Kapittel 1

Beskrivende statistikk

1.1 Innledning

Beskrivende statistikk kalles også **deskriptiv** statistikk, etter det engelske ordet *descriptive.* Kapitlet omhandler de vanligste metoder for sammenfatning og presentasjon av et statistisk tallmateriale. En kan gjerne si at statistikker i tradisjonell forstand, slik de blant annet kommer til uttrykk i Statistisk Årbok, sorterer under den del av statistikken som kalles beskrivende statistikk.

Beskrivende statistikk omhandler *ikke* det vi betegner som sannsynlighetsregning, statistiske metoder og statistisk inferens. En rekke *begreper* innen beskrivende statistikk er imidlertid felles med de andre delene av statistikken. Beskrivende statistikk utgjør ikke minst derfor en naturlig start på en innføringsbok i statistikk og sannsynlighetsregning.

1.2 Rådata

Med **rådata** (også kalt *primærdata*) skal vi mene det opprinnelige tallmaterialet vi skal behandle, inneholdende n observasjoner $x_1,\ldots,x_n$. Vi skal i første omgang se på beregning av følgende to svært viktige størrelser som karakteriserer tallmaterialet:

- (*Empirisk*) **middelverdi** som mål på tyngdepunkt (senter) i tallmaterialet.
- (*Empirisk*) **standardavvik** som mål på spredning rundt middelverdien.

Et annet ord for middelverdi er **gjennomsnitt**. Standardavvik er en størrelse som alltid er ikke-negativ og som får større og større verdi jo mer spredning (*avvik* fra middelverdien) det er i datamaterialet. Ordet *empirisk* betyr at tallmaterialet består av reelle data hentet fra virkelighetens verden, i motsetning til «teoretiske» data. Denne forskjellen vil komme klarere fram i senere kapitler. Formler for beregning av empirisk middelverdi og standardavvik er gitt i de neste rammer.

Empirisk middelverdi

Vi har n tall som vi betegner $x_1, x_2, \ldots, x_n$. Empirisk middelverdi, $\bar{x}$, til tallene er da definert ved følgende formel:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\left(x_1 + x_2 + \cdots + x_n \right) \qquad (1.1)$$

der den horisontale streken over x'en betyr middelverdi.

NB! Ordet gjennomsnitt er også ofte brukt, og det betyr akkurat det samme som middelverdi.

Eks. 1.1 **Beregning av middelverdi** (lign. (1.1))

a) Tre ørreter veier henholdsvis 2.1, 1.7 og 0.7 kg. Gjennomsnittsvekta, eller midlere vekt av de tre ørretene, blir:

$\bar{x} = \frac{1}{3}\cdot(2.1 + 1.7 + 0.7)$ kg $= \frac{1}{3}\cdot 4.5$ kg $= \underline{1.5 \text{ kg}}$.

b) I en liten bedrift med 5 ansatte har de ansatte følgende årslønn (i kr 1000):

240, 180, 270, 210 og 160.

Vi skal beregne totale årlige lønnskostnader for bedriften og gjennomsnittlig lønn. La x_i betegne lønning nr. i, dvs. $x_1 = 240$, $x_2 = 180$ osv. De totale lønnskostnadene finnes ved å summere alle lønningene:

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5 = 240 + 180 + 270 + 210 + 160 = 1060$$

dvs. totale lønnskostander er på kr $\underline{1\ 060\ 000}$

Gjennomsnittslønna $\bar{x}$ finner vi ved å dele summen av inntekter på antall ansatte:

$$\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = \frac{1}{5}\cdot \text{ kr } 1\ 060\ 000 = \underline{\text{kr } 212\ 000}$$

c) Gjennomsnittlig kubikkinnhold pr. tre i en homogen (ensartet) skog med 1000 trær er anslått til å være 1,1 m^3 pr. tre, og vi skal anslå totalt kubikk-innhold i skogen. Vi betegner kubikkinnholdet til trærne $x_1, \ldots, x_{1000}$, og er altså interessert i å finne summen av alle x-ene. Vi får da:

$$\bar{x} = \frac{1}{1000}\sum_{i=1}^{1000} x_i = 1{,}1 \text{ m}^3 \;\Rightarrow\; \sum_{i=1}^{1000} x_i = 1000 \cdot 1{,}1 \text{ m}^3 = \underline{1100 \text{ m}^3}$$

Totalt kubikkinnhold i skogen kan altså anslås til 1100 kubikkmeter. ☺

Empirisk standardavvik, *s*, og varians, s^2

Vi har *n* tall som vi betegner $x_1, x_2, \ldots, x_n$. Empirisk standardavvik, *s*, til tallene, er et mål på hvor spredt tallene er rundt middelverdien, $\bar{x}$, og beregnes ved en av følgende 3 formler:

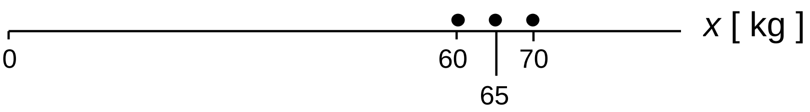
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \tag{1.2}$$

$$= \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\cdot\bar{x}^2\right)} \tag{1.3}$$

$$= \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)} \tag{1.4}$$

NB! Det er følgende entydige sammenheng mellom begrepene *varians* og *standardavvik*: Varians = kvadratet, s^2, av standardavviket, *s*.

Formlene for beregning av standardavviket *s* i forrige ramme kan synes noe abstrakte, så la oss prøve å se hva som kan ligge bak utgangsformelen i lign.(1.2). Vi tar utgangspunkt i dataene vist i figuren nedenfor, som viser 3 tilfeldige studentvekter: $x_1 = 60$ kg, $x_2 = 65$ kg og $x_3 = 70$ kg.



Vi beregner først middelverdien, og finner at $\bar{x} = 65$ kg. Dette stemmer vel rimelig bra med intuisjonen. Hva så med spredningen rundt middelverdien? De fleste vil vel være enige i at ± 5 kg vil være et rimelig anslag for denne spredningen. La oss prøve å bestemme en generell formel som gir et fornuftig anslag av spredningen Δx. Vi prøver 4 forskjellige alternativer:

1) Vi prøver først å ta gjennomsnittlig avvik fra middelverdien:

$$\Delta x = \tfrac{1}{3}\cdot\Sigma(x_i - \overline{x}) = \tfrac{1}{3}\big((60{-}65) + (65{-}65) + (70{-}65)\big)\text{ kg} = 0\text{ kg}$$

Dette stemmer dårlig, vi har jo helt klart en spredning forskjellig fra null.

2) For å komme unna fortegnsproblemet i 1), ser vi på gjennomsnittlig absoluttavvik fra middelverdien:

$$\Delta x = \tfrac{1}{3}\cdot\Sigma\,|\,x_i - \overline{x}\,| = \tfrac{1}{3}\cdot(|\,60{-}65\,| + |\,65{-}65\,| + |\,70{-}65\,|)\text{ kg} = 10/3\text{ kg}$$

Dette stemmer brukbart med intuisjonen, men absoluttverdier er ikke særlig attraktive å jobbe med rent matematisk.

3) Kvadrat er en hendigere matematisk funksjon enn absoluttverdi, så vi ser på gjennomsnittlig kvadratavvik fra middelverdien:

$$\Delta x^2 = \tfrac{1}{3}\cdot\Sigma(x_i - \overline{x})^2 = \tfrac{1}{3}\cdot\big((60{-}65)^2 + (65{-}65)^2 + (70{-}65)^2\big)\text{ kg}^2 = 50/3\text{ kg}^2.$$

Benevningen kg^2 er imidlertid en uhensiktsmessig og abstrakt benevning for å angi usikkerheten til data med benevning i kg.

4) Vi tar så kvadratrota av gjennomsnittlig kvadratavvik fra middelverdien, for å få en fornuftig benevning:

$$s = \Delta x = \big(\tfrac{1}{3}\cdot\Sigma(x_i - \overline{x})^2\big)^{½} = (50/3)^{½}\text{ kg} = 4\text{ kg}$$

som stemmer bra overens med intuisjonen.

Formelen i 4) minner svært om formelen i lign.(1.2). Eneste forskjell er at vi har brukt n ($n = 3$ i eksemplet) i nevner, mens det er brukt $n{-}1$ i lign.(1.2). Når det dreier seg om store n-verdier blir resultatet rimelig uavhengig av om vi bruker n eller $n{-}1$. I kap. 6 kommer vi tilbake til en begrunnelse for hvorfor det står $n{-}1$ og ikke n i nevner i lign.(1.2).

Før vi går løs på eksempler som viser beregning av standardavvik, skal vi angi følgende merknader til beregningsformlene i lign.(1.2)-(1.4) i forrige ramme:

1) NB! Særlig ved bruk av lign.(1.3) og (1.4) må en passe på å bruke tilstrekkelig mange siffer i mellomregningene. Hvis vi f.eks. har en stor positiv middelverdi, $\overline{x}$, og et lite standardavvik, s, fremkommer s ved å subtrahere to store tall fra hverandre. Hvis disse store tallene har for få siffer, får vi lett helt gale svar.

2) Lign.(1.4) ovenfor er trolig den formelen som generelt er mest nyttig ved beregning av empirisk varians og standardavvik. Alle formlene (1.2), (1.3) og (1.4) er imidlertid nyttige å beherske for å beregne varians og standardavvik. Eks. 1.2, 1.3 og 1.4 viser nytten ved de enkelte formlene.

3) Selv på rimelig enkle lommekalkulatorer er det lagt inn beregning av middelverdi og standardavvik, ofte angitt ved samme symboler som i formlene ovenfor. På enkelte kalkulatorer kan en velge hvorvidt en skal ha n eller $n-1$ i nevneren ved beregning av s. Vi kommer senere tilbake til denne forskjellen.

Eks. 1.2 **Beregning av standardavvik** (lign.(1.2))

$n = 3,\ x_1 = 10{,}0,\ x_2 = 9{,}9,\ x_3 = 10{,}1$

middelverdi: $\bar{x} = \frac{1}{3}\cdot(10{,}0 + 9{,}9 + 10{,}1) = 10{,}0$

varians: $s^2 = \frac{1}{2}\cdot\sum_{i=1}^{3}\left(x_i - \bar{x}\right)^2$

$= \frac{1}{2}\cdot\left((10{,}0-10{,}0)^2 + (10{,}0-9{,}9)^2 + (10{,}1-10{,}0)^2\right) = 0{,}01$

standardavvik: $s = \sqrt{0{,}01} = 0{,}1$ ☺

Eks. 1.3 **Beregning av standardavvik** (lign.(1.3))

$n = 3,\ x_1 = 0,\ x_2 = 1,\ x_3 = -1$

middelverdi: $\bar{x} = \frac{1}{3}\cdot(0 + 1 + (-1)) = 0$

varians: $s^2 = \frac{1}{2}\cdot\left(\sum_{i=1}^{3}x_i^2 - 3\bar{x}^2\right) = \frac{1}{2}\cdot\left(0^2 + 1^2 + (-1)^2 - 3\cdot 0^2\right) = 1$

standardavvik: $s = \sqrt{1} = 1$ ☺

Eks. 1.4 **Beregning av standardavvik** (lign.(1.4))

$n = 3,\ x_1 = 5,\ x_2 = 3,\ x_3 = 8$

$\Sigma x_i = 5 + 3 + 8 = 16$

$\Sigma x_i^2 = 25 + 9 + 64 = 98$

varians: $s^2 = \frac{1}{3-1}\left(\sum_{i=1}^{3}x_i^2 - \frac{1}{3}\left(\sum_{i=1}^{3}x_i\right)^2\right) = \frac{1}{2}\left(98 - \frac{1}{3}\cdot 16^2\right) = 6{,}33$

standardavvik: $s = \sqrt{6{,}33} = 2{,}52$ ☺

1.3 Rangordning av data

Dersom vi ordner våre observasjoner, $x_1, x_2, \ldots, x_n$, i stigende rekkefølge, skal vi bruke betegnelsen $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. $x_{(1)}$ er altså betegnelsen for minste verdi, $x_{(2)}$ for nest minste, opp til $x_{(n)}$ for største verdi. I tillegg til å beregne middelverdi og standardavvik, kan vi nå beregne følgende:

a) **median** (midtverdi) som (robust) mål på senter,
b) **interkvartilbredde** som (robust) mål på spredning og
c) 100p-**prosentiler** for ønsket verdi av p.

En fordel med medianen i forhold til middelverdien som mål på senter, er at medianen er lite påvirket av noen få «ekstreme» observasjonsverdier. Tilsvarende er interkvartilbredden et mer **robust** spredningsmål enn standardavviket.

Empirisk median (midtverdi), *m*

Her skiller vi mellom tilfellene hvor n er et **odde**tall ($n = 3,5,7,9,\ldots$) og hvor n er et **like**tall ($n = 2,4,6,8,\ldots$).

$n = 3,5,7,\ldots$ (*odde*): m = den midterste av observasjonene etter at alle observasjonene er ordnet i stigende rekkefølge:

(1.5) $$m = x_{\left(\frac{n+1}{2}\right)} \quad (n \text{ odde})$$

$n = 2,4,6,\ldots$ (*like*): m = gjennomsnittet av de 2 midterste observasjonene etter at alle observasjonene er ordnet i stigende rekkefølge:

(1.6) $$m = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)} \right) \quad (n \text{ like})$$

Eks. 1.5 **Beregning av median** (lign. (1.5))

$x_1 = 2{,}7,\ x_2 = 1{,}9,\ x_3 = 4{,}2,\qquad x_{(1)} = 1{,}9,\ x_{(2)} = 2{,}7,\ x_{(3)} = 4{,}2$

median: $m = x_{\left(\frac{3+1}{2}\right)} = x_{(2)} = 2{,}7$ ☺

Eks. 1.6 **Beregning av median** (lign.(1.6))

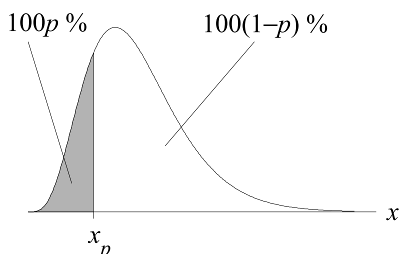
$x_1 = 2{,}7,\quad x_2 = 1{,}9,\quad x_3 = 4{,}2,\quad x_4 = -97{,}3$
$x_{(1)} = -97{,}3,\quad x_{(2)} = 1{,}9,\quad x_{(3)} = 2{,}7,\quad x_{(4)} = 4{,}2$

median: $m = \frac{1}{2}\left(x_{(4/2)} + x_{(4/2+1)}\right) = \frac{1}{2}\left(x_{(2)} + x_{(3)}\right) = 2{,}3$ ☺

Beregn selv middelverdien, $\bar{x}$, i eks. 1.5 og 1.6, og du vil se at du får stor forskjell i de 2 tilfellene, p.g.a. den «ville» verdien, −97,3. Medianen, *m*, gir imidlertid som du ser noenlunde samme resultat i begge tilfeller. Vi sier at medianen er **robust** med hensyn til «ville» (ekstreme) verdier, dvs. lar seg lite påvirke av disse.

100*p*-prosentilen, x_p (definisjon)
100*p*-prosentilen er en *x*-verdi som er slik at minst 100*p* % av observasjonene er mindre eller lik, og minst 100(1–*p*) % av observasjonene er større eller lik denne *x*-verdien. En enkel beregningsmåte som tilfredsstiller definisjonen over er:

$$x_p = \begin{cases} \frac{1}{2}\left(x_{(np)} + x_{(np+1)}\right), & np \text{ heltall} \\ x_{(j)}, & np \text{ ikke heltall} \end{cases}$$



der *j* er minste heltall større enn *np*.

Eks. 1.7 **Beregning av prosentiler**
$n = 5,\quad x_{(1)} = 1,\quad x_{(2)} = 2.3,\quad x_{(3)} = 3.4,\quad x_{(4)} = 3.7,\quad x_{(5)} = 4.9$

Beregning av 20-prosentilen:
$p = 0.2$: $np = 5 \cdot 0.2 = 1$ (heltall)
$\Rightarrow\ x_{0.2} = \frac{1}{2} \cdot \{x_{(1)} + x_{(2)}\} = \frac{1}{2} \cdot (1+2.3) = \underline{1.65}$, dvs. 20-prosentilen er lik 1.65.

Beregning av 70-prosentilen:
p = 0,7: $np = 5 \cdot 0{,}7 = 3{,}5$ (ikke heltall) $\Rightarrow$ forhøyer np til nærmeste heltall, dvs. $j = 4 \Rightarrow x_{0,7} = x_{(4)} = \underline{3{,}7}$, dvs. 70-prosentilen er lik 3,7. ☺

Kvartiler (definisjon):

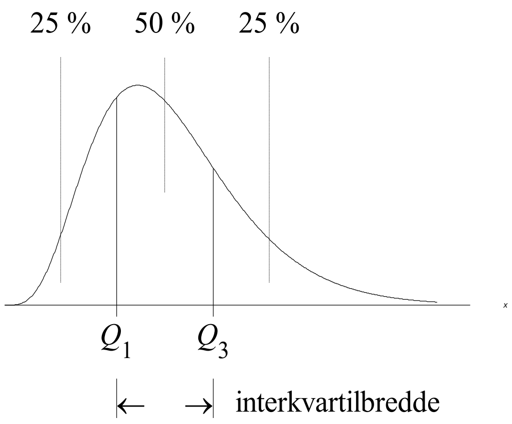
Nedre kvartil, Q_1, er identisk med 25-prosentilen (*en kvart* av dataene er mindre eller lik 25-prosentilen).

Øvre kvartil, Q_3, er identisk med 75-prosentilen (*tre kvart* av dataene er mindre eller lik 75-prosentilen).

Medianen er identisk med 50-prosentilen (*halvparten* av dataene er mindre eller lik medianen).

Interkvartilbredde, $Q_3 – Q_1$

Interkvartilbredden er et (robust) mål på hvor spredt dataene er rundt medianen (midtverdien). Den finnes ved å ta differansen mellom øvre kvartil, $Q_3 = x_{0.75}$, og nedre kvartil, $Q_1 = x_{0.25}$, slik det er illustrert i figuren til høyre.



Fordelen med interkvartilbredden i forhold til standardavviket er at interkvartilbredden er lite påvirket av noen få «ville» observasjonsverdier.

NB! Dersom vi ikke har «ville» verdier vil normalt interkvartilbredden gi en større verdi enn standardavviket. For normalfordelingen, som er en svært sentral og viktig statistisk fordeling (kap.5), er interkvartilbredden ca. en faktor på 1.35 større enn standardavviket.

Eks. 1.8 **Interkvartilbredde**
Data: $n = 5$, $x_1 = -1{,}2$, $x_2 = 0{,}7$, $x_3 = 0$, $x_4 = 172$, $x_5 = 1{,}0$
Vi rangordner dataene: $x_{(1)} = -1{,}2$, $x_{(2)} = 0$, $x_{(3)} = 0{,}7$, $x_{(4)} = 1{,}0$, $x_{(5)} = 172$

Q_1: $np = 5\cdot 0.25 = 1{,}25 \quad\Rightarrow\quad Q_1 = x_{(2)} = 0$
Q_3: $np = 5\cdot 0.75 = 3{,}75 \quad\Rightarrow\quad Q_3 = x_{(4)} = 1.0$
Interkvartilbredden $= Q_3 - Q_1 = 1.0 - 0 = \underline{1.0}$ ☺

Det bør understrekes at eksemplet ovenfor er et teknisk eksempel som viser hvordan vi beregner interkvartilbredden rent matematisk. I praksis bør vi generelt ha langt flere data enn $n = 5$ for at bruk av interkvartilbredde skal være fornuftig. Imidlertid viser eksemplet hvordan interkvartilbredden effektivt undertrykker den ekstreme verdien $x = 172$.

1.4 Grupperte data

For å oppnå en mer oversiktlig og informativ fremstilling av et innhentet datamateriale, er det vanlig å gruppere tallmaterialet. Det skilles mellom *to* forskjellige typer variabler:

1) **Diskrete** variabler: Observasjonene kan kun ha visse (diskrete) verdier som er adskilt fra hverandre (eks: antall øyne i et terningkast).

2) **Kontinuerlige** variabler: Observasjonene kan ha hvilke som helst verdier innenfor et begrenset eller ubegrenset definisjonsområde (eks: tiden, t).

NB! I prinsippet er det ofte en «glidende» overgang fra kontinuerlige til diskrete variabler. La oss som eksempel betrakte «pers'en» på 60m til studenter ved Høgskolen i Tromsø. I prinsippet er denne tida en kontinuerlig variabel som kan ha en hvilken som helst verdi mellom, la oss si, 6s og 15s. I praksis måles imidlertid tiden på nærmeste tidel eller hundredel. Vi har da med *diskrete* verdier å gjøre (6,0s, 6,1s, 6,2s,…,15,0s i tilfelle vi har tider på nærmeste tidel). Ved tilstrekkelig fin inndeling (diskretisering) vil det ikke gi noen praktisk forskjell om vi betrakter 60m-tidene som kontinuerlige eller diskrete.

Normal *fremgangsmåte* for gruppering av et tallmateriale bestående av enkelttall kan være:

a) Bestemmelse av maksimum og minimum

Vi finner minste og største observasjonsverdi, x_{min} og x_{maks}.

b) Klasseinndeling

Vi deler x-området inn i k *klasser*, som regel med like brede *klasseintervaller* (dvs. lik *klassebredde*). Klassene må ikke overlappe hverandre, og tilsammen må klassene dekke alle verdier fra og med x_{min} til og med x_{maks}.

Vi etterstreber å velge klasseintervaller der *klassebredden* (øvre klassegrense minus nedre klassegrense), *klassegrenser* og *klassemidtpunktet*, m_i, blir forholdsvis pene og runde tall, eller ligger i nærheten av slike. Dette gjør vi både for å lette regnearbeidet, og for å få mest mulig brukervennlige og informative tabeller.

NB! *Angivelse* av *klassegrensene* i en tabell er ofte ikke i samsvar med de *egentlige* klassegrensene. Eks: Dersom en vektklasse er angitt som [60,70> kg, så er nedre og øvre klassegrense angitt som henholdsvis 60 og 70 kg. Normalt vil vektdata være forhøyet: Hvis en person veier f.eks. 59,5 kg, og vekta skal angis i hele kg, forhøyes vekta til 60 kg. De egentlige klassegrensene i intervallet [60,70> er derfor 59,500.. og 69,499...

Øvre *klassegrense* i en klasse er lik nedre klassegrense i neste klasse.

For entydig å angi i hvilken klasse en observasjon som ligger i grenseland mellom 2 naboklasser tilhører, benyttes 1 av følgende 3 teknikker:

1) Vi føyer på en ekstra desimal i klassegrensene i tillegg til antall desimaler i observasjonene.

2) Vi bruker forskjellig parentes ved nedre og øvre klassegrense,

   <u>Eks:</u> < 5 , 10] betyr fra og med 6 til og med 10.

3) Vi lar det være et «sprang» mellom angivelse av øvre klassegrense i en klasse og nedre grense i neste klasse. Klasse*bredden* for en klasse blir da forskjellen på nedre klassegrense i neste klasse og nedre klassegrense i klassen selv.

   <u>Eks:</u> 5-9 betyr en klasse fra og med 5 til og med 9, neste klasse blir 10-14 (antar samme klassebredde), og klassebredden blir 10–5 = 5.

c) (Frekvens-) tabell

Vi lager en tabell med god plass til mange kolonner, og med plass til like mange rekker som det er antall klasser, pluss en summeringsrekke.

Første kolonne angir de ulike klasseintervallene, med de laveste verdiene først (øverst) og deretter klasser med stigende verdier.

Hva som skal stå i de neste kolonnene vil avhenge noe av oppgaven. Aktuelle kandidater er:

Klassemidtpunkt, m_i, dvs. midtpunktet mellom de *egentlige* klassegrensene.

Tellekolonne: Du merker av en strek i riktig klasserubrikk for hver av dine n observasjoner (tilsammen n streker).

Frekvens-kolonne (f_i): Du angir antall observasjoner (f.eks. ved opptelling fra tellekolonnen) innenfor hver klasse. Dette kalles *klassefrekvensen*.

Relativ frekvens-kolonne (f_i/n): Du tar klassefrekvensen og deler på n. Summen av alle relative frekvenser skal alltid være lik 1 (± avrundingsfeil).

Kumulativ frekvens-kolonne (F_i): Sum av klassefrekvensene fra og med den første klassen til og med den klassen du ser på. I siste klasse vil kumulativ frekvens alltid være lik n (dvs. antall data).

Relativ kumulativ frekvens-kolonne (F_i/n): Du tar kumulativ klassefrekvens og deler på n. I siste klasse skal da alltid relativ kumulativ frekvens være lik 1 (± avrundingsfeil).

$m_i \cdot f_i$-kolonne: For hver klasse beregner du produktet av klassemidtpunktet, m_i, og klassefrekvensen, f_i. Hensikt: Beregne gruppert middelverdi.

$m_i^2 \cdot f_i$-kolonne: For hver klasse beregner du produktet av kvadratet av klassemidtpunktet, m_i^2, og klassefrekvensen, f_i. Hensikt: Beregne gruppert standardavvik.

Rektangelhøyde: Relativ frekvens dividert på klassebredde. Hensikt: Finne høyden på hvert rektangel ved fremstilling av relativ frekvens-histogram (defineres litt senere).

d) Senter og spredning

På basis av de grupperte data kan vi f.eks. beregne *gruppert middelverdi,* x_g *, og gruppert standardavvik,* s_g, som mål på henholdsvis senter og spredning, etter oppskriften i neste ramme. Vi kan også beregne gruppert median, m_g, grupperte 100p-prosentiler og gruppert interkvartilbredde, slik angitt i neste ramme.

Bemerk at vi i uttrykket for gruppert standardavvik i neste ramme har n og ikke $n-1$ i nevner. Når vi har gruppert våre data har vi allerede fjernet så mye detaljinformasjon om rådataene, at forutsetningene for å velge $n-1$ ikke lenger er tilstede.

Gruppert middelverdi, $\bar{x}_g$

(1.9) $$\bar{x}_g = \frac{1}{n}\sum_{i=1}^{k} m_i \cdot f_i$$

Gruppert standardavvik, s_g:

(1.10) $$s_g = \sqrt{\frac{1}{n}\sum_{i=1}^{k}\left(m_i - \bar{x}_g\right)^2 f_i} = \sqrt{\frac{1}{n}\sum_{i=1}^{k} m_i^2 f_i - \bar{x}_g^2}$$

der m_i er klassemidtpunkt og f_i er frekvens til klasse nr. i ($i = 1,\ldots,k$). *Gruppert varians*, s_g^2, er identisk med uttrykket under rottegnet.

Gruppert median, m_g

(1.11) $$m_g = x_1 + \frac{n/2 - F_1}{f_m} \cdot \Delta x_m$$

der

x_1 = nedre klassegrense i medianklassen[1]
F_1 = kumulativ frekvens i klassen *forut for* medianklassen
f_m = frekvens til medianklassen
Δx_m = klassebredden til medianklassen

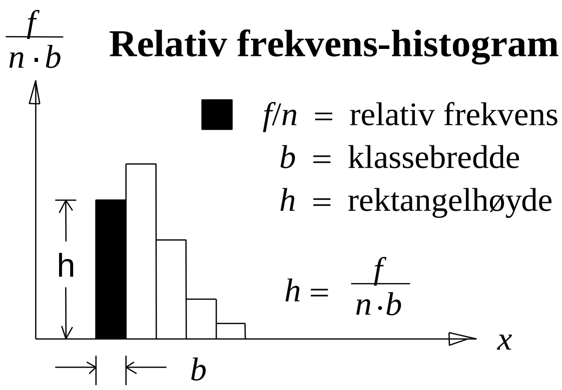
[1]Medianklassen er den første klassen der kumulativ frekvens, F_m, er større enn $n/2$.

(Gruppert 100p-prosentil, x_{gp}, finnes ved å erstatte $n/2$ i formelen ovenfor med np, og medianklassen med «p-klassen». Gruppert inter-kvartilbredde = $x_{g0,75} - x_{g0,25}$).

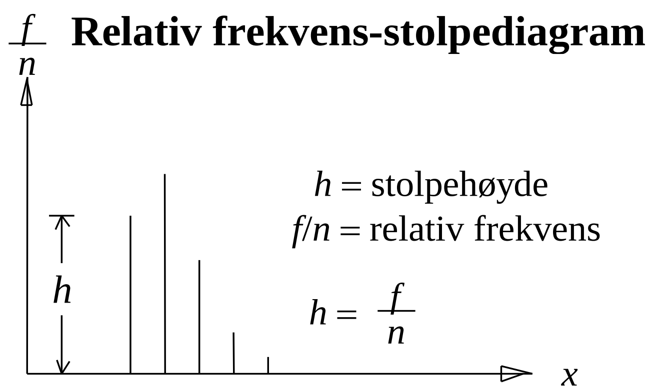
e) Grafisk fremstilling

De grupperte dataene fremstilles gjerne grafisk. Vi skal her betrakte 2 vanlige grafiske presentasjonsformer: *Relativ frekvens-histogram* og *relativ frekvens-stolpediagram* (det er også mange andre ord for stolpediagram: Linjediagram, stavdiagram, pinnediagram m.m.). Forskjellen på de 2 presentasjonsformene skulle gå fram av det følgende.

Relativ frekvens-histogram består av rektangler: Et rektangel pr. klasse og rektangelareal lik relativ klassefrekvens. Bredden av hvert rektangel er lik klassebredden. *Høyden av hvert rektangel* (y-verdien i x,y-diagrammet) blir derfor *relativ klassefrekvens dividert på klassebredden*.



Relativ *frekvens-stolpediagram* benyttes kun for diskrete variabler. Da tegnes en loddrett stolpe for hver diskrete verdi der vi har observasjoner. Høyden på hver stolpe (y-verdien) er normalt lik relativ frekvens.



Eks. 1.9 **Hovedeksempel, beskrivende statistikk for én variabel**

Følgende høydedata for studenter ved Høgskolen i Tromsø er gitt:

190	184	180	180	178	171	180	179	176	188
180	170	184	189	178	181	182	178	165	182
180	174	176	178	187	166	191	185	183	180
180	172	186	175	190	168	170	160	176	182
176	176								

Oppgave

a) Finn middelverdi, $\bar{x}$, og standardavvik, s, på basis av rådataene.

b) Rangordne dataene og bestem medianen og interkvartilbredden.

c) Gruppér dataene og bestem gruppert middelverdi, gruppert standardavvik og gruppert median.

d) Fremstill de grupperte høydedataene i et relativ frekvens-histogram.

Løsningsforslag

a) Vi beregner først middelverdi, $\bar{x}$, på basis av rådataene (se lign.(1.1)):

$$\bar{x} = \tfrac{1}{n}\sum_{i=1}^{n} x_i = \tfrac{1}{42}\left(190+184+180+\cdots+176\right) \text{ cm} = \tfrac{7506}{42}\text{ cm} = \underline{178.7 \text{ cm}}$$

Vi beregner så standardavviket, s, på basis av rådataene (se lign.(1.4)):

$$s^2 = \tfrac{1}{42-1}\left(\sum_{i=1}^{42} x_i^2 - \tfrac{1}{n}\left(\sum_{i=1}^{42} x_i\right)^2\right)\text{cm}^2 = \tfrac{1}{41}\left(1343458 - \frac{7506^2}{42}\right)\text{cm}^2$$

$$\approx 49.48 \text{ cm}^2 \Rightarrow s = \sqrt{49.48} \text{ cm} = \underline{7.0 \text{ cm}}$$

b) Vi rangordner dataene og får (stigende rekkefølge fra venstre mot høyre):

160	165	166	168	170	170	171	172	174	175
176	176	176	176	176	178	178	178	178	179
180	180	180	180	180	180	180	181	182	182
182	183	184	184	185	186	187	188	189	190
190	191								

Vi bestemmer nå medianen, m, på basis av de rangordnede dataene ovenfor. Her er n = 42 som er et like tall, og vi benytter lign.(1.6):

$$m = \tfrac{1}{2}\left(x_{(42/2)} + x_{(42/2+1)}\right) = \tfrac{1}{2}\left(x_{(21)} + x_{(22)}\right) = \tfrac{1}{2}(180+180) \text{ cm} = \underline{180 \text{ cm}}$$

For å finne interkvartilbredden, må vi først finne nedre og øvre kvartil, Q_1 og Q_3, respektive (se eks. 1.8).

Q_1 = 25-prosentilen (nedre kvartil):
p = 0,25, $np = 42 \cdot 0.25 = 10.5$, som *ikke* er et heltall. Forhøyer da til nærmeste heltall: $j = 11$, og får: $Q_1 = x_{0.25} = x_{(11)} = 176$

Q_3 = 75-prosentilen (øvre kvartil):
$p = 0.75$, $np = 42 \cdot 0.75 = 31.5$, som *ikke* er et heltall. Forhøyer da til nærmeste heltall: $j = 32$, og får: $Q_3 = x_{0.75} = x_{(32)} = 183$

$\Rightarrow$ interkvartilbredden = $Q_3 - Q_1$ = 183 cm − 176 cm = <u>7 cm</u>

c) Dataene skal nå grupperes, og vi følger «oppskriften»:

$x_{\min} = x_{(1)} = 160$, $x_{\text{maks}} = x_{(42)} = 191$. Vi velger derfor *første* høyde-klasse med *nedre intervallgrense høyst* lik *160*, og *siste* høydeklasse med *øvre klassegrense minst* lik *191*. Vi foretar følgende skjønnsmessige valg:

Vi velger én og samme klassebredde lik 4 for alle klasser, med nederste klassegrense lik 160 og øverste klassegrense lik 191. Klassemidtpunktene blir da 161.5, 165.5, 169.5,… osv. dersom vi antar at høydetallene er forhøyet (eks: 184.5 cm forhøyes til 185 cm). Vi kan lage følgende tabell, der tellekolonnen er fornuftig å ha dersom vi skal gruppere dataene fra råmaterialet *før* det er rangordnet. I vårt tilfelle er det lettere å fylle ut frekvenskolonnen i tabellen direkte fra de rangordnede dataene.

Høyde [cm]	fre-kvens f_i	relativ frekv. f_i/n	midt-pkt. m_i	m_if_i	$m_i^2f_i$	kum. frekv. F_i	rel. kum. frekv. F_i/n	rekt. høyde $\frac{f_i}{n\Delta x}$
160–163	1	1/42	161.5	161.5	26082.25	1	1/42	1/168
164–167	2	2/42	165.5	331.0	54780.5	3	3/42	2/168
168–171	4	4/42	169.5	678.0	114921	7	7/42	4/168
172–175	3	3/42	173.5	520.5	90306.75	10	10/42	3/168
176–179	10	10/42	177.5	1775	315062.5	20	20/42	10/168
180–183	12	12/42	181.5	2178	395307	32	32/42	12/168
184–187	5	5/42	185.5	927.5	172051.25	37	37/42	5/168
188–191	5	5/42	189.5	947.5	179551.25	42	1	5/168
Sum:	42	1		7519	1348062.5			

Vi finner gruppert middelverdi og standardavvik ved å benytte henholdsvis lign.(1.9) og (1.10), og benytte verdiene fra tabellen over:

$$\overline{x}_g = \tfrac{1}{n}\sum_{i=1}^{k} m_i f_i = \tfrac{1}{42}\cdot 7519 \text{ cm} = \underline{179.0 \text{ cm}}$$

Legg merke til at vi får nesten samme resultat som for de ugrupperte dataene, på tross av at vi har gruppert dataene i bare $k = 8$ klasser og tatt utgangspunkt i disses midtverdier, mot opprinnelig 42 tall.

Vi beregner så gruppert standardavvik:

$$s_g^2 = \tfrac{1}{n}\sum_{i=1}^{k} m_i^2 f_i - \overline{x}_g^2 = \tfrac{1}{42}\cdot 1348062.5 \text{ cm}^2 - \left(\tfrac{7519}{42}\right)^2 \text{cm}^2 = 47.20 \text{ cm}^2$$

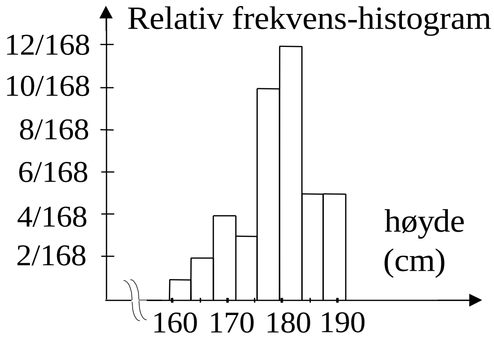
$$\Rightarrow\ s_g = \sqrt{47.20} \text{ cm} = \underline{6.9 \text{ cm}}$$

Også her får vi som vi ser godt samsvar med resultatene fra de ugrupperte dataene (s = 7,0 cm).

Deretter beregner vi gruppert median. Vi finner først medianklassen. $n/2 = 42/2 = 21 \Rightarrow$ medianklassen er den første klassen med kumulativ frekvens større enn 21. Vi ser at dette er klasse nummer 6 ovenfra i tabellen, med kumulativ klassefrekvens $F_m = 32$, klassefrekvens $f_m = 12$ og klassebredde $\Delta x_m = 4$. Videre er kumulativ klassefrekvens i klassen forut for median-klassen lik $F_1 = 20$. Nedre klassegrense i medianklassen skal her regnes som $x_1 = 179.5$. Vi får da i henhold til lign. (1.11):

$$m_g = x_1 + \frac{n/2 - F_1}{f_m}\cdot \Delta x_m = 179.5 \text{ cm} + \tfrac{21-20}{12}\cdot 4 \text{ cm} = \underline{179.8 \text{ cm}}$$

Som vi ser, er også dette nær medianen bestemt på basis av ugruppert datamateriale (m = 180 cm).



d) Vi skal til slutt fremstille de grupperte dataene i et relativ frekvenshistogram.

Siste kolonne i tabellen viser høyden på rektanglene.

Merk at totalt areal av alle histogramstolpene blir 1, eller 100%.

Høydedataene er lagt inn på statistikkprogrampakken **Minitab**. Sammenlign resultatene fra Minitab, gjengitt nedenfor, med de som er utført ovenfor.

MTB > describe 'hoyde'.

Descriptive Statistics

Variable	N	Mean	Median	StDev
hoyde	42	178.71	180.00	7.03

Variable	Min	Max	Q1	Q3
hoyde	160.00	191.00	175.75	183.25

MTB > histogram 'hoyde';
SUBC> start 161.5;
SUBC> increment 4.

Character Histogram

```
Histogram of hoyde   N = 42

Midpoint   Count
  161.50       1  *
  165.50       2  **
  169.50       4  ****
  173.50       3  ***
  177.50      10  **********
  181.50      12  ************
  185.50       5  *****
  189.50       5  *****
```

MINITAB

Høydedataene er lagt inn på kolonnen kalt ‘hoyde’.

«MTB >» er Minitab prompt, og det som følger etter er kommandoer lagt inn av bruker.

Hovedkommando legges inn av bruker etter «MTB >», og avsluttes med «;» dersom underkommandoer skal angis.

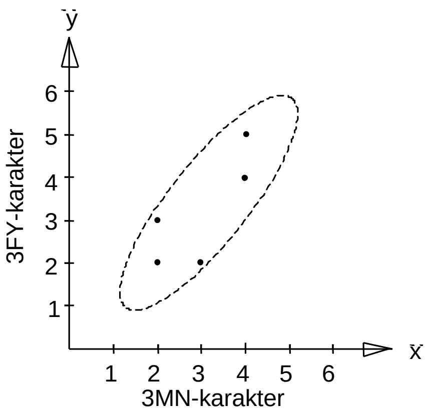
Kommandoene avsluttes med «.» tilslutt.

Utskriften til venstre er noe redigert. ☺

1.5 Spredningsdiagram

Hittil har vi sett på *enkelttall*. I de resterende avsnittene skal vi se på *tallpar*. La oss belyse denne forskjellen ut fra tabellen nedenfor over samhørende 3MN- (matematikk-) og 3FY- (fysikk-) karakterer til 5 tilfeldig valgte studenter. Dataene er tegnet inn i (x,y)-diagrammet til høyre, som er et eksempel på et *spredningsdiagram*.

Student nr:	3MN-karakter, x	3FY-karakter, y
1	3	2
2	2	2
3	4	5
4	2	3
5	4	4



Matematikkarakterene utgjør et tallmateriale bestående av enkelttall. Det samme gjør fysikkarakterene. Vi kan bruke teknikkene fra tidligere til å beregne f.eks. middelverdi og standardavvik til 3MN-karakterene såvel som til 3FY-karakterene. Slike mål sier imidlertid ingenting om hvordan x- og y-verdiene *samvarierer*.

Ser vi på matematikk- og fysikkarakteren til en og samme student samlet, får vi et *tallpar* for hver student. Tilsammen får vi følgende 5 tallpar:

(3,2), (2,2), (4,5), (2,3), (4,4)

Det er disse som er tegnet inn i spredningsdiagrammet ovenfor. Med tallpar kan vi belyse en del problemstillinger knyttet til hvordan x- og y-verdiene samvarierer:

- er det noen form for systematisk sammenheng mellom x- og y-verdiene?
- kan vi tallfeste i hvor stor grad det er en slik sammenheng?
- kan vi tilpasse en fornuftig funksjon som beskriver sammenhengen mellom x- og y-verdiene?
- kan vi forutsi en variabel dersom den andre er kjent?

Å studere enten x-målingene for seg selv eller y-målingene for seg selv vil ikke være til hjelp når det gjelder å svare på disse spørsmålene. Som vi ser fra spredningsdiagrammet ovenfor, er det en klar systematisk sammenheng mellom x- og y-verdiene: En student med god/dårlig karakter i 3MN ser ut til jevnt over å ha en rimelig god/dårlig karakter i 3FY (*positiv korrelasjon)*. Vi har ytterligere understreket denne sammenhengen med den stiplede ellipsen.

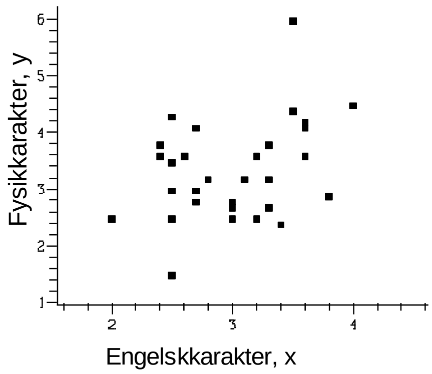
Å fremstille tallparene grafisk i form av et spredningsdiagram er et viktig første skritt i studiet av sammenhengen mellom 2 variabler. Vi avmerker (som det går fram av vårt innledningsdiagram) x-verdien langs horisontalaksen og den tilhørende y-verdien langs vertikalaksen. Parene (x,y), som består av observasjoner (målinger), blir da plottet som grafiske punkter. Det resulterende diagram er kalt et spredningsdiagram. Ved å se på spredningsdiagrammet, kan vi få et visuelt inntrykk av en eventuell (systematisk) sammenheng mellom variablene. For eksempel kan vi observere hvorvidt punktene ligger som et bånd rundt en rett linje, rundt en krum kurve eller om de simpelthen danner en mønsterløs samling av tilsynelatende tilfeldig spredte verdier.

NB! Ved tegning av spredningsdiagrammer bør skaleringen langs aksene (antall enheter pr. cm) velges slik at utstrekningen av $x_{maks} - x_{min}$ i cm er omtrent like stor som utstrekningen av $y_{maks} - y_{min}$ i cm.

Eks. 1.10 **Sammenheng mellom fysikk- og engelskkarakterer**
x = engelskkarakter, y = fysikkarakter

Tabellen nedenfor viser sammenhengen mellom engelsk- (x) og fysikk- (y) karakterene til $n = 30$ studenter.

x	y	x	y	x	y	x	y	x	y	x	y
3,3	3,8	2,5	3,5	3,3	3,2	2,5	3,0	3,4	2,4	2,7	2,8
2,7	4,1	3,0	2,5	3,1	3,2	3,6	4,1	3,3	2,7	3,0	2,8
2,6	3,6	2,7	3,0	2,5	1,5	3,6	4,2	2,4	3,6	3,0	2,7
2,8	3,2	4,0	4,5	2,5	2,5	3,5	4,4	2,4	3,8	3,8	2,9
3,2	3,6	3,5	6,0	2,0	2,5	3,2	2,5	2,5	4,3	3,6	3,6



Spredningsdiagrammet er vist i figuren til venstre. Sørvest til nordøst-mønsteret som punktene danner indikerer en positiv sammenheng mellom x og y: Studentene med gode/ dårlige karakterer i engelsk tenderer mot å ha gode/ dårlige karakterer i fysikk. Sammenhengen mellom x og y er imidlertid opplagt ikke gitt ved noen pen matematisk funksjon. ☺

1.6 Empirisk korrelasjonskoeffisient

Spredningsdiagrammet gir et visuelt inntrykk av sammenhengen mellom x- og y-verdiene i et tallmateriale som består av tallpar (bivariat datasett). I mange tilfeller synes punktene å ligge i et bånd rundt en rett linje. I varierende grad vil imidlertid tilfeldige variasjoner utelukke en perfekt lineær (rettlinjet) sammenheng.

Den empiriske **korrelasjonskoeffisienten**, som vi skal betegne med r, er et mål på graden av lineær sammenheng mellom x- og y-variablene. Før vi introduserer formelen for r, skal vi angi noen viktige egenskaper ved korrelasjonskoeffisienten, og diskutere på hvilken måte den kan brukes til å måle graden av lineær sammenheng.

a) r-verdiene ligger alltid mellom –1 og 1: $\quad -1 \le r \le 1$

b) Absoluttverdien til r indikerer graden av lineær sammenheng, mens fortegnet indikerer retning. Mer presist:

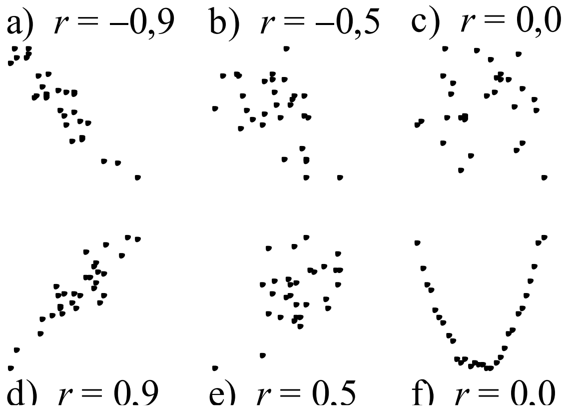
   $\boldsymbol{r > 0}$ hvis mønsteret til (x,y) verdiene er et bånd som løper fra nedre venstre til øvre høyre hjørne

   $\boldsymbol{r < 0}$ hvis mønsteret til (x,y) verdiene er et bånd som løper fra øvre venstre til nedre høyre hjørne

   $\boldsymbol{r = 1}$ hvis alle (x,y) verdiene ligger eksakt på en og samme rette linje med et positivt stigningstall

   $\boldsymbol{r = -1}$ hvis alle (x,y) verdiene ligger eksakt på en og samme rette linje med et negativt stigningstall

c) En r-verdi i nærheten av null betyr at det er liten grad av lineær sammenheng.



Figur: Eksempler på spredningsdiagram og tilhørende r-verdi

Korrelasjonskoeffisienten er nær null når det ikke er noe synlig mønster på sammenheng, dvs. y-verdiene synes ikke å variere i noen foretrukket retning når x-verdiene varierer. En r-verdi i nærheten av null kan også inntreffe fordi punktene ligger i et bånd rundt en kurve som er alt annet enn en rett linje. Husk at r er et mål på lineær sammenheng, og en svært bøyd kurve er langt fra lineær.

Forrige figur viser sammenhengen mellom ulike spredningsdiagram og den tilhørende r-verdien. Legg merke til at c) og f) begge tilsvarer situasjoner der $r = 0$. Null korrelasjon i Fig. 1.1c) skyldes fraværet av enhver sammenheng mellom x og y, mens null korrelasjon i Fig. 1.1f) skyldes at sammenhengen følger en (ikke-lineær) kurve som er mer eller mindre symmetrisk om middelverdien til x-verdiene.

Empirisk korrelasjonskoeffisient, *r*

r-verdien beregnes utifra n par av observasjoner $(x_1,y_1),\ldots,(x_n,y_n)$, ved følgende formel:

$$r = \frac{S_{xy}}{\sqrt{S_x^2}\sqrt{S_y^2}}$$

der

$$S_x^2 = \Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - \tfrac{1}{n}(\Sigma x_i)^2$$
$$S_y^2 = \Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - \tfrac{1}{n}(\Sigma y_i)^2$$
$$S_{xy} = \Sigma(x_i - \bar{x})(y_i - \bar{y}) = \Sigma x_i y_i - \tfrac{1}{n}(\Sigma x_i)(\Sigma y_i)$$

og alle summer er fra $i = 1$ til $i = n$. $S_x{}^2$ og $S_y{}^2$ er summene av kvadratiske avvik (fra middelverdien) til henholdsvis x-verdiene og y-verdiene. S_{xy} er summen av kryssproduktene mellom x-avvik og y-avvik fra de respektive middelverdier.

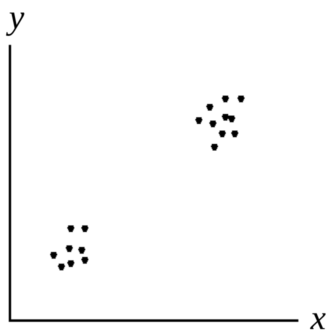
Eks. 1.11 **Beregne *r* for følgende *n* = 3 par av observasjoner:**
$(x,y) = (3,1),\ (1,0)$ og $(8,2)$

	x	y	x^2	y^2	xy
	3	1	9	1	3
	1	0	1	0	0
	8	2	64	4	16
Tot:	12	3	74	5	19
	$=\Sigma x$	$=\Sigma y$	$=\Sigma x^2$	$=\Sigma y^2$	$=\Sigma xy$

$$r = \frac{19 - \frac{12\cdot 3}{3}}{\sqrt{74 - \frac{12^2}{3}}\cdot\sqrt{5 - \frac{3^2}{3}}} = 0{,}97$$

☺

Vi minner leseren om at r måler hvor nær mønsteret til spredningen er en rett linje. Tilfelle f) i forrige figur presenterer en stor grad av samvariasjon mellom x og y, men en som ikke er lineær. Den lave verdien til r for disse data reflekterer ikke den store graden av ikke-lineær samvariasjon.



Figur: r ≈ 1, se tekst.

En annen situasjon der den empiriske korrelasjonskoeffisient r ikke er god, opptrer når spredningsplottet er delt i 2 adskilte punktsamlinger. I slike tilfeller kan det være best å forsøke å bestemme den underliggende årsak. Det kan f. eks. være at en del av utvalget kommer fra en populasjon mens en annen del kommer fra én annen populasjon (Populasjonsbegrepet vil bli nærmere definert i et senere kapittel).

Korrelasjon og årsak

NB! Det kan være lett å mistolke en observert korrelasjon (r i nærheten av -1 eller $+1$) mellom to variabler som et årsaksforhold mellom variablene. Et klassisk eksempel er at en har observert en høy positiv korrelasjon mellom antall storker og antall barnefødsler i europeiske byer. Årsaken til dette er ikke at babyene kommer med storken, men at det er en tredje variabel som ligger og «lurer» i bakgrunnen: Størrelsen på byene. Jo større byer, jo flere storker, og jo større byer jo flere babyer. Det er altså bystørrelsen som får antall storker og antall babyer til å variere i samme retning.

Observasjonen at 2 variabler synes å samvariere i en bestemt retning medfører altså ikke nødvendigvis at det er et direkte årsaksforhold mellom variablene. Det kan være variasjonen i en tredje variabel, som forårsaker at x og y varierer i samme retning, selv om de er uten sammenheng, eller til og med har motsatt sammenheng av den som indikeres av korrelasjonskoeffisienten. Den falske korrelasjonen som fremkommer på denne måten kan vi kalle *villedende* korrelasjon. Det er mer sunn fornuft enn statistisk resonnement som skal til for å bestemme hvorvidt en observert korrelasjon kan bli tolket praktisk eller om den er villedende.

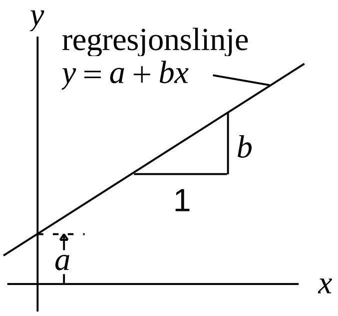
Advarsel: En observert korrelasjon mellom 2 variabler kan være villedende. Dvs. at den kan være forårsaket av en tredje variabel. Når vi bruker korrelasjonskoeffisienten som mål på sammenheng, må vi være oppmerksomme på muligheten for at en lurevariabel påvirker noen av variablene vi betrakter.

1.7 Lineær regresjon

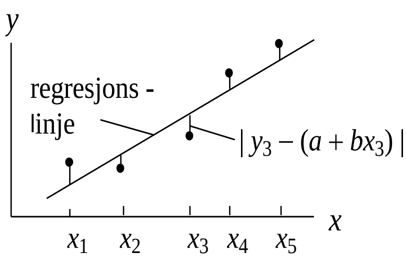
Studier av sammenhengen mellom to variabler ved målinger er ofte motivert ut fra et behov for å kunne forutsi den ene variabelen fra den andre. En leder for et jobbtreningsprogram kan ønske å studere sammenhengen mellom varigheten av treningen, og resultatet av treningen ved en påfølgende test. En skogeier kan ønske å anslå (estimere) tømmervolumet til et tre fra måling av stamme-diameteren 1 meter over bakken. En forsker innen medisin kan være interessert i å forutsi alkoholinnholdet i blod ut fra målinger fra et nylig oppfunnet pusteapparat.

I slike sammenhenger som disse, er det vanlig å la x betegne den *uavhengige* variabelen, også kalt *inn-variabelen*, og la y betegne *responsen*, eller *ut-variabelen*. Formålet er å finne hvilken form for sammenheng det er mellom x og y fra eksperimentelle data, og å bruke denne sammenhengen til å prediktere (forutsi) responsen til variabelen y (*responsvariabelen*) fra inn-variabelen, x (*prediktoren*).

Første skritt er å plotte og undersøke spredningsdiagrammet. Hvis en lineær sammenheng fremkommer, så vil beregningen av korrelasjonskoeffisienten, r, bekrefte styrken av lineær sammenheng. r-verdien indikerer hvor effektivt y kan forutsies fra x ved å tilpasse en rett linje til dataene.



En linje $y = a + bx$ er bestemt ved to konstanter: Høyden over origo (skjæringspunktet med y-aksen), a, og stigningstallet, b, dvs. mengden y øker med når x øker med en enhet (se figur til venstre).



Vi ønsker å bestemme de verdier a^* og b^* for henholdsvis a og b som gjør at regresjonslinja $y = a + bx$ er best mulig tilpasset våre observasjoner/data. Et mye benyttet prinsipp er her minste kvadraters metode. Prinsippet går ut på å minimere kvadratsummen

$$Q = \sum_{i=1}^{n} \left(y_i - (a + bx_i) \right)^2$$

med hensyn på a og b, dvs. summen av kvadratene av «vertikal» avstand fra punktene til linja, se forrige figur. Matematisk partiellderiverer vi Q med hensyn på a og b, setter de to ligningene vi får lik null, og løser disse med hensyn på a og b. Resultatet er gjengitt i neste ramme. Vi skal gå grundigere til verks i kap. 9 om lineær regresjon.

Eks. 1.12 viser et praktisk eksempel, som også belyser hva som menes med prediksjon ut fra den tilpassete rette linja.

Rett linje-tilpasning, $y^* = a^* + b^* \cdot x$

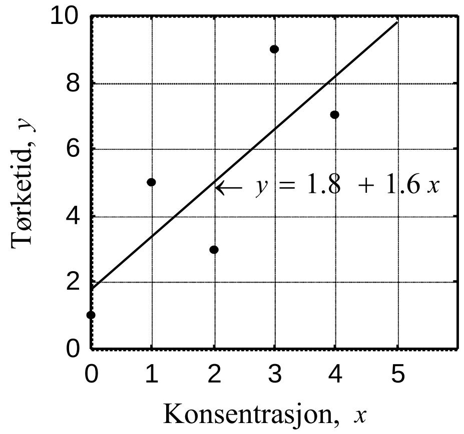
Vi har et tallmateriale $(x_1, y_1), \ldots, (x_n, y_n)$ og skal tilpasse en rett linje $y^* = a^* + b^* \cdot x$. Minste kvadraters metode gir da følgende formler til å bestemme a^* og b^*:

$$b^* = \frac{S_{xy}}{S_x^2}, \quad a^* = \bar{y} - b^* \cdot \bar{x}$$

der $\bar{x}$ og $\bar{y}$ er middelverdiene til henholdsvis x- og y-verdiene, og S_x^2 og S_{xy} er definert i ramme tidligere i kapitlet.

Eks. 1.12 **Maling og tørketid**.
En kjemiker ønsker å studere sammenhengen mellom tørketiden til en maling og konsentrasjonen av en basisk oppløsning som gjør det lettere å male. Data for konsentrasjonen (x) og den tilsvarende observerte tørketid (y) er gitt i de 2 første kolonnene i følgende tabell:

	Konsen-trasjon, x	Tørke-tid, y	x^2	y^2	xy
	0	1	0	1	0
	1	5	1	25	5
	2	3	4	9	6
	3	9	9	81	27
	4	7	16	49	28
tot:	10	25	30	165	66



Figur: Data for konsentrasjon, x, og tørketid, y (i minutter), og beregninger for regresjonslinja. Spredningsdiagram og tilpasset regresjonslinje til høyre.

Spredningsdiagrammet i figuren over gir et inntrykk av en viss grad av lineær sammenheng. For å beregne r og bestemme ligningen for den tilpassede linja, beregner vi først $\bar{x}$, $\bar{y}$, S_x^2, S_y^2 og S_{xy} fra tabellen ovenfor:

$$\bar{x} = \frac{10}{5} = 2, \quad \bar{y} = \frac{25}{5} = 5$$

$$S_x^2 = 30 - \frac{10^2}{5} = 10, \quad S_y^2 = 165 - \frac{25^2}{5} = 40, \quad S_{xy} = 66 - \frac{10 \cdot 25}{5} = 16$$

$$r = \frac{16}{\sqrt{40 \cdot 10}} = 0.8, \quad b^* = \frac{16}{10} = 1.6, \quad a^* = 5 - 1.6 \cdot 2 = 1.8$$

Ligningen for den tilpassede linja blir da:

$$y^* = 1.8 + 1.6x$$

Linja er vist i spredningsdiagrammet i figuren ovenfor. Dersom vi ønsker å forutsi tørketiden y som tilsvarer konsentrasjonen 2.5, setter vi bare inn for $x = 2.5$ i ligningen over og får resultatet:

Prediktert tørketid for $x = 2.5$ er $y = 1.8 + 1.6 \cdot 2.5 = 5.8$, dvs. 5.8 min. Grafisk finner vi denne verdien ved å lese av på y-aksen den verdien vår tilpassede linje har når $x = 2.5$. Merk imidlertid at dette anslaget er ganske usikkert, vi har ikke en svært sterk lineær sammenheng i dette tilfellet. ☺

I kap. 9 skal vi se nærmere på hvordan vi kan tallfeste usikkerheten til prediksjoner.

1.8 *Oppgaver*

1.1 Beregn empirisk middelverdi og standardavvik av følgende tall:

a) −1, 3, 8, 4, 13
b) −8, −9, −14, −11, 0

1.2 Gitt følgende tallmateriale:

1.43438 1.43443 1.43442 1.4344

a) La tallene ovenfor være *x*-verdier, og finn de tilsvarende *z*-verdier ved transformasjonen

$z = (x - 1.4344)/0.00001$

b) Beregn empirisk middelverdi, $\bar{z}$, og standardavvik, s_z, til *z*-verdiene.

c) Beregn middelverdi, $\bar{x}$, og standardavvik, s_x, til *x*-verdiene ved transformasjonen $s_x = 0{,}00001 s_z$ og

$\bar{x} = 1.4344 + 0.00001 \cdot \bar{z}$

1.3 Følg fremgangsmåten i oppgave 1.2 og beregn middelverdi og standardavvik til tallene

1221.3 1220.9 1221.7 1220.8 1221

1.4 I en bedrift med 14 ansatte er gjennomsnittsinntekten kr 158 000 pr. ansatt. Hvor store lønnsutgifter har bedriften?

1.5 a) Finn empirisk middelverdi og median for følgende tilfeldig målte utetemperaturer (^oC) som en person har notert i sin dagbok i løpet av ett år:

x: 14, −12, 0, 16, 7

b) En annen person har følgende måleresultater fra sin dagbok (samme sted og samme år):

y: 14, −12, 0, 16, 7, 112

Beregn empirisk middelverdi og median også i dette tilfellet. Ville du stole mest på middelverdien eller medianen?

1.6 Gitt følgende *x*-verdier:

1, −3, 7, 12, 6

Beregn 20- og 70-prosentilen.

1.7 Gitt *x*-verdiene:

21, 17, 18, 17, 22, 21, 78, 22, 24

Beregn standardavvik og interkvartilbredde. Hvilket av de to målene på spredning ville du stole mest på, dersom du i ettertid fikk opplyst at en av verdiene var feil?

1.8 Tabellutsnittet til høyre viser 3 vektklasser. Angi nedre og øvre klassegrense, samt klassemidtpunkt og klassebredde.

Vekt [kg]
⋮
[10-20>
[20-30>
[30-40>
⋮

1.9 Neste tabell viser daglig inntekt av CD-salg i en musikkforretning, fordelt på 4 inntektgrupper. Anta at et dagssalg på kr. 9999 kommer i den første klassen.

Videosalg (kr 1000)	Frekvens (dager)
0-9	37
10-19	148
20-29	123
30-39	57

a) Bestem klassemidtpunkt i hver klasse.

b) Bestem gruppert middelverdi.
c) Bestem gruppert standardavvik.
d) Bestem gruppert median.
e) Bestem gruppert interkvartilbredde.

1.10 Gitt tallparene

(0.9 , 1.1), (2.1 , 1.8) og (2.9 , 3.3)

a) Velg 1 cm/enhet langs x- og y-aksen, og lag et spredningsdiagram.
b) Beregn korrelasjonskoeffisienten, r.
c) Beregn en regresjonslinje tilpasset dataene, og tegn den inn i spredningsdiagrammet.

1.11 Gitt følgende samhørende x- og y-verdier:

x	.68	.52	.58	.72	.62
y	11	5	1	4	6

a) Beregn korrelasjonskoeffisienten, r.
b) Beregn en regresjonslinje tilpasset dataene, $y^* = a^* + b^*x$.
c) Tegn (x,y)-verdiene i et spredningsdiagram med 1 cm pr. enhet ***både*** langs x- og y-aksen, og tegn inn regresjonslinja. Hvorfor står linja «på tvers» av dataene?
d) Gjør c) på nytt, men velg selv en fornuftig skalering langs aksene.

1.12 Størrelsen av en dyrebestand måles en gang i året. De 4 siste årene ble følgende observert:

år:	87	88	89	90
bestand:	123	237	471	982

La x = 1, 2, 3 og 4 betegne årstallene 87, 88, 89 og 90, og la y betegne de tilsvarende bestandene.

a) Beregn korrelasjonskoeffisienten.
b) Utfør transformasjonen $z = \ln y$, og beregn z-verdiene.
c) Bestem korrelasjonskoeffisienten for (x,z)-verdiene.
d) Tilpass en rett linje til (x,z)-dataene.
e) Bestem på basis av d) en regresjonsfunksjon, $y = ce^{dx}$, tilpasset (x,y)-dataene.

1.13 (Vekt-data i tabell lenger ned).

a) Finn middelverdi og standardavvik til vekt-dataene på basis av rådataene.
b) Rangordne vekt-dataene og bestem median og interkvartilbredde.
c) Gruppér vekt-dataene og bestem gruppert middelverdi, gruppert standardavvik og gruppert median.
d) Fremstill de grupperte vekt-dataene i et relativ frekvens-histogram.

1.14 (data i tabell lenger ned).

a) Tegn et spredningsdiagram for samhørende vekt- og høyde-data.

b) Beregn korrelasjonskoeffisienten for vekt- og høyde-dataene, og kommenter spredningsdiagrammet i a).

c) Beregn korrelasjonskoeffisienten for høyde- og alders-dataene, og kommenter resultatet.

Tabell for oppgave 1.13 og 1.14

høyde [cm]	vekt [kg]	alder [år]	høyde [cm]	vekt [kg]	alder [år]	høyde [cm]	vekt [kg]	alder [år]
167	55	32	180	64	24	185	83	25
169	59	23	175	73	21	160	54	26
163	55	25	170	52	20	172	75	22
177	70	23	172	65	21	167	58	19
180	76	24	170	66	23	190	82	25
174	61	25	175	60	19	172	66	21
172	60	19	185	90	23	185	74	21
174	59	19	178	67	21	182	69	21
174	80	21	187	85	28	181.5	82	21
182	55	31	185	82	27	176	72	21
158	50	23	183	72	19	187	85	28
168	60	21	185	70	19			

1.15 I et utvalg på 20 fire-barnsfamilier er antall gutter følgende:

3 2 3 2 3 1 3 3 2 3
1 3 0 3 2 1 4 2 3 2

a) Regn ut middelverdien for utvalget.
b) Regn ut utvalgsstandardavviket.
c) Sett opp en fordelingstabell for de absolutte og relative frekvensene.
d) Bestem median og interkvartilbredde.
e) Lag relativ frekvens stolpediagram.

1.16 Aldersfordelingen i en bedrift er:

Alder [år]	15-25	25-35	35-45	45-55	55-65
Frekvens	2	7	6	4	2

a) Tegn histogram over aldersfordelingen.
b) Beregn aritmetisk middelverdi, standardavvik og median.
c) Avmerk de beregnede størrelsene i histogrammet.

1.17 En måleserie resulterte i følgende klassedelte observasjonsmateriale:

intervall	frekvens
[195, 200>	2
[200, 205>	4
[205, 210>	10
[210, 215>	11
[215, 220>	9
[220, 225>	4

a) Utvid tabellen med relative frekvenser, kumulative frekvenser og relative kumulative frekvenser.

Lag et histogram for observasjonsmaterialet.

b) Beregn tilnærmet aritmetisk middelverdi og standardavvik for observasjonsmaterialet.

c) Beregn tilnærmet median og interkvartilbredde for observasjonsmaterialet.

1.9 *Formelsamling*

Ugrupperte data

Empirisk middelverdi

$$\bar{x} = \tfrac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\cdot\left(x_1 + \cdots + x_n\right)$$

Empirisk standardavvik

$$s = \sqrt{\tfrac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

$$= \sqrt{\tfrac{1}{n-1}\left(\Sigma x_i^2 - \tfrac{1}{n}\left(\Sigma x_i\right)^2\right)}$$

$$= \sqrt{\tfrac{1}{n-1}\left(\Sigma x_i^2 - n\bar{x}^2\right)}$$

Empirisk median

$$m = \begin{cases} \tfrac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right), & n \text{ like} \\ x_{\left(\frac{n+1}{2}\right)}, & n \text{ odde} \end{cases}$$

Empirisk 100*p*-prosentil

$$x_p = \begin{cases} \tfrac{1}{2}\left(x_{(np)} + x_{(np+1)}\right), & n \text{ like} \\ x_{(j)}, & n \text{ odde} \end{cases}$$

der j er minste heltall større enn np.

Interkvartilbredde

$Q_3 – Q_1 = x_{0.75} - x_{0.25}$

Grupperte data

Betegnelser

k er antall klasser, n er totalt antall observasjoner, m_i er klassemidtpunkt, f_i er klassefrekvens, F_i er kumulativ klassefrekvens i klasse nr. i, $i = 1,\ldots, k$.

Gruppert middelverdi

$$\bar{x}_g = \tfrac{1}{n}\sum_{i=1}^{k} m_i \cdot f_i$$

Gruppert standardavvik

$$s_g = \sqrt{\tfrac{1}{n}\sum_{i=1}^{k} m_i^2 \cdot f_i - \bar{x}_g^2}$$

Gruppert median

$$m_g = x_1 + \frac{n/2 - F_1}{f_m}\cdot \Delta x_m$$

der x_1 er nedre klassegrense i median-klassen, F_1 er kumulativ frekvens i klassen *forut* for medianklassen, f_m er frekvensen til medianklassen og Δx_m er bredden til medianklassen.

Bivariate data

Summasjonsvariabler

$$S_x^2 = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \Sigma x_i^2 - \tfrac{1}{n}\left(\Sigma x_i\right)^2$$

$$S_y^2 = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \Sigma y_i^2 - \tfrac{1}{n}\left(\Sigma y_i\right)^2$$

$$S_{xy} = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$

$$= \Sigma x_i y_i - \tfrac{1}{n}\Sigma x_i \Sigma y_i = \Sigma x_i y_i - n\,\bar{x}\cdot\bar{y}$$

Empirisk korrelasjonskoeffisient

$$r = \frac{S_{xy}}{S_x S_y} = \frac{S_{xy}}{\sqrt{S_x^2}\cdot\sqrt{S_y^2}}$$

Regresjonslinje

$y^*(x) = a^* + b^* \cdot x,$ der

$b^* = S_{xy} / S_x^2, \quad a^* = \bar{y} - b^* \cdot \bar{x}$