0.1 Introduction

In a **survey** we collect information. This information is often words or numbers, and is called **data**. A collection of data is called a **data set**.

For example, say you ask two people whether they like caviar. The one answers ”yes”, the other ”no”. Then ”yes” and ”no” are the data (answers) you have collected, and {”yes”, ”no”} is your data set.

Roughly speaking, statistics involves two things; *presenting* and *interpreting* data sets. For both purposes we have some terms which we will study in the following sections, helped by different examples of studies. These examples are found on page 2.

There are no universal laws telling you how to present or interpret data sets, however, you should follow these two guidelines:

- Let it always be clear exactly what you have studied, and what data you have collected.
- Always be aware the methods you use to interpret the data.

The language box

Persons participating in a survey where they are asked to answer questions are called **respondents**.

Survey 1

10 persones tested how many seconds they could hold their breath. These were the results (in seconds):

47 124 61 38 97 84 101 79 56 40

Survey 2

15 persons were asked how many apples they eat during a week. The answers were these:

7 4 5 4 1 0 6 5 4 8 1 6 8 0 14

Survey 3

300 persones where asked to name their favorite animal.

- 46 persons answered tiger
- 23 persons answered lion
- 17 persons answered crocodile
- 91 persons answered dog
- 72 persons answered cat
- 51 persons answered other animals

Survey 4

Mobile phones with smart-functions (app-based) came to the Norwegian market in 2009. The table[1] below shows the total sale of mobile phones during the time period 2009-2014, and the share with and without smart-functions. The numbers express the amount of 1 000 phones.

År	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**
totalt	2 365	2 500	2 250	2 200	2 400	2 100
wtho. sm.f.	1 665	1 250	790	300	240	147
wth. sm.f.	700	1 250	1 460	1 900	2 160	1 953

[1] Numbers imported from medienorge.uib.no.

0.2 Ways of presenting

When presenting data sets, it should be easy to see for others what we have found. This can be accomplished by using frequency tables, bar charts, sector graphs, or line graphs.

0.2.1 Frequency table

In a **frequency table** the data set are organized in a table showing the amount of times each unique answer appears. This amount is called the **frequency**.

Example - Survey 2

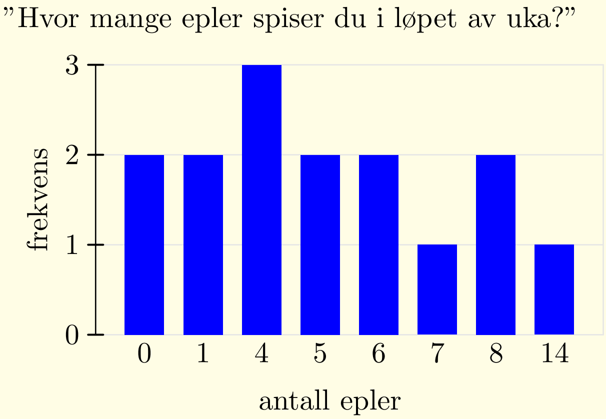
In this survey we have two 0's, two 1's, three 4's, two 5's, two 6's, one 7, two 8's and one 14. In a frequency table we then write

amount of apples	frequency
0	2
1	2
4	3
5	2
6	2
7	1
8	2
14	1

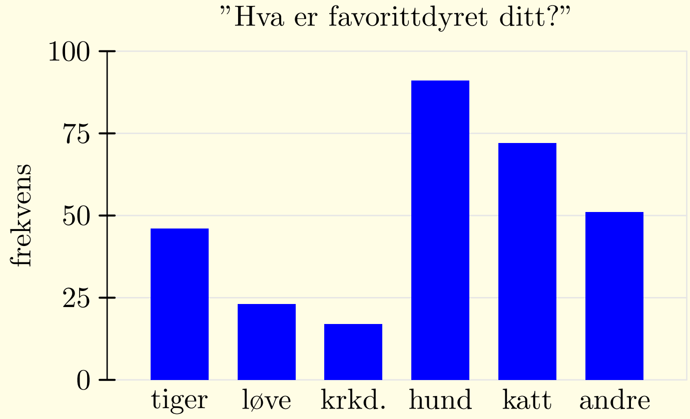
0.2.2 Søylediagram (stolpediagram)

In a bar chart the frequencies are represented by bars.

Survey 2



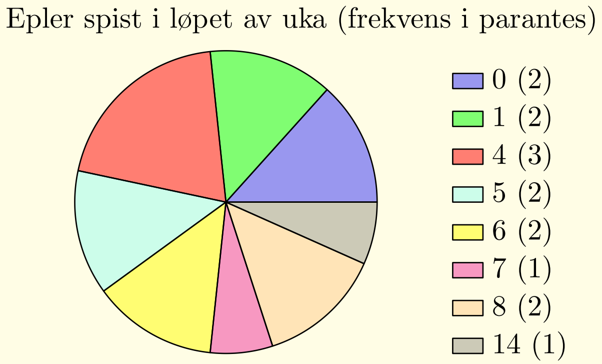
Survey 3



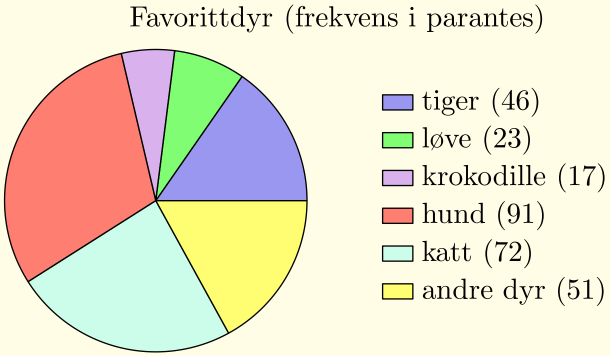
0.2.3 Sektordiagram (kakediagram)

In a sector graph the frequencies are represented by sectors in a circle.

Survey 2



Survey 3



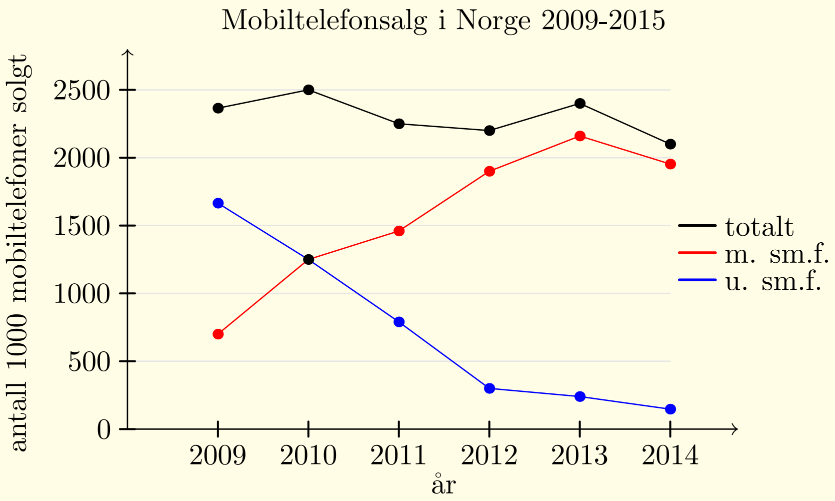
Making a sector graph from scratch

There are a lot of software available that generates sector graphs. However, if you were to make one from scratch, you will need basic knowledge of angles and fraction shares (see MB).

0.2.4 Linjediagram

In a line graph the data is represented as points in a coordinate system, with lines drawn between the points. Line graphs is typically used for describing evolving data.

Survey 4



0.3 Interpretation; central tendencies

In a data set there are often answers which are totally or approximately equal, and which reoccur. This means we can tell something about things that apply to the many; a **central tendeny**. The most common measures of central tendencies are the mode, the mean and the median.

0.3.1 Mode

0.1 Mode

The **mode** is the value occurring the most in the data set.

Example - Survey 2

In this dataset 4 has the highest occurrence (three), so 4 is the mode.

Multiple modes

If multiple values have the highest occurance in the data set, the data set has multiple modes.

0.3.2 Mean

When a data set includes numbers, we can find the sum of their values. When raising the question what the **mean** is, we ask this:

"If all the numbers had equal value, and the sum were still the same, what would the value be?"

The question is answered by the aid of division[1]:

0.2 Gjennomsnitt

$$\text{mean} = \frac{\text{sum of the values of the data set}}{\text{amount of values}}$$

The language box

The mean, as defined here, is also called the **average**. Also, there are multiple types of means. More specifically, the mean from Rule 0.2 is called the **arithmetic mean**.

Example - Survey 1

We sum the values from the data set, and divide by the amount of values:

$$\begin{aligned} \text{mean} &= \frac{47 + 124 + 61 + 38 + 97 + 84 + 101 + 79 + 56 + 40}{10} \\ &= \frac{727}{10} \\ &= 72.7 \end{aligned}$$

Hence, the 10 participators held their breath for 72.7 seconds on average.

[1] See MB, side 23.

Example - Survey 2

Method 1

$$\begin{aligned}\text{mean} &= \frac{7+4+5+4+1+0+6+5+4+8+1+6+8+0+14}{15}\\ &= \frac{73}{15}\\ &\approx 4.87\end{aligned}$$

Method 2

We expand our frequency table 3 to find the sum of the values from the data set. (we have also included the sum of the frequencies):

Amount of apples	Frequency	amount · frequency
0	2	$0 \cdot 2 = 0$
1	2	$1 \cdot 2 = 2$
4	3	$4 \cdot 3 = 12$
5	2	$5 \cdot 2 = 10$
6	2	$6 \cdot 2 = 12$
7	1	$7 \cdot 1 = 14$
8	1	$8 \cdot 2 = 16$
14	1	$14 \cdot 1 = 14$
sum	15	73

Now

$$\begin{aligned}\text{mean} &= \frac{73}{15}\\ &\approx 4.87\end{aligned}$$

Hence, on average, the 15 respondents ate 4.87 apples a week.

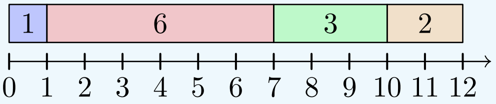
Example - Survey 4

(Calculations omitted. The values are rounded off to the nearest one.)

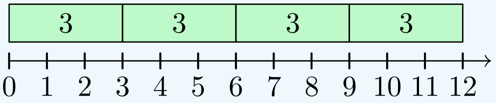
- Mean of total sale of mobiles: 2302
- Mean of sale of mobiles without smart-functions: 732
- Mean of sale of mobiles with smart-functions: 1570

Equal distribution

Note that the mean is about equal distribution. If we have 4 rectangles with respective lengths 1, 6, 3 and 2, their joint length is $1 + 6 + 3 + 2 = 12$.



Therefore, their mean length is $\frac{12}{4} = 3$. Thus, if we could reshape the rectangles so that their lengths were equal, with their joint length unchanged, they would each have length 3.



The average rate of change

Say you go for a jog, and measure your speed three times. Also, say that the data set you end up with is

$$10\,\mathrm{m/s} \qquad 10\,\mathrm{m/s} \qquad 10\,\mathrm{m/s}$$

Then your **average speed** was

$$\frac{10 + 10 + 10}{3}\,\mathrm{m/s} = 10\,\mathrm{m/s}$$

In other words; if your speed is the same alle the time[1], this speed is also you average speed. Consequently, the formula for the speed from Definition **??** is also the formula for the average speed. Alternatively stated, it is the formula for the average rate og change of length per time.

0.3 Average rate of change

If we *assume* or *hold* two quantities to be proportional, the proportionality constant from (**??**) is called the **averate rate of change**.

[1] In other words, your speed is *constant*.

Example - Survey 4

- For the years 2009 and 2010, the difference of smartphones sold to the difference of years is

  $$\frac{1\,260 - 700}{2010 - 2009} = \frac{550}{1} = 550$$

  Therefore, between 2009 and 2010 the sale of smartphones have *increased* by 550 000 smartphones per year.

- For the years 2010 and 2014, the difference of smartphones sold to the difference of years is

  $$\frac{1\,953 - 1\,250}{2014 - 2010} = \frac{703}{4} = 175{,}75$$

  Therefore, between 2010 and 2014 the sale of smartphones have *increased* by ca. 176 000 smartphones per year.

- For the years 2013 and 2014, the difference of smartphones sold to the difference of years is

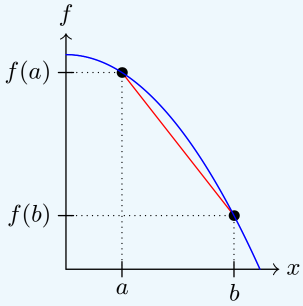
  $$\frac{1\,953 - 2\,160}{2014 - 2013} = \frac{-207}{1} = -207$$

  Therefore, between 2013 and 2014 the sale of smartphones have *decreased* by ca. 207 000 smartphones per year.

The slope of the line through two points

Given a function $f(x)$. In MB we have seen that the slope of the line through the points $(a, f(a))$ og $(b, f(b))$ is

$$\frac{f(b) - f(a)}{b - a}$$



Comparing this expression with the calculations made on page 11, we realize that the expressions are, in general, identical. Hence, the slope of a line through two points yields the average rate of change between the two points.

0.3.3 Median

0.4 Median

The **median** is the number that ends up in the middle when the data set is arranged in an increasing order.

If the data set has an even amount of values, the median equals the mean of the two values in the middle (after the arranging).

Example - Survey 1

We arrange the data set in an increasing order:

38 40 47 56 61 79 84 97 101 124

The two numbers in the middle are 61 and 79. The mean of these is

$$\frac{61 + 79}{2} = 70$$

Hence, the median is 70.

Example - Survey 2

We arrange the data set in an increasing order:

0 0 1 1 4 4 4 5 5 6 6 7 8 8 14

The value in the middle is 5, and thus the median is 5.

Example - Survey 4

(Calculations omitted. The values are rounded off to the nearest one).

- The median for the total sale of mobiles: 2307
- The median for the sale of mobiles without smart-functions: 545
- The median for the sale of mobiles with smart-functions: 1570

0.4 Interpretation; variations

Often there will be large differences between collected data. The collective term for various type of differences is **variation**. The most common measures of variation are variation width, quartile width, variance and standard deviation.

0.4.1 Variation width

0.5 Variasjonsbredde

The difference between the data with the largest value and the data with the smallest value yields the **variation width**.

Example Survey 1

The datas with, respectively, the highest and lowest values are 124 and 38. Thus

$$\text{variation width} = 124 - 38 = 86$$

Example - Survey 2

The datas with, respectively, the highest and lowest values are 14 and 0. Hence

$$\text{variation width} = 14 - 0 = 14$$

Example - Survey 4

- Variation width for mobiles in total:

$$2\,500 - 2\,100 = 400$$

- Variation width e for mobiles without smart functions:

$$1\,665 - 147 = 518$$

- Variation width for mobiles with smart functions:

$$2\,160 - 700 = 1460$$

0.4.2 Quartile width

0.6 Quartile width and upper and lower quartile

The **quartile width** of a data set can be found doing the following:

1. Arrange the data set in increasing order.
2. Divide the arranged data set, on the middle, into two new data sets (if the number of values in the set is even, the median is excluded).
3. Find the respective medians in the two new data sets.
4. Find the difference between the medians from point 3.

Regarding the medians from point 3: The one with the largest value is called the **upper quartile**, and the one with the lowest value is called the **lower quartile**.

Example - Survey 1

1. 38 40 47 56 61 79 84 97 101 124
2. 38 40 47 56 61 79 84 97 101 124
3. The median in the blue set is 47 (lower quartile), and the median in the red set is 97 (upper quartile).

   38 40 47 56 61 79 84 97 101 124

4. Quartile width $= 97 - 47 = 50$

Example - Survey 2

1. 0 0 1 1 4 4 4 5 5 6 6 7 8 8 14
2. 0 0 1 1 4 4 4 5 5 6 6 7 8 8 14
3. The median in the blue set is 1 (lower quartile), and the median in the red set is 7 (upper quartile).

   0 0 1 1 4 4 4 5 6 6 7 8 8 14

4. Quartile width $= 7 - 1 = 6$

Example - Survey 4

(Calculations omitted)

- For mobiles in total the quartile with is 200
- For mobiles without smart functions the quartile width is 1010
- For mobiles with smart functions the quartile width is 703

The language box

The lower quartile, the median and the upper quartile is also called the **1st. quartile**, the **2nd quartile** and the **3rd quartile**, respectively.

0.4.3 Deviations, variance and standard deviation

0.7 Variance

The difference between a value in a data set and the mean of the dataset is called the **deviation** of the value.

The **variance** of a data set can be found in the following way:

1. Square the deviation of every value in the data set, and add these.
2. Divide by the amount of values in the data set.

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

Example

Given the data set

$$2 \quad 5 \quad 9 \quad 7 \quad 7$$

Then

$$\text{gjennomsnitt} = \frac{2+5+9+7+7}{5} = 6$$

Moreover,

$$\begin{aligned}\text{variance} &= \frac{(2-6)^2+(5-6)^2+(9-6)^2+(7-6)^2+(7-6)^2}{5}\\ &= 5\end{aligned}$$

Then standard deviation $= \sqrt{5} \approx 2.23$.

Example - Survey 1

(Calculations omitted)

The variance is 754.01. The standard deviation is $\sqrt{754.01} \approx 27.46$

Example - Survey 2

We found the mean of the data set on page 9. We expand our table from page 3:

apples	frequency	frequency · squared deviation
0	2	$2 \cdot \left(0 - \frac{73}{15}\right)^2$
1	2	$2 \cdot \left(1 - \frac{73}{15}\right)^2$
4	3	$3 \cdot \left(4 - \frac{73}{15}\right)^2$
5	2	$2 \cdot \left(5 - \frac{73}{15}\right)^2$
6	2	$2 \cdot \left(6 - \frac{73}{15}\right)^2$
7	1	$1 \cdot \left(7 - \frac{73}{15}\right)^2$
8	2	$2 \cdot \left(8 - \frac{73}{15}\right)^2$
14	1	$1 \cdot \left(9 - \frac{73}{15}\right)^2$
sum	15	$189.7\bar{3}$

Hence

$$\text{variance} = \frac{189{,}7\bar{3}}{15} \approx 12.65$$

Thus, the standard deviation is $\sqrt{12.65} \approx 3.57$

Example - Sruvey 4

(Calculations omitted)

- For mobiles in total the variance is 17 781,25, and the standard deviation ca. 133.4.
- For mobiles without smart functions the variance is $318\,848.\bar{3}$, and the standard deviation ca. 17.87.
- For mobiles with smart functions the variance is $245\,847.91\bar{6}$, and the standard deviation ca. 495,83.

Why squaring when finding the variance?

Let us see what happens if we repeat our calculations from *Example 1* on page 17, but without squaring:

$$\frac{(2-6)+(5-6)+(9-6)+(7-6)+(7-6)}{5} = \frac{2+5+9+7+7}{5} - 6 \qquad (1)$$

But, by definition, the fraction $\frac{2+5+9+7+7}{5}$ is the mean of the data set, and thus the above expression equals 0. This applies for all data sets, and, in this context, 0 provides no further information. If we however square the deviations, we avoid expressions deemed to equal 0.