and not for the population - we simply don't know all the values in the population! But we want to learn about the population from the sample. For example when we have a random sample from a population we say that the *sample mean* ($\bar{x}$) is an *estimate* of the *mean* of the population, often then denoted μ, as illustrated in Figure 1.1.

|||| Remark 1.3

Notice, that we put 'sample' in front of the name of the statistic, when it is calculated for the sample, but we don't put 'population' in front when we refer to it for the population (e.g. we can think of the *mean* as the true mean).

HOWEVER we don't put *sample* in front of the name every time it should be there! This is to keep the text simpler and since traditionally this is not strictly done, for example the median is rarely called the sample median, even though it makes perfect sense to distinguish between the sample median and the median (i.e. the population median). Further, it should be clear from the context if the statistic refers to the sample or the population, when it is not clear then we distinguish in the text. Most of the way we do distinguish strictly for the *mean*, *standard deviation*, *variance*, *covariance* and *correlation*.

1.4.1 Measures of centrality

The sample mean is a key number that indicates the centre of gravity or centring of the sample. Given a sample of n observations $x_1, \ldots, x_n$, it is defined as follows:

Definition 1.4 Sample mean

The sample mean is the sum of observations divided by the number of observations

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{1-1}$$

Sometimes this is refereed to as the *average*.

The median is also a key number indicating the center of sample (note that to be strict we should call it 'sample median', see Remark 1.3 above). In some cases, for example in the case of extreme values or skewed distributions, the median can be preferable to the mean. The median is the observation in the middle of the sample (in sorted order). One may express the ordered observations as $x_{(1)}, \ldots, x_{(n)}$, where then $x_{(1)}$ is the smallest of all $x_1, \ldots, x_n$ (also called the minimum) and $x_{(n)}$ is the largest of all $x_1, \ldots, x_n$ (also called the maximum).

Definition 1.5 Median

Order the n observations $x_1, \ldots, x_n$ from the smallest to largest: $x_{(1)}, \ldots, x_{(n)}$. The median is defined as:

- If n is odd the median is the observation in position $\frac{n+1}{2}$:

$$Q_2 = x_{(\frac{n+1}{2})}. \tag{1-2}$$

- If n is even the median is the average of the two observations in positions $\frac{n}{2}$ and $\frac{n+2}{2}$:

$$Q_2 = \frac{x_{(\frac{n}{2})} + x_{(\frac{n+2}{2})}}{2}. \tag{1-3}$$

The reason why it is denoted with Q_2 is explained below in Definition 1.8.

|||| Example 1.6 Student heights

A random sample of the heights (in cm) of 10 students in a statistics class was

$$168 \quad 161 \quad 167 \quad 179 \quad 184 \quad 166 \quad 198 \quad 187 \quad 191 \quad 179\,.$$

The sample mean height is

$$\bar{x} = \frac{1}{10}(168 + 161 + 167 + 179 + 184 + 166 + 198 + 187 + 191 + 179) = 178.$$

To find the sample median we first order the observations from smallest to largest

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$
161	166	167	168	179	179	184	187	191	198

.

Note that having duplicate observations (like e.g. two of 179) is not a problem - they all just have to appear in the ordered list. Since $n = 10$ is an even number the median becomes the average of the 5th and 6th observations

$$\frac{x_{(\frac{n}{2})} + x_{(\frac{n+2}{2})}}{2} = \frac{x_{(5)} + x_{(6)}}{2} = \frac{179 + 179}{2} = 179.$$

As an illustration, let's look at the results if the sample did not include the 198 cm height, hence for $n = 9$

$$\bar{x} = \frac{1}{9}(168 + 161 + 167 + 179 + 184 + 166 + 187 + 191 + 179) = 175.78.$$

then the median would have been

$$x_{(\frac{n+1}{2})} = x_{(5)} = 179.$$

This illustrates the robustness of the median compared to the sample mean: the sample mean changes a lot more by the inclusion/exclusion of a single "extreme" measurement. Similarly, it is clear that the median does not depend at all on the actual values of the most extreme ones.

The median is the point that divides the observations into two halves. It is of course possible to find other points that divide into other proportions, they are called quantiles or percentiles (note, that this is actually the *sample quantile* or *sample percentile*, see Remark 1.3).

Definition 1.7 Quantiles and percentiles

The p *quantile* also called the $100p\%$ quantile or $100p$'th *percentile*, can be defined by the following procedure: [a]

1. Order the n observations from smallest to largest: $x_{(1)}, \ldots, x_{(n)}$
2. Compute pn
3. If pn is an integer: average the pn'th and $(pn+1)$'th ordered observations. Then the p quantile is

$$q_p = \left(x_{(np)} + x_{(np+1)}\right)/2 \tag{1-4}$$

4. If pn is a non-integer: take the "next one" in the ordered list. Then the p'th quantile is

$$q_p = x_{(\lceil np \rceil)}, \tag{1-5}$$

where $\lceil np \rceil$ is the *ceiling* of np, that is, the smallest integer larger than np

[a] There exist several other formal definitions. To obtain this definition of quantiles/percentiles in R use quantile(..., type=2). Using the default in R is also a perfectly valid approach - just a different one.

Often calculated percentiles are the so-called *quartiles* (splitting the sample in quarters, i.e. 0%, 25%, 50%, 75% and 100%):

- q_0, $q_{0.25}$, $q_{0.50}$, $q_{0.75}$ and q_1

Note that the 0'th percentile is the minimum (smallest) observation and the 100'th percentile is the maximum (largest) observation. We have specific names for the three other quartiles:

Definition 1.8 Quartiles

Q_1	$= q_{0.25}$	= "lower quartile"	= "0.25 quantile"	= "25'th percentile"
Q_2	$= q_{0.50}$	= "median"	= "0.50 quantile"	= "50'th percentile"
Q_3	$= q_{0.75}$	= "upper quartile"	= "0.75 quartile"	= "75'th percentile"

|||| Example 1.9 Student heights

Using the $n = 10$ sample from Example 1.6 and the ordered data table from there, let us find the lower and upper quartiles (i.e. Q_1 and Q_3), as we already found $Q_2 = 179$.

First, the Q_1: with $p = 0.25$, we get that $np = 2.5$ and we find that

$$Q_1 = x_{(\lceil 2.5 \rceil)} = x_{(3)} = 167,$$

and since $n \cdot 0.75 = 7.5$, the upper quartile becomes

$$Q_3 = x_{(\lceil 7.5 \rceil)} = x_{(8)} = 187.$$

We could also find the 0'th percentile

$$q_0 = \min(x_1, \dots, x_n) = x_{(1)} = 161,$$

and the 100'th percentile

$$q_1 = \max(x_1, \dots, x_n) = x_{(10)} = 198.$$

Finally, 10'th percentile (i.e. 0.10 quantile) is

$$q_{0.10} = \frac{x_{(1)} + x_{(2)}}{2} = \frac{161 + 166}{2} = 163.5,$$

since $np = 1$ for $p = 0.10$.

1.4.2 Measures of variability

A crucial aspect to understand when dealing with statistics is the concept of variability - the obvious fact that not everyone in a population, nor in a sample, will be exactly the same. If that was the case they would all equal the mean of the population or sample. But different phenomena will have different degrees of variation: An adult (non dwarf) height population will maybe spread from around 150 cm up to around 210 cm with very few exceptions. A kitchen scale measurement error population might span from -5 g to $+5$ g. We need a way to quantify the degree of variability in a population and in a sample. The most commonly used measure of sample variability is the sample variance or its square root, called the sample standard deviation:

Definition 1.10 Sample variance

The *sample variance* of a sample $x_1, \ldots, x_n$ is the sum of squared differences from the sample mean divided by $n - 1$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{1-6}$$

Definition 1.11 Sample standard deviation

The *sample standard deviation* is the square root of the sample variance

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{1-7}$$

The sample standard deviation and the sample variance are key numbers of absolute variation. If it is of interest to compare variation between different samples, it might be a good idea to use a relative measure - most obvious is the coefficient of variation:

Definition 1.12 Coefficient of variation

The *coefficient of variation* is the sample standard deviation seen relative to the sample mean

$$V = \frac{s}{\bar{x}}. \tag{1-8}$$

We interpret the standard deviation as the *average absolute deviation from the mean* or simply: the *average level of differences*, and this is by far the most used measure of spread. Two (relevant) questions are often asked at this point (it is perfectly fine if you didn't wonder about them by now and you might skip the answers and return to them later):

|||| Remark 1.13

Question: Why not actually compute directly what the interpretation is stating, which would be: $\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$?

Answer: This is indeed an alternative, called the *mean absolute deviation*, that one could use. The reason for most often measuring "mean deviation" NOT by the *Mean Absolute Deviation* statistic, but rather by the sample standard deviation s, is the so-called theoretical statistical properties of the sample variance s^2. This is a bit early in the material for going into details about this, but in short: inferential statistics is heavily based on probability considerations, and it turns out that it is theoretically much easier to put probabilities related to the sample variance s^2 on explicit mathematical formulas than probabilities related to most other alternative measures of variability. Further, in many cases this choice is in fact also the optimal choice in many ways.

|||| Remark 1.14

Question: Why divide by $n-1$ and not n in the formulas of s and s^2? (which *also* appears to fit better with the stated interpretation)

Answer: The sample variance s^2 will most often be used as an estimate of the (true but unknown) population variance σ^2, which is the average of $(x_i - \mu)^2$ in the population. In doing that, one should ideally compare each observation x_i with the population mean, usually called μ. However, we do not know μ and instead we use $\bar{x}$ in the computation of s^2. In doing so, the squared differences $(x_i - \bar{x})^2$ that we compute in this way will tend to be slightly smaller than those we ideally should have used: $(x_i - \mu)^2$ (as the observations themselves were used to find $\bar{x}$ so they will be closer to $\bar{x}$ than to μ). It turns out, that the correct way to correct for this is by dividing by $n-1$ instead of n.

Spread in the sample can also be described and quantified by quartiles:

Definition 1.15 Range

The *range* of the sample is

$$\text{Range} = \text{Maximum} - \text{Minimum} = Q_4 - Q_0 = x_{(n)} - x_{(1)}. \tag{1-9}$$

The Inter Quartile Range (IQR) is the middle 50% range of data defined as

$$IQR = q_{0.75} - q_{0.25} = Q_3 - Q_1. \tag{1-10}$$

Example 1.16 Student heights

Consider again the $n = 10$ data from Example 1.6. To find the variance let us compute the $n = 10$ differences to the mean, that is $(x_i - 178)$

-10 -17 -11 1 6 -12 20 9 13 1 .

So, if we square these and add them up we get

$$\sum_{i=1}^{10}(x_i - \bar{x})^2 = 10^2 + 17^2 + 11^2 + 1^2 + 6^2 + 12^2 + 20^2 + 9^2 + 13^2 + 1^2 = 1342.$$

Therefore the sample variance is

$$s^2 = \frac{1}{9}1342 = 149.1,$$

and the sample standard deviation is

$$s = 12.21.$$

We can interpret this as: people are on average around 12 cm away from the mean height of 178 cm. The Range and Inter Quartile Range (IQR) are easily found from the ordered data table in Example 1.6 and the earlier found quartiles in Example 1.9

$$\text{Range} = \text{maximum} - \text{minimum} = 198 - 161 = 37,$$

$$IQR = Q_3 - Q_1 = 187 - 167 = 20.$$

Hence 50% of all people (in the sample) lie within 20 cm.

Note, that the standard deviation in the example has the physical unit cm, whereas the variance has cm^2. This illustrates the fact that the standard deviation has a more direct interpretation than the variance in general.

1.4.3 Measures of relation: correlation and covariance

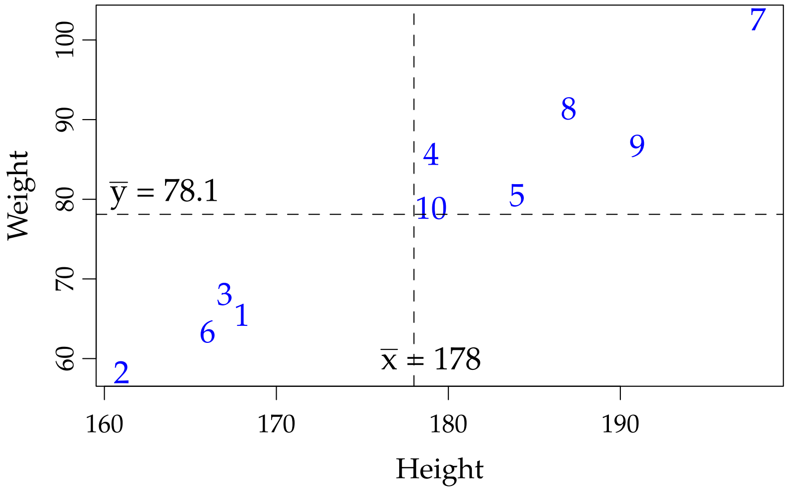
When two observational variables are available for each observational unit, it may be of interest to quantify the relation between the two, that is to quantify how the two variables *co-vary* with each other, their *sample covariance* and/or *sample correlation*.

|||| Example 1.17 Student heights and weights

In addition to the previously given student heights we also have their weights (in kg) available

Heights (x_i)	168	161	167	179	184	166	198	187	191	179
Weights (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9 .

The relation between weights and heights can be illustrated by the so-called scatterplot, cf. Section 1.6.4, where e.g. weights are plotted versus heights:



Each point in the plot corresponds to one student - here illustrated by using the observation number as plot symbol. The (expected) relation is pretty clear now - different wordings could be used for what we see:

- Weights and heights are related to each other

- Higher students tend to weigh more than smaller students
- There is an increasing pattern from left to right in the "point cloud”
- If the point cloud is seen as an (approximate) ellipse, then the ellipse clearly is horizontally upwards ”tilted”.
- Weights and heights are (positively) *correlated* to each other

The sample covariance and sample correlation coefficients are a summary statistics that can be calculated for two (related) sets of observations. They quantify the (linear) strength of the relation between the two. They are calculated by combining the two sets of observations (and the means and standard deviations from the two) in the following ways:

|||| Definition 1.18 Sample covariance

The sample covariance is

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right). \tag{1-11}$$

|||| Definition 1.19 Sample correlation

The sample correlation coefficient is

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{s_{xy}}{s_x \cdot s_y}, \tag{1-12}$$

where s_x and s_y is the sample standard deviation for x and y respectively.

When $x_i - \bar{x}$ and $y_i - \bar{y}$ have the same sign, then the point (x_i, y_i) give a positive contribution to the sample correlation coefficient and when they have opposite signs the point give a negative contribution to the sample correlation coefficient, as illustrated here:

|||| Example 1.20 Student heights and weights

The sample means are found to be

$$\bar{x} = 178 \text{ and } \bar{y} = 78.1.$$

Using these we can show how each student deviate from the average height and weight (these deviations are exactly used for the sample correlation and covariance computations)

Student	1	2	3	4	5	6	7	8	9	10
Height (x_i)	168	161	167	179	184	166	198	187	191	179
Weight (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9
$(x_i - \bar{x})$	-10	-17	-11	1	6	-12	20	9	13	1
$(y_i - \bar{y})$	-12.6	-19.8	-10	7.6	2.4	-14.7	24.5	13.3	8.6	0.8
$(x_i - \bar{x})(y_i - \bar{y})$	126.1	336.8	110.1	7.6	14.3	176.5	489.8	119.6	111.7	0.8

Student 1 is below average on both height and weight $(-10 \text{ and } -12.6)$. Student 10 is above average on both height and weight $(+1 \text{ and } +0.8)$.s

The sample covariance is then given by the sum of the 10 numbers in the last row of the table

$$\begin{aligned} s_{xy} &= \frac{1}{9}(126.1 + 336.8 + 110.1 + 7.6 + 14.3 + 176.5 + 489.8 + 119.6 + 111.7 + 0.8) \\ &= \frac{1}{9} \cdot 1493.3 \\ &= 165.9 \end{aligned}$$

And the sample correlation is then found from this number and the standard deviations

$$s_x = 12.21 \text{ and } s_y = 14.07.$$

(the details of the s_y computation is not shown). Thus we get the sample correlation as

$$r = \frac{165.9}{12.21 \cdot 14.07} = 0.97.$$

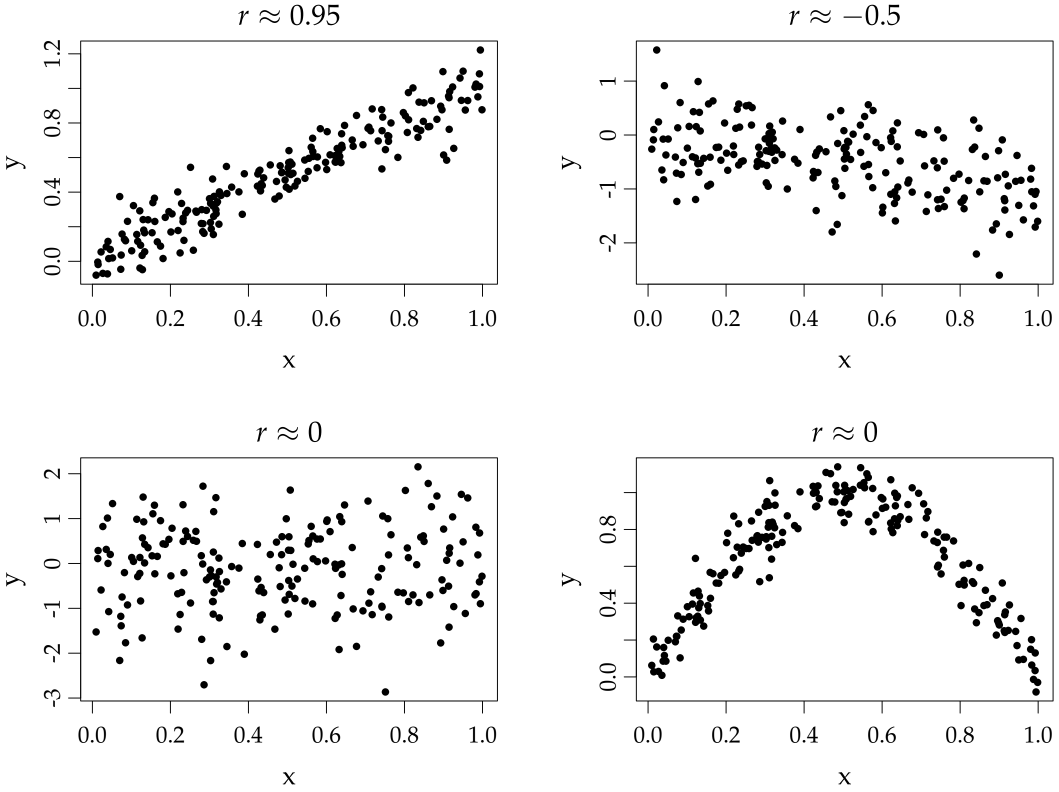
Note how all 10 contributions to the sample covariance are positive in the example case - in line with the fact that all observations are found in the first and third quadrants of the scatter plot (where the quadrants are defined by the sample means of x and y). Observations in second and fourth quadrant would contribute with negative numbers to the sum, hence such observations would be from students with below average on one feature while above average on the other. Then it is clear that: had all students been like that, then the covariance and the correlation would have been negative, in line with a negative (downwards) trend in the relation.

We can state (without proofs) a number of properties of the sample correlation r:

Remark 1.21 Properties of the sample correlation, r

- r is always between -1 and 1: $-1 \leq r \leq 1$
- r measures the degree of linear relation between x and y
- $r = \pm 1$ if and only if all points in the scatterplot are exactly on a line
- $r > 0$ if and only if the general trend in the scatterplot is positive
- $r < 0$ if and only if the general trend in the scatterplot is negative

The sample correlation coefficient measures the degree of linear relation between x and y, which imply that we might fail to detect non-linear relationships, illustrated in the following plot of four different point clouds and their sample correlations:



The sample correlation in both the bottom plots are close to zero, but as we see from the plot this number itself doesn't imply that there no relation between y and x - which clearly is the case in the bottom right and highly non-linear case.

Sample covariances and correlation are closely related to the topic of linear regression, treated in Chapter 5 and 6 , where we will treat in more detail how we can find the line that could be added to such scatter-plots to describe the relation between x and y in a different (but related) way, as well as the statistical analysis used for this.

1.5 Introduction to R and RStudio

The program R is an open source software for statistics that you can download to your own laptop for free. Go to http://mirrors.dotsrc.org/cran/ and select your platform (Windows, Mac or Linux) and follow instructions to install.

RStudio is a free and open source integrated development environment (IDE) for R. You can run it on your desktop (Windows, Mac or Linux) or even over the web using RStudio Server. It works as (an extended) alternative to running R in the basic way through a terminal. This will be used in the course. Download it from http://www.rstudio.com/ and follow installation instructions. To use the software, you only need to open RStudio (R will then be used by RStudio for carrying out the calculations).

1.5.1 Console and scripts

Once you have opened RStudio, you will see a number of different windows. One of them is the console. Here you can write commands and execute them by hitting Enter. For instance:

```
> # Add two numbers in the console
> 2+3

[1] 5
```

In the console you cannot go back and change previous commands and neither can you save your work for later. To do this you need to write a script. Go to `File->New->R Script`. In the script you can write a line and execute it in the console by hitting `Ctrl+Enter` (Windows) or `Cmd+Enter` (Mac). You can also mark several lines and execute them all at the same time.

1.5.2 Assignments and vectors

If you want to assign a value to a variable, you can use = or `<-`. The latter is the preferred by R-users, so for instance:

```
> # Assign the value 3 to y
> y <- 3
```

It is often useful to assign a set of values to a variable like a vector. This is done with the function `c` (short for concatenate):

```
# Concatenate numbers to a vector
x <- c(1, 4, 6, 2)
x

[1] 1 4 6 2
```

Use the colon :, if you need a sequence, e.g. 1 to 10:

```
> # A sequence from 1 to 10
> x <- 1:10
> x

 [1]  1  2  3  4  5  6  7  8  9 10
```

You can also make a sequence with a specific step-size different from 1

```
> # Sequence with specified steps
> x <- seq(0, 1, by=0.1)
> x

 [1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
```

If you are in doubt of how to use a certain function, the help page can be opened by typing ? followed by the function, e.g. ?seq.

If you know Matlab then this document Hiebeler-matlabR.pdf can be very helpful.

1.5.3 Descriptive statistics

All the summary statistics measures presented in Section 1.4 can be found as functions or part of functions in R:

- mean(x) - mean value of the vector x
- var(x) - variance
- sd(x) - standard deviation
- median(x) - median
- quantile(x,p) - finds the pth quantile. p can consist of several different values, e.g. quantile(x,c(0.25,0.75)) or quantile(x,c(0.25,0.75), type=2)
- cov(x, y) - the covariance of the vectors x and y
- cor(x, y) - the correlation

Please again note that the words *quantiles* and *percentiles* are used interchangeably - they are essentially synonyms meaning exactly the same, even though the formal distinction has been clarified earlier.

Example 1.22 Summary statistics in R

Consider again the $n = 10$ data from Example 1.6. We can read these data into R and compute the sample mean and sample median as follows:

```
# Sample Mean and Median
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
mean(x)

[1] 178


median(x)

[1] 179
```

The sample variance and sample standard deviation are found as follows:

```
# Sample variance and standard deviation
var(x)

[1] 149.1


sqrt(var(x))

[1] 12.21


sd(x)

[1] 12.21
```

The sample quartiles can be found by using the `quantile` function as follows:

```
# Sample quartiles
quantile(x, type=2)

  0%  25%  50%  75% 100%
 161  167  179  187  198
```

The option "type=2" makes sure that the quantiles found by the function is found using the definition given in Definition 1.7. By default, the quantile function would use another definition (not detailed here). Generally, we consider this default choice just as valid as the one explicitly given here, it is merely a different one. Also the quantile function has an option called "probs" where any list of probability values from 0 to 1 can be given. For instance:

```
# Sample quantiles 0%, 10%,..,90%, 100%:
quantile(x, probs=seq(0, 1, by=0.10), type=2)

   0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100% 
161.0 163.5 166.5 168.0 173.5 179.0 184.0 187.0 189.0 194.5 198.0 
```

1.5.4 Use of R in the course and at the exam

You should bring your laptop with R installed with you to the teaching activity and to the exam. We will need access to the so-called probability distributions to do statistical computations, and the values of these distributions are not otherwise part of the written material: These probability distributions are part of many different software, also Excel, but it is part of the syllabus to be able to work with these within R.

Apart from access to these probability distributions, the R-software is used in three ways in our course

1. As a pedagogical learning tool: The random variable simulation tools inbuilt in R enables the use of R as a way to illustrate and learn the principles of statistical reasoning that are the main purposes of this course.

2. As a pocket calculator substitute - that is making R calculate "manually" - by simple routines - plus, minus, squareroot etc. whatever needs to be calculated, that you have identified by applying the right formulas from the proper definitions and methods in the written material.

3. As a "probability calculus and statistical analysis machine" where e.g. with some data fed into it, it will, by inbuilt functions and procedures do all relevant computations for you and present the final results in some overview tables and plots.

We will see and present all three types of applications of R during the course. For the first type, the aim is not to learn how to use the given R-code itself but rather to learn from the insights that the code together with the results of applying it is providing. It will be stated clearly whenever an R-example is of this type. Types 2 and 3 are specific tools that should be learned as a part of the course and represent tools that are explicitly relevant in your future engineering activity. It is clear that at some point one would love to just do the last kind of applications. However, it must be stressed that even though the program is able to calculate things for the user, understanding the details of the calculations must NOT be forgotten - understanding the methods and knowing the formulas is an important part of the syllabus, and will be checked at the exam.

|||| Remark 1.23 BRING and USE pen and paper PRIOR to R

For many of the exercises that you are asked to do it will not be possible to just directly identify what R-command(s) should be used to find the results. The exercises are often to be seen as what could be termed "problem mathematics" exercises. So, it is recommended to also bring and use pen and paper to work with the exercises to be able to subsequently know how to finally finish them by some R-calculations. (If you adjusted yourself to some digital version of "pen-and-paper", then this is fine of course.)

|||| Remark 1.24 R is not a substitute for your brain activity in this course!

The software R should be seen as the most fantastic and easy computational companion that we can have for doing statistical computations that we could have done "manually", if we wanted to spend the time doing it. All definitions, formulas, methods, theorems etc. in the written material should be known by the student, as should also certain R-routines and functions.

A good question to ask yourself each time that you apply en inbuilt R-function is: "Would I know how to make this computation "manually"?". There are few exceptions to this requirement in the course, but only a few. And for these the question would be: "Do I really understand what R is computing for me now?"

1.6 Plotting, graphics - data visualisation

A really important part of working with data analysis is the visualisation of the raw data, as well as the results of the statistical analysis – the combination of the two leads to reliable results. Let us focus on the first part now, which can be seen as being part of the explorative descriptive analysis also mentioned in Section 1.4. Depending on the data at hand different types of plots and graphics could be relevant. One can distinguish between *quantitative* vs. *categorical* data. We will touch on the following type of basic plots:

- Quantitative data:
  - Frequency plots and histograms
  - box plots
  - cumulative distribution
  - Scatter plot (xy plot)
- Categorical data:
  - Bar charts
  - Pie charts

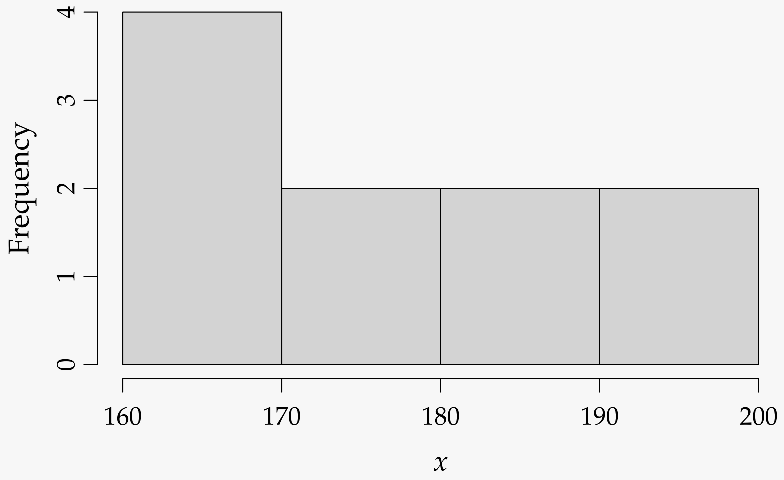
1.6.1 Frequency distributions and the histogram

The frequency distribution is the count of occurrences of values in the sample for different classes using some classification, for example in intervals or by some other property. It is nicely depicted by the histogram, which is a bar plot of the occurrences in each classes.

|||| Example 1.25 Histogram in R

Consider again the $n = 10$ sample from Example 1.6.

```
# A histogram of the heights
hist(x)
```



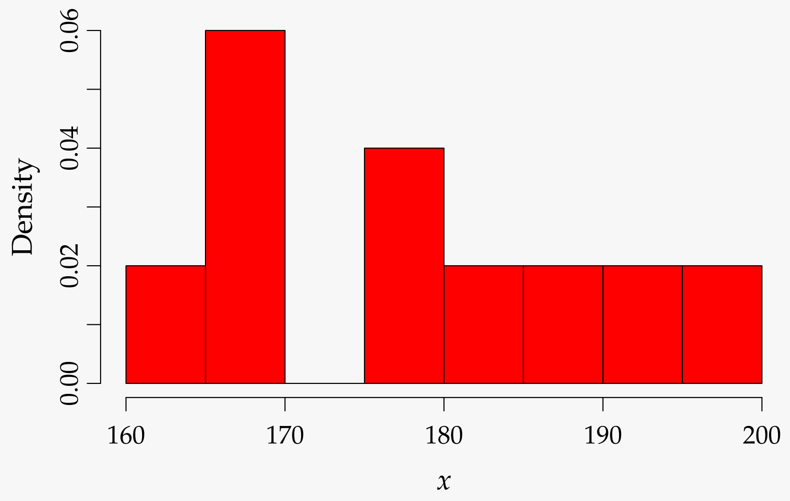
The default histogram uses equidistant interval widths (the same width for all intervals) and depicts the raw frequencies/counts in each interval. One may change the scale into showing what we will learn to be *densities* by dividing the raw counts by n and the interval width, i.e.

$$\frac{\text{"Interval count"}}{n \cdot (\text{"Interval width"})}.$$

By plotting the densities a density histogram also called the empirical density the area of all the bars add up to 1:

|||| Example 1.26 Empirical density in R

```
# A density histogram or empirical density of the heights
hist(x, prob=TRUE, col="red", nclass=8)
```



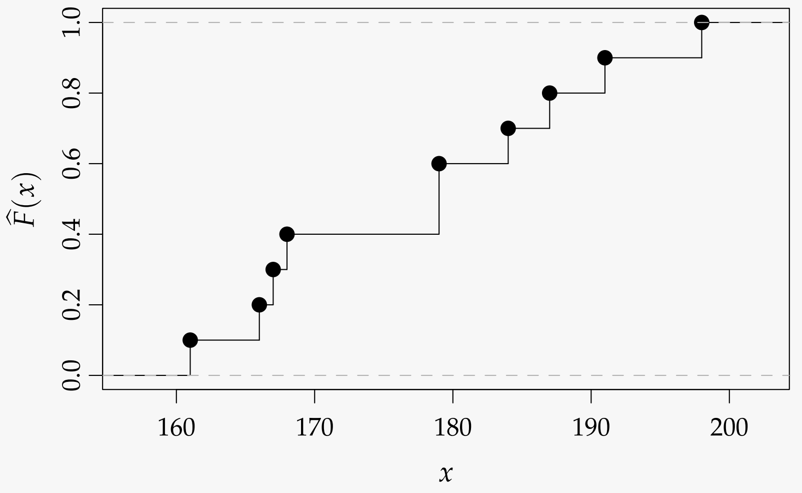
The R-function `hist` makes some choice of the number of classess based on the number of observations - it may be changed by the user option `nclass` as illustrated here, although the original choice seems better in this case due to the very small sample.

1.6.2 Cumulative distributions

The cumulative distribution can be visualized simply as the cumulated relative frequencies either across classes, as also used in the histogram, or individual data points, which is then called the *empirical cumulative distribution function:*

Example 1.27 Cumulative distribution plot in R

```r
# Empirical cumulative distribution plot
plot(ecdf(x), verticals=TRUE)
```



The empirical cumulative distribution function F_n is a step function with jumps i/n at observation values, where i is the number of identical(tied) observations at that value.

For observations $(x_1, x_2, \ldots, x_n)$, $F_n(x)$ is the fraction of observations less or equal to x, that mathematically can be expressed as

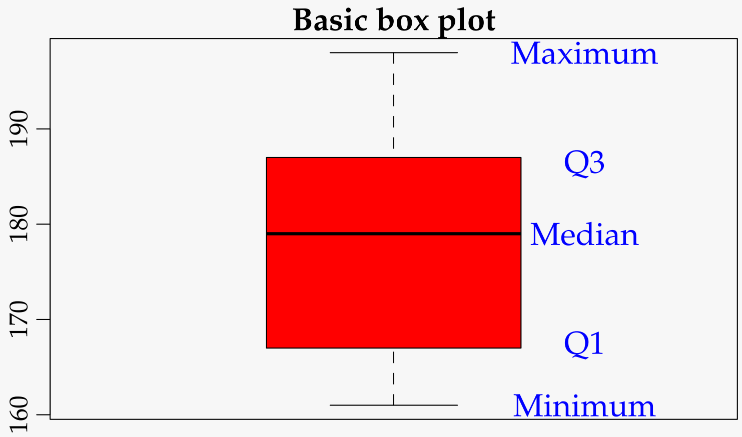
$$F_n(x) = \sum_{j \text{ where } x_j \leq x} \frac{1}{n}. \tag{1-13}$$

1.6.3 The box plot and the modified box plot

The so-called box plot in its basic form depicts the five quartiles (min, Q_1, median, Q_3, max) with a box from Q_1 to Q_3 emphasizing the Inter Quartile Range (IQR):

|||| Example 1.28 Box plot in R

```
# A basic box plot of the heights (range=0 makes it "basic")
boxplot(x, range=0, col="red", main="Basic box plot")
# Add the blue text
text(1.3, quantile(x), c("Minimum","Q1","Median","Q3","Maximum"),
     col="blue")
```

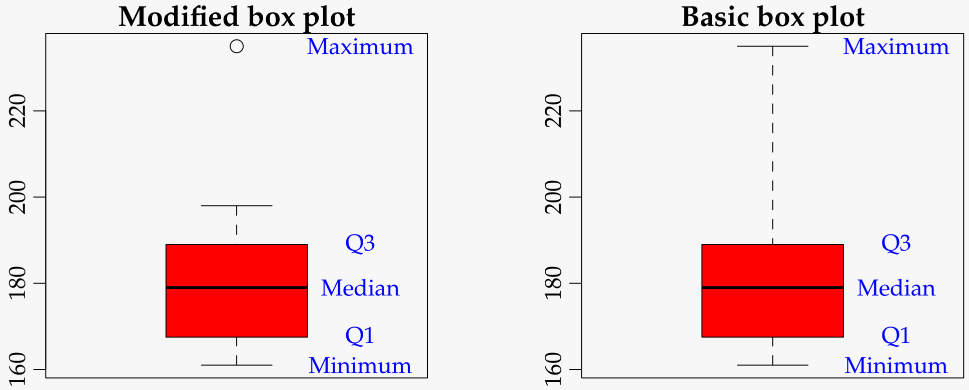


In the modified box plot the whiskers only extend to the min. and max. observation if they are not too far away from the box: defined to be 1.5 × IQR. Observations further away are considered as *extreme observations* and will be plotted individually - hence the whiskers extend from the smallest to the largest observation within a distance of 1.5 × IQR of the box (defined as either 1.5 × IQR larger than Q_3 or 1.5 × IQR smaller than Q_1).

|||| Example 1.29 Box plot in R

If we add an extreme observation, 235 cm, to the heights sample and make the *modified box plot* - the default in R- and the basic *box plot*, then we have:

```
# Add an extreme value and box plot
boxplot(c(x, 235), col="red", main="Modified box plot")
boxplot(c(x, 235), col="red", main="Basic box plot", range=0)
```



Note that since there was no extreme observations among the original 10 observations, the two ”different” plots would be the same if we didn’t add the extreme 235 cm observation.

The box plot hence is an alternative to the histogram in visualising the distribution of the sample. It is a convenient way of comparing distributions in different groups, if such data is at hand.

|||| Example 1.30 Box plot in R

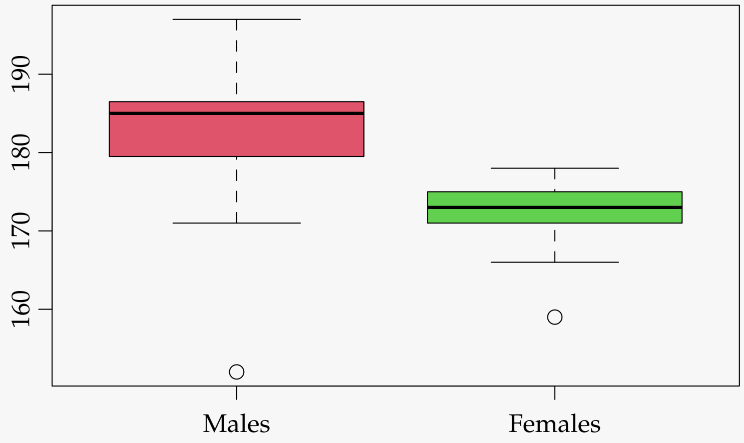
This example shows some ways of working with R to illustrate data.

In another statistics course the following heights of 17 female and 23 male students were found:

Males	152	171	173	173	178	179	180	180	182	182	182	185
	185	185	185	185	186	187	190	190	192	192	197	
Females	159	166	168	168	171	171	172	172	173	174	175	175
	175	175	175	177	178							

The two modified box plots of the distributions for each gender can be generated by a single call to the `boxplot` function:

```r
# Box plot with two groups
Males <-  c(152, 171, 173, 173, 178, 179, 180, 180, 182, 182, 182, 185,
            185 ,185, 185, 185 ,186 ,187 ,190 ,190, 192, 192, 197)
Females <-c(159, 166, 168 ,168 ,171 ,171 ,172, 172, 173, 174 ,175 ,175,
            175, 175, 175, 177, 178)
boxplot(list(Males, Females), col=2:3, names=c("Males", "Females"))
```



At this point, it should be noted that in real work with data using R, one would generally not import data into R by explicit listings in an R-script as here. This only works for very small data sets. Usually the data is imported from somewhere else, e.g. from a spread sheet exported in a `.csv` (*comma separated values*) format as shown here:

|||| Example 1.31 Read and explore data in R

The gender grouped student heights data used in Example 1.30 is available as a .csv-file via http://www2.compute.dtu.dk/courses/introstat/data/studentheights.csv. The structure of the data file, as it would appear in a spread sheet program (e.g. LibreOffice Calc or Excel) is two columns and 40+1 rows including a header row:

```
1 Height Gender
2    152 male
3    171 male
4    173 male
.      . .
.      . .
24   197 male
25   159 female
26   166 female
27   168 female
.      . .
.      . .
39   175 female
40   177 female
41   178 female
```

The data can now be imported into R with the read.table function:

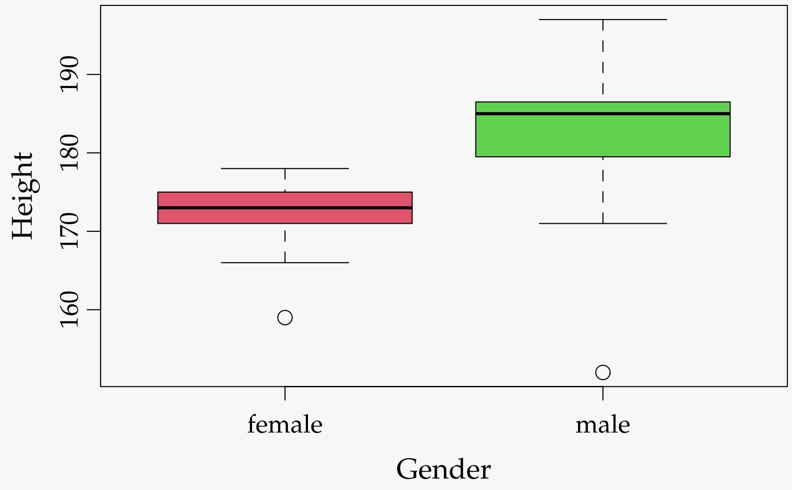
```r
# Read the data (note that per default sep="," but here semicolon)
studentheights <- read.table("studentheights.csv", sep=";", dec=".",
                             header=TRUE, stringsAsFactors=TRUE)
```

The resulting object studentheights is now a so-called data.frame, which is the class used for such tables in R. There are some ways of getting a quick look at what kind of data is really in a data set:

```
# Have a look at the first 6 rows of the data
head(studentheights)

  Height Gender
1    152   male
2    171   male
3    173   male
4    173   male
5    178   male
6    179   male


# Get an overview
str(studentheights)

’data.frame’: 40 obs. of  2 variables:
 $ Height: int  152 171 173 173 178 179 180 180 182 182 ...
 $ Gender: Factor w/ 2 levels "female","male": 2 2 2 2 2 2 2 2 2 2 ...


# Get a summary of each column/variable in the data
summary(studentheights, quantile.type=2)

     Height          Gender
 Min.   :152.0   female:17
 1st Qu.:172.5   male  :23
 Median :177.5
 Mean   :177.9
 3rd Qu.:185.0
 Max.   :197.0
```

For quantitative variables we get the quartiles and the mean from `summary`. For categorical variables we see the category frequencies. A data structure like this is commonly encountered (and often the only needed) for statistical analysis. The gender grouped box plot can now be generated by:

```
# Box plot for each gender
boxplot(Height ~ Gender, data=studentheights, col=2:3)
```



The R-syntax Height ~ Gender with the tilde symbol "~" is one that we will use a lot in various contexts such as plotting and model fitting. In this context it can be understood as "Height is plotted as a function of Gender".

1.6.4 The Scatter plot

The scatter plot can be used for two quantitative variables. It is simply one variable plotted versus the other using some plotting symbol.

|||| Example 1.32 Explore data included in R

Now we will use a data set available as part of R itself. Both base R and many add-on R-packages include data sets, which can be used for testing and practising. Here we will use the mtcars data set. If you write:

```
# See information about the mtcars data
?mtcars
```

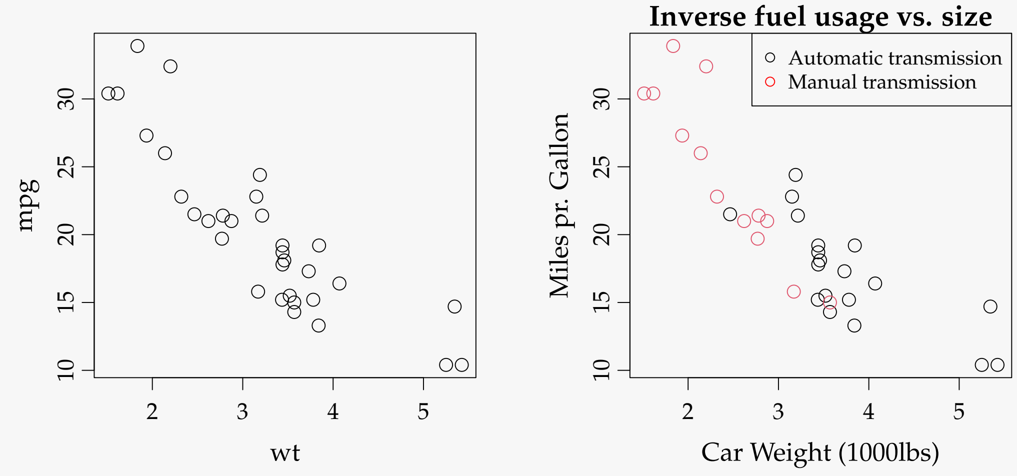
you will be able to read the following as part of the help info:

"The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74

models). A data frame with 32 observations on 11 variables. Source: Henderson and Velleman (1981), Building multiple regression models interactively. Biometrics, 37, 391-411.”

Let us plot the gasoline use, (mpg=miles pr. gallon), versus the weight (wt):

```
# To make 2 plots
par(mfrow=c(1,2))
# First the default version
plot(mtcars$wt, mtcars$mpg, xlab="wt", ylab="mpg")
# Then a nicer version
plot(mpg ~ wt, xlab="Car Weight (1000lbs)", data=mtcars,
     ylab="Miles pr. Gallon", col=factor(am),
     main="Inverse fuel usage vs. size")
# Add a legend to the plot
legend("topright", c("Automatic transmission","Manual transmission"),
       col=c("black","red"), pch=1, cex=0.7)
```



In the second plot call we have used the so-called `formula` syntax of R, that was introduced above for the grouped box plot. Again, it can be read: “`mpg` is plotted as a function of `wt`”. Note also how a color option, `col=factor(am)`, can be used to group the cars with and without automatic transmission, stored in the data column `am` in the data set.

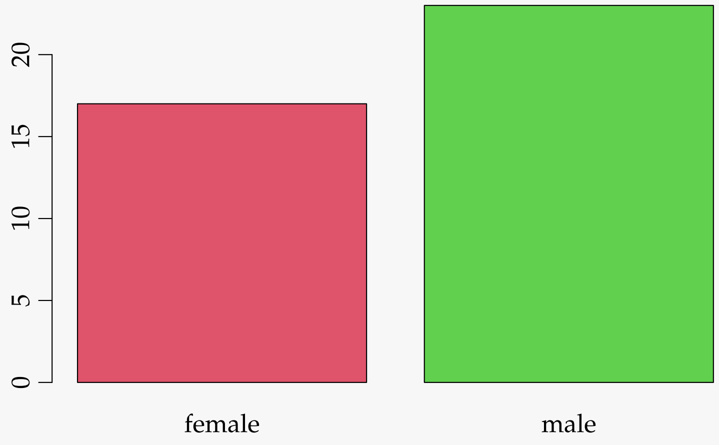
1.6.5 Bar plots and Pie charts

All the plots described so far were for quantitative variables. For categorical variables the natural basic plot would be a bar plot or pie chart visualizing the relative frequencies in each category.

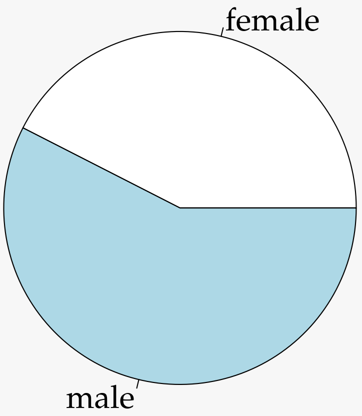
|||| Example 1.33 Bar plots and Pie charts in R

For the gender grouped student heights data used in Example 1.30 we can plot the gender distribution by:

```
# Barplot
barplot(table(studentheights$Gender), col=2:3)
```



```
# Pie chart
pie(table(studentheights$Gender), cex=1, radius=1)
```



1.6.6 More plots in R?

A good place for getting more inspired on how to do easy and nice plots in R is: http://www.statmethods.net/.

1.7 Exercises

Exercise 1.1 Infant birth weight

In a study of different occupational groups the infant birth weight was recorded for randomly selected babies born by hairdressers, who had their first child. The following table shows the weight in grams (observations specified in sorted order) for 10 female births and 10 male births:

Females (x)	2474	2547	2830	3219	3429	3448	3677	3872	4001	4116
Males (y)	2844	2863	2963	3239	3379	3449	3582	3926	4151	4356

Solve at least the following questions a)-c) first "manually" and then by the inbuilt functions in R. It is OK to use R as alternative to your pocket calculator for the "manual" part, but avoid the inbuilt functions that will produce the results without forcing you to think about how to compute it during the manual part.

a) What is the sample mean, variance and standard deviation of the female births? Express in your own words the story told by these numbers. The idea is to force you to interpret what can be learned from these numbers.

b) Compute the same summary statistics of the male births. Compare and explain differences with the results for the female births.

c) Find the five quartiles for each sample — and draw the two box plots with pen and paper (i.e. not using R.)

d) Are there any "extreme" observations in the two samples (use the *modified box plot* definition of extremness)?

e) What are the coefficient of variations in the two groups?

Exercise 1.2 Course grades

To compare the difficulty of 2 different courses at a university the following grades distributions (given as number of pupils who achieved the grades) were registered:

	Course 1	Course 2	Total
Grade 12	20	14	34
Grade 10	14	14	28
Grade 7	16	27	43
Grade 4	20	22	42
Grade 2	12	27	39
Grade 0	16	17	33
Grade -3	10	22	32
Total	108	143	251

a) What is the median of the 251 achieved grades?

b) What are the quartiles and the IQR (Inter Quartile Range)?

Exercise 1.3 Cholesterol

In a clinical trial of a cholesterol-lowering agent, 15 patients' cholesterol (in mmol L^{-1}) was measured before treatment and 3 weeks after starting treatment. Data is listed in the following table:

Patient	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Before	9.1	8.0	7.7	10.0	9.6	7.9	9.0	7.1	8.3	9.6	8.2	9.2	7.3	8.5	9.5
After	8.2	6.4	6.6	8.5	8.0	5.8	7.8	7.2	6.7	9.8	7.1	7.7	6.0	6.6	8.4

a) What is the median of the cholesterol measurements for the patients before treatment, and similarly after treatment?

b) Find the standard deviations of the cholesterol measurements of the patients before and after treatment.

c) Find the sample covariance between cholesterol measurements of the patients before and after treatment.

d) Find the sample correlation between cholesterol measurements of the patients before and after treatment.

e) Compute the 15 differences (Dif = Before − After) and do various summary statistics and plotting of these: sample mean, sample variance, sample standard deviation, boxplot etc.

f) Observing such data the big question is whether an average decrease in cholesterol level can be "shown statistically". How to formally answer this question is presented in Chapter 3, but consider now which summary statistics and/or plots would you look at to have some idea of what the answer will be?

|||| Exercise 1.4 Project start

a) Go to CampusNet and take a look at the first project and read the project page on the website for more information (02323.compute.dtu.dk/projects or 02402.compute.dtu.dk/projects). Follow the steps to import the data into R and get started with the explorative data analysis.

Glossaries

Box plot [Box plot] The so-called boxplot in its basic form depicts the five quartiles (min, Q1 , median, Q3 , max) with a box from Q1 to Q3 emphasizing the IQR 26, 29–32, 34, 36

Categorical data [Kategorisk data] A variable is called categorical if each observation belongs to one of a set of categories 1, 26

Class The frequency distribution of the data for a certain grouping of the data 26, 28

Correlation [Korrelation] The sample correlation coefficient are a summary statistic that can be calculated for two (related) sets of observations. It quantifies the (linear) strength of the relation between the two. See also: Covariance 16–20, 22

Covariance [Kovarians] The sample covariance coefficient are a summary statistic that can be calculated for two (related) sets of observations. It quantifies the (linear) strength of the relation between the two. See also: Correlation 16–20, 22

Descriptive statistics [Beskrivende statistik] Descriptive statistics, or explorative statistics, is an important part of statistics, where the data is summarized and described 1, 4, 7

Empirical cumulative distribution [Empirisk fordeling] The empirical cumulative distribution function F_n is a step function with jumps i/n at observation values, where i is the number of identical observations at that value 28, 29

Frequency [Frekvens] How frequent data is observed. The frequency distribution of the data for a certain grouping is nicely depicted by the histogram, which is a barplot of either raw frequencies or for some number of classes 26–28, 34, 37

Histogram [Histogram] The default histogram uses the same width for all classes and depicts the raw frequencies/counts in each class. By dividing the raw counts by n times the class width the density histogram is found where the area of all bars sum to 1 26–28, 31

(Statistical) Inference [Statistisk inferens (følgeslutninger baseret på data)] 5

Inter Quartile Range [Interkvartil bredde] The Inter Quartile Range (IQR) is the middle 50% range of data 15

Linear regression [Lineær regression (-sanalyse)] 1, 20

Median [Median, stikprøvemedian] The median of population or sample (note, in text no distinguishment between *population median* and *sample median*) 7, 9, 10, 23

Multiple linear regression [Multipel lineær regression (-sanalyse)] 1

Quantile [Fraktil, stikprøvefraktil] The quantiles of population or sample (note, in text no distinguishment between *population quantile* and *sample quantile*) 11

Quartile [Fraktil, stikprøvefraktil] The quartiles of population or sample (note, in text no distinguishment between *population quartile* and *sample quartile*) 11

Sample variance [Empirisk varians, stikprøvevarians] 13

Sample mean [Stikprøvegennemsnit] The average of a sample 8, 10, 13, 23

Acronyms

ANOVA Analysis of Variance *Glossary:* Analysis of Variance

cdf cumulated distribution function *Glossary:* cumulated distribution function

CI confidence interval *Glossary:* confidence interval

CLT Central Limit Theorem *Glossary:* Central Limit Theorem

IQR Inter Quartile Range 8, 15, 29, 30, *Glossary:* Inter Quartile Range

LSD Least Significant Difference *Glossary:* Least Significant Difference

pdf probability density function *Glossary:* probability density function