0.1 Introduksjon

I en *undersøkelse* henter vi inn informasjon. Denne informasjonen kan gjerne være tall eller ord, og kalles *data*. En samling av innhentet data kalles et *datasett*.

For eksempel, tenk at du spør to mennesker om de liker kaviar. Den éne svarer "ja", den andre "nei". Da er "ja" og "nei" datene du har samlet inn, og ["ja", "nei"] er datasettet ditt.

Statistikk handler grovt sett om to ting; *å presentere* og *å tolke* innsamlet data. For begge disse formålene har vi noen verktøy som vi i kommende seksjoner skal studere ved hjelp av noen forskjellige undersøkelser på side 2.

Det er ikke noen fullstendige fasitsvar på hvordan man presenterer eller tolker data, men to retningslinjer bør du alltid ta med deg:

- La det alltid komme tydelig fram hva du har undersøkt og hvilke data som er skaffet.
- Tenk alltid over hvilke metoder du bruker for å tolke dataene.

Undersøkelse 1

10 personer testet hvor mange sekunder de kunne holde pusten. Resultatene ble disse:

47 124 61 38 97 84 101 79 56 40

Undersøkelse 2

15 personer ble spurt hvor mange epler de spiser i løpet av en uke. Svarene ble disse:

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

Undersøkelse 3

300 personer ble spurt hva deres favorittdyr er.

- 46 personer svarte tiger
- 23 personer svarte løve
- 17 personer svarte krokodille
- 91 personer svarte hund
- 72 personer svarte katt
- 51 personer svarte andre dyr

Undersøkelse 4

Mobiltelefoner med smartfunksjoner (app-baserte) kom på det norske markedet i 2009. Tabellen[1] under viser det totale salget mobiltelefoner i tidsperioden 2009-2014 og andelen med og uten smartfunkskjoner.

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

[1]Tallene er hentet fra medienorge.uib.no.

0.2 Presentasjonsmetoder

Skal vi presentere våre undersøkelser, bør vi vise datasett slik at det er lett for andre å se hva vi har funnet. Dette kan vi gjøre blant annet ved hjelp av *frekvenstabeller*, *søylediagram*, *sektordiagram* eller *linjediagram*.

0.2.1 Frekvenstabell

I en frekvenstabell setter man opp dataene i en tabell som viser hvor mange ganger hver unike data dukker opp. Dette antallet kalles *frekvensen*.

Undersøkelse 2

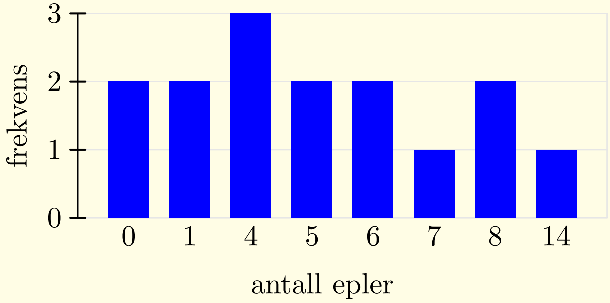
I vår undersøkelse har vi to 0, to 1, tre 4, to 5, to 6, én 7, to 8 og én 14. I en frekvenstabell skriver vi da

antall epler	frekvens
0	2
1	2
4	3
5	2
6	2
7	1
8	2
14	1

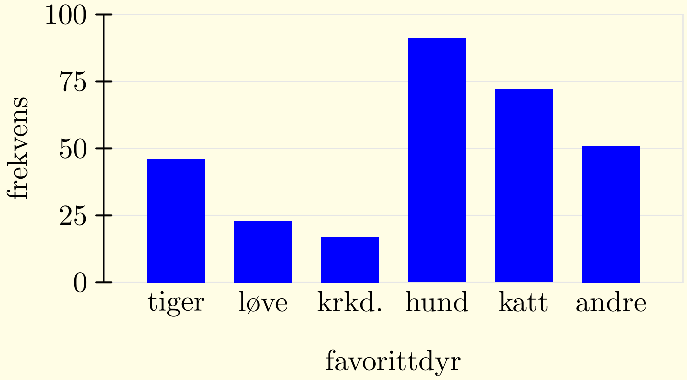
0.2.2 Søylediagram (stolpediagram)

Med et søylediagram presenterer vi dataene med søyler som viser frekvensen.

Undersøkelse 2



Undersøkelse 3

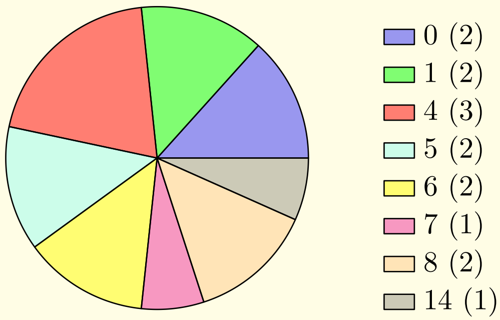


0.2.3 Sektordiagram (kakediagram)

I et sektordiagram vises frekvensene som sektorer av en sirkel:

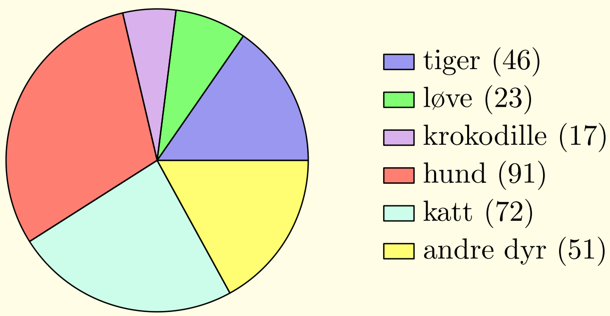
Undersøkelse 1

Epler spist i løpet av uka (frekvens i parantes)



Undersøkelse 2

Favorittdyr (frekvens i parantes)



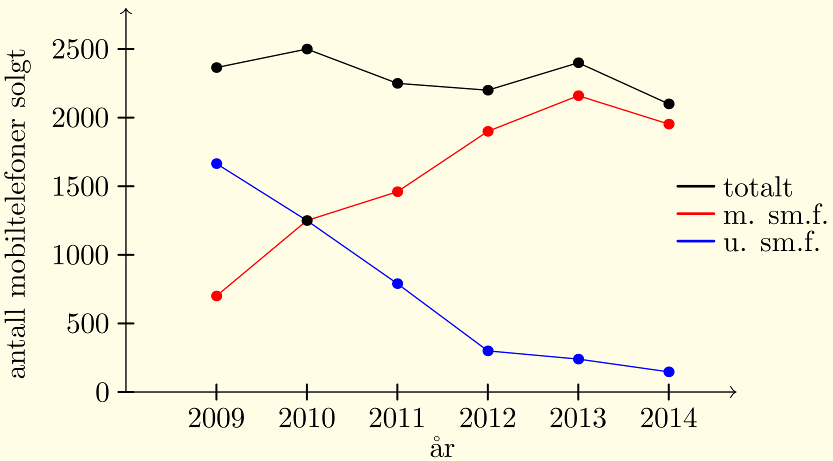
Å lage et sektordiagram for hand

Skal du selv tegne et sektordiagram, trenger du kunnskaper om vinkler og om brøkandeler. Se s. ??, MB, s. ?? og oppgave **??**.

0.2.4 Linjediagram

I et linjediagram legger vi inn dataene som punkt i et koordinatsystem, og trekker en linje mellom dem. Linjediagram brukes oftest når det er snakk om en form for utvikling.

Undersøkelse 4



0.3 Tolking av tendenser; sentralmål

I datasett vil det ofte være svar som er helt eller veldig like, og som gjentar seg. Dette betyr at vi kan si noe om hva som gjelder for mange. De matematiske begrepene som forteller noe om dette kalles *sentralmål*. De vanligste sentralmålene er *typetall*, *gjennomsnitt* og *median*.

0.3.1 Typetall

0.1 Typetall

Typetallet er verdien det er flest eksemplarer av i datasettet.

Undersøkelse 1

I datasettet er det tallet 4 som opptrer flest (tre) ganger. Dette kan vi se både fra selve datasett på s ?? eller frekvenstabellen på s ??, søylediagrammet på s ?? eller sektordiagrammet ??.

4 er altså typetallet.

0.3.2 Gjennomsnitt

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

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

Dette er jo ingenting annet enn divisjon (se MB, s. 23):

0.2 Gjennomsnitt

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

Undersøkelse 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.

Undersøkelse 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 ?? 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.

Undersøkelse 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

0.3.3 Median

0.3 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).

Undersøkelse 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.

Undersøkelse 2

Vi rangerer datasettet fra lavest til høyest 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.

Undersøkelse 4

(Utregning 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 matematiske begrepene som forteller noe om dette er *variasjonsbredde*, *kvartilbredde*, *varians* og *standardavvik*.

0.4.1 Variasjonsbredde

0.4 Variasjonsbredde

Differansen mellom svarene med henholdsvis høyest og lavest verdi.

Undersøkelse 1

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

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

Undersøkelse 2

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

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

Undersøkelse 4

- Variasjonsbredde for mobiltelefoner:

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

- Variasjonsbredde for mobiltelefoner uten smartfunksjoner:

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

- Variasjonsbredde for mobiltelefoner med smartfunksjoner:

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

0.4.2 Kvartilbredde

0.5 Kvartilbredde og øvre og nedre kvartil

Kvartilbredden til et datasett kan finnes på følgende måte:

1. Ranger datasettet fra høyest til lavest.
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 høyest verdi kalles *øvre kvartil* og den med lavest verdi kalles *nedre kvartil*.

Undersøkelse 2

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 $= 97 - 47 = 50$

Undersøkelse 1

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 $= 7 - 1 = 6$

Undersøkelse 4

(Utregning utelatt)

- For mobiltelefoner er kvartilbredden: 200
- For mobiltelefoner uten smartfunksjoner er kvartilbredden: 1010
- For mobiltelefoner med smartfunksjoner er kvartilbredden: 703

0.4.3 Avvik og varians

0.6 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.

Eksempel

Gitt datasettet

$$2 \quad 5 \quad 9 \quad 6 \quad 8$$

Da har vi at

$$\text{gjennomsnitt} = \frac{2+5+9+6+8}{5} = 6$$

Og videre er

$$\text{variansen} = \frac{(2-6)^2 + (5-6)^2 + (9-6)^2 + (6-6)^2 + (8-6)^2}{5}$$

Undersøkelse 1

(Utregning utelatt)

Variansen er 754,01

Undersøkelse 2

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

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	$189{,}7\overline{3}$

Altså er variansen

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

Undersøkelse 4

(Utregning utelatt)

- For mobiltelefoner er variansen 17 781,25
- For mobiltelefoner uten smartfunksjoner er variansen $318\,848.\overline{3}$
- For mobiltelefoner med smartfunksjoner er variansen $245\,847.91\overline{6}$

Hvorfor innebærer variansen kvadrering?

La oss se hva som skjer hvis vi gjentar utregningen fra *Eksempel 1* på side 13, men uten å kvadrere:

$$\frac{(2-6)+(5-6)+(9-6)+(6-6)+(8-6)}{5} = \frac{2+5+9+6+8}{5} - 6$$

Men brøken $\frac{2+5+9+6+8}{5}$ 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 datasettet. Om vi derimot kvadrerer avvikene, unngår et uttrykk som alltid blir lik 0.