STATS201.stat 202 10 Experimental design and descriptive statistics

Anna Helga Jónsdóttir
Sigrún Helga Lund

December 11, 2012

Contents

1 The basics of statistics

1.1 Sample and population

Population
The **population** of a study is the set of all subjects that inference should be made about.

Sample
A **sample** is a set of subjects that are sampled from a given population.

- Every sample can only be sampled from one population.
- Different samples can be sampled from the same population.

1.2 Categorical and numerical variables

Variable
A **Variable** is a certain property that is noted or measured on the subjects in the sample.

Categorical variables
Categorical variables do not have numerical values but, as the name suggests, indicate which category the subject belongs to.

Numerical variables
Numerical variables have numerical values that are measured in some units.

1.3 Continuous and discrete variables

Continuous variables
When a numerical variable can have any numerical value on some interval it is referred to as **continuous**. Only numerical variables can be continuous.

Discrete variables
If variables are not continuous they are referred to as **discrete**. All categorical variables are discrete and some numerical variables.

Exploratory and response variables
For every subject, the value of an **explanatory variable** influences the value that its **response variable** will obtain. A response variable can be influenced by several explanatory variables.

1.4 Randomness

We apply statistics because our measurements are influenced by some randomness:

- We measure only a sample of the whole population.
- The phenomena to be measured are random by nature.

This property is described with the concept **random variable**.

A random variable describes the outcome of a variable before it is measured.

2 Experimental design

2.1 Bias

Bias
Bias occurs when the applied methods give a systematically biased view of the population to be analyzed.

- Subjects are chosen in a systematically biased manner: **Sample bias**.
- A good **sampling design** minimizes sample bias.
- Interfering influences from researchers and subjects: **Experimenters bias** and **placebo effects**.
- **Blinding** minimizes researchers bias and placebo effects.

2.2 Variability

Variability
Variability occurs because the phenomena under investigation are influenced by some randomness and therefore can the outcome of the measurements change each time the experiment is conducted.

- This variability causes that our results can change every time the experiment is repeated.
- Repetitions allow us to estimate the variability of the measurements.
- The more repetitions - the better we can draw conclusions from our experiment!

2.3 Controlled experiment

The objective of many researches is to demonstrate the effect of applying certain **interventions** to our subjects. In order to do so an **controlled experiment** needs to be conducted.

Controlled experiment
In order for an investigation to be classified as an **controlled experiment** two requirements need to be fulfilled:

1 The investigator can control which subjects obtain which interventions.

2 Measurements on the subjects are made both before and after the intervention is applied.

2.4 Random sampling

Random samling
A **random sampling** is made when subjects are chosen randomly from the population and all subjects have the same probability of being selected in the sample.

A sample that is chosen with random sampling is called a **random sample**.

We will look at three types of random samples:

- **simple random sample**
- **stratified random sample**
- **paired random sample**

2.5 Simple and stratified random sample

Simple random sample
When a **simple random sample** is chosen, subjects are chosen randomly from the whole population.

Stratified random sample
When a **stratified random sample** is chosen, the population is first divided into a few groups or strata and subsequently subjects are chosen by random sampling from each stratum.

The number of subjects chosen from each stratum need to be decided beforehand, but it can vary between strata.

Good to use when the number of subject within strata varies greatly.

2.6 Paired random sample

Paired random sample
When a **paired random sample** is chosen, the subjects are paired into groups of two and two and a fixed number of pairs of subjects is randomly sampled.

Good to use when the measurements are influenced by many uncontrolled variables. Then subjects that are influenced in a similar manner paired together.

2.7 What if a random sample cannot be chosen?

Sometimes difficulties in implementation of the experiment make random sampling impossible. Then one of two actions need to be chosen:

1. To redefine the population such that random sampling will be possible.
   - Then conclusions can only by made about the "new population". Is that feasible?
2. To accept the induced bias.
   - We not the sampling bias in our discussion.
   - Discuss thoroughly which consequences it can result.
   - Can it be assumed that the bias is small compared to the phenomena to be investigated?

2.8 Volunteer samples

- Volunteer samples are only gathered when the subjects are human beings and then measurements are only made on the subjects that volunteer to participate in the research.
- This induces a sampling bias because certain subjects can be more likely to volunteer than others.
- This bias can be so large that no inference can be made on the population from the measurements gathered.

2.9 Convenience samples

- Convenience samples are gathered when measurements are only made on subjects that are (conveniently) accessible to the researchers.
- This induces a sampling bias because certain subjects are more likely to be accessible to the researchers than others.
- This bias can be so large that no inference can be made on the population from the measurements gathered.

2.10 Missing values

Missing values
Often successful measurements are not made on every subject in the sample. Then we have **missing values** for these subjects.

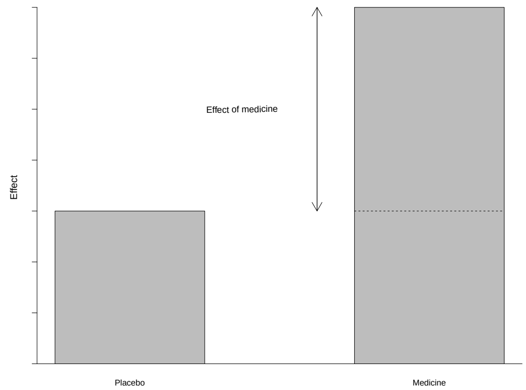
- One cannot simply use the successful measurements and overlook the missing values.
- Some subjects are often more likely to have missing values than others.
- These subjects will, as a consequence, be less likely to be chosen to the reduced "sample".
- That causes a samling bias!

2.11 Experimenters bias and placebo effects

Experimenters bias
Experimenters bias occurs when the anticipations of the researchers influence the measurements of the effect of the interventions on the subjects.

Placebo effect
Placebo is every intervention that the subject wrongly assumes that is the intervention under investigation.

Placebo effect
The difference in the measurements of the subjects before and after the placebo intervention is called a **placebo effect**.



2.12 Placebo effects

2.13 Single-blinded and double-blinded experiments.

Double blinded experiments
When an experiment is **double-blind** neither the investigator nor the subjects know which intervention they will receive. notice that an intervention can be a placebo intervention.

Single blind experiments
When an experiment is **single-blind** either the subjects don't know which intervention they received but the investigator does or vice versa.

2.14 Repetitions

- We assume that there is always some variability in our measurements - they are influenced by some randomness.
- The results can change every time a new sample is chosen and the investigation repeated.
- As soon as we have measured more than one subject, that is we have **repetitions**, we have some idea of the variability in the measurements.
- The more measurements we have, the better idea we have of this variability.

2.15 Drawing conclusions

- Statistical inference is drawing conclusions about a population based on investigations of a sample from that population.
- The more repetitions of subjects, the more likely it is that we can draw conclusions about the population.
- The main rule is: "The more repetitions, the better".

2.16 Causation

Causation
Causation between two variables is when the outcome of one variable influences the outcome of the other. Causation can only be shown with controlled experiments.

MYND Causal relationship /pictures/Orsakasamband

2.17 Good experimental design

Which requirements should a controlled experiment fulfill?
Every investigator should seek that his experiment fulfills the following conditions:

1 Sampling design.

The subjects are chosen by random sampling and/or divided into groups by random sampling.

2 Blinding.

The experiment is by all means double-blind but at least single-blind if that is impossible.

3 Repetitions.

The intervention is applied to a repeated number of subjects.

If a controlled experiment fulfills these conditions, claims of causal relationship can be made.

3 Graphical representation

3.1 Graphical representation

- The first step in any statistical analysis should be to look at the data visually.
- The essence of statistical analysis is to understand the nature of the measurements that are investigated.
- The variability of the data is a key issue - how are the measurements distributed?
  - How much difference do we notice in the outcomes of our subjects?
  - How are the outcomes distributed?
- Graphical representation is the best way to understand the nature of the distribution of the measurements.
- The graphs that we will see in this lecture will only show the measurements of one variable at a time and we make a distinction whether the variable is continuous or discrete.

3.2 Graphical representation of discrete variables

- The most common form of graphs for discrete variables are the **bar chart** and the **pie chart**.
- Pie charts are commonly used in business and media but are rare in journals and books in natural sciences.
- Bar charts are commonly seen and they are better suited then pie charts to show the measurements of discrete variables. than pie charts.

3.3 Bar chart

Bar chart
A bar chart consists of two or more bars. The number of bars is determined by the number of categories/values that the discrete variable takes. Every bar represents one category/value and they may not lie close to each other. The height of the bars shows the frequency of the corresponding category. The bars shall be ordered in an informative way, often by size.

Before one draws a bar chart it is often convenient to make a little table that shows the categories of the variable and how many subjects belong to each category.

3.4 Pie chart

Pie chart
When making a pie chart, it is important that all categories/values of the variable under investigation are pictured on the chart. The number of slices in the pie chart is determined by the number of categories/values of the variable. The size of the slice is determined by the proportional number of subjects in the corresponding category compared to the whole sample. Watch out that the ratios add up to 100 %.

Before one draws a bar chart it is often convenient to make a little table that shows the categories of the variable, how many subjects belong to each category and the corresponding percentage of subjects in that category as a proportion of the whole sample.

PIE CHARTS ARE NOT DRAWN BY HAND!

3.5 Graphical representation of continuous variables

- The most common method to visualize continuous variables are **histograms**.
- A **Box plot** is also a good method to visualize continuous variables and they will be introduced in the lecture about descriptive statistics (Lecture 40).
- **Scatter plots** will be introduced in the lecture about linear regression (Lecture170) but they are used to explore the relationship between two continuous variables.

3.6 Histograms

- Histograms are similar to bar charts but the main difference in their appearance is that there is no gap between the bars in a histogram.
- It is slightly more difficult to make a histogram then a bar chart as continuous variables do not contain real categories or groups.
- First one has to define groups (intervals) before one counts how many measurements belong to each group.
- When the groups have been made it is often useful to make a table that contains the groups and how many measurements lie within each group.

3.7 Histograms

Histogram
A histogram consists of bars that are lined continuously one by another. The number of bars is determined by the number of groups (intervals) that the continuous variable is split up into. When the groups are made is is often good to keep in mind the following criteria:

- Lower and upper limits should be simple and easily understood.
- The intervals may not overlap and must cover all measurements.
- The intervals should be equally wide.
- The number of intervals should be appropriate. A rule of thumb is that the number of intervals should be approx. 5 times the logarithm of the total number of measurements.

When the intervals have been made, one bar is drawn for each interval and the height of the bar is determined by the number of measurements within the corresponding interval.

3.8 Shapes of distributions

Shapes of distributions
The following concepts are used to describe the distribution of measurements.

- The distribution of the smallest measurements are called the *left-tail* of the distribution. The distribution of the largest measurements is called the *right-tail* of the distribution.
- A distribution is *symmetric* if its right-tail is distributed as the mirror image of the left-tail.
- A distribution that is not symmetric is *skewed*. A distribution is *skewed to the right* if its right-tail is longer then the left-tail and *skewed to the left* if the left one is longer then the right one.
- If a distribution has one peak it is referred to as *unimodal*.
- If a dis-

3.9 Outliers

Outliers
Outliers are measurements that differ greatly from other measurements in the sample. There can be various reasons for outliers and it is important to look at them specifically and consider their cause.

4 Descriptive statistics

4.1 Statistic

- Descriptive statistics are used to describe certain properties of our measurements.
- A **statistic** is a good way to do so, but it is a number that describes a certain property of our measurements.

Statistic
A **statistic** is a number that is calculated in a some particular way from our measurements.

- There are many different types of statistics that describe different properties of measurements.
- Let us look closer at two types of statistics:
  - Statistics that describe the **central tendency** of the measurements.
  - Statistics that describe the **spread** of the measurements.
- The two most common statistics are the mean and the standard deviation.

4.2 Statistics that describe central tendency

Now look at five different statistics that all describe the central tendency of measurements

1. Mid range
2. Mode
3. Median
4. Mean (arithmetic mean)
5. Weighted mean

4.3 Mid range

Mid range
Assume that we have n measurements $x_1, x_2, \ldots x_n$. Let $x_{\min}$ denote the smallest one and $x_{\max}$ denote the largest one. The **Mid range** is calculated with

$$\text{Mid range} = \frac{x_{\min} + x_{\max}}{2}.$$

We have the following measurements: `1, 2, 3, 5, 9, 9, 15`. Find the mid range.

$$\text{Mid range} = \frac{x_{\min} + x_{\max}}{2} = \frac{1+15}{2} = 8.$$

4.4 Mode

Mode
Assume that we have n measurements, $x_1, x_2, \ldots x_n$. The **mode** is the most frequent outcome among the measurements. It is the only statistic that describes central tendency that can be used for categorical data. It is on the other hand inappropriate to use the mode to describe continuous variables.

We have the following measurements: `1, 2, 2, 3, 5, 9, 9, 15`. What is the mode?

The mode is 2 and 9.

4.5 Median

Median
Assume that we have n measurements, $x_1, x_2, \ldots x_n$. Arrange the measurements by order from the smallest measurement to the largest. Then calculate

$$\text{index number} = 0.5 \cdot (n+1).$$

The **median** is often denoted by M. It depends on whether n is an odd or an even number how we calculate the median.

- If n is an odd number, then the median is the measurement with the index number $0.5 \cdot (n+1)$.
- If n is an even number then the median is the average of the two numbers that have index numbers next to $0.5 \cdot (n+1)$

CAUTION: $0.5 \cdot (n+1)$ is the index number for the measurement, not the median itself!

We have the following measurements: `1, 2, 3, 5, 9, 9, 15`. Find the median.

The measurements are ranked from the smallest value to the largest value. We need to find the placement of the median:

$$\text{Placement} = 0.5 \cdot (n+1) = 0.5 \cdot 8 = 4.$$

so the median is number 4 in the ranked sequence. The median is thus $M = 5$.

4.6 Mean

Mean (arithmetic mean)
Assume that we have n measurements, $x_1, x_2, \ldots x_n$. The **mean** is calculated by adding all of the measurements together and divide by their number.

$$\bar{x} = \frac{x_1 + x_2 + \ldots x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

We have the following measurements: `1, 2, 3, 5, 9, 9, 15`. Find the mean.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1+2+3+5+9+9+15}{7} = \frac{44}{7} = 6.29.$$

4.7 Weighted mean

When the mean is calculated as described in the previous slide, all measurements have the same weight. In some cases we want to give different weights to the measurements. Then we calculate **weighted mean**.

Weighted mean
Assume that we have n measurements, $x_1, x_2, \ldots x_n$ and their weights $w_1, w_2, \ldots w_n$. The weighted mean is calculated as

$$\bar{x}_w = \frac{w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots w_n \cdot x_n}{w_1 + w_2 + \ldots + w_n} = \frac{\sum_{i=1}^{n} w_i \cdot x_i}{\sum_{i=1}^{n} w_i}$$

A geography student has finished the following courses with the given grades. The number of credits each courses gives is also shown.

Course	Grade	Credits
Course 1	7	8
Course 2	9	8
Course 3	7	8
Course 4	8	6
Course 5	6	8
Course 6	9	8
Course 7	9	6
Course 8	10	8

What is the weighted mean of the grades?

The x-is in the following equation are the grades and the w-s the number of credits:

$$\overline{x_w} = \frac{\sum_{i=1}^{n} w_i \cdot x_i}{\sum_{i=1}^{n} w_i} = \frac{8\cdot 7+8\cdot 9+8\cdot 7+6\cdot 8+8\cdot 6+8\cdot 9+6\cdot 9+8\cdot 10}{8+8+8+6+8+8+6+8} = 8.10.$$

4.8 Comparison of statistics that describe central tendency

- All of the statistics previously mentioned are interesting and good to calculate when looking at new data
- Before deciding which statistic is appropriate to use it is good to look at graphical representations of the data in order to get a picture of the distribution of the measurements
- If the distribution is skewed bimodal or multimodal the median shall be used rather then the mean
- The median should also be preferred if there are outliers in the measurements

4.9 Comparison of the median and the meani

- If the distribution is skewed to the right, the mean is higher then the median.
- If the distribution is symmetric, the median and the mean are the same.
- If the distribution is skewed to the left, the mean is less then the median.

4.10 Statistics that describe spread

Spread
The **spread** of measurements describes how spread out the measurements are.

We will discuss 6 statistics that describe the spread of measurements

1. Range
2. Quartiles
3. Interquartile range
4. Percentiles
5. Variance
6. Standard deviation
7. Coefficient of variation

4.11 Range

Range
Assume that we have n measurements, $x_1, x_2, \ldots x_n$ and let x_{min} denote the smallest one and x_{max} the largest one. The **range** of the measurements is calculated with

$$\text{Range} = x_{\text{max}} - x_{\text{min}}$$

We have the following measuremetns: `1, 2, 3, 5, 9, 9, 15.` Find the range.

$$\text{Range} = x_{\text{max}} - x_{\text{min}} = 15 - 1 = 14.$$

4.12 Quartiles

There are three **quartiles** and they are commonly named Q_1, Q_2 og Q_3. They are often sometimes denoted with $Q_{25\%}, Q_{50\%}$ og $Q_{75\%}$. The former notation will be used.

Q_1: The first quartile is such that 25% of the measurements are lower then Q_1. Q_1 is therfore the median of the lower half of the measurements, excluding the median.

Q_1: The second quartile is such that 50% of the measurements are lower then Q_2. Q_2 is therefore the median, $Q_2 = M$.

Q_1: The third quartile is such that 75% of the measurements are lower then Q_3. Q_3 is therfore the median of the upper half of the measurements, excluding the median.

We have the following measurements: `1, 2, 3, 5, 9, 9, 15, 17`. Find the quartiles.

We start by finding the placement in the ranked sequence:

$$\begin{aligned}
Q_1 \text{ - sæti í röð: } 0.25 \cdot (n+1) &= 0.25 \cdot 9 = 2.25 \\
Q_2 \text{ - sæti í röð: } 0.50 \cdot (n+1) &= 0.50 \cdot 9 = 4.50 \\
Q_3 \text{ - sæti í röð: } 0.75 \cdot (n+1) &= 0.75 \cdot 9 = 6.75
\end{aligned}$$

So Q_1 is the mean of the measurements in place 2 and 3:

$$Q_1 = \frac{2+3}{2} = 2.5.$$

Q_2 is the mean of the numbers in place 4 and 5:

$$Q_2 = \frac{5+9}{2} = 7.$$

and Q_3 is the mean of the numbers in place 6 and 7:

$$Q_3 = \frac{9+15}{2} = 12.$$

4.13 Quartiles

Quartiles
Assume that we have n measurements, $x_1, x_2, \ldots x_n$. Arrange the measurements by order from the smallest measurement to the largest. Then calculate

$$\begin{aligned} Q_1 \text{ - index number:} &= 0.25 \cdot (n+1) \\ Q_2 \text{ - index number:} &= 0.50 \cdot (n+1) \\ Q_3 \text{ - index number:} &= 0.75 \cdot (n+1) \end{aligned}$$

- Q_1 is the measurement with index number $0.25 \cdot (n+1)$ or the average of the two measurements that have index numbers next to $0.25 \cdot (n+1)$.
- Q_2 is the measurement with index number $0.50 \cdot (n+1)$ or the average of the two measurements that have index numbers next to $0.50 \cdot (n+1)$.
- Q_3 is the measurement with index number $0.75 \cdot (n+1)$ or the average of the two measurements that have index numbers next to $0.75 \cdot (n+1)$.

4.14 Interquartile range

Interquartile range
Assume that we have n measurements, $x_1, x_2, \ldots x_n$ and let Q_1 denote the first quartile and Q_3 denote the third quartile. The **Interquartile range** of the measurements is denoted with IQR and calculated with

$$IQR = Q_3 - Q_1.$$

We have the following measurements: `1, 2, 3, 5, 9, 9, 15, 17`. Find the interquartile range.

We have $Q_1 = 2.5$ and $Q_3 = 12$.

$$IQR = 12 - 2.5 = 9.5.$$

4.15 Percentiles

The idea behind **percentiles** is similar to that of quartiles, but instead of only looking at the borders at 25%, 50% or 75% of the measurements any proportion at all can be used.

Percentiles
The a% percentile is the numbers that has the property that a% of the measurements have values less then that number.

As with the quartiles, there are several different ways to calculate percentiles, and is almost never done "by hand" but a statistical software used to do so.

4.16 Variance

Variance
Assume that we have n measurements, $x_1, x_2, \ldots x_n$. The **variance** of the measurements is denoted with s^2 and calculated with

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

$s^2 = 0$ if and only if all of the measurements are equal, if not, s^2 is always greater then 0. The further the measurements lie from the mean, the higher s^2 will be.

We have the following measurements: `2, 2, 3, 5, 8`. Find the variance.

We need the mean value:

$$\bar{x} = \frac{2+2+3+5+8}{5} = \frac{20}{5} = 4.$$

Let us make a small table. The first column is the data. The second column has the difference between the data and the mean. In the third column the the number in column two is squared and in the last line, the numbers in the corresponding column are added.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2	2-4 = -2	4
2	2-4 = -2	4
3	3-4 = -1	1
5	5-4 = 1	1
8	8-4 = 4	16
$\sum_{i=1}^{n}(x_i - \bar{x})^2$	=	26

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{26}{4} = 6.5.$$

4.17 Standard deviation

> **Standard deviation**
> Assume that we have n measurements, $x_1, x_2, \ldots x_n$. The **standard deviation** of the measurements is denoted with s and calculated with
>
> $$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$
>
> s = 0 if and only if all of the measurements are equal, if not, s is always greater then 0. The further the measurements lie from the mean, the higher s will be.

We have the the following measurements: `2, 2, 3, 5, 8`. Find the standard deviation.

We need to start by calculating the variance. We have already done that (example of variance), $s^2 = 6.25$.

$$s = \sqrt{s^2} = \sqrt{6.25} = 2.25.$$

4.18 Coefficient of variation

> - One should be careful when comparing the standard deviation of measurements that have different units or have very different means.
> - In these cases the **coefficient of variation** is used to compare the spread of two or more groups. It is denoted with CV.
>
> **Coefficient of variation**
> The **coefficient of variation** is calculated with
>
> $$CV = \frac{s}{\bar{x}}$$
>
> As CV is higher, the more spread are the measurements.

The following table shows the mean, standard deviation and the coefficient of variation of measurements in three cities in the US:

City	$\bar{x}$	s	CV
Buffalo, N.Y	35.47	4.70	0.13
St. Louis, Mo	35.56	6.62	0.19
San Diego, Calif.	9.62	4.42	0.46

Where is the largest variability.

The CV is largest in San Diego so the variability is largest there.

4.19 Comparison of statistics for spread

- All of the previously noticed statistics that describe spread are interesting and good to calculated when looking at new measurements.
- Variance and standard deviation are used to describe the spread of measurements around the mean and should therefore only be used when the mean is used to describe the central tendency.
- Standard deviation is normally preferred to variance as the unit of the standard deviation is the same as for the measurements.
- Standard deviation is sensitive to skewness and outliers. Only few outliers can increase the standard deviation greatly.
- If the measurements are skewed or if there are outliers in the dataset, a **five-number summary** and a **box plot** are the best descriptives for the spread of the measurements.

4.20 Five-number summary

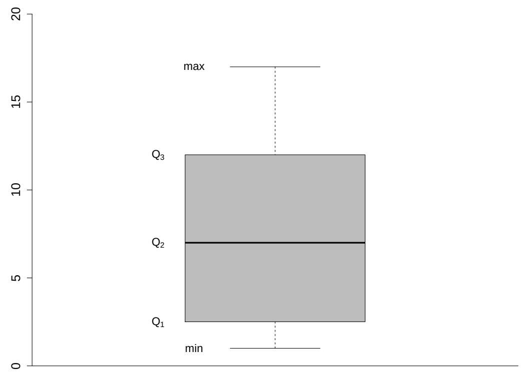
Five-number summary
A **Five-number summary** consists of the smallest value (min), the interquartile ranges and the largest value (max), that is

$$\min,\ Q_1,\ Q_2,\ Q_3,\ \max$$

We have the following measurements: `1, 2, 3, 5, 9, 9, 15, 17`. We found the quartiles in the quartiles example: og er þau: $Q_1 = 2.5$, $Q_2 = 7$ og $Q_3 = 12$.

The five number summary is:

$$\min = 1,\ Q_1 = 2.5,\ Q_2 = 7,\ Q_3 = 12 \text{ og } \max = 17.$$



- **Boxplot** is used to look at the central tendency and spread of the data.
- They reflect the measurements well and show clearly whether the distribution is symmetric or skewed.
- If the distribution is skewed, bimodal or multimodal, the five-number summary is the best method to describe central tendency and spread.
- There are a few different versions. The one seen here is the simples one.

4.21 Boxplot

4.22 Boxplot

Boxplot

- A **boxplot** consists of a box and two lines that go out of the ends of the box. These lines are often called whiskers.
- The box may lie (horizontal) or stand (vertical), we let it stand in our explanation. Then the y-axis shall cover both the smallest and the largest measurements.
- The lower end of the box shall be at Q_1 and the upper end at Q_3 a line should be drawn through the box in Q_2.
- The lower whisker should touch the lowest measurement (min) and the upper whisker should touch the tallest measurement (max).

4.23 1.5 · IQR rule for outliers

- Outliers are measurements that differ greatly from other measurements and are therefore important to identify.
- One method of determining whether a measurement is an outlier is to compare the distance of its value to the next quartile (Q_1 or Q_3).

1,5 · IQR rule for outliers

- First we calculate the distance of the potentially outlying measurement to the next quartile (Q_1 or Q_3).
- This distance is then compared to the IQR. If the distance of the measurement to the next quartile is more then $1.5 \cdot IQR$ the measurement is considered an outlier.

4.24 IQR rule

- Many statistical programs use the 1.5 · IQR rule when drawing boxplots and these boxplots are often called **modified boxplots**.
- If the highest and/or lowest value reaches further then one and half of the length of the box, the lines that go out of the box, the whiskers, stop at that value, but do not reach all the way to the highest and/or lowest value.
- The measurements outside the whiskers are outliers and labelled with a circle.