

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 table¹ 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

¹Numbers 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**.

Survey 2

I 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

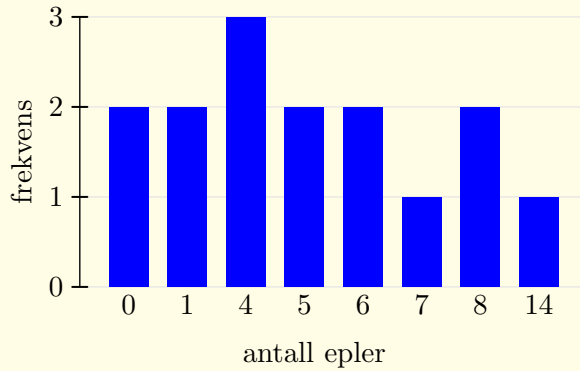
amount of apples	frequency
0	2
1	2
4	3
5	2
6	2
7	1
8	2
14	1

0.2.2 Søylediagram (stolpediagram)

In a bar chart the frequencies are represented by bars.

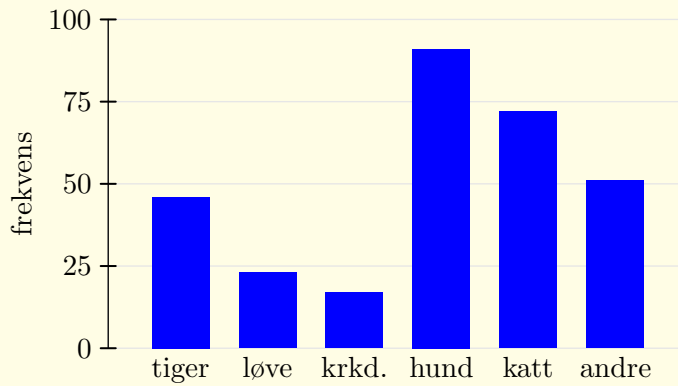
Survey 2

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



Survey 3

”Hva er favorittdyret ditt?”

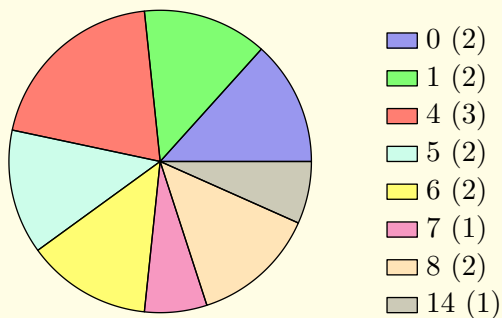


0.2.3 Sektordiagram (kakediagram)

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

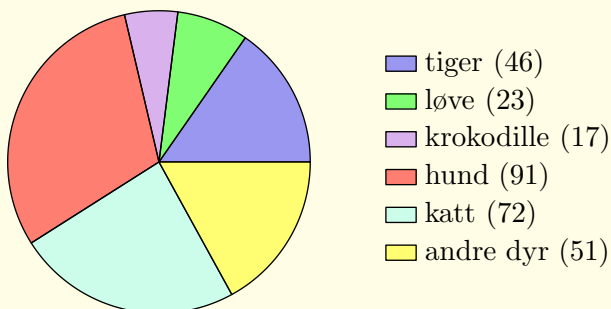
Survey 2

Epler spist i løpet av uka (frekvens i parantes)



Survey 3

Favorittdyr (frekvens i parantes)

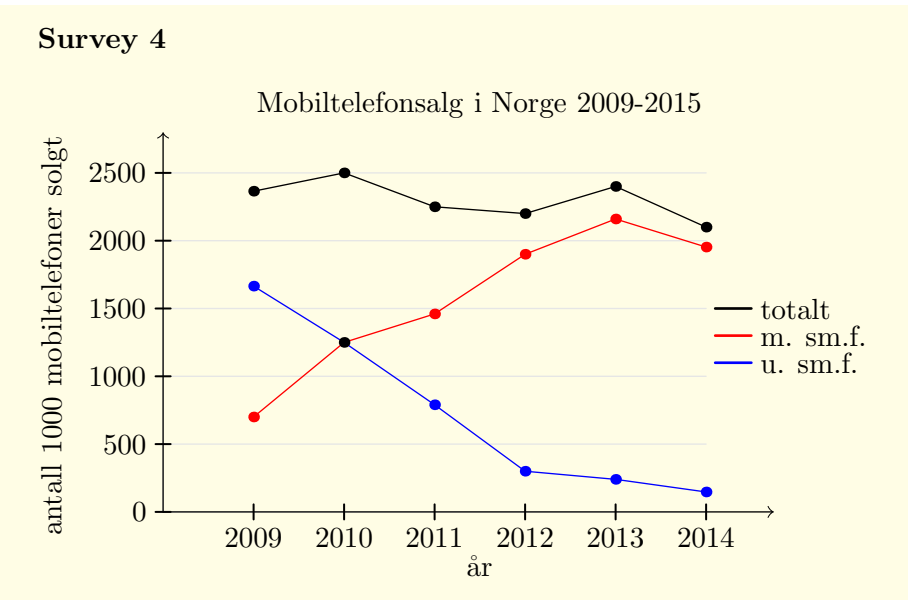


Making a sector graph from scratch

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.

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

Når et datasett består av svar i form av tall kan vi finne summen av svarene. Når vi spør hva gjennomsnittet er, spør vi om dette:

”Hvis alle svarene var like, og summen den samme, hvilken verdi måtte alle svarene da ha hatt?”

Dette er jo ingenting annet enn divisjon¹:

0.2 Gjennomsnitt

$$\text{gjennomsnitt} = \frac{\text{summen av verdiene fra datasettet}}{\text{antall verdier}}$$

Survey 1

Vi summerer verdiene fra datasettet, og deler med antall verdier:

$$\begin{aligned}\text{gjennomsnitt} &= \frac{47 + 124 + 61 + 38 + 97 + 84 + 101 + 79 + 56 + 40}{10} \\ &= \frac{727}{10} \\ &= 72,7\end{aligned}$$

Altså, i gjennomsnitt holdt de 10 deltakerne pusten i 72,7 sekunder.

¹se [MB](#), side 23.

Survey 2

Metode 1

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

Metode 2

Vi utvider frekvenstabellen fra side 3 for å finne summen av verdiene fra datasettet (vi har også tatt med summen av frekvensene):

Antall epler	Frekvens	antall · frekvens
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

Nå har vi at

$$\begin{aligned}\text{gjennomsnitt} &= \frac{73}{15} \\ &\approx 4,87\end{aligned}$$

Altså, i gjennomsnitt spiser de 15 respondentene 4,87 epler i uka.

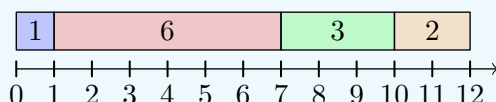
Survey 4

(Utregning utelatt. Verdiene er rundet ned til nærmeste éner).

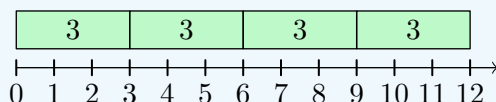
- Gjennomsnitt for totalt salg av mobiler: 2302
- Gjennomsnitt for salg av mobiler uten smartfunksjon: 732
- Gjennomsnitt for salg av mobiler med smartfunksjon: 1570

Lik fordeling

Legg merke til at gjennomsnitt handler om lik fordeling. Hvis vi har 4 rektangler med respektive lengder 1, 6, 3 og 2, blir den samlede lengden $1 + 6 + 3 + 2 = 12$.



Dette betyr at gjennomsnittslengden er $\frac{12}{4} = 3$. Hvis vi kunne omformet rektanglene slik at de ble like lange, ville altså hver av dem hatt lengde 3.



Gjennomsnittlig endring per enhet

Tenk deg at du på en løpetur gjør tre målinger av farten din. Datasettet du ender opp med er

$$10 \text{ m/s} \quad 10 \text{ m/s} \quad 10 \text{ m/s}$$

Gjennomsnittsfarten din var da

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

Siden alle målinger av farten din hadde samme verdi, kan det være rimelig å anta at farten din var konstant. Og hvis den virkelig var det, ville alle målinger av farten din hatt samme verdi, uansett hvor mange målinger du tok. Dette gjør at en konstant fart fra [Definisjon ??](#) i dagligtale også kalles **gjennomsnittsfart**. Sagt på en annen måte er dette den gjennomsnittlige endringen i antall meter per sekund.

0.3 Gjennomsnittlig endring per enhet

Hvis vi *antar* at to størrelser er proporsjonale, kaller vi proporsjonalitetskonstanten fra (??) den **gjennomsnittlige endringen per enhet**.

Survey 4

- For årene 2009 og 2010 er differansen mellom smarttelefoner solgt delt på differansen mellom år gått lik

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

Mellom 2009 og 2010 har altså salget av smarttelefoner i gjennomsnitt *økt* med 550 000 smarttelefoner per år.

- For årene 2010 og 2014 er differansen mellom smarttelefoner solgt delt på differansen mellom år gått lik

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

Mellom 2010 og 2014 har altså salget av smarttelefoner i gjennomsnitt *økt* med ca. 176 smarttelefoner per år.

- For årene 2013 og 2014 er differansen mellom smarttelefoner solgt delt på differansen mellom år gått lik

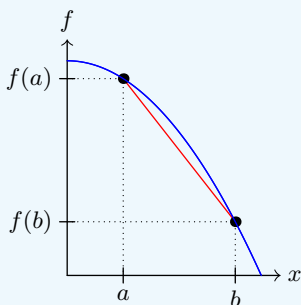
$$\frac{1\,953 - 2\,160}{2014 - 2013} = \frac{-207}{1} = -207$$

Mellom 2013 og 2014 har altså salget av smarttelefoner *sunket* med ca. 207 000 smarttelefoner per år.

Stignistallet til linja mellom to punkt

Gitt en funksjon $f(x)$. I MB har vi sett at stigningstallet til linja mellom punktene $(a, f(a))$ og $(b, f(b))$ er gitt som

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



Sammenlikner vi dette uttrykket med utregningene fra side 11, ser vi at utregningene er identiske. Stigningstallet mellom to punkt på en graf gir oss dermed den gjennomsnittlige endringen per enhet.

0.3.3 Median

0.4 Median

Medianen er tallet som ender opp i midten av datasettet når det rangeres fra tallet med lavest til høyest verdi.

Hvis datasettet har partalls antall verdier, er medianen gjennomsnittet av de to verdiene i midten (etter rangering).

Survey 1

Vi rangerer datasettet fra lavest til høyest verdi:

38 40 47 56 61 79 84 97 101 124

De to tallene i midten er 61 og 79. Gjennomsnittet av disse er

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

Altså er medianen 70.

Survey 2

Vi rangerer datasettet fra lavest til høgest verdi:

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

Tallet i midten er 5, altså er medianen 5.

Survey 4

(Utrekning utelatt. Verdiene er rundet ned til nærmeste éner).

- Median for totalt salg av mobiler: 2307
- Median for salg av mobiler uten smartfunksjon: 545
- Median for salg av mobiler med smartfunksjon: 1570

0.4 Tolking av forskjeller; spredningsmål

Ofte vil det også være store forskjeller (stor spredning) mellom dataene som er samlet inn. De vanligste matematiske begrepene som forteller noe om dette er **variasjonsbredde**, **kvartilbredde**, **varians** og **standardavvik**.

0.4.1 Variasjonsbredde

0.5 Variasjonsbredde

Differansen mellom svarene med henholdsvis høyest og lavest verdi.

Survey 1

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

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

Survey 2

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

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

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

(Utgregning 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 18, 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.