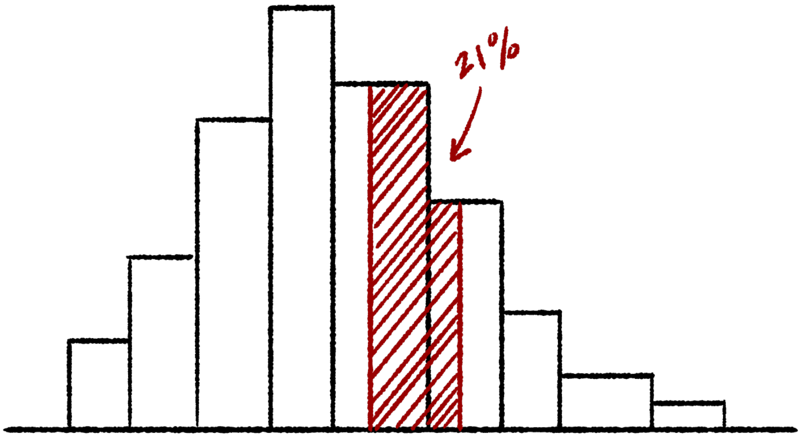
Statistics

Version 0.9
March 2, 2020



Mike Vandal Auerbach
www.mathematicus.dk

Statistics
Version 0.9, 2020

These notes are a translation of the Danish “Statistik” written for the Danish stx.

The notes do not include probability theory, but most of the sections do not require understanding of probability; the only exceptions are the closing remarks about residuals and the normal distribution, and the section on confidence intervals.

This document is written primarily for the Danish stx, but may be used freely for non-commercial purposes.

The document is written using the document preparation system LaTeX, see www.tug.org and www.miktex.org. Figures and diagrams are produced using *pgf/TikZ*, see www.ctan.org/pkg/pgf.

This document and other resources are available for download from www.mathematicus.dk

Contents

What is statistics? 1

Statistics is an area of mathematics in which we investigate data sets to describe them or to find relationships between observations. The data set which we investigate is called the *population*. So, when we talk about the *population* in statistics, we mean the entire set of persons, items or abstract objects, we wish to investigate.

The quantity we measure is called a (statistical) variable. A variable in statistics is not necessarily a number. If the population consistsof a certain group of people (e.g. Danish citizens), we can measure their height or weight, and this statistical variable is a number—but we can also write down their hair colour, and this cannot be described by a number. A variabel given by a number is called a *quantitative* variable, while a variable which is not a number is called a *qualitative* variable.

We can use statistics to simple describe different measurements to give an overview. We call this *descriptive statistics*. Here, we try to present an overview of data which at a glance might seem impossible to get an overview of.

We can create this overview in different ways, we might

- write down a table of the data set, and maybe group some of the data,
- calculate some descriptors, i.e. numbers which describe the data set, or
- draw diagrams which display the data.

In many circumstances, a data set will only be a *sample*. If we want to make a poll of voters before an election, we cannot call every single voter and ask them what they will vote in the upcoming election. Therefore, we instead take a sample and ask maybe 1000 people about their political views. Here, we need to make sure that the sample is *representative*, i.e. that the results we get from the sample correspond to the entire population. The relationship between population and sample is shown in the next two examples (see also figure 1.1).

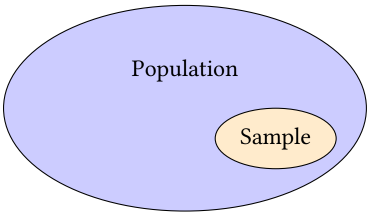


Figure 1.1: The sample is a subset of the population we want to investigate.

Example 1.1 We want to investigate voter support for a certain political party. In this case, the population is every registered voter. The sample consists of those voters, we ask in the poll.

Example 1.2 A candy company wants to investigate whether their bags of mixed candy contain equal amounts of every kind of candy. In this case, the population is all of the bags of candy which the company produces. A sample might be a random selection of bags from the comany warehouse.

Statistics is also sometimes used to find statistical models from given data. If we measure two different statistical variables, we might do e.g. a regression analysis to search for a mathematical relationship. When we analyse data in this way, it is important to remember that an apparent connection could be the result of a third so-called *hidden* variable (see section 1.2 below).

1.1 Representativity and systematic errors

When we choose a sample, it is important that the sample is *representative*. I.e. that the sample is put together in such a way that the characteristics of the sample correspond to the characteristics of the population as a whole. If the sample is not *representative*, we talk about *systematic errors*.

Example 1.3 A newspaper want to investigate the populations attitude towards public digitisation. So, they post a questionnaire on their web site.

Here, the problem is that those people who oppose digitisation do not necessarily read the newspaper on the internet. They will therefore be underrepresented (or completely misssing) in the poll. Therefore, the sample is not representative.

We get systematic errors when certain positions or properties are over- or underrepresented in the sample compared with the population. An often cited example of a systematic error in sample selection is the Literary Digest's prediction of the winner of the American presidential election in 1936:

Example 1.4 In 1936, the American magazine Literary Digest predicted that Alfred Landon would win the American presidential election with 57% of the votes. Instead, the president in office, Franklin D. Roosevelt, won the election with 62% of the votes. The magazine came to the wrong conclusion even though they had posted questionnaires to 10 million Americans and received 2.4 million answers.[2]

Two things went wrong in this poll. First of all, the magazine found the addresses of the 10 million Americans via automobile clubs, telephone books, and their own list of subscribers. In 1936, during the height of the depression, Americans who owned a car of a telephone, or subscribed to a magazine probably belonged to the wealthiest part of society.

But it was probably the second design flaw, which contributed most to the wrong conclusion:[5] The poll was based on the answers, the magazine received—so it is entirely possible that a certain voting position was overrepresented among those who actually took their time to answer.

As the examples above show, we need to consider quite carefully how to choose a sample. In voting polls and similar investigations, where we examine the position of a population on some subject, we usually choose samples of around 1000 people. This is usually enough to ensure that the sample mirrors the populations—but we need to be careful how we select the sample, and we need to make sure that the position of those people who do not want to answer a poll is still represented.

1.2 Hidden variables

We might also use statistics to look for a relationship between different quantitites. Here, we have to make sure that the the relationship we see is actually a relationship between these two quantities and not the result of some common cause. If this is the case, we talk about *hidden variables*.

Example 1.5 If we look at ice cream sales and drowning accidents, we find that when ice cream sales are high, more drowning accidents happen. We might therefore draw the conclusion that eating ice cream increases the risk of drowning.

This is of course nonsense. If we instead look at both variables (ice cream sales and drowning accidents) and compare them with the weather, we quickly determine that on warm days, ice cream sales increase and so does the number of people going to the beach—which in turn increases the number of drowning accidents.

In this case, it is the heat which is the hidden variable on which the others depend.

2 Ungrouped statistics

In ungrouped statistics, we describe separate data. As an example, we might consider asking a high school class of 25 students how many times they have been to the cinema during the last year. The answers might look like table 2.1.

Table 2.1: Visits to the cinema (unsorted).

1	0	3	2	4
2	3	4	6	5
4	2	3	4	0
4	4	5	3	3
1	0	0	4	5

It is hard to get an overview of these numbers. The first thing we might do, therefore, is to sort them. This is done in table 2.2.

Table 2.2: Visits to the cinema (sorted).

0	0	0	0	1
1	2	2	2	3
3	3	3	3	4
4	4	4	4	4
4	5	5	5	6

As table 2.2 shows, some of the numbers occur several times. We might therefore construct a table of the different numbers and their *frequencies* (i.e. how many times they occur). The table might look like this:

Observation, x	Frequency, $h(x)$	Rel. freq., $f(x)$	Cum. fr., $F(x)$
0	4	16%	16%
1	2	8%	24%
2	3	12%	36%
3	5	20%	56%
4	7	28%	84%
5	3	12%	96%
6	1	4%	100%
I alt	25	100%	

The first two columns in the table show the observation, i.e. the number of visits to the cinema, and the frequency. The next column shows the relative frequency, i.e. how large a fraction this number makes up of the entire data set.

The last column shows the *cumulative relative frequency* which shows how large a fraction of the data set is made up of this observations *up to and including* the observation in question. The cumulative relative frequency for 3 visits to the cinema is 56%, because 56% have been to the cinema at most 3 times—in other words: if we count up to and including 3 visits to the cinema, we will have counted 56% of the students.

The following definition provides an overview of the three quantities connected to the observation:

Definition 2.1

For a data set with n observations $x_1, x_2, \dots, x_n$, we define the following quantities:

1. The *frequency* $h(x)$ is the number of times the observation x occurs in the data set.
2. The *relative frequency* $f(x)$ is the frequency as a fraction of the number of observations, i.e. $f(x) = \frac{h(x)}{n}$.
3. The *cumulative relative frequency* $F(x)$ is the sum of the relative frequencies *up to and including* the relative frequency of the observation in question, i.e.[1]

$$F(x) = \sum_{t \le x} f(t) \ .$$

So, the frequencies in the table above show how large a fraction of the students have been to the cinema 0 times, 1 time, etc. This is useful if we want to compare two classes that do not have the same number of students. We often write the relative frequency as a percentage, but we do not have to.

The cumulative relative frequency shows how many students have been to the cinema x times *or less.* The cumulative relative frequency for the observation $x = 2$ is 36%. This means that 36% of the students have been to the cinema 2 times or less. We find the number by adding the relative frequencies for the observations $x = 0$, $x = 1$, and $x = 2$:

$$F(2) = f(0) + f(1) + f(2) = 16\% + 8\% + 12\% = 36\% \ .$$

2.1 Range, mode, and mean

Even though a table, such as the one above, gives an overview of a data set, it is sometimes easier to compare data sets if we can describe them via a few numbers, so-called *descriptors.*

In the table, we see that the smallest value is 0, and the largest is 6. This allows us to calculate the so-called *range* which is the difference between the smallest and the largest value. In this case, the range is

$$x_{\text{max}} - x_{\text{min}} = 6 - 0 = 6 \ .$$

The *mode* is the observation which occurs the largest number of times. In our case the mode is 4—i.e. most students have been to the cinema 4 times.

A descriptor requiring a bit more calculation is the mean which tells us what the average observation is. We find the mean by adding all of the observations and dividing by the number of observations. For the numbers in table 2.2, the mean is

$$\overline{x} = \frac{0 + 0 + 0 + 0 + 1 + \cdots + 5 + 5 + 5 + 6}{25} = 2.88 \ .$$

[1] In this context, the symbol $\sum_{t \le x}$ means that we take the sum of all values less than or equal to x.

As we have already found the frequencies of all of the observations (in the table we see, e.g., that the observation “2” occurs 5 times), we can use the frequencies to instead write the calculation as

$$\overline{x} = \frac{0 \cdot 4 + 1 \cdot 2 + 2 \cdot 3 + 3 \cdot 5 + 4 \cdot 7 + 5 \cdot 3 + 6 \cdot 1}{25} = 2.88 \ .$$

This, of course, does not change the result.

Because we get the relative frequencies by dividing all of the frequencies by the number of observations, we might also start out by dividing all of the frequencies by 25 before we calculate the mean[2]

$$\overline{x} = 0 \cdot 0.16 + 1 \cdot 0.08 + 2 \cdot 0.12 + 3 \cdot 0.20 + 4 \cdot 0.28 + 5 \cdot 0.12 + 6 \cdot 0.04 = 2.88 \ .$$

Definition 2.2

For a data set with n observations $x_1, x_2, \ldots, x_n \in X$, we define the mean $\overline{x}$ as[3]

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{x \in X} x \cdot h(x)}{n} = \sum_{x \in X} x \cdot f(x) \ .$$

The mean shows the average observation. When $\overline{x} = 2.88$ for our data set, it means that the 25 students have been to the ciname 2.88 times on average.

We sometimes use the symbol μ for the mean of the entire population. We might view the 25 students as a sample of the entire population of Danish high school students (or young people between the ages of 15 and 20). In this case, $\overline{x}$ is an *estimate* of the true average μ of the population (and this value is unknown).

2.2 Quartiles

The mean of a data set changes drastically if extreme values occur. E.g. if a single student had been to the cinema 40 times, the mean would have been a lot larger. Therefore, we sometimes describe a data set using the *median*, which is the observation in the middle.

If we list all of the 25 numbers in table 2.2, the median will be the number in the middle, i.e. number 13:[4]

$$0, 0, 0, 0, 1, 1, 2, 2, 2, 3, 3, 3, \overset{\text{median}}{\boxed{3,}} 3, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 6$$

So, the median of the students’ visits to the cinema is 3. This means that half of the students have been to the cinema 3 times or less. The other half has been to the ciname 3 times or more. It is important to note that this number has nothing to do with the mean, and we see that the two numbers are actually different.

[2] Note that we write the relative frequencies as decimals. E.g. the frequency of the first observation is not 16, but 16%, which is the same as 0.16.

[3] The symbol $\sum_{i=1}^{n}$ shows that we add every observation from 1 to n, while $\sum_{x \in X}$ shows that we add all of the *different* observations.
In this context, X is the set of all possible observations.

[4] If we have an even number of observations, the median is the average of the two middle observations.

Sometimes, we wish for more information than the median alone can provide. We find the median by splitting the data set into two halves. We acquire more information if we split the data set into four quarters. When we do this, we find the so-called *quartiles*:

$$0, 0, 0, 0, 1, 1, \overset{\text{lower quartile}}{|} 2, 2, 2, 3, 3, 3, \overset{\text{median}}{\boxed{3,}} 3, 4, 4, 4, 4, 4, \overset{\text{upper quartile}}{|} 4, 4, 5, 5, 5, 6$$

The *lower quartile* is the median of the lower half of the data. Because the lower half of the data contains an even amount of observations (12), the lower quartile is the average of the to middle observations (number 6 and 7). Therefore, the lower quartile is

$$Q_1 = \frac{1+2}{2} = 1.5 \ .$$

The median is teh same number as before, i.e.

$$m = 3 \ .$$

The upper quartile is the median of the upper half. Here, we again take the average of two numbers, i.e.

$$Q_3 = \frac{4+4}{2} = 4 \ .$$

The three numbers Q_1, m, and Q_3 make up the quartiles, and in our case, the quartiles for the number of visits to the cinema are $(1.5, 3, 4)$.

We sometimes refer to the three numbers as the “first, second and third quartile” instead of “upper quartile, median, and lower quartile”.

Definition 2.3

For an ungrouped data set, we define

- The *medianen* (or *second quartile*) m, as the middle observation. If the data set has an even number of observations, the median is the average of the two middle observations.
- The *lower* (or *first*) *quartile* Q_1, as the median of the lower half of the observations.
- The *upper* (or *third*) *quartile* Q_3, as the median of the upper half of the observations.

The quartiles are the numbers (Q_1, m, Q_3).

The set of numbers $(x_{\text{min}}; Q_1; m; Q_3; x_{\text{max}})$ containing the smallest observation, the quartiles, and the largest observation, is known as the *five-number summary.*

Because the lower quartile in our case was 1.5, we know that a quarter (25%) of the students went to the cinema 1.5 times or less, while three quarters (75%) went to the cinema 1.5 times or more.

The upper quartile $Q_3 = 4$ shows that three quarters of the students went to the cinema 4 times or less, while a quarter went to the cinema 4 times or more.

If we want a complete overview of the distribution, we sometimes write down the *five-number summary*, which (as described) contains the quartiles, and the smallest and largest observation. In this case, the five-number summary is

$$(0, 1.5, 3, 4, 6) ,$$

i.e. the smallest observation is 0, the lower quartile is 1.5, the median is 3, the upper quartile is 4, and the largest observation is 6.

Another quantity we might calculate is the *interquartile range*, which is the distance between Q_1 and Q_3. In our case of cinema visits, the interquartile range is

$$Q_3 - Q_1 = 4 - 1.5 = 2.5 .$$

So, the sample of visits to the cinema can now be described with the following descriptors

Descriptor		Value	
Minimum	$x_{\min}$	0	Five-number summary
Lower quartile	Q_1	1.5	
Median	m	3	
Upper quartile	Q_3	4	
Maximum	$x_{\max}$	6	
Mean	$\overline{x}$	2.88	
Mode		4	
Interquartile range	$Q_3 - Q_1$	2.5	
Range	$x_{\max} - x_{\min}$	6	

2.3 Outliers

We can imagine asking 10 new students about how many times they have been to the cinema and receiving the answers in table 2.3. This set of observations has the mean

$$\overline{x} = 2.8 ,$$

and the five-number summary is

$$(0, 2, 2.5, 4, 8) .$$

Table 2.3: New sample of ciname visits

4	3	2	0	2
3	2	4	8	0

If we look at this data set, we see that a single observation (the student who has been to the ciname 8 times) is quite large compared to the others. In this case, we might have a so-called *outlier*, i.e. an observation which is far from the typical observation. We have the following definition:

Definition 2.4

In a set of observations, an observation x is called an *outlier* if it more than 1.5 times the interquartile range below the lower quartile or above the upper quartile.

In other words, x is an outlier when

$$x < Q_1 - 1.5 \cdot (Q_3 - Q_1) \quad \text{or} \quad x > Q_3 + 1.5 \cdot (Q_3 - Q_1) \, .$$

In the case above, the interquartile range is

$$Q_3 - Q_1 = 4 - 2 = 2 \, .$$

So, 1.5 times the interquartile range below or above the median corresponds to

$$\begin{aligned} Q_1 - 1.5 \cdot (Q_3 - Q_1) &= 2 - 1.5 \cdot 2 = -1 \\ Q_3 + 1.5 \cdot (Q_3 - Q_1) &= 4 + 1.5 \cdot 2 = 7 \end{aligned}$$

Because the observation 8 is larger than 7, it is an outlier. In the same way, every observation below −1 is an outlier (but in this case we cannot get negative numbers as observations).

2.4 Skewness

For the first 25 students, we found a mean of 2.88 and a median of 3. Such a distribution where the mean is less than the median is called a *left-skewed* distribution.

In the previous section, we looked at a data set (10 students) where the mean was 2.8, and the median was 2.5. Here, the mean is larger than the median. This distribution is therefore *right-skewed*.

Definition 2.5

A data set has a

- *left-skewed distribution* when the mean is less than the median, $\overline{x} < m$,
- *non-skewed distribution* when the mean is equal to the median $\overline{x} = m$, or a
- *right-skewed distribution* when the mean is larger than the median, $\overline{x} > m$.

If the distribution is left- or right-skewed, a graphical illustration of the distribution will show this, cf. section 2.6.

2.5 Standard deviation

The *standard deviation* is a measurement which shows how far the observations are on average from the mean. The standard deviation is defined to be

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}} = \sqrt{\frac{\sum_{x \in X}(x - \mu)^2 \cdot h(x)}{n}} \ . \tag{2.1}$$

Our problem is that we cannot calculate this quantity based only on a sample. If we only have a sample, we cannot determine the true mean μ of the population, but only an estimate given by the sample mean $\overline{x}$. But if we just use $\overline{x}$ instead of μ, we will always come up with a too small estimate of the standard deviation.

Example 2.6 At a certain high school, the mean of the boys' height is 173 cm. Now, we take a sample to estimate this mean. We measure the height of 3 boys to be 168, 176 and 181 cm. The mean of these height is then

$$\overline{x} = \frac{168 + 176 + 181}{3} = 175 \ .$$

This quantity is an estimate of the true mean, which we know in this case to be 173 cm.

If we use the true mean to calculate the standard deviation, we get

$$\sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}} = \sqrt{\frac{(168 - 173)^2 + (176 - 173)^2 + (181 - 173)^2}{3}} = 5.72 \ .$$

If we did not know the true mean, we would have to use the estimate $\overline{x}$, and we would instead get

$$\sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}} = \sqrt{\frac{(168 - 175)^2 + (176 - 175)^2 + (181 - 175)^2}{3}} = 5.35 \ .$$

So, we get a too small estimate for σ when we use $\overline{x}$ as an estimate for μ.

If we instead of dividing by 3 in the calculation above divide by 1 less (i.e. 2), we get

$$\sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n - 1}} = \sqrt{\frac{(168 - 175)^2 + (176 - 175)^2 + (181 - 175)^2}{2}} = 6.56 \ .$$

This number is a too large estimate for the true standard deviation; but we always prefer to have a too larger rather than a too small estimate.

Because we get a too small estimate when we use $\overline{x}$ instead of μ in the formula (2.1), we divide by $n - 1$ instead of n; in this way we get a better estimate for the standard deviation:

Definition 2.7

For a sample containing the elements $x_1, x_2, \dots, x_n$, we define the *sample standard deviation* to be the number

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = \sqrt{\frac{\sum_{x \in X}(x - \overline{x})^2 \cdot h(x)}{n-1}} \ .$$

If we look again at the sample of 25 students' visits to the cinema, we find a sample standard deviation of

$$\begin{aligned} s &= \sqrt{\frac{\sum_{x \in X}(x - \overline{x})^2 \cdot h(x)}{n-1}} \\ &= \sqrt{\frac{(0-2.88)^2 \cdot 4 + (1-2.88)^2 \cdot 2 + \cdots + (6-2.88)^2 \cdot 6}{25-1}} \\ &= 1.76 \ . \end{aligned}$$

For the following sample of 10 students, the sample standard deviation is $s = 2.30$. The standard deviation is larger here because this data set contains an outlier.

2.6 Diagrams

In this section, we show 3 different types of diagrams, which can be used to describe an ungrouped data set:

- A bar chart, which is useful if we want to illustrate a single data set.
- A cumulative relative frequency graph.
- A box plot, which is useful when we want to compare different data sets.

Bar chart

The sample of 25 students' visits to the cinema yielded the following table:

Observation, x	Frequency, $h(x)$	Rel. fr., $f(x)$	Cum. rel. fr., $F(x)$
0	4	16%	16%
1	2	8%	24%
2	3	12%	36%
3	5	20%	56%
4	7	28%	84%
5	3	12%	96%
6	1	4%	100%
Total	25	100%	

This table enables us to draw a bar chart. The x-axis represents the individual observations, and at each observation we draw a bar, whose height equals the frequency of the observation.

In figure 2.4, we see a bar chart where the height of the columns indicates the frequency. Figure 2.5 shows the same bar chart, but here the heights indicate the relative frequencies. The two charts are identical except for the numbers on the y-axis.

If we just want a quick description of the data set, we might just as well use the frequencies. But if we want to compare two data sets, it is easier when we use the relative frequencies—especially if the two data sets contain a different number of observations. This might be the case if we were comparing two high school classes with a different number of students.

The distribution of the observations in this data set is—as previously mentioned—left-skewed. If we look at the bar chart, we see that the "weight" of the diagram appears to be shifted towards the left. If the distribution were right-skewed, the bars would instead have seemed to be shifted to the right.

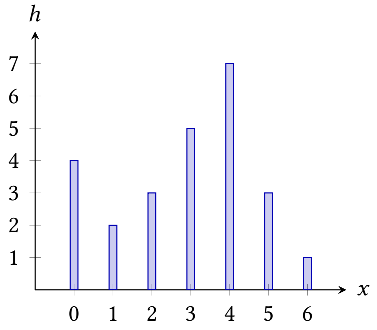


Figure 2.4: The students' visits to the cinema as a bar chart using the frequencies.

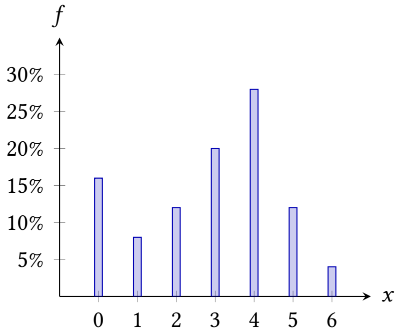


Figure 2.5: The students' visits to the cinema as a bar chart using the relative frequencies.

Cumulative relative frequency graph

A cumulative relative frequency graph is a graph of the cumulative relative frequencies. We plot the cumulative relative frequency at the corresponding observation, and the we move horizontally until we get to the next observation, where we jump to the next cumulative relative frequency. In this way, we get a graph that looks a bit like a set of steps—a function, which has a such a graph is called a step function, see figure 2.6.

It is possible to use cumulative relative frequency graphs to compare different data sets. But the box plot, which we describe below, is a much easier tool to use for comparisons.

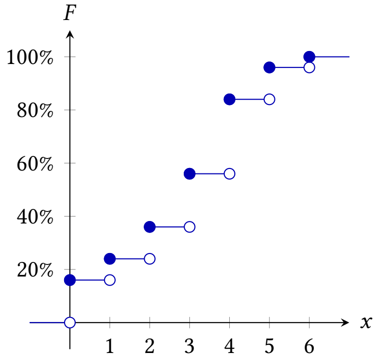


Figure 2.6: Cumulative relative frequency graph of the students' visits to the cinema.

Box plot

A box plot is a diagram drawn using only the quartiles. When we do this, we discard a lot of information. But in return, we get a diagram which shows us how the numbers are distributed in way that is easy to read.

A box plot of our data set can be seen in figure 2.7. We draw vertical lines at the minimum value (0), the lower quartile (1.5), the median (3), the upper quartile (4), and at the maximum value (6). Then we connect the vertical lines as shown in the figure.

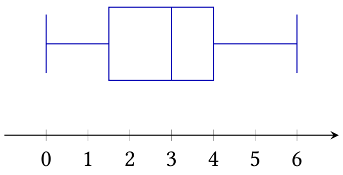


Figure 2.7: Box plot of the students' visits to the cinema.

The box contains the middle half of the observations, while the horizontal lines at both ends show the hours of TV watched for the lower and the upper quarter of the class.

When we draw box plots of different distributions, they are easy to compare. If we have the five-number summary for the first data set of 25 students (A) and the following data set of 10 students (B), we get the following table:

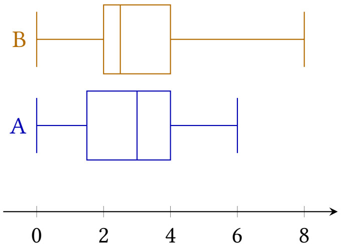


Figure 2.8: Comparing visits to the cinema.

Data set	Minimum	Q_1	m	Q_3	Maximum
A	0	1.5	3	4	6
B	0	2	2.5	4	8

If we just look at the numbers, it is hard to tell what the difference is between the two classes. If, however, we draw box plots of both in the same diagram (see figure 2.8), they are much easier to compare.

Here we see that even though B has the student with the largest number of visits to the cinema, the lower 50% of B have been to the cinema a little less often than the lower 50% of A. The middle half of B is closer than the middle half of A, which means that the interquartile range is less here (although B still has a larger sample standard deviation than A, cf. the previous section).

Grouped statistics 3

We talked about grouped statistics when the data set is grouped in intervals. This might be the case if the data set is very large, or if we measure data with a lot of decimals. Here, we will typically have a large number of observations, and it makes sense to group them into intervals. The intervals will usually be adjacent; but this is not strictly necessary. However, the intervals must always be separate—i.e. the same observation cannot belong to different intervals.

Table 3.1 shows a sample from a company producing bags of sugar. The table shows a the weight of 500 bags. Here, we have so many observations that nothing would be gained by listing all of the individual weights making up the table. Therefore, the different weights are instead grouped into intervals.

Table 3.1: Sample of the weight of bags of sugar.

Interval (grammes)	Number
800-850	11
850-900	17
900-950	53
950-1000	208
1000-1050	125
1050-1100	86

When we look at the table, we cannot immediately see whether a weight of 850 g should be counted in the first or the second interval. It would therefore be a good idea to use mathematical interval notation to describe whether the weights bordering two intervals belong to one or the other.

We can also see that the frequencies are quite large numbers. The corresponding relative frequencies are these:

Interval	Frequency, h	Rel. freq., f	Cum. rel. fr., F
$[800; 850[$	11	2.2%	2.2%
$[850; 900[$	17	3.4%	5.6%
$[900; 950[$	53	10.6%	16.2%
$[950; 1000[$	208	41.6%	57.8%
$[1000; 1050[$	125	25.0%	82.8%
$[1050; 1100[$	86	17.2%	100.0%
I alt	500	100.0%	

3.1 Mean and standard deviation

We cannot calculate the mean and the sample standard deviation as we did with ungrouped statistics. This is because we do not know how the weights are distributed within the individual intervals; we do not have the raw data for the table.

Instead, we assume that the weights are evenly distributed in the intervals.

This allows us to use the midpoints of the intervals as a substitute for the individual observations.

Definition 3.1

For a data set grouped into n intervals, $[a_1; b_1[\,, [a_2; b_2[\,, \ldots, [a_n; b_n[\,$, where the total number of observations is N, the mean is

$$\overline{x} = \frac{\sum_{i=1}^{n} m_i \cdot h_i}{N} = \sum_{i=1}^{n} m_i \cdot f_i \ ,$$

and the sample standard deviation is

$$s = \sqrt{\frac{\sum_{i=1}^{n} (m_i - \overline{x})^2 \cdot h_i}{N-1}} \ .$$

h_i is the frequency of the interval, f_i is the relative frequency, and m_i is the midpoint of the interval, $m_i = \frac{a_i+b_i}{2}$.

To determine the mean of the above data set, we add another column of interval midpoints:

Interval	Interval midpoint, m	Rel. frequency, f
[800; 850[	825	2.2%
[850; 900[	875	3.4%
[900; 950[	925	10.6%
[950; 1000[	975	41.6%
[1000; 1050[	1025	25.0%
[1050; 1100[	1075	17.2%

The mean is then

$$\overline{x} = 825 \cdot 0.022 + 875 \cdot 0.034 + \cdots + 1075 \cdot 0.172 = 992.7 \ .$$

So, the average weight in the table is 992.7 g.

3.2 Diagrams

In this section, we describe three ways of illustrating grouped data:

- Histograms, which correspond to bar charts of ungrouped data.
- Cumulative relative frequency graphs which can be used to determine the quartiles.
- Box plots, which are exactly the same type of diagram as a box plot of ungrouped data.

Histograms

In a histogram, the relative frequencies of the intervals are drawn as columns. For ungrouped data, we could draw a bar chart—and the height of the bars corresponded to the relative frequencies. Here, we cannot do that, since then wider intervals would carry more weight than narrow intervals.

Instead, we let the *area* of the columns correspond to their (relative) frequency, see figure 3.2.

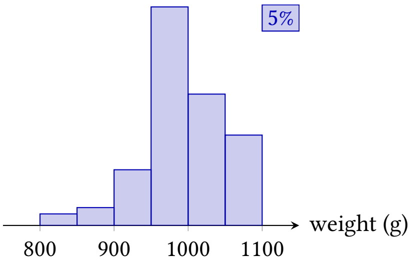


Figure 3.2: Histogram for the distribution of weights. The area corresponds to the frequency.

When the relative frequency is given by the area, we need to show which area corresponds to a certain percentage. This is illustrated in the figure, where the rectangle in the upper right hand corner shows, which area corresponds to 5%.

Since the area shows the relative frequency, we have no use for a y-axis, so this is usually omitted.

If, however, all of the intervals are of equal width, we can let the height correspond the relative frequency. A lot of CAS tools illustrate data this way. But when we draw histogram, we need to remember that the intervals *have to be* equally wide.

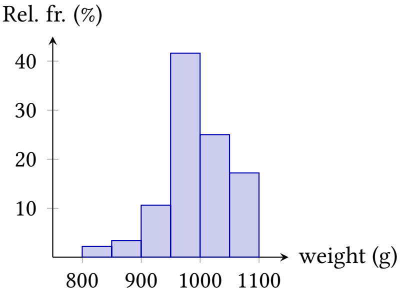


Figure 3.3: Histogram for the distribution of weights. The height corresponds to the frequency.

In our case, the intervals are actually of equal with, i.e. it is allowed to draw the histogram as in figure 3.3.

So, for a histogram it is important to remember that

> the interval's relative frequency is the *area* of the corresponding column—unless all of the intervals are of *equal width*.

Cumulative relative frequency graphs

A cumulative relative frequency graph illustrates how many percent of a data set falls below a given value. Because the curve shows the percentage *below* a given value, we use the *right* interval end points as x-values and the cumulative relative frequencies as y-values.

Therefore, we add a column of interval end points to the table above:

Interval	Right interval end point	Cum. rel. frequency
[800; 850[	850	2.2%
[850; 900[	900	5.6%
[900; 950[	950	16.2%
[950; 1000[	1000	57.8%
[1000; 1050[	1050	82.8%
[1050; 1100[	1100	100.0%

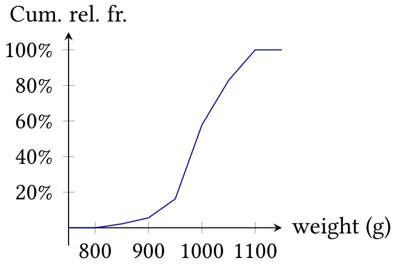
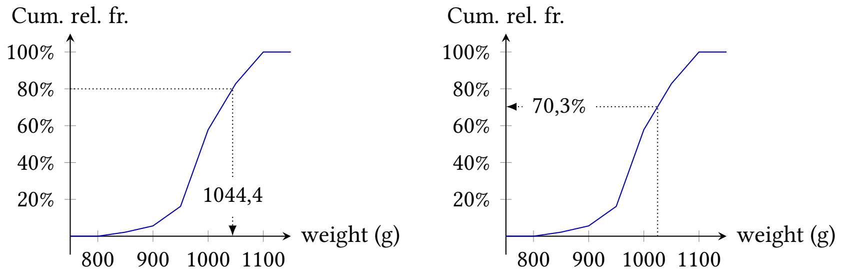


Figure 3.4: Cumulative relative frequency graph of the weight distribution.

After we have plotted the cumulative relative frequencies against the interval end points, we connect the points with straight lines.

This curve which is also called a distribution curve shows how many percent of the bags' weights are below a given value. This means that we can use the curve to answer questions such as how many percent of the



(a) The 80th percentile

(b) How many bags weigh less than 1025 g?

Figure 3.5: In the figure on the left, we find the 80th percentile. The number shows that 80% of the bags weigh less than 1044.4 g. On the right, we find 1025 on the x-axis. The corresponding cumulative relative frequency shows that 70.3% of the bags weigh 1025 g or less.

bags weigh less than 1025 g, or what the largest weight is for the lightest 80% of the bags. The last number is called the *80th percentile*. We have the following definition:

Definition 3.2

For a data set, the pth percentile is the value in the data set for which the cumulative relative frequency is $p\%$.

In figure 3.5, we find the 80th percentile. We find 80% on the y-axis and then the corresponding value on the x-axis. This number (1044.4) shows that 80% of the bags in the sample weigh 1044.4 g or less. Similarly, 20% of the bags weigh 1044.4 g or more.

The figure also shows how to find the percentile corresponding to a weight of 1025 g. Here, we find 1025 on the x-axis and then find the corresponding number on the y-axis. We find 70.3%, which means that 70.3% of the bags weigh less than 1025 g, and 29.7% weigh more than 1025 g.

We also use the distribution curve to define the quartiles.

Definition 3.3

From a cumulative relative frequency graph, we find the quartiles (Q_1, Q_2, Q_3).

1. The *lower quartile*, Q_1 is the 25th percentile.
2. The *median*, Q_2 is the 50th percentile.
3. The *upper quartile*, Q_3 is the 75th percentile.

Figure 3.6 shows how to find the quartiles. We find 25%, 50% and 75% on the y-axis, and then find the corresponding values on the x-axis. Here, we see that the quartiles are

$$(960.6, 990.6, 1034.4)\ .$$

These numbers show that

- 25% of the bags weigh 960.6 g or less,
- 50% of the bags weigh 990.6 g or less, and
- 75% of the bags weigh 1034.4 g or less.

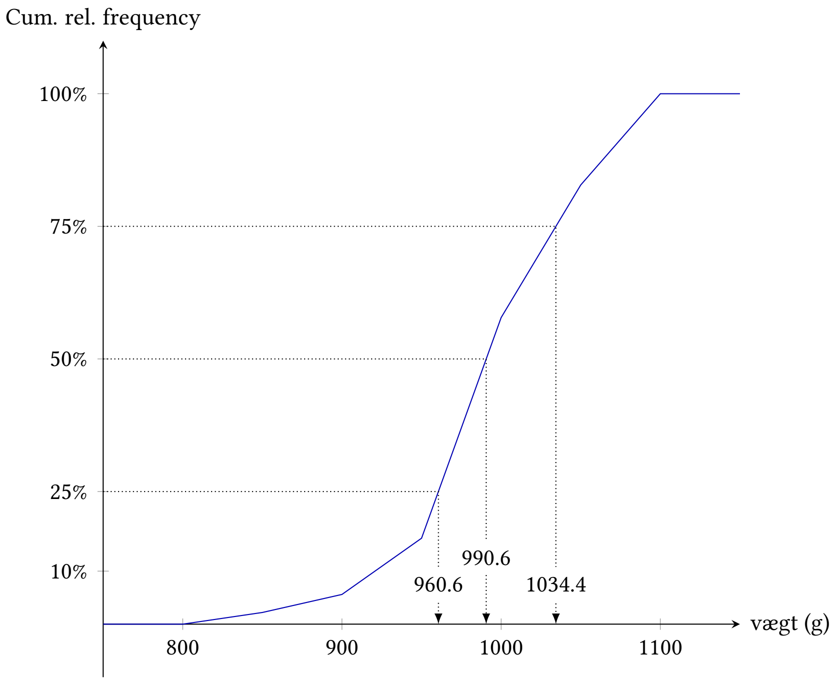


Figure 3.6: Finding the quartiles on the distribution curve for the weights.

Box plot

A box plot for a grouped data set is exactly the same as a box plot of an ungrouped data set. The only difference between the two is how we determine the quartiles. When we have this, we do exactly the same.

For the distribution of weighs above, the quartiles were

$$(960.6, 990.6, 1034.4) .$$

The lowest value was 800 and the largest was 1100. The five-number summary is therefore

$$(800, 960.6, 990.6, 1034.4, 1100) ,$$

and a box plot for this distribution will look like figure 3.7.

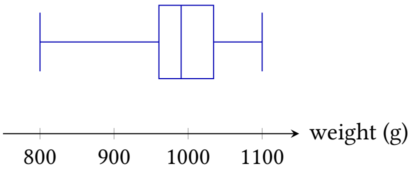


Figure 3.7: Box plot for the distribution of weights.

Linear regression 4

If we measure a series of corresponding values of two variables where one is related to the other, we can sometimes set up a model of the relationship between the two variables.

When we have a data set $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, we can try to model the relationship with a function f, so that the graph of f is as close to the data points as possible. Because there is always measurement errors, such a graph will never pass through all of the data points. The difference between the model's y-value $\hat{y}_i = f(x_i)$ (also known as the *estimated* value) and the measured y-value y_i is called the *residual*. For the data point (x_i, y_i), the residual is

$$r_i = y_i - \hat{y}_i \,.$$

A way to determine the function is to look for the function f which minimises the residuals in total. A measurement for the total difference is given by the summed squares of the residuals[1]

$$SSE = r_1^2 + r_2^2 + \cdots + r_n^2 \,.$$

We want to minimise this quantity. Because we look at the squares of the residuals, the method is also called the *method of least squares*.

If the function f we are looking for, is a linear function, $f(x) = ax + b$, the method will give us the straight line which best fits the n points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. The residuals will then have the form

$$r_i = y_i - (ax_i + b) \,.$$

The sum of squares SSE of the residuals is then

$$SSE = r_1^2 + r_2^2 + \cdots + r_n^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2 \,. \tag{4.1}$$

The straight line we are looking for, is the line which minimises the sum of squares SSE.

In this case, it turns out that we can calculate the numbers a and b in the following way:

[1] SSE is an abbreviation of "sum of squares of error of prediction", the errors in this case are the residuals.

Theorem 4.1

For the data points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, we find the best straight line $y = ax + b$, where

$$a = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} \ ,$$

and

$$b = \overline{y} - a \cdot \overline{x} \ .$$

Here, $\overline{x}$ is the average of the x-values, $\overline{y}$ is the average of the y-values, $\overline{x \cdot y}$ is the average of $x \cdot y$, etc.

Table 4.1: Corresponding values of x and y.

x	y
0	1
2	3
4	6
6	8

Example 4.2 Table 4.1 shows corresponding values of the independent variable x and the dependent variable y. To use the formulas, we need a series of averages. These have been calculated in table 4.2.

Table 4.2: x, y, $x \cdot y$ and x^2. The bottom row lists the averages.

x	y	$x \cdot y$	x^2
0	1	0	0
2	3	6	4
4	6	24	16
6	8	48	36
$\overline{x}$	$\overline{y}$	$\overline{x \cdot y}$	$\overline{x^2}$
3	4.5	19.5	14

We can now calculate

$$a = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{19.5 - 3 \cdot 4.5}{14 - 3^2} = 1.2 \ .$$

and

$$b = \overline{y} - a \cdot \overline{x} = 4.5 - 1.2 \cdot 3 = 0.9 \ .$$

So, the best-fit straight line has the equation

$$y = 1.2x + 0.9 \ .$$

The points and the line are shown in figure 4.3.

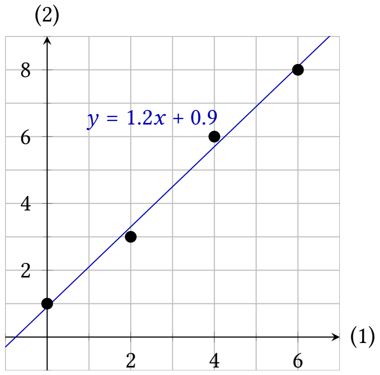


Figure 4.3: The best-fit straight line through the 4 points.

As the example shows, it involves a lot of work to use the formulas in theorem 4.1 to calculate the numbers a and b—especially if we have many data points. Fortunately, most CAS's have the method built in, which means we can enter the points and have the tool calculate the numbers.

Proof of the formulas

To prove the formulas in theorem 4.1, we need to know where a sum of squares has its minimum.

Theorem 4.3

The sum of squares $q(c) = \sum_{i=1}^{n} (z_i - c)^2$ has its minimum where $c = \overline{z}$.

Proof
We have the sum of squares

$$q(c) = \sum_{i=1}^{n} (z_i - c)^2 \ .$$

So, $q(c)$ is a function of c given by

$$q(c) = (z_1 - c)^2 + (z_2 - c)^2 + \cdots + (z_n - c)^2 \ .$$

We can now rewrite the expression $q(c)$ in the following way,[2]

$$\begin{aligned} q(c) &= \sum_{i=1}^{n}(z_i - c)^2 \\ &= \sum_{i=1}^{n}\left(z_i^2 + c^2 - 2z_i c\right) \\ &= nc^2 - \left(2\sum_{i=1}^{n} z_i\right)c + \sum_{i=1}^{n} z_i^2 \\ &= nc^2 - (2n\overline{z})c + \sum_{i=1}^{n} z_i^2 \ . \end{aligned}$$

Therefore $q(c)$ is a quadratic function of c. A quadratic function $y = Ac^2 + Bc + C$ where $A > 0$ has its minimum at the vertex, here $c = -\frac{B}{2A}$. In $q(c) = nc^2 - (2n\overline{z})c + \sum_{i=1}^{n} z_i^2$, the coefficients are

$$A = n \ , \quad B = -2n\overline{z} \quad \text{og} \quad C = \sum_{i=1}^{n} z_i^2 \ .$$

So, $q(c)$ has its minimum where

$$c = -\frac{-2n\overline{z}}{2n} = \overline{z} \ . \qquad \blacksquare$$

Theorem 4.3 tells us that S has a minimum when[3]

$$b = \overline{y - ax} = \overline{y} - a \cdot \overline{x} \ . \tag{4.2}$$

Now, we have an expression for the line's y-axis intercept. To find an expression for the slope a, we insert the expression (4.2) into the expression for the SSE from (4.1):

$$\begin{aligned} SSE &= \sum_{i=1}^{n}(y_i - ax_i - b)^2 \\ &= \sum_{i=1}^{n}(y_i - ax_i - \overline{y} + a\overline{x})^2 \\ &= \sum_{i=1}^{n}((y_i - \overline{y}) - a(x_i - \overline{x}))^2 \\ &= \left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)a^2 - \left(2\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})\right)a + \sum_{i=1}^{n}(y_i - \overline{y})^2 \ . \end{aligned}$$

This is a quadratic function of a, which has its minimum where

$$a = -\frac{-2\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{2\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \ .$$

Through quite a lot of calculations, we can show that this fraction can also be written as

$$a = \frac{\sum_{i=1}^{n} x_i y_i - n \cdot \overline{x} \cdot \overline{y}}{\sum_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2} \ ,$$

which we can reduce further to arrive at the formulas in theorem 4.1.

[2] We use that $\sum_{i=1}^{n} c^2 = nc^2$, and that $\sum_{i=1}^{n} z_i = n\overline{z}$.

[3] We set $z_i = y_i - ax_i$ and $c = b$ in the expression from the theorem.

4.1 Coefficient of determination

We can always use the formulas in theorem 4.1 to calculate the best-fit straight line, but this in no way guarantees that the points are on a line to any degree of accuracy. Therefore, we define the so-called *coefficient of determination* which measures how well the calculated line fits the data points.

If there is no relationship between the x- and y-values of the data points, we would expect the y-values to vary randomly around the average $\overline{y}$ independently of the x-value. We can calculate the sum of the squares of the differences between the y-values and the expected y-value (in this case $\overline{y}$); this is

$$S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2 \ .$$

S_{yy} is the sum of the squares of the errors when we assume that there is no relationship between x and y.

But actually, we do expect a relationship between x and y, and in this case the errors are described by the sum of squares SSE which is the sum of the squares of the errors when we model the given data with a linear function. This will be less than S_{yy}. The coefficient of determination is then defined to be the number

$$R^2 = \frac{S_{yy} - SSE}{S_{yy}} \ .$$

So, the number R^2 shows how many percent SSE is less than S_{yy}. If SSE is very small compared to S_{yy}, this number will be close to 1, while it will be close to 0 when SSE is almost as large as S_{yy}, i.e. when the linear function is not much closer to the points than a vertical line through the average of the y-values.

The coefficient of determination is a good measure for how well the line fits the given data points, but it cannot stand on its own. When we perform linear regression to find the best-fit straight line, we can do so without drawing the graph. So, we can have a CAS calculate the equation of the line and the coefficient of determination R^2. We can the use the coefficient of determination to decide whether it makes sense to model the relationship with a straight line.

But it is always a good idea to draw the graph because, as it turns out, we can get the same straight line and coefficient of determination from very different data sets.

In 1973, the statistician Francis Anscombe described in an article four different data sets which all had the same regression equation and coefficient of determination, but which were very different.[1] The four data sets are shown in table 4.4.

If we plot each of the four data sets in a coordinate system, we get diagrams in figure 4.5. Here we can clearly see that the four data sets are distributed very differently. The first data set looks like it could be modelled with a

x	y	x	y	x	y	x	y
4	4.26	4	3.1	4	5.39	8	6.58
5	5.68	5	4.74	5	5.73	8	5.76
6	7.24	6	6.13	6	6.08	8	7.71
7	4.82	7	7.26	7	6.42	8	8.84
8	6.95	8	8.14	8	6.77	8	8.47
9	8.81	9	8.77	9	7.11	8	7.04
10	8.04	10	9.14	10	7.46	8	5.25
11	8.33	11	9.26	11	7.81	8	5.56
12	10.84	12	9.13	12	8.15	8	7.91
13	7.58	13	8.74	13	12.74	8	6.89
14	9.96	14	8.1	14	8.84	19	12.5

Table 4.4: Anscombe's four data sets. From [1].

straight line. The next data set (top right) shows a clear relationship—but is is certainly not linear. The last two data sets both include an outlier.

Despite their differences, all of the data sets have the same regression line and coefficient of determination,

$$y = 0.50 \cdot x + 3.00\ , \quad R^2 = 0.67\ .$$

So, the coefficient of determination is not enough on its own to determine whether a linear model is a "good" fit for the given data. It is therefore a good idea to draw the graph so that we can see the distribution of the points before we perform a linear regression.

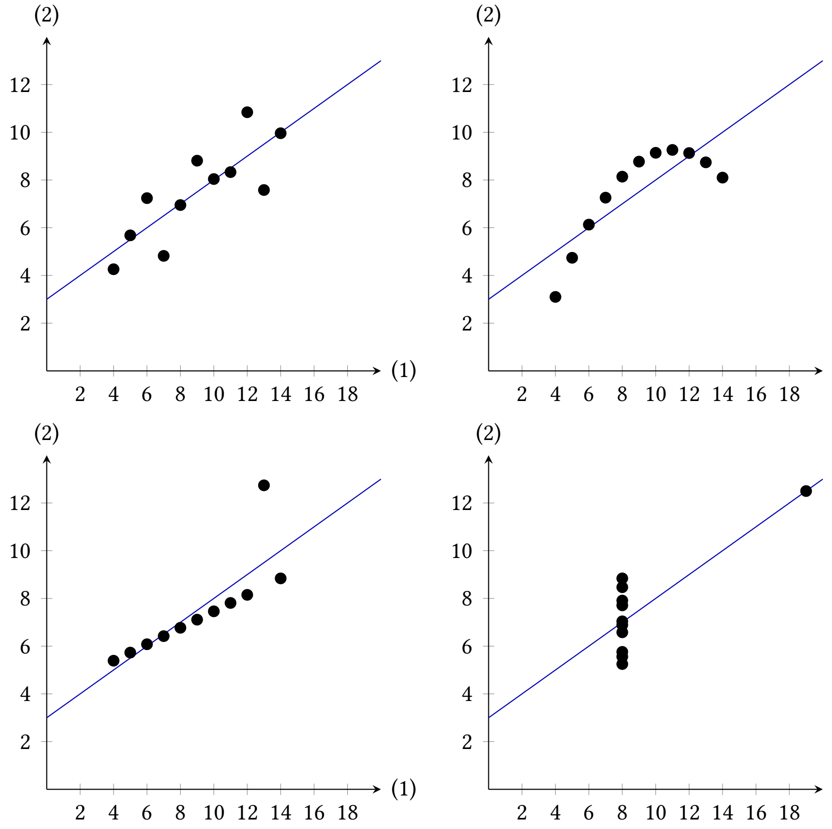


Figure 4.5: Anscombe's four data sets plotted in four coordinate systems.
It is clear that the distributions are very different.

In the case of outliers, it also makes sense to investigate this data point further. Could it be the result of a measurement error? And if the graph curves in a characteristic way, we might need to use a completely different type of regression.

4.2 Residual plot and residual standard deviation

Beacuse the coefficient of determination is not always a good measure, it makes sense to also investigate the graph. But it can be difficult to determine with the naked eye whether the points are close to the graph, or diverge from the line in some characteristic way.

We should therefore always look at the residual plot, i.e. a plot of the residuals as a function of the corresponding x-values. These should be small when compared to the measured y-values, and they cannot have any form of pattern.

But it is possible to analyse the residuals further. If the data shows a linear relationship, the errors (i.e. the residuals) will be a result of measurement errors. We can therefore analyse the residuals statistically to determine if this is the case.

Table 4.6: The population of Denmark 2010–2019.[4]

Årstal	Indbyggertal
2010	5 534 738
2011	5 560 628
2012	5 580 516
2013	5 602 628
2014	5 627 235
2015	5 659 715
2016	5 707 251
2017	5 748 769
2018	5 781 190
2019	5 806 081

When we perform a linear regression, the residuals are

$$r_i = y_i - ax_i - b\ ,$$

i.e. the average of the residuals must be

$$\overline{r} = \overline{y} - a\overline{x} - b\ .$$

But $b = \overline{y} - a\overline{x}$, so $\overline{r} = 0$. Therefore, the average of the residuals is 0.

The standard deviation of the residuals can be estimated by the so-called residual standard deviation, which we calculate in the following way:

Definition 4.4

If a series of data points $(x_1, y_1), \ldots, (x_n, y_n)$ is modelled by the straight line $y = a \cdot x + b$, the *residual standard deviation* is given by

$$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{r_1^2 + r_2^2 + \cdots + r_n^2}{n-2}}\ ,$$

where $r_1, r_2, \ldots, r_n$ are the residuals.

Measurement errors are usually normally distributed. So, if use a linear model to describe the data, then the residuals should be normally distributed with mean 0 and standard deviation s.[3]

Example 4.5 Table 4.6 shows the population of Denmark during the years 2010–2019. If we use linear regression on this data set, we get the graph in figure 4.7. The regression equation is

$$y = 0.031x + 5.520\ ,$$

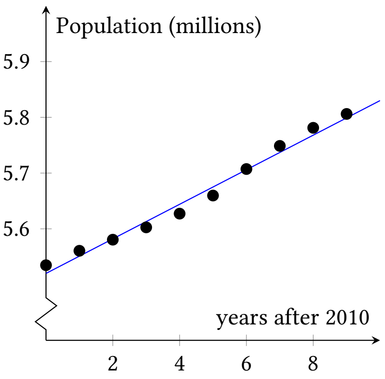
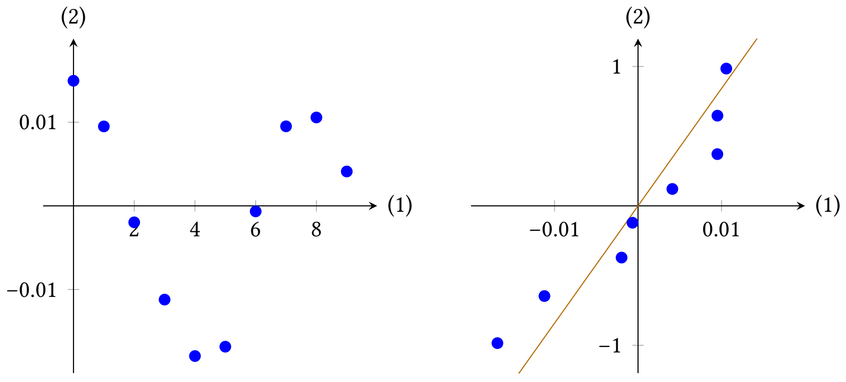


Figure 4.7: Regression of the population of Denmark 2010–2019.



(a) Residual plot.

(b) Normal probability plot.

Figure 4.8: Residual plot and normal probability plot of the residuals of the population of Denmark 2010–2019.

where x is the number of years after 2010, and y is the population in millions.

The coefficient of determination and the residual standard deviation are

$$R^2 = 0.985 \quad \text{and} \quad s = 0.013 \, .$$

The coefficient of determination shows that the points are quite close to the straight line, and the residual standard deviation is small compared to the y-values, which are all between 5 and 6.

Figure 4.8 shows the residual plot and a normal probability plot of the residuals. From the residual plot, we might argue that there seems to be a pattern in the residuals, but without more data this we cannot rule out that this could be a coincidence.

The normal probability plot shows that the residuals are approximately normally distributed because the points are close to the straight line. Using the two parameters, the graph, and the two plots in figure 4.8, we can now argue that the population of Denmark can be described quite well by a linear model in the given time period.

4.3 Confidence intervals

If a set of data can be described well by a linear model, the calculated slope and y-axis intercept become an estimate for the real values of the model behind the data. In this case, we distinguish between the real values of the two parameters a and b, and the estimated numbers $\hat{a}$ and $\hat{b}$ which are calculated from the sample.

We are usually interested in assessing how good the estimate actually is. Therefore, we calculate a so-called *confidence interval* for the slope a in which we can be sure to find the slope with some given probability. We often calculate the so-called 95% confidence interval: If the data is a sample which expresses a linear relationship, 95% of the slopes calculated from a sample will be in this interval. It is therefore reasonable to claim that the 95% confidence interval shows where the “real” slope is with a probability of 95%.

Because measurement errors are expected to be normally distributed, we can deduce that the slopes we calculate from samples are normally distributed with mean a and standard deviation

$$\sigma_a = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = \frac{\sigma}{\sqrt{S_{xx}}} \ ,$$

where σ is the theoretical standard deviation of the residuals. We can then find the 95% confidence interval by calculting the interval limits

$$a \pm n_{0.025} \cdot \frac{\sigma}{\sqrt{S_{xx}}} \ .$$

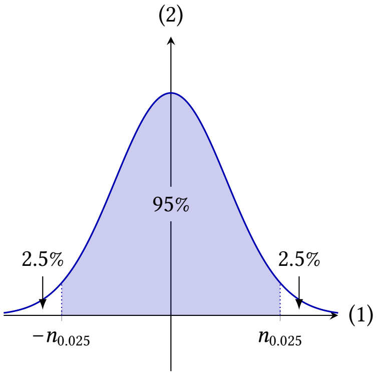


Figure 4.9: For the standard normal distribution, 95% of the distribution falls between $\pm n_{0.025}$.

The number $n_{0.025}$ shows where we find the top 2.5% of the standard normal distribution (i.e. with $\mu = 0$, $\sigma = 1$). So, 95% of the normal distribution will fall between $\pm n_{0.025}$ (see figure 4.9). Because the normal distribution scales nicely, 95% of any normal distribution will fall between the two values $\mu \pm n_{0,025} \cdot \sigma$.

95% of the possible samples will yield an estimated value $\hat{a}$ which falls in the interval

$$a \pm 1,96 \cdot \frac{\sigma}{\sqrt{S_{xx}}} \ .$$

There is just one problem. We know neither the theoretical residual standard deviation σ nor the true value of a, but only the estimated residual standard deviation s and the estimated slope $\hat{a}$. And when we use the estimated standard deviation, the estimated $\hat{a}$-values are no longer normally distributed—they follow instead the so-called t-distribution.[4] The estimated standard deviation of a is then[7]

$$s_a = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = \frac{s}{\sqrt{S_{xx}}} \ ,$$

where s is the estimated residual standard deviation.

As previously mentioned, the quantity $\hat{a}$ is t-distributed. The t-distribution describes the distribution of a normally distributed parameter estimated from a sample. Because we only know a sample and not the entire set of underlying data, we get a distribution of frequencies which looks like the normal distribution, but with "thicker tails" because a larger percentage of measured values will be further from the mean when the mean and the standard deviation are only estimates.

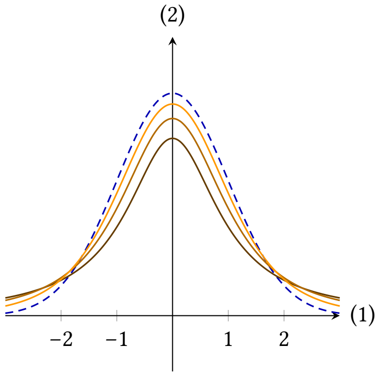


Figure 4.10: The standard normal distribution (dashed) and the t-distribution with 1, 2 and 5 degrees of freedom.

Furthermore, the t-distribution depends on the size of the sample—the so-called *degrees of freedom*. As you can see in figure 4.10, the t-distribution looks more and more like the normal distribution, the more degrees of freedom it has, i.e. how larger the sample is. The reason for this is that the more data we have, the closer the estimated mean and standard deviation will be to the true values—and the estimated distribution will then be closer to the theoretical normal distribution.

[4] The t-distribution is the distribution we get when we investigate normally distributed data by using a mean and a standard deviation estimated from a sample.[6]

If we have n data points, we can calculate a 95% confidence interval for the parameter a which is

$$\hat{a} \pm t_{0.025} \cdot \frac{s}{\sqrt{S_{xx}}} \, .$$

where $t_{0.025}$ corresponds to the number $n_{0.025}$, but for the t-distribution with $n - 2$ degrees of freedom.

In the same way, we can determine a confidence interval for the parameter b (the line's y-axis intercept). Here, we find the confidence interval[7]

$$\hat{b} \pm t_{0.025} \cdot \frac{s}{\sqrt{n}} \, ,$$

where s is the estimated residual standard deviation, and n is the number of data points.

All of the arguments above can be put together to form this theorem:

Theorem 4.6

If we perform linear regression on a data set, we can determine the $(1 - \alpha)$ confidence intervals for the parameters a and b in the linear model $y = ax + b$ as

$$\hat{a} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{S_{xx}}} \, ,$$

and

$$\hat{b} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \, .$$

The number α in the theorem is 5% (i.e. 0.05) for a 95% confidence interval, 1% (i.e. 0.01) for a 99% confidence interval, etc.

4.4 Other types of regression

The relationship between x and y in a data set $(x_1, y_1), \ldots, (x_n, y_n)$ is not necessarily linear. E.g. it could also be a power, an exponential or a polynomial relationship.

Power and exponential regression are done by most CAS's by transforming the data set and then performing linear regression on the transformed data. E.g. if the relationship between x and y is exponential, we have

$$\begin{aligned} y &= b \cdot a^x \\ \log(y) &= \log(b \cdot a^x) \\ \log(y) &= \log(a) \cdot x + \log(b) \, , \end{aligned}$$

i.e. when the relationship between x and y is exponential, the relationship between x and $\log(y)$ is linear. We can therefore perform linear regression on the data set $(x_1, \log(y_1)), \ldots, (x_n, \log(y_n))$. Then we find $\log(a)$ and $\log(b)$ which we can transform back into a and b.

In a similar way, we can transform a power relation into a linear relation by taking a logarithm to both the x- and the y-values in the data set.

The point here is this: If we perform regression on transformed data, R^2 will also be calculated from transformed data. This means that we have to be extra careful when we interpret the results, and therefore it is very important that we also look at the graph and the residual plot in these cases.

Bibliography

[1] F. J. Anscombe. “Graphs in Statistical Analysis”. In: *The American Statistician* 27.1 (Feb. 1973), pp. 17–21.

[2] Dan Bobkoff. *A magazine once polled millions on the presidential election – and got the results dead wrong*. Business Insider. Aug. 23, 2016. URL: http://nordic.businessinsider.com/magazines-presidential-poll-was-dead-wrong-2016-8 (visited on 06/07/2018).

[3] Per Bruun Brockhoff, Claus Thorn Ekstrøm, and Ernst Hansen. “Lineær regression – lidt mere tekniske betragtninger om R^2 og et godt alternativ”. In: *LMFK-bladet* nr. 2 (2017).

[4] Danmarks Statistik. *FOLK2: Folketal 1. januar efter køn, alder, herkomst, oprindelsesland og statsborgerskab*. URL: https://statistiksbanken.dk.

[5] Dominic Lusinchi. “‘President’ Landon and the 1936 Literary Digest Poll: Were Automobile and Telephone Owners to Blame?” In: *Social Science History* 36.1 (2012), pp. 23–54.

[6] Christian Walck. *Hand-book on statistical distributions for experimentalists*. University of Stockholm, Sept. 10, 2007.

[7] Thomas H. Wonnacott and Ronald J. Wonnacott. *Introductory Statistics for Business and Economics*. 2nd ed. John Wiley & Sons, Inc., 1977.