

## 0.1 Introduction

In a **undersøkelse** 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 personnes 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 personnes 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<sup>1</sup> 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

<sup>1</sup>Numbers imported from [medienorge.uib.no](http://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

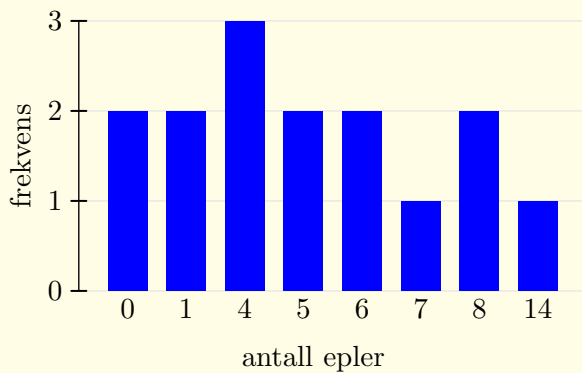
amount of apples	frequency
0	2
1	2
4	3
5	2
6	2
7	1
8	2
14	1

### 0.2.2 Søylediagram (stolpediagram)

In a bar chart the frequencies are represented by bars.

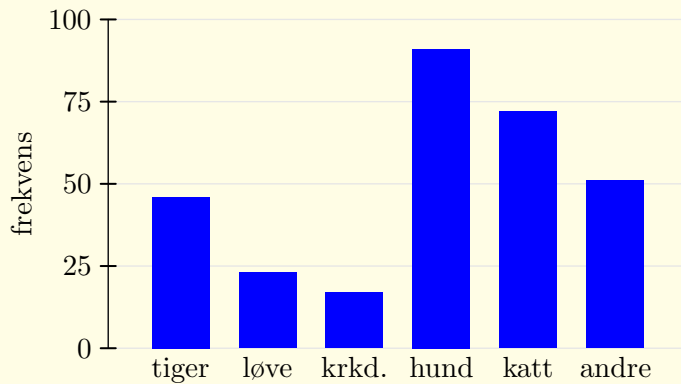
#### Survey 2

”Hvor mange epler spiser du i løpet av uka?”



#### Survey 3

”Hva er favorittdyret ditt?”

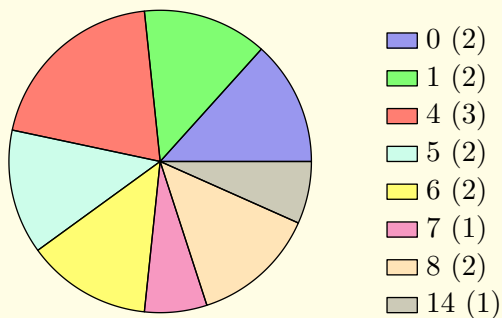


### 0.2.3 Sektordiagram (kakediagram)

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

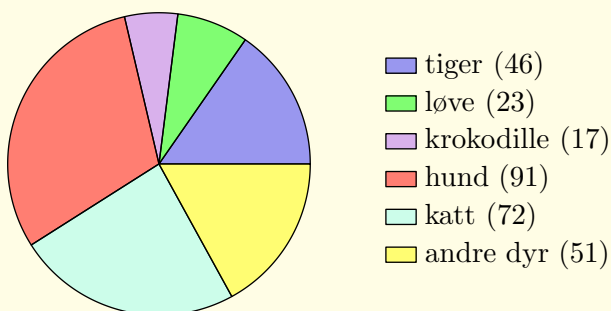
#### Survey 2

Epler spist i løpet av uka (frekvens i parantes)



#### Survey 3

Favorittdyr (frekvens i parantes)

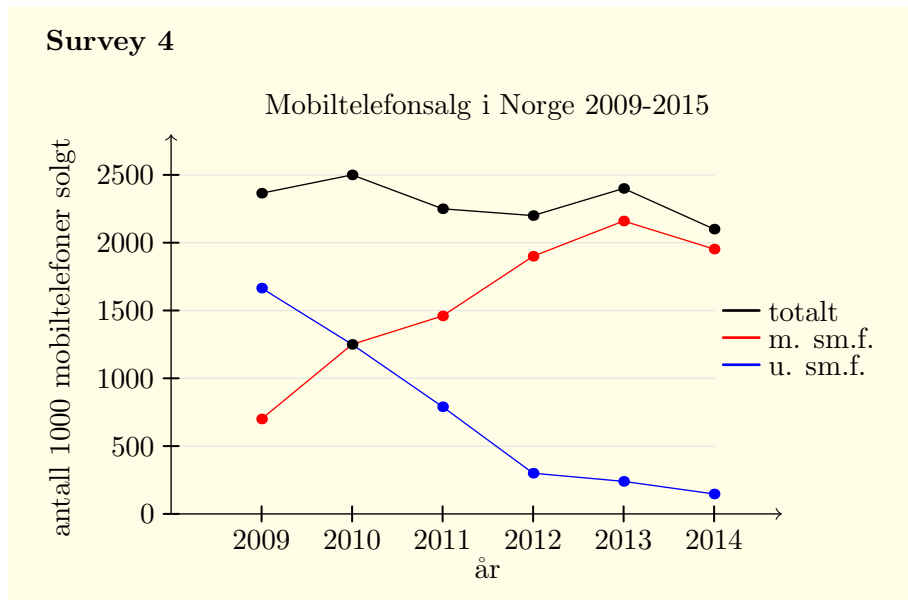


#### Making a sector graph from scratch

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.



## 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 tendency**. 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 occurrence 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<sup>1</sup>:

#### 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.

---

<sup>1</sup>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	$7 \cdot 1 = 7$
8	1	$8 \cdot 1 = 8$
14	1	$14 \cdot 1 = 14$
<b>sum</b>	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.

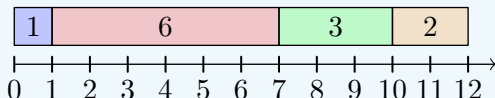
## Example - Survey 4

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

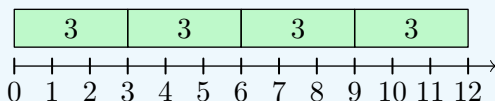
- Mean of total sale of mobiles: 2302
- Mean of sale of mobiles without smart-functions: 732
- Mean of sale of mobiles with smart-functions: 1570

## 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 re-shape 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 m/s      10 m/s      10 m/s

Then your **average speed** was

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

In other words; if your speed is the same all the time<sup>1</sup>, this speed is also your 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 of 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 **average rate of change**.

---

<sup>1</sup>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

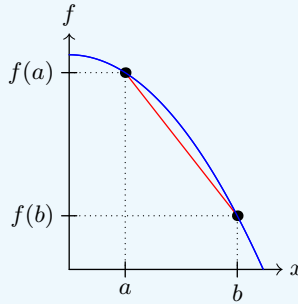
$$\frac{1\,953 - 2\,160}{2014 - 2013} = \frac{-207}{1} = -207$$

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 deviation, 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

Svaret med henholdsvis høyest og lavest verdi er 124 og 38. Altså er

$$\text{variasjonsbredde} = 124 - 38 = 86$$

#### Example - Survey 2

Svaret med henholdsvis høyest og lavest verdi er 14 og 0. Altså er

$$\text{variasjonsbredde} = 14 - 0 = 14$$

#### Example - Survey 4

- Variasjonsbredde for mobiler totalt:

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

- Variasjonsbredde for mobiler uten smartfunksjoner:

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

- Variasjonsbredde for mobiler med smartfunksjoner:

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

## 0.4.2 Kvartilbredde

### 0.6 Kvartilbredde og øvre og nedre kvartil

Kvartilbredden til et datasett kan finnes på følgende måte:

1. Ranger datasettet fra høgest til lavest verdi.
2. Skill det rangerte datasettet på midten, slik at to nye sett oppstår. (Viss det er oddetalls antall verdier i datasettet, utelates medianen).
3. Finn de respektive medianene i de to nye settene.
4. Finn differansen mellom medianene fra punkt 3.

Om medianene fra punkt 3: Den med høgest verdi kalles **øvre kvartil** og den med lavest verdi kalles **nedre kvartil**.

#### 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. Medianen i det blå settet er 47 (nedre kvartil) og medianen i det røde settet er 97 (øvre kvartil).

38 40 47 56 61 79 84 97 101 124

4. Kvartilbredde =  $97 - 47 = 50$

#### 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. Medianen i det blå settet er 1 (nedre kvartil) og medianen i det røde settet er 7 (øvre kvartil).

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

4. Kvartilbredde =  $7 - 1 = 6$

## Survey 4

(Utrekning utelatt)

- For mobiler totalt er kvartilbredden: 200
- For mobiler uten smartfunksjoner er kvartilbredden: 1010
- For mobiler med smartfunksjoner er kvartilbredden: 703

## The language box

Nedre kvartil, medianen og øvre kvartil blir også kalt henholdsvis **1. kvartil**, **2. kvartil** og **3. kvartil**.



### 0.4.3 Avvik, varians og standardavvik

#### 0.7 Varians

Differansen mellom en verdi og gjennomsnittet i et datasett kalles **avviket** til verdien.

Variansen til et datasett kan finnes på følgende måte:

1. Kvadrer avviket til hver verdi i datasettet, og summer disse.
2. Divider med antall verdier i datasettet.

**Standardavviket** er kvadratroten av variansen.

#### Example

Gitt datasettet

2 5 9 7 7

Da har vi at

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

Og videre er

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

Da er standardavviket  $= \sqrt{5} \approx 2,23$ .

#### Survey 1

(Utgning utelatt)

Variansen er 754,01. Standardavviket er  $\sqrt{754,01} \approx 27,46$

## Survey 2

Gjennomsnittet fant vi på side 9. Vi utvider frekvenstabellen vår fra side 3:

antall epler	frekvens	frekvens · kvadrert avvik
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 $\bar{3}$

Altså er variansen

$$\frac{189,7\bar{3}}{15} \approx 12,65$$

Da er standardavviket  $\sqrt{12,65} \approx 3.57$

## Survey 4

(Utgning utelatt)

- For mobiler totalt er variansen 17 781,25 og standardavviket ca. 133,4.
- For mobiler uten smartfunksjoner er variansen 318 848. $\bar{3}$  og standardavviket ca. 17,87
- For mobiler med smartfunksjoner er variansen 245 847.91 $\bar{6}$  og standardavviket ca. 495,83.

## Hvorfor innebærer variansen kvadrering?

La oss se hva som skjer hvis vi gjentar utregningen fra *Eksempel* på side 17, men uten å kvadrere:

$$\begin{aligned} & \frac{(2 - 6) + (5 - 6) + (9 - 6) + (7 - 6) + (7 - 6)}{5} \\ &= \frac{2 + 5 + 9 + 7 + 7}{5} - 6 \quad (1) \end{aligned}$$

Men brøken  $\frac{2+5+9+7+7}{5}$  er jo per definisjon gjennomsnittet til datasettet, og dermed blir uttrykket over lik 0. Dette vil gjelde for alle datasett, så i denne sammenhengen gir ikke tallet 0 noen ytterligere informasjon. Om vi derimot kvadrerer avvikene, unngår vi et uttrykk som alltid blir lik 0.