

Jon Anjer

STATISTIKK

3. utgave, ny versjon

Høgskolen i Oslo
Avdeling for journalistikk, bibliotek- og informasjonsfag
2005

Innhold

1. Hvorfor statistikk for bibliotekarer?	4
1.1. Hva er statistikk? (4); 1.2. Beskrivende statistikk (4); 1.3. Slutningsstatistikk (5); 1.4. Dette kompendiet (5)	
2. Statistiske grunnbegreper	6
2.1. Data (6); 2.2. Undersøkelsesenheter (6); 2.3. Variabler (7); 2.4. Verdier (7); 2.5. Verdier og klasser (8); 2.6. Frekvenser (9)	
3. Målenivå	10
3.1. Nominalskala (10); 3.2. Ordinalskala (10); 3.3. Intervallskala (10); 3.4. Forholdstalls- skala (11); 3.5. Kontinuerlig vs. diskontinuerlig skala (11); 3.6. Dikotome variabler (11)	
4. Datamatriser og regneark	12
4.1. Datamatriksen (12); 4.2. Regneark (12); 4.3. EXCEL (13); 4.4. Knapperader i EXCEL (14); 4.5. Formater i EXCEL (15); 4.6. Inntasting av datamatrikse i EXCEL (15)	
5. Datapresentasjon: Tabeller	17
5.1. Frekvenstabeller (17); 5.2. Relative (prosentvise) tabeller (17); 5.3. Gruppering av data (18); 5.4. Funksjoner i EXCEL (19); 5.5. EXCEL for frekvenstabeller (20); 5.6. Kumulative tabeller (21); 5.7. EXCEL og kumulerte frekvenser (22)	
6. Datapresentasjon: Grafiske fremstillinger	23
6.1. Frekvenspolygon (23); 6.2. Stolpediagram (23); 6.3. Histogram (24); 6.4. Sektordiagram (Kakediagram) (24); 6.5. Diagram og målenivå (24); 6.6. Kumulativt diagram (25); 6.7. Diagrammer i EXCEL (26)	
7. Sentraltendens	28
7.1. Modus (28); 7.2. Aritmetisk gjennomsnitt (28); 7.3. Median (30); 7.4. Skjeve fordelinger (31); 7.5. Sentraltendens og EXCEL (32); 7.6. Sentraltendens og målenivå (32)	
8. Spredning	34
8.1. Variasjonsbredde (34); 8.2. Standardavvik (34); 8.3. Kvartiler (36); 8.4. Median og kvartiler ved frekvensfordelinger (37); 8.5. Kvartilavvik (38); 8.6. Spredningsmål og EXCEL (39); 8.7. Spredningsmål og målenivå (39)	
9. Korrelasjon	40
9.1. Punktdiagram (spredningsdiagram) (40); 9.2. Produkt-moment-korrelasjon (41); 9.3. Utregning av korrelasjonskoeffisienten (42); 9.4. Tolkning av korrelasjonskoeffisienten (43); 9.5. Ikke-lineær sammenheng (44); 9.6. Korrelasjon og målenivå (44)	
10. Regresjon	45
10.1. Formelen for en linje (46); 10.2. Formelen for regresjonslinjen (46); 10.3. Tegning av regresjonslinjen (47); 10.4. Prediksjon (47); 10.5. Regresjonslinje for x uttrykt ved y (49); 10.6. Korrelasjon, regresjon og EXCEL (49)	

11. Normalfordelingen	51
11.1. Standardskårer (z-verdier) (53); 11.2. Fra andeler til antall (55); 11.3. Fra andeler til verdier (56)	
12. Samplingfordelingen for gjennomsnitt	57
12.1. Sammenheng mellom utvalg og populasjon (57); 12.2. Samplingfordelingen (58); 12.3. Samplingfordelingen er en abstrakt fordeling (60); 12.4. Sentraltendens og spredning i de tre fordelingene (60)	
13. Konfidensintervall for gjennomsnitt	61
13.1. Sikkerhetsnivå (62); 13.2. Formelen for konfidensintervallet (62); 13.3. Andre modeller og andre konfidensintervall (63)	
14. Utvalgets størrelse	64
14.1. Konfidensintervall og utvalgsstørrelse (64); 14.2. Feilmargin (65); 14.3. Valg av utvalgsstørrelse (65)	
15. Hypotesetesting	66
15.1. Hypotese belyst ved konfidensintervall (66); 15.2. Testing av $H_0: \mu = k$ (67); 15.3. Signifikansnivå (67); 15.4. Kritisk verdi (68); 15.5. Andre hypotesetester (69); 15.6. Begrensninger ved bruk av hypotesetesting (69)	
16. Formler	71
17. Tabell over normalfordelingen	72
18. Register	73

1. Hvorfor statistikk for bibliotekarer?

1.1. Hva er statistikk?

Vi kan kort beskrive statistikk som beskrivelse og tolkning av kvantitative data. I bibliotek brukes statistikk først og fremst til å gi oversikt over data. Vi kan oppgi tall vi har samlet inn (f. eks. over utlån fra dag til dag: utlånsstatistikk). Vi kan analysere tallene vi har funnet (f. eks. regne ut gjennomsnittlig utlån pr. dag). Vi kan sette opp tabeller der vi viser utlån pr. måned fordelt på type bøker. Vi kan beskrive sammenhenger, og vi kan fremstille det vi finner i figurer.

I tillegg brukes statistikk til å generalisere fra et gitt datamateriale. Typiske tilfeller av dette er meningsmålinger, der det blir gjort slutninger fra et gitt utvalg av spurte personer til alle aktuelle personer. Brukerundersøkelser i bibliotek kan sjelden omfatte alle brukerne. I stedet velger en ut visse brukere. Ut fra hva svarene til disse brukerne er det mulig å trekke slutninger som med en viss grad av sikkerhet vil gjelde for den gjennomsnittlige brukeren.

Ut fra disse to formålene med statistikk skiller en gjerne mellom *beskrivende statistikk* (deskriptiv statistikk) og *slutningsstatistikk* (induktiv statistikk). I dette kompendiet legges hovedvekten på beskrivende statistikk, som de fleste bibliotekarer vil møte i det praktiske liv.

Statistikk er en av de viktigste samfunnsfaglige metodene. Bruk av statistikk i bibliotekfag er ofte av samme type som bruk av statistikk i samfunnsfag. Statistikk betydde opprinnelig beskrivelser av samfunnsforhold (Store norske leksikon).

Statistikk kan vise sammenhenger. Vi kan f.eks. finne ut at bruken av folkebibliotek har økt i en gitt periode. Samtidig kan vi se på andre endringer i samme periode, som endringer i arbeidslivet, bruk av fritid osv. Vi kan ut fra dette prøve å finne årsaker til endret bibliotekbruk. Her må vi bruke teori og kunnskaper vi har fra andre områder, statistikk alene kan ikke bevise årsakssammenhengene.

Statistikk bør knyttes til klare problemstillinger. Når vi skal samle eller bruke statistikk er det svært viktig å ha klart for seg på forhånd hva vi er ute etter, og hva det skal brukes til!

1.2. Beskrivende statistikk

Med ordet statistikk forestiller vi oss gjerne tabeller eller figurer som skal beskrive tallmessige fenomener. Mest kjent er Statistisk sentralbyrås statistiske serier. I biblioteksektoren er de viktigste statistiske publikasjonene Statistisk sentralbyrås «Undersøkelse om bruk av folkebibliotek», og Statens

Tabell I Fra årsrapporten til biblioteket ved SBIH (Statens bibliotek- og informasjonshøgskole). Tilvekstdata

	1992	1991	1990	1989
Antall bind	1517	833	1646	2036
Antall titler	603	457	961	742
Kassert	483	579	308	928
Tilvekst bøker	1034	254	1338	1108
Løpende periodika	358	341	348	338
Aviser	16	16	15	15

bibliotektilsyn og Riksbibliotektjenesten sine årlige statistiske oversikter over henholdsvis folkebibliotek og fagbibliotek., etc.

Beskrivende statistikk gir oss metoder til å presentere de data vi har samlet inn på en oversiktlig måte.

Denne formen for statistikk kalles også *deskriptiv statistikk*.

1.3. Slutningsstatistikk

Statistikk gir oss også metoder for å trekke mer generelle slutninger fra et innsamlet materiale. Vi kan stille spørsmålet: Hvor sikkert er det at data fra et utvalg er representative for hele befolkningen? Vi kan også ut fra tidligere observasjoner forutsi om et eller annet vil inntreffe.

Denne formen for statistikk kalles også *induktiv statistikk*.

For å kunne foreta statistiske slutninger, må vi ha kontroll over hva slags utvalg vi bruker. Utvalgsmetoder blir i dette kurset dekket i metoddelen.

1.4. Dette kompendiet

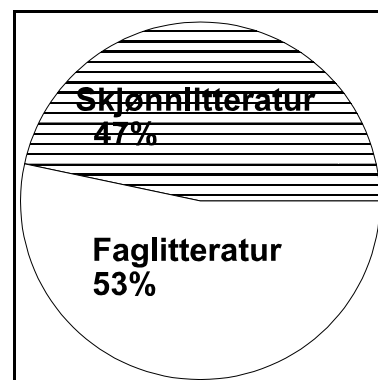
Dette kompendiet ble utarbeidet for innføring i statistikk i faget «2A2 Samfunnsfaglige metoder m/ statistikk» ved Høgskolen i Oslo, bibliotek- og informasjonsstudiene. En foreløpig utgave ble brukt i undervisningen høstsemesteret 1993, mens 1. utgave ble brukt i 1994 og 1995, 2. utgave i 1996. Med ny struktur har nå faget betegnelsen «Emne BoS21 Undersøkellesmetoder».

I denne utgaven er avsnittene om regneark oppgradert til EXCEL 97, og senere i en viss grad til 2003-utgaven. Avsnittene om regneark kan være vanskelige å studere på egen hånd. Du kan gjerne vente med å lese avsnittene om EXCEL, uten at dette går ut over forståelsen av de andre avsnittene. EXCEL-filer knyttet til eksemplene i boka ligger tilgjengelige på utdanningens nettverk for studentene.

Stor takk til Tord Høivik som har gått gjennom store deler av manuskriptet, og gitt verdifulle råd og kommentarer. Studentene som fulgte kurset høsten 1993, og også seinere, skal ha stor honnør: De har vist en høy overbærenhet og hurtig melding om trykkfeil, og gitt inspirerende tilbakemeldinger.

I denne versjonen er alle feil som hittil er oppdaget rettet. Det er likevel sikkert uklarheter og feil i kompendiet. Feil blir rettet fortløpende etter hvert som de blir oppdaget. Jeg er svært takknemlig for melding om feil og uklarheter, slik at det også kan komme nye lesere av kompendiet til gode.

Juli 2005 Jon Anjer



Figur 1 Fra årsrapporten til biblioteket ved SBIH (Statens bibliotek- og informasjons-høgskole). Fordelingen mellom fag- og skjønnlitteratur i samlingen

2. Statistiske grunnbegreper

*Hvem (eller hvilke ting) skal vi måle eller kartlegge? Hva skal vi måle?
Hvordan skal vi måle?*

2.1. Data

Data er et vanskelig begrep å definere. Halvorsen (1993, s. 15) definerer data som innsamlet informasjon. En mer presis definisjon av data er:

«Representasjon av informasjon ordnet med tanke på kommunikasjon, tolkning eller behandling»
(Ordbok for dokumentasjon, 1993).

I statistikk er det *behandling* av data som er det sentrale. Ved behandling av data kan vi lage oversikter i tabeller og figurer, finne typiske data, og trekke konklusjoner fra de data vi har funnet.

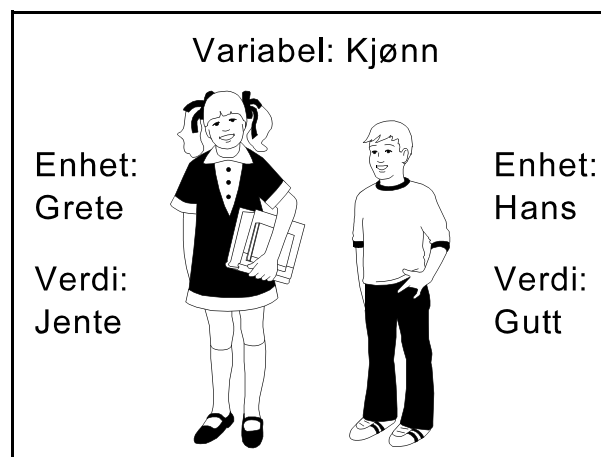
Data beskriver det materialet vi har undersøkt. For at vi skal kunne behandle data, må informasjonen være systematisk, og vi må ha klart definerte kategorier som gjør det mulig å beskrive materialet med tall.

Eksempel:

Vi kan undersøke kjønn og alder til lånerne i et bibliotek. Skal vi beskrive en gruppe lånerne, kan vi ha følgende data:

Hans, en gutt, er 10 år gammel
Grete, en jente, er 12 år gammel

Her er det 2 undersøkte individer. Vi kaller disse for *undersøkelsesenheter*. De to egenskapene som vi har kartlagt, kalles *variabler*. Variabelen «Kjønn» har to mulige *verdier*, nemlig «Mann» og «Kvinne». Variabelen «Alder» har som regel hele tall som verdier.



Figur 2 Enheter, variabel, verdier

2.2. Undersøkelsesenheter

En *undersøkelsesenhet* (kalles ofte bare enhet) eller et *undersøkelsesobjekt* er et individ, gjenstand, eller hendelse som vi vil beskrive eller kartlegge. Ved statistiske undersøkelser er det ikke hver enkelt undersøkelsesenhet som er interessant, men en samling av undersøkelsesenheter.

Eksempler:

Hvis vi vil si noe om lånerne i et bibliotek, er lånerne undersøkelsesenheter. Undersøker vi studentene i klassen, utgjør studentene enhetene. Tilsvarende er bøkene i biblioteket enheter når vi katalogiserer og klassifiserer dem. Hvis vi sammenligner utlån fra dag til dag, er det *dagene* som er enhetene.

Alle enheter som oppfyller et gitt kriterium kalles en **populasjon**. Alle studentene i 2. klasse ved bibliotek- og informasjonsstudiet studieåret 1998/99 utgjør en populasjon. Tilsvarende kan vi kartlegge populasjoner av forskjellige typer: Innbyggerne i en kommune, bøkene i et gitt bibliotek, dagene i et gitt år.

Hvis antall enheter er lite, er det praktisk mulig å undersøke alle aktuelle enheter. F. eks. kan det la seg gjøre å kartlegge pris for innkjøpte bøker i et bibliotek i en kort periode. Da er samlingen av de aktuelle bøkene populasjonen.

I andre tilfeller er det for dyrt, eller praktisk umulig, å undersøke en hel populasjon. En må da velge ut visse undersøkelsesenheter, et **utvalg**, ut fra gitte kriterier.

I de fleste beregninger og oversikter trenger vi å vite hvor mange enheter som er undersøkt. Som symbol for antall enheter i en undersøkelse bruker vi bokstaven **n**.

2.3. Variabler

Det vi vil kartlegge, måle eller undersøke, kalles **variabler**. Mens undersøkelsesenheterne er *enkelte objekter* vi skal undersøke, er variabelen *hva* ved de enkelte enhetene som skal undersøkes. En variabel er et kjennetegn eller en egenskap ved undersøkelsesenheterne.

Vi kan måle en eller flere variabler. Hvis vi vil vite lånernes alder, er det én variabel som blir undersøkt. Men det kan være praktisk å måle andre variabler samtidig: Kjønn, sosial klasse, besøkshyppighet i biblioteket osv.

Undersøker vi bøker, kan de aktuelle variablene være utlånsfrekvens, fagområde, innkjøpsår, slitasje.

Katalogisering av bøker kan betraktes som en kartlegging av disse variablene: Forfatter, tittel, utgiver, sidetall, høyde, ISBN osv.

2.4. Verdier

Variabler er egenskaper, men de konkrete data som brukes til å beskrive egenskapene hos hver enkelt enhet, kalles **verdier**.

Eksempler:

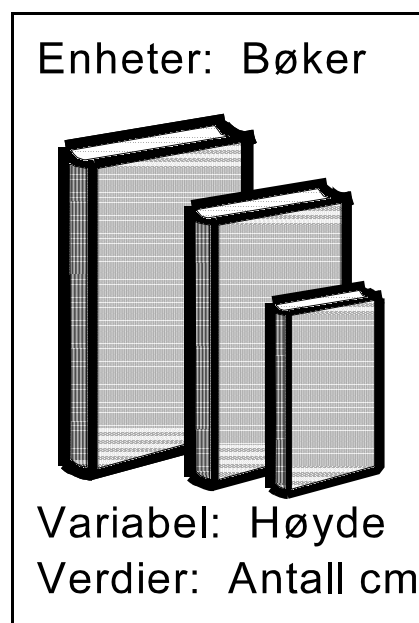
Lånetid (variabel) måles i antall dager (verdier).

Bokhøyde (variabel) måles i antall cm. (verdier).

Alder måles i antall år.

Variabelen «Kjønn» kan anta to verdier: mann, kvinne.

Ofte er verdiene tall, men som vi ser av siste eksempel, behøver det ikke være tilfelle.



Figur 3 Enheter, variabel, verdier

Det må være minst to verdier for at variabel-begrepet skal ha noen mening. Med bare én verdi er det ikke noe som varierer!

Eksempel:

La oss tenke oss at vi måler et bokmagasins temperatur ved gitte tidspunkter for å undersøke om temperaturen er stabil. 9 målinger kan gi resultatene:
16°, 22°, 20°, 23°, 17°, 20°, 17°, 20°, 25°.

Her ser vi at samme verdi opptrer flere ganger, men at vi bare har observert noen få av alle mulige verdier. De teoretiske verdier er alle mulige temperaturer innen visse grenser.

Det vil ofte lønne seg å skille mellom de **teoretiske verdiene** (de verdiene en variabel kan anta) og de **observerte verdiene**. Hvis vi f.eks. kartlegger alderen (variabel) til bibliotekarstudenter (enheter) i antall år (verdier), vil verdiene teoretisk kunne ligge i et nokså stort intervall. I praksis vil de observerte verdiene på langt nær fylle hele dette intervallet.

Symbolet **x** brukes ofte for å betegne verdier. Ønsker vi å skille, kan vi bruke **x** for de observerte verdier og **v** for de teoretiske.

Enheter, variabler og verdier er vanskelige begreper, men hvis en blander blir de statistiske oversiktene lett meningsløse. En praktisk huskeregel kan knyttes til følgende uttrykk (legg merke til genitivs-s'en knyttet til enhetene):

Enheterens variabel (abstrakt) måles i *verdier* (konkret).

Eksempler:

Brukernes besøksfrekvens i antall besøk pr. år skal kartlegges.
Enheter: brukerne. Variabel: besøksfrekvens. Verdier: hele tall.

Bøkenes høyde skal måles.
Enheter: bøkene. Variabel: høyde i cm. Verdier: Tall.

2.5. Verdier og klasser

De verdiene som en variabel kan anta, bør utelukke hverandre gjensidig, og være uttømmende, dvs. dekke alle tenkelige muligheter. Verdiene deler da enhetene inn i **klasser**. Ordet «klasser» kjenner vi også fra klassifikasjon av bøker, der hvert dokument skal tilordnes ett og bare ett klassenr. (f.eks. i DDK). Hvis verdiene ikke utelukker hverandre gjensidig, kan grunnen ofte være at man vil kartlegge *flere* variabler.

Eksempler:

Variabelen kjønn, med de to verdiene «Mann» og «Kvinne», deler personer inn i to klasser.

Variabelen alder, hvor alderen oppgis i antall år med klare avrundingsregler, deler også personer inn i klasser. Hvis derimot alderen oppgis som: 0-5 år, 5-10 år, 10-15 år etc., er det ikke entydighet, og

derfor ikke klasser. 10-åringene havner i 2 klasser. Problemet kan løses ved å la klassene være «fra og med 0 til 5 år»; «f. o. m. 5 til 10 år» osv.

Da bibliotekarstudentene på statistikk-kurset 1989 ble spurt om grunnen til at de begynte på dette studiet, ble det oppgitt flere alternativer som de skulle krysse av. Variabel var «Grunn til å begynne på bibliotekarstudiet», men her delte verdiene ikke inn i klasser, siden de fleste studentene krysset av flere svaralternativer.

2.6. Frekvenser

En verdis *frekvens* forteller oss antall ganger den opptrer i et materiale. Det kan være at vi vil vite antall lånere som er menn, eller hvor mange bøker som blir lånt ut 4 ganger i løpet av et år. Som symbol for frekvenser bruker vi vanligvis **f**.

Eksempel:

Vi går tilbake til magasinet med temperaturene: 16°, 22°, 20°, 23°, 17°, 20°, 17°, 20°, 25°.

Temperaturen 16° har frekvensen 1, siden vi finner den 1 gang i vårt tallmateriale. Verdien 20° har frekvensen 3.

Hvorfor valgte du bibliotekarstudiet?

- Interessert i bøker
- Yrkesrettet utdanning
- Gode jobbmuligheter
- Ønsker å arbeide med informasjon
- Liker å arbeide med mennesker
- Liker ikke å arbeide med mennesker
- Interesse for orden og systematikk
- Ønsker et stille og rolig arbeidssted
- Lyst til å bli bibliotekar
- Kom ikke inn andre steder

Figur 4 Svarkategorier som *ikke* utgjør klasser



Figur 5 Hvilken variabel blir undersøkt her?

3. Målenivå

Hvilke regneoperasjoner kan vi utføre på data vi har samlet inn?

Verdiene som vi får når vi har kartlagt en variabel, er av forskjellig type. Ofte er verdier tall, som vi kan regne på. F.eks. kan vi legge sammen 21 år og 23 år, og finne gjennomsnittlig alder. Derimot kan vi ikke legge sammen verdien «mann» og verdien «kvinne».

3.1. Nominalskala

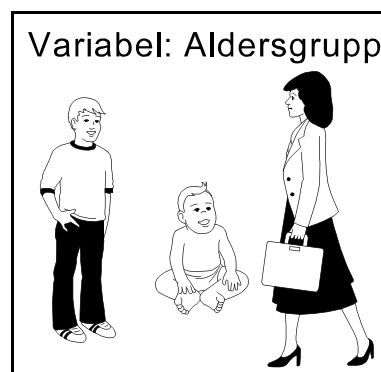
Så sant verdiene utgjør klasser, dvs. verdiene er *uttømmende* og *gjensidig utelukkende*, er kravet til en *nominalskala* oppfylt. Dvs. at hver enhet skal ha én og bare én verdi for vedkommende variabel.

Ved en nominalskala er verdiene vanligvis ikke tall, men gjensidig utelukkende kategorier. Likevel kan kategoriene være *kodet* som tall, og disse tallene vil fremdeles utgjøre en nominalskala (f. eks. i et numerisk klassifikasjons-skjema).

Det kravet som stilles til nominalskala er med andre ord et krav om entydighet. Vi kaller skalaen for nominalskala bare dersom kravene til neste målenivå *ikke* er oppfylt:

3.2. Ordinalskala

Hvis verdiene kan rangordnes, dvs. at det eksisterer en naturlig rekkefølge, kalles skalaen en *ordinalskala*. Eksempel på ordinalskala er variabelen «aldersgruppe», med verdiene «barn», «ungdom», «voksen», «eldre». Rekkefølgen mellom verdiene er ikke tilfeldig, men følger livsløpet.

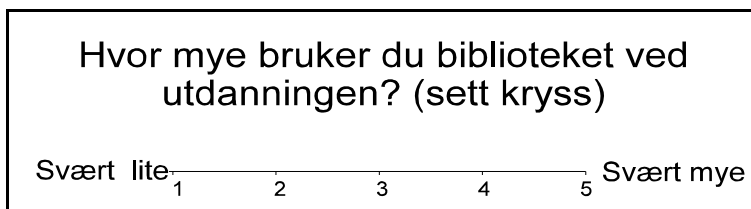


Figur 6 Ordinalskala

3.3. Intervallskala

Neste målenivå er *intervallskala*: En intervallskala er en skala der avstanden (intervallet) mellom verdiene lar seg måle, og gir mening. Samme avstand skal kunne tolkes som det samme overalt.

Dette gjelder de fleste skalaer der en benytter tall-verdier. F.eks. alder (vi kan regne aldersforskjell), høyde, utlånstid. Hvis en derimot har en skala der en skal krysse av ut fra subjektivt skjønn hvor enig en er i et utsagn, er det ikke sikkert at det er like langt mellom verdiene «1» og «2» som mellom «3» og «4» for en person.



Figur 7 Spørreskjema i metodekurset 1992

Også dette er en ordinalskala

3.4. Forholdstalls-skala

Et enda sterkere krav stilles til *forholdstalls-skala* (kalles også ratio-skala). Forholdstall betyr at størrelser kan sammenlignes, og at det er mulig å angi forholdet mellom dem. En person på 24 år er dobbelt så gammel som en på 12 år. Tilsvarende er en 24 cm høy bok halvannen gang så høy som en på 16 cm.

Likevel kan en ikke bestandig dividere tall-verdier med hverandre med fornuft. Forutsetningen for meningsfull divisjon er et meningsfullt 0-punkt.

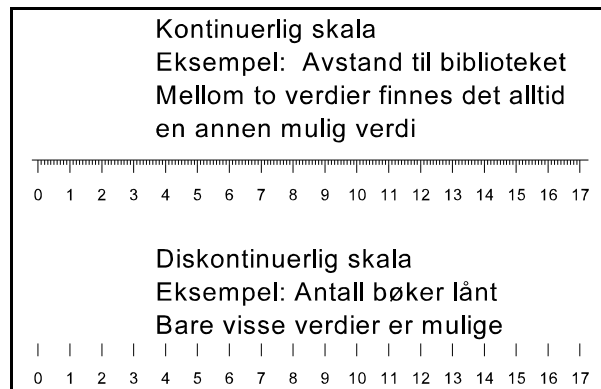
Et eksempel på en skala som *ikke* er forholdstalls-skala, er temperatur. Selv om 2 er det dobbelte av 1, er det meningsløst å si at det er dobbelt så varmt når det er 2° som når det er 1°. Vårt 0-punkt, vannets frysepunkt, er tilfeldig valgt. Fahrenheit-skalaen, der 0° C = 32° F, gir andre forholdstall mellom temperaturer. Absolutt temperatur målt etter Kelvin-skalaen (der 0 K er lavest tenkelig temperatur, - 273° C), er en forholdstalls-skala.

Tilsvarende til *nominalskala*, *ordinalskala* osv., sier vi også at variabelen måles på henholdsvis *nominalnivå*, *ordinalnivå*, *intervallnivå* og *forholdstallnivå*. Forholdstallnivået regnes som det *høyeste* nivået, fordi vi kan utføre flest regneoperasjoner på det nivået. Nominalnivået er det *laveste* nivået. Når vi skal oppgi en variabels målenivå, bruker vi det høyeste nivået variabelen tilfredsstiller. Disse fire nivåene kommer vi tilbake til i kapittel 7 og 8.

3.5. Kontinuerlig vs. diskontinuerlig skala

Noen variabler kan anta alle tenkelige verdier i et intervall. Vi kaller da variabelen for *kontinuerlig*. Eksempler: bokhøyde, temperatur, avstand til biblioteket. Også alder, hvis denne ikke oppgis avrundet, er i prinsippet kontinuerlig.

Selv om en variabel er kontinuerlig, oppgir en ofte verdiene avrundet, slik at de noterte verdier ikke utgjør en kontinuerlig skala. Kontinuitetsbegrepet er matematisk komplisert, og blir ikke presisert nærmere her.



Figur 8 Kontinuerlig og diskontinuerlig skala

En variabel som bare kan oppnå visse verdier, kalles *diskontinuerlig* eller *diskret*. Eksempler: Antall bøker, antall dager.

3.6. Dikotome variabler

Dersom en variabel bare kan anta to verdier, kalles den *dikotom*. Eksempel: Kjønn, med verdiene kvinne og mann, Bibliotekbruk, med verdiene bruker og ikke-bruker.

4. Datamatriser og regneark

Hvordan holde oversikt over de data vi har samlet inn?

4.1. Datamatrisen

Som regel undersøker vi flere enheter. Hvis vi samtidig undersøker flere variabler, må vi holde orden på verdiene, gjerne i tabell-form. Nå er det vanlig å bruke regneark (program for personlig datamaskin), men *datamatriser* ble brukt lenge før slike programmer var tilgjengelige.

Vanligvis settes datamatriser opp slik at linjene vil stå for de enkelte enhetene. Kolonnene brukes til variablene. I de enkelte rutene (cellene) lagrer vi opplysningene om de enkelte verdiene som tilsvarer enhetenes egenskaper.

Eksempel:

I **Tabell II** er det gitt en oversikt over kjønn og alder til bibliotekarstudentene som var til stede på en gitt forelesning.

De to første linjene er brukt for overskrifter. I de neste linjene er opplysninger om de enkelte studentene. I første kolonne er det identifikasjon av studentene (her som nr., men navn, kode e.l. kan brukes). De neste kolonnene brukes til variablene. For student nr. 3 kan vi lese av verdiene i linjen, og får vite at det er en kvinne på 25 år.

Tilsvarende kunne vi tenke oss registrert mange slags data, f.eks. aktivitetene i biblioteket fra dag til dag. Her vil da de enkelte dager være undersøkelsesenheter (datoer plasseres i 1. kolonne), mens variablene vil være f.eks. utlån, referansespørsmål, fjernlån, databasesøkinger osv.

De enkelte oppnådde verdiene til hver enkelt enhet kalles ofte for *råskårer*, dvs. de er ikke gruppert, omformet eller satt opp i tabell-form.

Datamatrisen vil som regel være et internt arbeidsredskap for å arbeide med data før materialet blir lagt fram i en mer oversiktlig form. Likevel må råskårene inn på en ryddig måte, før de blir bearbeidet.

4.2. Regneark

I en undersøkelse er det som regel mange enheter og flere variabler som blir undersøkt. Den enkleste databehandlingen skjer ved å legge data inn i et såkalt *regneark* (engelsk: Spreadsheet).

Tabell II Fordeling av menn og kvinner i en 1. klasse bibliotekarstudenter

1. klasse		
Nr.	Alder	Kjønn
1	34	M
2	32	K
3	25	K
4	20	K
5	22	K
6	19	M
7	20	K
8	20	K
9	21	K
10	19	K
11	21	K
12	26	M
13	27	K
14	31	K
15	23	K
16	22	M
17	20	K
18	46	K
19	18	K
20	34	K
21	31	K
22	19	K

Regneark kan forstås som ruteark på skjermen. Rutene, som kalles *celler*, blir bestemt ved hvilken rad og hvilken kolonne de ligger i. Raden symboliseres gjerne ved et tall, kolonnen ved en bokstav. Celle B3 er f.eks. cellen som bestemmes av 2. kolonne (siden B er andre bokstav i alfabetet), og av 3. rad.

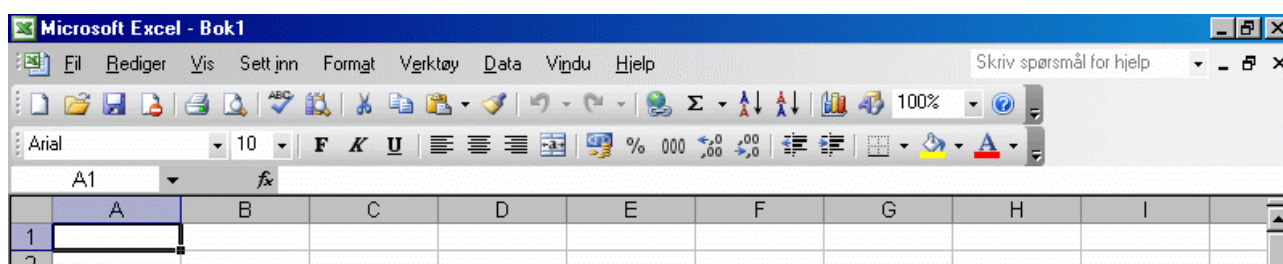
4.3. EXCEL

Det finnes mange regneark på markedet, men EXCEL later til å være det mest utbredte regnearket for PC. I vår utdannings nettverk finnes EXCEL (EXCEL 2003) i Windows XP under startknappen, deretter «Programmer» «Microsoft Office». En tom arbeidsbok, som består av 3 regneark, kommer fram klart til bruk.



Figur 9
Ikone for
EXCEL

Når EXCEL startes, ser toppen av skjermen omtrent slik ut:



Figur 10 Excel (nytt regneark)

Som i andre regneark betyr f.eks. B4 eller b4 cellen i 2. kolonne og 4 rad (store og små kolonnebokstaver har samme betydning). En sammenhengende rekke celler eller et rektangulært område av celler angis ved kolon mellom øverste venstre celle og nederste høyre celle. F.eks. betyr **a3:b5** området som består av cellene a3, a4, a5, b3, b4, b5. Tilsvarende betyr **c2:c15** cellene c2, c3, ... t.o.m. celle c15.

I cellene kan vi sette inn f.eks. tall, tekst eller formler. Markøren flyttes i regnearket ved hjelp av piltaster eller mus. Trykker du <Enter> etter inntasting, kommer innholdet inn i cellen, og markøren flytter seg nedover, mens <Tab> flytter markøren til høyre.

Formler begynner med likhetstegn (=). Enkle regnestykker utføres ved * for multiplikasjon, / for divisjon. Ved divisjon og multiplikasjon må vi bruke parenteser rundt summer eller differanser som skal samles, f. eks. skriver vi $\frac{1 + 2}{16 - 7}$ som **=(1+2)/(16-7)**

Venstre musetast brukes til å *merke* elementer (celler, områder, rader, kolonner, diagrammer osv.) i regnearket. Et område merkes ved å holde musetasten nede, deretter trekke musen gjennom området til de aktuelle cellene er merket. En kolonne merkes ved å trykke med venstre musetast på kolonnebokstaven, mens en rad merkes ved å trykke på linjenummeret. Et ark merkes ved å trykke på feltet over radnummer 1 og til for venstre kolonnebokstav A. Formatering, f.eks. midtstilling av innholdet i cellene, vil gjelde de cellene som er merket. Det er også mulig å merke områder som ikke henger sammen, ved å merke først et element (venstre musetast), og deretter *holde kontrolltasten nede mens de neste elementene merkes med venstre musetast*.

Dobbelklikk vil *aktivere* elementene, og innholdet kan endres direkte.

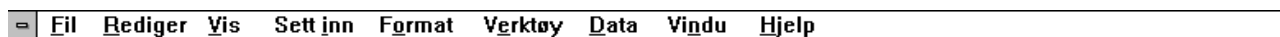
Trykk med høyre musetast får fram en hurtigmeny. Innholdet i denne menyen vil variere etter hva som er merket, og kan f.eks. omfatte format-type, kopiering, kolonnebredde o.l.

Sette inn kolonne eller rad, slette kolonne eller rad:

Gjennom menyen Sett inn kan du sette inn en rad over den cellen som er merket, eller en kolonne til venstre for cellen. Slette kolonne eller rad er enklest hvis kolonnen eller raden er merket (klikk på bokstav eller nummer). Høyre musetast vil blant sine valg også gi muligheten for å sette inn og slette rader eller kolonner.

4.4. Knapperader i EXCEL

Valgmulighetene for visning av knapperader ligger under menyen Vis på menylinja øverst i regnearket. Skulle knapperadene bli borte, er det lett å få dem fram igjen:



Figur 11 Menylinje

Formellinje: Her har du muligheten for å redigere formler og inntastede verdier. Dette valget anbefales sterkt. Til venstre ligger en *navneboks* der navnet på den merkede cellen kommer fram. I navneboksen kan vi også skrive inn navn på en celle eller et område.

	A1				
	A	B	C	D	E
1	20				
2					

Figur 12 Formellinje viser formelen for cellen

Statuslinje: Linje nederst som gir mer opplysning om knapper du peker på, hva som er ditt neste skritt osv.

Verktøylinjer: Her vil det lønne seg å vise *Standard* og *Formatering*.



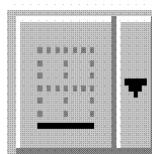
Figur 13 Viktig verktøylinje: Standard



Figur 14 Verktøylinje for formatering

4.5. Formater i EXCEL

Det finnes spesielle *formater* (innstillinger) for hvilken type informasjon som ligger i en celle, og hvordan innholdet vises fram (justering, utheving, kanter osv.). De vanligste formatene er tekst, tall, formler. I tillegg finnes f.eks. dato, tid, vitenskapelig, valuta, prosent.



Figur 15
Kanter



Figur 16 Boksen for valg av kantlinjer kan trekkes ut i regnearket

Hvis du ønsker et spesielt format for visse celler, kan det velges med høyre musetast. Det kan gjelde en enkelt celle eller et område (merk området først), kolonne eller rad (merk kolonne-bokstaven eller rad-nummeret). Du kan også velge menyen **Format** øverst i regnearket, hvor du har et valg for formatering av celler. Det kan noen ganger være nødvendig å velge format i celler, hvis du har skrevet inn noe som blir oppfattet som et spesielt format (særlig hvis regnearket har oppfattet et uttrykk som dato eller tid, er det vanligvis umulig å få fram tall uten å formatere cellen).

Hvis formatet for enkelte celler er uhensiktsmessig, kan du bruke formateringspenselen som kopierer format fra en celle nedover eller bortover.

Bredden på kolonnene og høyden på radene kan lett justeres ved å gripe fast i høyre kant på kolonne-bokstaven eller nedre kant på rad-nummeret. Et dobbelt-klikk på disse kantene vil gi beste tilpasning av bredde eller høyde.



Figur 17 Midtstilling på tvers av kolonner



Figur 18
Pensel for formatering

4.6. Inntasting av datamatrikse i EXCEL

Eksempel: Datamatriksen i **Tabell II**.

Overskriften skrives inn i celle A1. Mens vi skriver teksten, ser vi det vi skriver på formellinja. Kolonne-overskriftene kan legges inn i linje 2 (celle A2, B2 og C2). En kan så legge inn nr. og verdier i de riktige cellene.

Som standard plasseres tekst til venstre i cellene, mens tall plasseres til høyre. Ønskes en annen justering, finnes det ikoner for justering. De cellene, radene eller kolonnene som skal justeres merkes med venstre musetast. Eventuelt kan en merke hele regnearket (trykke venstre musetast i skjæringen mellom kolonnebokstaver og

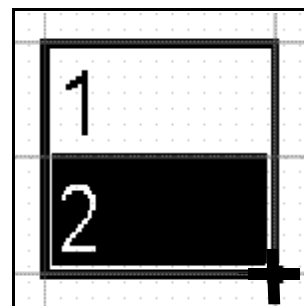
	A	B	C	D
1	1. Klasse			
2	Nr	Alder	Kjønn	
3	1	34	M	
4	2	32	K	
5	3	25	K	
6	4	20	K	
7	5	22	K	
8				

Figur 19 Datamatriksen etter inntasting av celle C7. Markøren står nå i celle C8

radnumre), og deretter f.eks. bruke formateringsikonet for midtstilling.

Snarvei for utfylling av tallrekke:

I området a3:a24 skal vi ha numrene 1 til 22. Det er nok å skrive inn de to første tallene (1 og 2). Deretter merkes disse to cellene ved å trykke inn venstre musetast i a3, trekke markøren til celle a4, og slippe musetasten. Flytter vi nå markøren til nederste høyre hjørne i celle a4 kommer det et «fyllhåndtak» (lite kors) til syne. Dette korset kan trekkes videre med venstre musetast ned gjennom det området som skal fylles ut. Deretter slippes musetasten, og området er utfylt. Området er fremdeles merket, vi kan tilpasse området med fyllhåndtaket på ny hvis vi har fylt inn for mange eller få celler. Flytter vi markøren til et annet sted på regnearket, og klikker (venstre musetast), forsvinner merkingen av området.



Figur 20 Fyllhåndtak nederst til høyre

Regnearket lagres gjennom Fil (menylinja), deretter Lagre som. Alternativt (for de fleste operasjonene i EXCEL finnes flere alternativer: Trykk ikonet på verktøylinja som forestiller en diskett. Vi gir arbeidsboka navnet «alder», filnavnet blir da «alder.xls». Alle EXCEL-filer får etternavnet «xls», dersom ikke annet etternavn oppgis. Videre bearbeiding av regnearket kommer vi tilbake til i seinere kapitler.

5. Datapresentasjon: Tabeller

Hvordan presentere data på en oversiktlig måte?

5.1. Frekvenstabeller

Frekvenstabeller viser hvor mange undersøkelsesenheter (f) som er tilordnet hver verdi (x) på variabelen. Det gir en langt bedre oversikt å se direkte hvor mange 18-, 19-, 20-åringer osv. det er enn å gi opp hver enkelt persons alder, slik det er satt opp i **Tabell II**.

Frekvensene bør summeres. Det gir bedre oversikt!

I frekvenstabeller oppgir vi gjerne verdiene på hver sin linje, med frekvensene til høyre. I **Tabell III** er satt opp en enkel frekvenstabell som viser antall menn og kvinner i **Tabell II**, side 12.

I **Tabell IV** er det tilsvarende satt opp en frekvenstabell over temperaturer i bokmagasinet i eksemplet på s. 8.

Tabell III Fordeling av kvinner og menn

Kjønn (x)	Antall (f)
Mann	4
Kvinne	18
Sum (n)	22

Tabell IV
Fordeling av temperatur
9 forskjellige dager

Temperatur (x)	Frekvens (f)
16°	1
17°	2
20°	3
22°	1
23°	1
25°	1
Sum	9

5.2. Relative (prosentvise) tabeller

Når vi leser tabeller, særlig hvis det er mange, eller store, tall, er det vanskelig å sammenligne. Vi har lettere for å holde oversikt hvis vi får antallet oppgitt i prosent av alle enhetene.

Definisjon:

En **relativ tabell** viser frekvensene som er knyttet til de enkelte verdier som brøkdeler (andel) av det samlede antall enheter.

Som regel angir vi andeler i %. Vi kaller da tabellen en prosentvis tabell.

Regneprosedyre:

Andelen finnes ved å dividere frekvensen (f) på antallet enheter (n)

Prosent-tallet utregnes ved å multiplisere andelen med 100 %: $\frac{f}{n} \cdot 100\%$

Eksempel på prosentregning:

Av 30 elever er 15 gutter. Dette er halvparten, eller mer matematisk: 0,5. Vi multipliserer med 100 og får en andel på 50 %.

Av de 9 dagene i **Tabell IV** har 2 temperaturen 17°. Dette blir

$$\frac{f}{n} \cdot 100\% = \frac{2}{9} \cdot 100\% \approx 22\%$$

Kommentar: Tegnet « \approx » betyr «omtrent lik». Prosent-tall bør sjelden oppgis med desimaler (særlig når total-antallet er så lite som her).

Prosentvise tabeller gir frekvensene oppgitt i % av totalsummen. En kan både oppgi absolutte og relative frekvenser (**Tabell V**), eller bare de relative (**Tabell VI**).

Tabell V Fordeling av kvinner og menn.
Absolutte og relative frekvenser

Kjønnsfordeling		
Kjønn (x)	Antall (f)	Andel (%)
Menn	4	18 %
Kvinner	18	82 %
Sum	22	100 %

Tabell VI Fordeling av kvinner og menn. Relative frekvenser

Kjønnsfordeling	
Kjønn (x)	Andel (%)
Menn	18 %
Kvinner	82 %
Sum	100 %

5.3. Gruppering av data

Ofte vil en frekvenstabell være så stor at en mister oversikten. I **Tabell VII** er det satt opp en frekvenstabell over aldersfordelingen i **Tabell II** s. 12. Tabellen er «komprimert», slik at bare de oppnådde verdiene vises. Alle observerte verdier er med, men det er lett å drukne i detaljer.

Tabellen blir lettere å lese hvis vi *grupperer* data. Vi oppgir da ikke alle oppnådde verdier, men intervaller som verdiene ligger innenfor.

Gruppering av data gir som nevnt økt oversiktighet, men mindre informasjon. Spørsmålet blir: Hvordan skal en gruppere? Som regel vil ca. 10-15 klasser gi god informasjon, men hvis antallet enheter er lite, kan en gjerne vurdere å bruke færre klasser. En må angi hvor hvert intervall begynner og slutter. Det er vanlig å gruppere slik at hvert intervall har bredde delelig med 2, 3, 5 (evt. 10, 20, 30, 50 etc.).

I **Tabell VIII**, **Tabell VIII** er aldersfordelingen i **Tabell VII** satt opp i en gruppert frekvenstabell. Her er 5 og 5 årsklasser gruppert sammen. I dette tilfellet er også tomme grupper tatt med for å lette oversikten.

Tabell VII Aldersfordeling

1. klasse	
Alder (x)	Antall (f)
18	1
19	3
20	4
21	2
22	2
23	1
25	1
26	1
27	1
31	2
32	1
34	2
46	1
Sum	22

Legg merke til at for alder, avrunder vi annerledes enn ved andre tall. Vi avrunder *nedover*, slik at du er 19 år helt til den dagen du fyller 20. Det betyr at «16-20» i tabellen betyr «fra og med 16 til 21»!

Tabell VIII Frekvensfordeling
Grupperte data

1. klasse	
Alder (x)	Antall (f)
16-20	8
21-25	6
26-30	2
31-35	5
36-40	0
41-45	0
46-50	1
Sum	22

5.4. Funksjoner i EXCEL

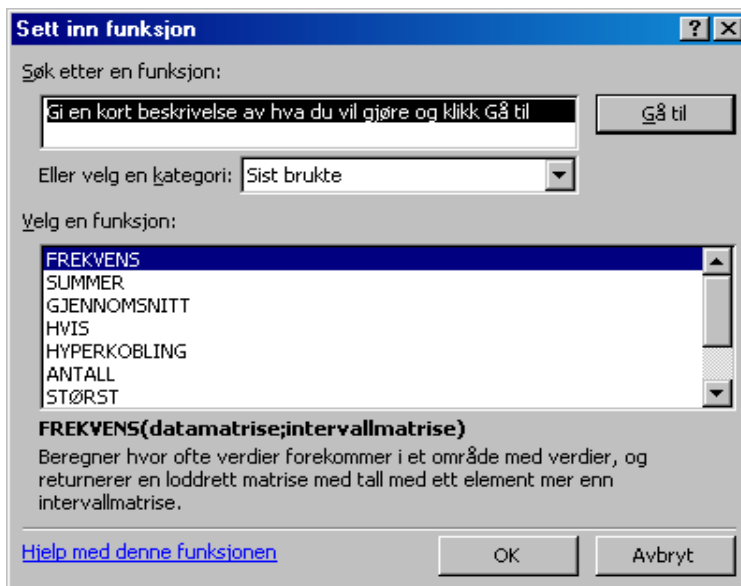
Som de fleste andre regneark, har EXCEL visse ferdigdefinerte funksjoner. Funksjonene kan være matematiske, statistiske, økonomiske, logiske, knyttet til tekst o.l. Hvis en ikke kjenner funksjonene, gir funksjonsveiviseren (ikonet f_x ved siden av formellinja) hjelp.



Figur 22
Funksjonsveiviser

Eksempel:

Vi vil finne kvadratroten av innholdet i celle b9. Resultatet skal stå i celle c9, og cellen merkes (et klikk). Vi klikker funksjonsveiviseren, velger matematiske funksjoner, funksjonen ROT (dobbelklikk eller trykk OK), og skriver b9 i stedet for tallet vi skal trekke ut kvadratroten av. Når vi så trykker **Fullført**, kommer resultatet inn i c9, mens formelen =ROT(B9) blir stående på formellinja.



Figur 21 Funksjonsveiviseren i EXCEL

Eksempel:

Viktigste funksjon i et regneark er å summere innholdet i celler inntil hverandre, enten i celle eller i kolonne. Her finnes det et eget ikon, Σ . Regnearket vil foreslå hvilket område som skal summeres. Hvis dette ikke passer, kan et annet område skrives inn i formelen, eller merkes med venstre musetast.



Kopiering av funksjoner:

Ofte er det aktuelt å bruke samme funksjon flere ganger. Dette vil særlig gjelde der samme funksjon brukes flere ganger under hverandre. Kopiering av celle med en formel vil tilpasse argumentene (de leddene som inngår i formelen) til den nye plasseringen. Dette betyr at hvis celle a4 står for summen av innholdet i celle a1 t.o.m. a3, ved formelen **=SUMMER(A1:A3)**, tilpasses formelen ved kopiering fra a4 til b4 til **=SUMMER(B1:B3)**. Det vil si at det brukes *relative referanser til celler* ved kopiering.

Hvis vi ønsker å angi at en adresse ikke skal endres, må vi bruke *absolutte referanser*. En adresse gjøres absolutt ved å sette inn dollartegn foran kolonnebokstav og radnummer, f.eks. **\$A\$1**. Vi kan evt. sette dollartegn foran bare den delen av referansen som ikke skal endres. Hvis vi skriver **A\$1** endres kolonnebokstav relativt, mens radnummer holdes fast.

Kopiering av formel rett nedover eller rett bortover gjøres enklest ved å markere cellen med formelen som skal kopieres, og deretter trekke fyllhåndtaket i riktig retning (jfr. s. 16)

5.5. EXCEL for frekvenstabeller

Når datamatriksen er tastet inn, kan vi lage frekvenstabeller. Arbeidsboka med datamatriksen må åpnes («alder.xls», jfr. **Figur 19** side 15).

Telle antall enheter for en variabel med gitt verdi:

Vi vil finne antall kvinner (kodet K i matrisen).

Vi skriver inn K i en celle (f. eks. f3), og flytter markøren til høyre for denne (celle g3). Gjennom funksjonsveiviseren finner vi den funksjonen «ANTALL.HVIS» i funksjonskategorien «Statistisk». Vi velger «fortsett» og oppgir området c3:c24 ved å skrive det inn eller merke det med mus (venstre musetast trykkes inn mens markøren står i celle c3, deretter holdes musetasten inne mens markøren trekkes til celle c24). Vilkåret kan skrives inn som k eller som a27 (fordi innholdet av denne cellen er bokstaven k).

Når vi trykker «fullført» kommer følgende formel inn i formel-linja:
=ANTALL.HVIS(C3:C24;"k") eller **=ANTALL.HVIS(C3:C24;F3)**

Tilsvarende kan vi finne antallet menn, med symbolet M i celle f4 og resultatet i celle g4.

*Kommentar: Hvis vi stadig skal henvise til samme område, kan det være nyttig å gi det et navn. Hvis vi f.eks. merker området c3:c24 med venstre musetast, kan vi gi dette navnet **kjønn** ved å skrive inn navnet i navneboksen til venstre på formel.. Med dette navnet blir formelen **=ANTALL.HVIS(kjønn;F3)***

I **Tabell III** s. 17 er det satt opp en enkel frekvenstabell, som viser frekvensen for to verdier. En vanlig type frekvenstabell grupperer verdiene i *intervaller* med tilsvarende frekvenser. De fleste regneark forutsetter at intervallenes *øvre grenser* legges inn i et område. Området med verdiene som skal grupperes kalles *datamatrikse*, mens området med grensene kalles *intervallmatrikse*. Området der

frekvenstabellen skal plasseres må ha én celle mer enn intervallmatrisa, for å få plass til eventuelt antall enheter med større verdi enn øverste grense.

Eksempel:

Vi skal lage frekvenstabell for alder, gruppert i intervallene 16-20, 21-25 osv. t.o.m. intervallet 46-50.

I området a33:a39 fyller vi inn de øvre grensene for intervallene. Enkleste metode er å først taste inn tallet 20 i celle a33, tallet 25 i celle a34. De to cellene merkes (mus), og fyllhåndtaket trekkes nedover og slippes i celle a39.

Deretter merker vi området der vi vil ha frekvensene (b33:b40). Igjen brukes venstre musetast. Vi leter nå opp funksjonen «Frekvens» blant de statistiske funksjonene i funksjonsveiviseren. Ved å dobbeltklikke eller be om «Neste», får vi oppfordring om å angi datamatrikse (kan angis ved å skrive inn b3:b24 eller ved å merke dette området) og intervallmatrise (angis ved område; a33:a39 skrives inn eller merkes). Vi må ikke trykke «Enter» mellom de to operasjonene, men velge de to inntastingsområdene med markøren. Når områdene er riktig angitt, trykker vi «OK», og formelen =FREKVENS(B3:B24;A33:A39) kommer inn i redigeringsområdet.

For å få frekvenstabellen opp må du ha markøren i **redigeringsområdet**, (mot toppen av skjermen) og klikke der med venstre musetast. Deretter trykker du <Ctrl>+<Skift>+<Enter> (kontroll- og skift-tasten holdes mens enter trykkes).

Lage kolonne som viser intervaller i frekvenstabell:

Noen ganger ønsker vi å få inn intervaller i en kolonne, ut fra at vi har intervallenes topp-punkter i en annen kolonne. Vi må ha en ledig kolonne, og merker cellen der vi vil ha et intervall. Vi bruker tekst-funksjonen **kjede.sammen**. Funksjonsveiviseren ber om tekst1, her kan vi sette cellenummeret hvor topp-punktet står - bredden på intervallet (eksempel: a33-4). Vi klikker på tekst2, og setter inn en bindestrek (evt med blankt tegn foran og etter). I tekst3 setter vi inn cellenummeret for topp-punktet, og har fullført formelen. Hvis topp-punktet er 20 og intervallet 4, bør vi få **16 - 20** som resultat. Cellen kan kopieres nedover på vanlig måte med fyllhåndtaket.

5.6. Kumulative tabeller

Kumulative tabeller viser antallet enheter som har en gitt verdi eller lavere verdi, de såkalte **kumulerte frekvenser** (kumulativ betyr «som fører til opphoping, stigning, økning, opphopende», kommer av det latinske ordet cumulus «haug, hop»). En verdis kumulerte frekvens er da summen av frekvensene til alle verdier lavere eller lik verdien.

Tabell IX Kumulativ tabell

1. klasse		
Alder (x)	Antall (f)	Kumulerte frekvenser (kf)
18	1	1
19	3	4
20	4	8
21	2	10
22	2	12
23	1	13
25	1	14
26	1	15
27	1	16
31	2	18
32	1	19
34	2	21
46	1	22
Sum	22	

Som symbol for kumulerte frekvenser bruker vi **kf**.

Når vi lager kumulative tabeller, tar vi utgangspunkt i frekvenstabellen. Vi starter med den laveste verdien. For den laveste verdien er den kumulerte frekvensen lik frekvensen for verdien. For neste verdi finnes den kumulerte frekvensen ved å legge verdiens frekvens til forrige kumulerte frekvens, og slik fortsetter vi til vi har kommet til høyeste verdi. Denne verdien skal være summen av antall enheter (**n**), fordi det er n enheter som har høyeste verdi eller lavere.

Eksempel:

I **Tabell IX** er satt opp kumulativ tabell for aldersfordelingen i **Tabell VII** på side 18.

5.7. EXCEL og kumulerte frekvenser

Eksempel:

Vi fortsetter med eksemplet i avsnitt **5.5**. Hvis alt er gått bra ved står frekvensene i celle b33:b39. I celle c33 kan vi skrive =c32+b33 (det samme som å skrive =b33, fordi celle c32 er tom). Deretter griper vi fast i fyllhåndtaket nederst til høyre i cellen, og fyller cellene nedover t.o.m. c39, der den kumulerte tabellen bør slutte. Formelen i cellen *kopieres*, men referansene til celler forskyves.

Igjen er det noe spesielt med alder: Det at regnearket har funnet 8 studenter som har verdien 20 eller lavere, betyr at enhetene som er telt opp ennå ikke har fylt 21 år!

6. Datapresentasjon: Grafiske fremstillinger

Hvordan presentere data på en visuelt oversiktlig måte?

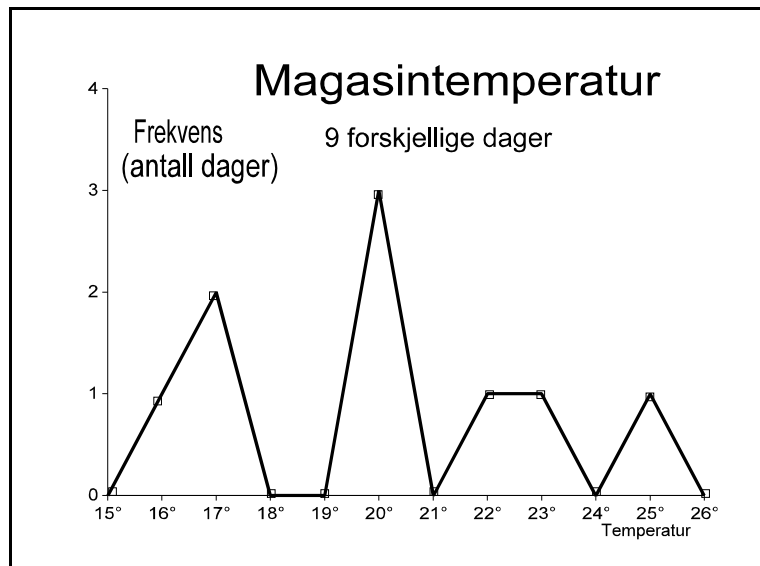
Selv om tabeller vil gi et godt uttrykk for hvordan verdier er fordelt, har de fleste av oss vanskelig for å få et bilde av fordelinger ved litt større tabeller.

For oversiktens skyld presenterer vi data i figurer. De vanligste typene er frekvenspolygon eller linjediagram, stolpediagram og kakediagram:

6.1. Frekvenspolygon

Frekvenspolygon (linjediagram) er en kurve som viser frekvensene for hver verdi. En rett strek trekkes mellom punktene. Kurven skal trekkes ned til 0 for de verdiene som ikke forekommer i materialet. Diagrammet tegnes ved å sette verdiene på den vannrette (horisontale) akse, x-aksen, og frekvensene på den loddrette (vertikale) akse, y-aksen.

Figur 24 viser et frekvenspolygon for temperaturene vist i **Tabell IV** side 17. Legg merke til at også her er *alle* verdier (temperaturer) i det aktuelle intervallet tegnet inn, også de som ikke er oppnådde verdier (f. eks. 18°, 19° osv.). Hvis vi ikke tok med disse, ville diagrammet bli misvisende, og avstanden mellom verdiene på figuren ville ikke gi mening. Det er vanlig å trekke kurven ned til x-aksen ($f = 0$) også for verdien *under* den laveste oppnådde og *over* den høyeste. Ellers blir figuren «hengende i lufta», og en får inntrykk av at den kan fortsette oppover.



Figur 24 Frekvenspolygon

6.2. Stolpediagram

Stolpediagram eller søylediagram angir frekvensene som søyler (stolper) som plasseres oppover fra de aktuelle verdiene. Høyden vil angi frekvensene.

Figur 25 viser et enkelt stolpediagram for temperaturene vist i **Tabell IV** side 17, dvs. samme fordeling som i **Figur 24**. Også her må en ta med verdier med frekvens 0, ellers blir skalaen fortegnet.

6.3. Histogram

Et *histogram* er et søylediagram hvor søylene er satt inntil hverandre, mens antallet (frekvensen) avleses av arealet på søylene. Det normale er at søylene er like brede. Da kan høyden brukes som frekvensangivelse.

Eksempel:

I **Figur 26** er gjengitt et histogram for den grupperte fordelingen over alder (**Tabell VIII, Tabell VIII** s. 19, 19). På x-aksen er her for oversiktens skyld tegnet opp grensene for de enkelte intervallene.

Hvis søylene ikke er like brede (f.eks. en aldersgruppering hvor klassene er 0-7 år, 8-17, 18-35 etc.), er det *arealet* og ikke høyden som gir uttrykk for frekvensene.

6.4. Sektordiagram (Kakediagram)

Et sektordiagram (kake-) er et relativt diagram, hvor andelen av totalsummen angis som delen av en sirkelskive.

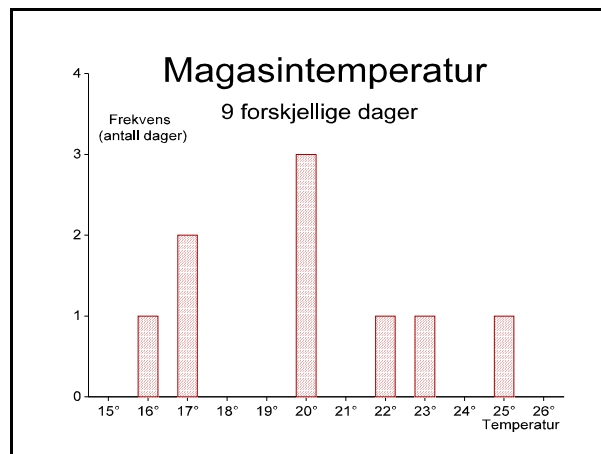
For å tegne et sektordiagram manuelt, må vi bruke en såkalt transportør (halvsirkelformet skive med gradtall langs randen. Vi må da multiplisere andelen med 360° for å finne gradtallet for et «kakestykke». De fleste bruker data-verktøy for å tegne slike figurer.

Eksempel:

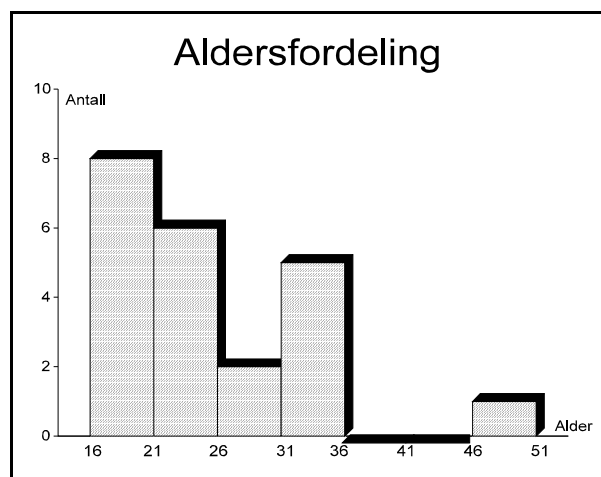
I **Figur 27** ser vi et sektordiagram som svarer til **Tabell III** s. 17.

6.5. Diagram og målenivå

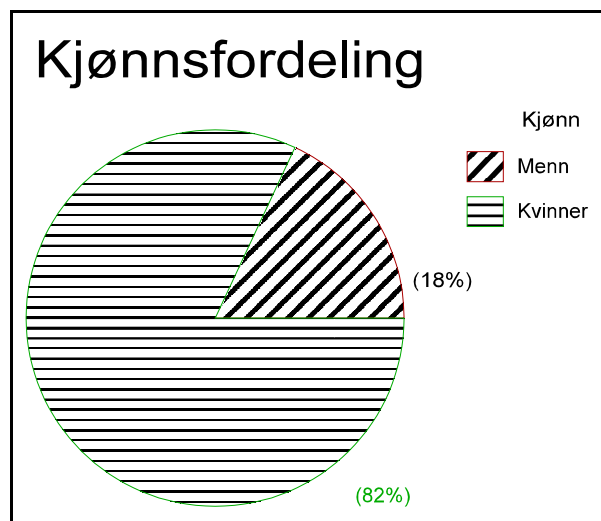
Det er ikke entydige regler for hva slags diagrammer vi bør benytte for de forskjellige former for målenivå. Følgende retningslinjer kan likevel benyttes med egen vurdering i tillegg:



Figur 25 Stolpediagram



Figur 26 Histogram



Figur 27 Kakediagram
En klasse bibliotekarstudenter

Nominalskala: Her kan kakediagram få frem et godt bilde av fordelingen. F.eks. vil fordeling av utgiftsposter ved et bibliotek gjerne fremstilles ved et kakediagram. Søylediagram kan også benyttes, men søylene bør ikke settes inntil hverandre.

Ordinalskala: For data på ordinalt nivå bør en bruke søylediagram. Heller ikke i dette tilfellet bør vi sette søylene inntil hverandre. Hvis søylene står inntil hverandre, får vi lettere inntrykk av at det er en veldefinert avstand mellom verdiene.

Intervallskala og forholdstallsskala: Her kan linjediagram gi et godt bilde av fordelingen. Også histogram vil gi god oversikt ved kontinuerlige variabler. Vi kan med fordel bruke søylediagram ved diskontinuerlige (diskrete) variabler for å få fram at det er sprang mellom verdiene. Ved grupperte verdier vil histogram gi bedre bilde enn søylediagram.

Tabell X Kumulativt fordeling over temperaturer

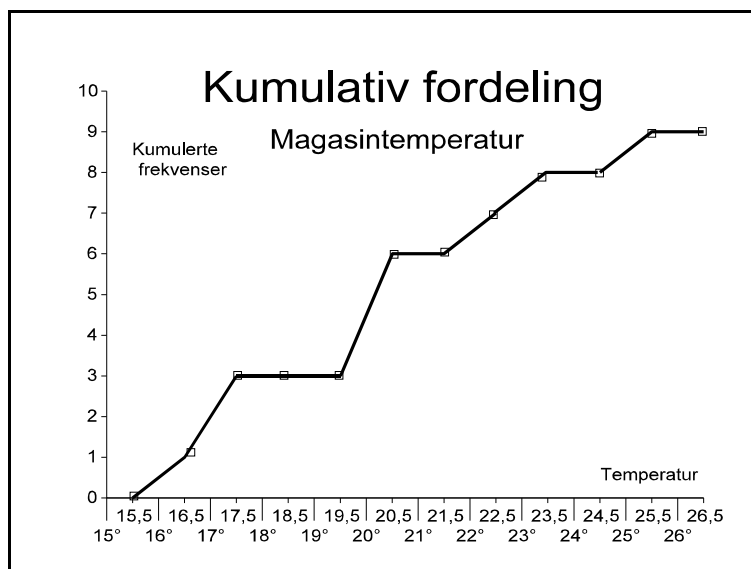
Temperatur (x)	Frekvens (f)	Kumulerte frekvenser (kf)
16°	1	1
17°	2	3
18°	0	3
19°	0	3
20°	3	6
21°	0	6
22°	1	7
23°	1	8
24°	0	8
25°	1	9
Sum	9	

6.6. Kumulativt diagram

Et *kumulativt diagram* er et diagram over de kumulerte frekvensene. Her kan en velge mellom søyle- og linje-diagram, men linje-diagram er vanligst.

Eksempel:

I **Tabell X** er satt opp en kumulativ tabell over temperaturene i **Tabell IV** side 17. Her er også satt inn alle temperaturer i det aktuelle intervallet med frekvens 0.



Figur 28 Kumulativt diagram over temperaturer

I **Figur 28** er det tegnet et kumulativt diagram som viser de kumulerte frekvensene. Legg spesielt merke til at her er verdiene satt til $0,5^\circ$ høyere enn det vi skulle vente. I stedet for 20° er det $20,5^\circ$ som har kumulert frekvens 6.

Årsaken til dette er at 20° tenkes som avrunding av temperaturer, dvs. alle temperaturer mellom $19,5^\circ$ og $20,5^\circ$. Ut fra denne avrunding er det 6 dager som har temperatur $20,5^\circ$ eller lavere, mens det blir noe ukorrekt å si at det er 6 dager som har temperatur 20° eller lavere.

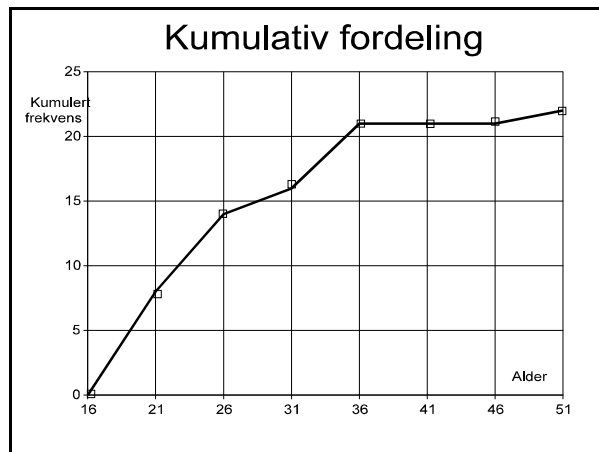
Ved grupperinger velges vanligvis *midtpunktene mellom* intervallene som verdier når vi skal tegne kumulative diagrammer. Når intervallene møtes, velges grenseverdien mellom intervallene.

Eksempel:

I **Figur 29** er tegnet et kumulativt diagram over aldersfordelingen for 1. klasse bibliotekarstudenter (tilsvarer den grupperte frekvensfordelingen i **Tabell IX**). Grensene her er bestemt slik at det går fram hvor mange som er *ynge* enn en gitt alder. Dette er et spesialtilfelle, fordi når alder oppgis, avrunder man ikke, men regner seg som 20 år helt til den dagen man fyller 21.

Frekvenspolygon med forskjellig type streker kan gi et godt bilde når vi skal sammenligne to fordelinger.

Hvis vi skal lage en god figur, lønner det seg ofte å la høyden være omtrent 2/3 av bredden.



Figur 29 Kumulativt diagram med grupperte verdier

6.7. Diagrammer i EXCEL

Siden dette kompendiet i utgangspunktet er skrevet før utdanningen hadde adgang til EXCEL, er PlanPerfect brukt for å lage de fleste diagrammene. PlanPerfect er et regneark-program for DOS. Tilsvarende diagrammer kan tegnes med andre regneark, med statistikk-pakker, eller med tegneprogrammer.

Diagrammer er lette å få fram i EXCEL. Området som skal inn i diagrammet, gjerne med overskriftene merkes (med mus). For å få fram diagrammet brukes så *diagramveiviseren*. Ikonet har form av et søylediagram.



Figur 30
Diagramveiviser

Diagramveiviseren har 4 trinn:

- Trinn 1 Diagramtype: her velges diagramtyper med undertyper. De vanligste typene er *linjediagram*, *stolpediagram*, *punktdiagram* og *sektordiagram* (kalles også *kakediagram*).
- Trinn 2 Kildedata for diagrammet: valg (eller bekreftelse) av dataområdet (cellereferanser som skal inn i figuren). Justering av hva som skal inn i diagrammet og hva som skal stå på x-aksen (kalles «etiketter for kategoriakse») velges gjennom fanen «Serie»
- Trinn 3 Diagramalternativer: Blant annet tittel på diagrammet, titler på aksene o.l.
- Trinn 4 Plassering av diagrammet: her gis muligheten for å plassere diagrammet i regnearket det bygger på, eller som eget ark. Ofte vil det øke oversiktligheten å plassere diagram som nye ark.

En stor styrke med diagramveiviseren er at resultatene av hvert valg vises ved et eksempel-diagram. Dessuten er hjelpefunksjonen tilgjengelig, og kan gi ytterligere forklaring.

Ut fra de foreliggende data kan vi f.eks. lage:

Sektordiagram:

Utgangspunktet er tabellen over antall kvinner og menn i klassen. Området f3:g4 merkes, og med diagramveiviseren velges 3-dimensjonalt sektordiagram. Den typen som viser prosent-andelen virker god, og til slutt velger vi tittelen «Kjønnfordeling».

Når diagrammet er ferdig, kan det lett flyttes (venstre musetast holdes), redigeres (ved dobbeltklikk aktiveres diagrammet, og vi kan endre elementer i diagrammet, som å trekke ut en sektor osv.)

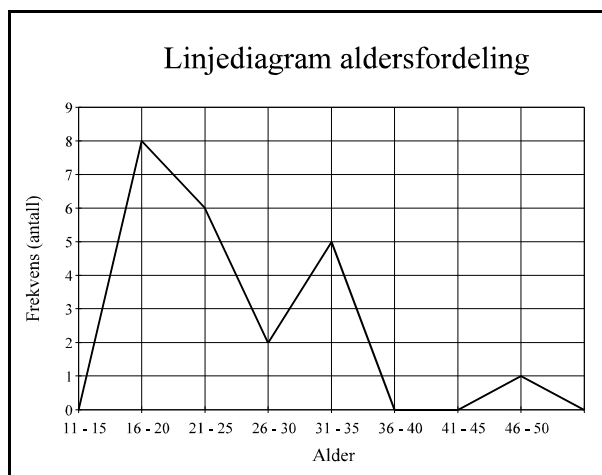
Søylediagram:

Frekvensfordelingen over alder i cellene b33:b39. Området a32:b39 merkes, deretter velges stolpe, med format hvor stolpene står inntil hverandre. Vi må angi at 1 kolonne brukes som kategoriakse-etiketter (ellers oppfattes de øverste grensene for intervallene som tall som skal framstilles som søyler). Vi kan gjerne sette inn litt tekst (tittel f.eks. «Aldersfordeling», langs x-aksen f.eks. «Øvre grense for intervallet», langs y-aksen «Antall».

Linjediagram:

Her kan vi gjerne også fremstille aldersfordelingen. For å få et godt diagram bør diagrammet ned til x-aksen, ellers får vi inntrykk av at fordelingen er ukjent utenfor de verdiene som oppgis. Derfor setter vi verdien 15 inn i celle a32, frekvensen 0 inn i celle b32. Området som danner utgangspunkt for diagrammet blir nå a32:b40, og vi kan velge type osv. videre.

Når diagrammet er ferdig, kan det redigeres: ved dobbeltklikk hvis det er et objekt i et regneark, eller direkte hvis det er satt inn som eget ark. Vi merker enkelte elementer (f.eks. akser, titler osv.) med enkeltklikk, og får fram en hurtigmeny som gir videre redigeringsmuligheter med høyre musetast.



Figur 31 Linjediagram laget i EXCEL

7. Sentraltendens

Hva er den typiske verdien i det materialet vi har samlet inn?

En fordeling har 2 svært viktige mål: Hvor er «midten», og hvilken spredning har fordelingen? Både «midtpunkt» og «spredning» er uklare begreper, og må defineres. *Sentraltendens* er det statistiske begrepet for «midtpunkt». Her finnes det tre vanlige mål:

7.1. Modus

Modus, som også kalles modalverdi eller typetall, er det enkleste målet for sentraltendens. Modus er definert som den verdien som har størst frekvens. Når vi ser på figuren over en fordeling, er modus den verdien hvor toppen ligger. I en tabell må vi finne den verdien som har høyest frekvens. Denne verdien er det «mest typiske tilfellet».

Eksempel:

I eksemplet over temperaturer kan vi avlese modus enten fra frekvenstabellen (**Tabell IV**) på s. 17, eller fra figurene over den samme fordelingen (**Figur 24** s. 23 og **Figur 25** s. 24). Den verdien som har høyest frekvens er 20°, med frekvens 3. Fordelingens modus er derfor 20°.

7.2. Aritmetisk gjennomsnitt

Det mest brukte mål for sentraltendens er *aritmetisk gjennomsnitt* eller kort og godt gjennomsnittet. Dette beregnes som summen av de enkelte observerte verdiene (måleresultatene) delt på antallet.

Formelen for aritmetisk gjennomsnitt er $m = \frac{\sum x}{n}$

I formelen betegner x de enkelte verdier og n antallet enheter. Symbolet for aritmetisk gjennomsnitt er m .

Tabell XI Utregning av gjennomsnitt ved summering av råskårer

Vi vil seinere støte på et annet symbol for aritmetisk gjennomsnitt, nemlig μ (liten gresk m, uttales «my»). Dette symbolet bruker vi for det aritmetiske gjennomsnittet i en populasjon, slik at vi kan skille mellom et utvalgs gjennomsnitt (m) og gjennomsnittet i populasjonen (μ) som utvalget er hentet fra.

Legg merke til at det er n antall x -er som skal legges sammen. For å gjøre formelen klarere at det er n verdier som summeres, kan formelen skrives slik:

$$m = \frac{\sum_{i=1}^n x_i}{n}$$

Uttrykket $\sum_{i=1}^n x_i$ leses «summen av x i fra 1 til n ».

Enhet nr.	Verdi (°C)
1	16
2	22
3	20
4	23
5	17
6	20
7	17
8	20
9	25
Σ	180

Symbolet Σ , er den bokstaven «Sigma» i det greske alfabetet, som betegner «stor S». Det er et summeringstegn, som betyr *legg sammen*. Vi skal legge sammen alle verdiene, fordi x er symbolet for verdier.

Eksempel:

I **Tabell XI** er satt opp mellomregningen for å regne gjennomsnittlig temperatur. At råskårene er brukt, betyr at de enkelte verdiene er summert, uten noen form for gruppering eller bearbeiding. Det aritmetiske gjennomsnittet blir:

$$m = \frac{\sum x}{n} = \frac{180^\circ}{9} = 20^\circ$$

Hvis vi har en tabell med frekvenser, må vi huske på at enkelte verdier er oppnådd av flere enheter. Vi kan derfor ikke legge verdiene i frekvenstabellen sammen uten videre. Hver verdi forekommer så mange ganger som frekvensen angir. Formelen blir da: $m = \frac{\sum f \cdot x}{n}$

I denne formelen er n antall enheter totalt. Det lønner seg å kontrollere at $n = \sum f$. Summen av frekvensene må nemlig omfatte alle enhetene.

Tabell XII Utregning av gjennomsnitt ved frekvenstabeller

Temperatur (x)	Frekvens (f)	f · x
16°	1	16
17°	2	34
20°	3	60
22°	1	22
23°	1	23
25°	1	25
Σ	9	180

Eksempel:

Ved frekvensfordeling av temperaturene bør resultatet bli det samme som om vi regner ut ved hver enkelt verdi. Vi ser at her blir:

$$m = \frac{\sum f \cdot x}{n} = \frac{180^\circ}{9} = 20^\circ$$

Resultatet stemmer! Vi kan evt. sette $n = \sum f = 9$ inn i formelen.

Hvis data er grupperte, regnes *midtpunktet* av intervallet som verdien i intervallet. Vi har gitt avkall på den nøyaktige informasjonen om hver enkelt enhets verdi. Det beste anslaget vi da kan få for verdien, er *midt* i intervallet, fordi denne verdien gjør feilen ved gruppering så liten som mulig.

Eksempel:

I **Tabell XIII** s. 30 er nødvendig forarbeid for å regne ut aritmetisk gjennomsnitt for den grupperte aldersfordelingen i **Tabell VIII**, **Tabell VIII** s. 19, 19.

Gjennomsnittet regnes ut slik:

$$m = \frac{\sum f \cdot x_0}{n} = \frac{551}{22} = 25,05 \approx \underline{25 \text{ (år)}}$$

Her er x erstattet med x_0 , som blir brukt som symbol for intervallets midtpunkt.

7.3. Median

Medianen er den verdien som deler en ordnet fordeling i to like store deler, slik at det er like mange observasjoner på hver side. For medianen bruker vi symbolet **md**.

For å kunne finne medianen, må verdiene ordnes i sin naturlige rekkefølge. Er det et like antall verdier (2, 4, 6, 8, ...), ligger medianen midt mellom de to midterste verdiene. Er det et ulike antall (1, 3, 5, 7, ...), blir det den midterste.

Har vi oppgitt verdier med frekvenser, er det ikke nok å regne slik. Vi må her merke oss at det er et visst antall av hver verdi, og derfor nummerere undersøkelsesenheterne fra den laveste til den høyeste. Medianen er den verdien som enhet nr. q_2 oppnår, der

$$q_2 = \frac{n + 1}{2}$$

Intuitivt vil en kanskje tro at den midterste enheten har nr. tilsvarende halvparten av enhetene. Grunnen til at det er at midterste enhet ligger midt mellom 1. enhet (nr. 1) og siste enhet (enhet nr. n).

Når medianen skal finnes, må enhetene ordnes (gjøres vanligvis i stigende rekkefølge, fra enheten med laveste verdi til enheten med høyest verdi).

Eksempel:

Vi har undersøkt 5 enheter, med verdiene 2, 10, 8, 7, 5. Vi ordner verdiene i stigende rekkefølge. Midterste enhetsnr. blir 3 (som er midt mellom enhet nr. 1 og enhet nr. 5); mens medianen blir verdien til enhet nr. 3, nemlig 7.

I **Tabell XV** er aldersopplysningene fra **Tabell II** s. 12 satt opp i ordnet rekkefølge, fra lavest til høyest verdi. Enhetene er nummerert fra 1 (laveste verdi) til 22 (høyeste verdi). Midterste enhet er nr.

Tabell XIII Utregning av gjennomsnitt ved grupperte data

1. klasse			
Alder (x)	Midtpunkt (x_0)	Antall (f)	$f \cdot x_0$
16-20	18	8	144
21-25	23	6	138
26-30	28	2	56
31-35	33	5	165
36-40	38	0	0
41-45	43	0	0
46-50	48	1	48
Σ		22	551

Tabell XIV Oppsett for utregning av median

			q_2 ↓		
Enhet nr.	1	2	3	4	5
Verdi	2	5	7	8	10
			↑ md		

$$q_2 = \frac{n + 1}{2} = \frac{22 + 1}{2} = \frac{23}{2} = 11,5$$

Siden dette ikke er et helt tall, men midt mellom 11 og 12, velger vi som median verdien midt mellom verdiene til enhet nr. 11 og enhet nr. 12. Siden begge verdiene er 22, blir medianen 22 år.

For utregning av median ved frekvensfordelinger vises til s. 37.

7.4. Skjeve fordelinger

Hvis en fordeling er symmetrisk, vil de tre målene for sentraltendens være like (med ett unntak: hvis det ikke er én klar topp, vil det ikke gjelde for modus). Men ved skjeve fordelinger (fordelinger som er brattere mot den ene siden enn mot den andre), blir målene forskjellige.

Eksempel:

I **Figur 32** er en (omtrentlig) inntekts-statistikk for bibliotekarer. Mange vil være grunn-plassert på det offentlige regulativet, og her er det en klar topp (modus). Gjennomsnittet vil bli mest påvirket av de høye lønningene, mens medianen vil ligge midt mellom.

Hva som er typisk, kommer an på øynene som ser, og hva sentralverdien skal brukes til. For en nyutdannet bibliotekar vil kanskje modus være mest interessant på kort sikt, median på litt lengre sikt (gir beskjed om lønnsnivået dersom det er like mange som tjener mere og mindre enn dette nivået). Arbeidsgiverforeningen bruker aritmetisk gjennomsnitt, som gir høyest verdi!

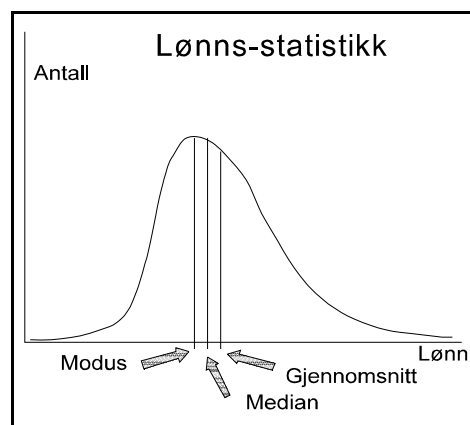
Modus brukes helst når vi eksplisitt sier at dette er verdien som «de fleste» enhetene oppnår. «De fleste» bibliotekarer i staten står i grunnstigen lønnsmessig.

Median brukes ofte ved skjeve fordelinger. Her vil ofte ekstremverdiene dra gjennomsnittet uforholdsmessig mye opp.

Gjennomsnitt er det mest brukte målet. Brukes når vi skal sammenligne fordelinger (her brukes også median), når vi skal bruke verdier til budsjettering o.l., og når vi skal regne videre på data.

Tabell XV Sortering av enheter for å finne median

1. klasse	
Nr.	Alder
1	18
2	19
3	19
4	19
5	20
6	20
7	20
8	20
9	21
10	21
11	22
12	22
13	23
14	25
15	26
16	27
17	31
18	31
19	32
20	34
21	34
22	46



Figur 32 En skjev fordeling

7.5. Sentraltendens og EXCEL

Alle de tre vanlige målene for sentraltendens har ferdigdefinerte formler, nemlig:

- =gjennomsnitt(område)
- =median(område)
- =modus(område),

der område vanligvis vil ha formen cellereferanse:cellereferanse, f. eks. b5:b26. Har du gitt et område navn, kan du bruke navnet.

Funksjonsveiviseren (via Statistiske funksjoner) kan gjerne brukes fra den cellen der resultatet skal ligge.

Siden EXCEL har funksjon for median, er det ikke nødvendig å sortere verdiene for å finne medianen. Likevel kan det gi oversikt over fordelingen å sortere verdier. Hvis du skal sortere et område i bare én kolonne, er det knapper for å sortere i stigende eller synkende rekkefølge. Er sorteringsoppgaven mer avansert, som å sortere et større område etter gitt kolonne eller gitt rad, gir Sorter under menyen Data en rekke valgmuligheter.

Husk at celler eller områder som er merket, kan gis navn i feltet til venstre i formel-linja (der celle-referansen vanligvis er angitt). Avslutt med <Enter> når navnet er skrevet inn!

Hvis en fordeling har flere topper, får du likevel bare én verdi for modus, nemlig den laveste av verdiene til topp-punktene.

Hvis gjennomsnittet skal beregnes på bakgrunn av en frekvenstabell, finnes det *ingen* ferdig definert formel. Vi må selv sette opp den formelen som skal brukes. Dette kan gjøres slik:

- Legge inn verdier og frekvenser i to kolonner
- Multiplisere verdi og tilhørende frekvens for første verdi, og kopiere nedover med fyllhåndtaket.
- Summere frekvenser, slik at vi finner $\sum f$; deretter summere produktet mellom frekvenser og verdier., så vi får $\sum f \cdot x$. Bruk formelen «=summer», eller autosummer-ikonet.
- Utføre den divisjonen som gir gjennomsnittet.

7.6. Sentraltendens og målenivå

Modus, som er verdien med tilhørende høyest frekvens, kan finnes i alle tilfeller. Det er nok at målingen er på nominalnivå. I kjønnsfordelingen over bibliotekarstudentene er modus verdien «kvinne», fordi det er flest kvinner ved utdanningen.

For å finne *medianen*, er det nødvendig å kunne sortere enhetene etter stigende verdi i forhold til den variabelen som er målt. Denne sorteringen krever ordinalnivå, og medianen kan derfor finnes ved dette og høyere målenivåer.

Gjennomsnittet derimot, forutsetter at en kan regne med verdiene. Verdiene må kunne adderes (legges sammen) og deles på antallet, slik at resultatet gir mening. Dette forutsetter at variabelen minst er målt på intervallnivå.

8. Spredning

Hvor forskjellige er de verdiene vi har funnet?

Helt tilsvarende som for gjennomsnitt, er det tre mål for spredning som er mye brukt. Hver av disse har sammenheng med et av målene for sentraltendens. Målene er:

1. Variasjonsbredde, som i likhet med modus er et mål som er lett å finne.
2. Standardavvik, som har sammenheng med gjennomsnitt
3. Kvartilavvik, hvor utregningsmetoden ligner på utregning av median

8.1. Variasjonsbredde

Variasjonsbredden i en fordeling er forskjellen mellom høyeste verdi og laveste verdi

Hvis vi kaller høyeste verdi x_{\max} og laveste verdi x_{\min} , blir formelen for variasjonsbredden:

$$x_{\max} - x_{\min}$$

Eksempel:

I vårt gjennomgangseksempel med temperaturer (side 8), er høyeste verdi $x_{\max} = 25^\circ$, laveste verdi

$x_{\min} = 16^\circ$, og variasjonsbredden blir

$$x_{\max} - x_{\min} = 25^\circ - 16^\circ = \underline{9^\circ}$$

8.2. Standardavvik

Standardavvik er et mål for spredningen rundt det aritmetiske gjennomsnittet. Standardavvik er definert ved en matematisk formel, nemlig

$$s = \sqrt{\frac{\sum (x - m)^2}{n}}$$

Symbolet for standardavvik er s .

En alternativ formel, som vi vil bruke her, er:

$$s = \sqrt{\frac{\sum (x - m)^2}{n - 1}}$$

Denne formelen gir en *større* verdi for standardavviket. Hvis n er stor, er forskjellen mellom de to målene relativt liten.

For å skille mellom de to formlene, kan en bruke symbolene s_n og s_{n-1} . I denne fremstillingen blir den *siste* formelen brukt (s_{n-1}), og vi nøyer oss med symbolet s .

Vi vil seinere støte på et annet symbol for standardavvik, nemlig σ (uttales «sigma», liten gresk s). Dette symbolet brukes om det standardavviket vi vil anta at en *populasjon* har, når vi slutter fra et utvalg til en populasjon.

Begrunnelse for å bruke s_{n-1} er at hvis en skal slutte fra standardavviket i et utvalg til standardavviket i populasjonen, σ , blir standardavviket for lite når vi bruker n . For lite, fordi en i utgangspunktet regner ut avvikene fra *utvalgets* gjennomsnitt (m), ikke fra populasjonens gjennomsnitt (μ). Slutningen blir derimot riktig hvis du benytter $(n-1)$. Beviset bygger på matematisk statistikk.

Hvis det standardavviket vi har regnet ut, gjelder et utvalg, vil det være fornuftig å bruke s_{n-1} , men hvis vi regner ut standardavviket for populasjonen, bør vi bruke s_n . At s_{n-1} er valgt her (under sterk tvil), skyldes et ønske om å forenkle, slik at studentene bare har ett standardavvik å forholde seg til.

Eksempel:

I **Tabell XVI** er satt opp de nødvendige mellomregningene for å finne standardavvik for temperaturen i et bokmagasin.

Standardavviket blir:

$$s = \sqrt{\frac{\sum (x - m)^2}{n - 1}} = \sqrt{\frac{72}{8}} = \sqrt{9} = 3(\text{°C})$$

Når vi skal finne standardavvik ved frekvenstabeller, er det nødvendig å ta hensyn til antallet som har lik verdi. Dette er helt analogt til at vi må justere formelen når vi regner ut gjennomsnitt for frekvenstabeller. Uttrykket blir da:

$$s = \sqrt{\frac{\sum f \cdot (x - m)^2}{n - 1}}$$

Eksempel:

I **Tabell XVII** er satt opp samme data som i **Tabell XVI**, men her i frekvensfordeling, slik den er gitt i **Tabell IV** s. 17.

Vi ser at summen til høyre i tabellen, nemlig $\sum f \cdot (x - m)^2 = 72$, tilsvarende til $\sum (x - m)^2 = 72$ i **Tabell XVI**.

Standardavviket blir derfor det samme som i forrige eksempel:

$$s = \sqrt{\frac{\sum f \cdot (x - m)^2}{n - 1}} = \sqrt{\frac{72}{8}} = 3(\text{°C})$$

Tabell XVI Utregning av standardavvik ut fra råskårer

Enhet nr.	Verdi x (°C)	x-m	(x-m) ²
1	16	-4	16
2	22	2	4
3	20	0	0
4	23	3	9
5	17	-3	9
6	20	0	0
7	17	-3	9
8	20	0	0
9	25	5	25
Σ	180	0	72

Tabell XVII Utregning av standardavvik ved frekvensfordeling

x	f	x-m	(x-m) ²	f·(x-m) ²
16	1	-4	16	16
17	2	-3	9	18
20	3	0	0	0
22	1	2	4	4
23	1	3	9	9
25	1	5	25	25
Σ				72

Standardavvik kan tolkes som et slags gjennomsnitt (men ikke aritmetisk gjennomsnitt) av avvikene fra gjennomsnittet. Standardavviket kan gjerne tegnes ut fra gjennomsnittet i begge retninger. Ved regelmessige, såkalte normalfordelte fordelinger, som vi kommer til seinere, ligger ca. 68% av fordelingen innenfor et standardavvik på begge sider av gjennomsnittet.

Utrekning av standardavvik

1. Trekk gjennomsnittet fra verdiene
2. Opphøy i annen potens
3. Ved frekvensfordelinger: Multipliser med frekvens
4. Legg sammen
5. Divider på én mindre enn antallet enheter
6. Trekk ut kvadratrot

8.3. Kvartiler

Kvartiler ligner på medianen ved at de er verdier som deler en fordeling i like store deler. Mens medianen deler fordelingen i to like store deler, deler kvartilene sammen med medianen en fordeling i 4 like store deler.

Regnearbeidet for å finne kvartilene er svært likt det som skal til for å finne medianen. Vi ordner enhetene etter stigende verdi. Så må vi finne numrene til to enheter:

- Den enheten som er slik at $\frac{1}{4}$ av enhetene har *lavere* verdi, mens $\frac{3}{4}$ har høyere verdi. Nummeret til denne enheten, når enhetene er ordnet i stigende rekkefølge, kaller vi q_1 , som regnes ut slik:

$$q_1 = \frac{n}{4} + \frac{1}{2}$$

- Den enheten hvor $\frac{3}{4}$ av enhetene har lavere verdi, mens $\frac{1}{4}$ har høyere verdi. Her bruker vi symbolet q_3 for enhetsnummeret:

$$q_3 = \frac{3n}{4} + \frac{1}{2}$$

Når vi har funnet enhetsnumrene, kan vi finne kvartilene:

Vi finner verdien til enhet nr. q_1 . Denne verdien kalles *nedre kvartil*. Som symbol bruker vi Q_1 .

Tilsvarende kalles verdien til enhet nr. q_3 *øvre kvartil*. Her bruker vi symbolet Q_3 .

Tabell XVIII Hvordan finne kvartilene

			q_1 ↓					q_3 ↓		
Enhet nr.	1	2	3	4	5	6	7	8	9	10
Verdi	2	3	4	5	6	6	7	9	10	12
			↑ Q_1					↑ Q_3		

Eksempel:

I **Tabell XVIII** er det satt opp $n = 10$ verdier.

Vi finner først

$$q_1 = \frac{n}{4} + \frac{1}{2} = \frac{10}{4} + \frac{1}{2} = 3$$

og

$$q_3 = \frac{3n}{4} + \frac{1}{2} = \frac{30}{4} + \frac{1}{2} = 8$$

Av tabellen kan vi da lese av $Q_1 = 4$ og $Q_3 = 9$.

8.4. Median og kvartiler ved frekvensfordelinger

Ved frekvensfordelinger gir den kumulative tabellen oversikt over hvordan enhetene fordeler seg, fra enheten med lavest verdi, til enheten med høyest verdi.

Vi tar utgangspunkt i den kumulative tabellen, og regner ut enhetsnumre for kvartilene, q_1 og q_3 på vanlig måte. For å finne de tilhørende verdiene, tar vi utgangspunkt i den kumulative tabellen for å finne hvilke enheter de enkelte frekvenser omfatter.

Eksempel:

I **Tabell XIX** er det tatt utgangspunkt i **Tabell VIII**. Vi finner:

$$q_1 = \frac{n}{4} + \frac{1}{2} = \frac{22}{4} + \frac{1}{2} = 6$$

$$q_2 = \frac{n+1}{2} = \frac{22+1}{2} = 11,5$$

$$q_3 = \frac{3 \cdot n}{4} + \frac{1}{2} = \frac{3 \cdot 22}{4} + \frac{1}{2} = 17$$

Tabell XIX Kvartiler og median ved grupperte fordelinger

	Alder (x)	Antall (f)	Kumulerte frekvenser (kf)	Enhetsnr.	
Q ₁ →	16-20	8	8	1 - 8	← q ₁
md →	21-25	6	14	9 - 14	← q ₂
	26-30	2	16	15 - 16	
Q ₃ →	31-35	5	21	17 - 21	← q ₃
	36-40	0	21		
	41-45	0	21		
	46-50	1	22	22	

Vi ser at aldersintervallet 16-20 omfatter de første 8 enhetene, dvs. enhet nr. 1-8. Neste aldersintervall, 21-25, dekker f.o.m. enhet nr. 9, t.o.m. enhet nr. 14, osv. Når dette er satt opp, finner vi enhetene tilsvarende q₁, q₂ og q₃ til høyre i tabellen. Q₁, md og Q₃ finnes ved å følge linjene bort til verdiene i venstre kolonne.

Vi har ikke nok informasjon til å angi nøyaktig median. Vi angir da hvilket *intervall* kvartilene og medianen ligger innenfor. Dersom vi må velge en konkret verdi, f.eks. for å regne ut kvartilavviket, velger vi *midtpunktene* i intervallene.

Q₁ er i intervallet 16 til 20 år, md er i intervallet 21 til 25 år, og Q₃ er i intervallet 31 til 35 år

8.5. Kvartilavvik

Når vi har funnet kvartilene, er kvartilavviket definert ved formelen: $Q = \frac{Q_3 - Q_1}{2}$

For kvartilavvik brukes symbolet **Q**.

Eksempel:

Fordelingen i **Tabell XVIII** har kvartilavviket $Q = \frac{Q_3 - Q_1}{2} = \frac{9 - 4}{2} = 2,5$

Kvartilavvik kan forstås som avstand mellom kvartiler og median (hvis kurven er skjev blir det gjennomsnitt av avstanden mellom kvartilene og medianen). Kvartilavviket kan gjerne tegnes ut fra

medianen i begge retninger. Ved symmetrisk fordeling ligger da halvparten av enhetene innenfor en avstand på Q fra medianen. Med en annen skrivemåte:

Av enhetene ligger 50 % i intervallet fra $md-Q$ t.o.m. $md+Q$.

Utregning av kvartilavvik

1. Ordne enhetene etter stigende verdier
Ved frekvensfordelinger: Finn kumulerte frekvenser, og enhetsnumre
2. Regn ut enhetsnumrene q_1 og q_3
3. Finn de tilhørende verdiene Q_1 og Q_3 (kvartilene)
4. Finn halvparten av differansen mellom Q_3 og Q_1 (kvartilavviket)

8.6. Spredningsmål og EXCEL

EXCEL har vi disse formlene tilgjengelige:

- =stdav(område) Resultatet blir standardavviket, s
- =kvartil(område;1) Resultatet blir nedre kvartil, Q_1
- =kvartil(område;3) Resultatet blir øvre kvartil, Q_3
- =min(område) Resultatet blir minste verdi, x_{\min}
- =størst(område) Resultatet blir største verdi, x_{\max}

For kvartiler er det en liten forskjell på formelen til EXCEL og den som er brukt i kompendiet

For å finne kvartilavviket, må vi bruke formelen $Q = \frac{Q_3 - Q_1}{2}$

Variasjonsbredden blir $x_{\max} - x_{\min}$.

Hvis verdiene er satt opp som en frekvensfordeling, finnes det ingen ferdig-definert funksjon. Vi må da lage en tabell tilsvarende det manuelle oppsettet, men all mellomregningen utføres av regnearket. Oppsettet blir analogt til utregning av gjennomsnitt ved frekvensfordelinger, men en må legge inn kolonner med formler tilsvarende kolonnene i **Tabell XVI**.

8.7. Spredningsmål og målenivå

Variasjonsbredden kan måles hvis verdier kan trekkes fra hverandre. Dette betyr at variabelen må være på intervallnivå (eller høyere nivå, dvs. forholdstallsnivå).

Også **kvartilavvik** og **standardavvik** krever minst intervallnivå. Også her må vi kunne subtrahere (trekke fra), og også dividere med et tall.

Hvis målenivået er lavere, kan vi evt. oppgi:

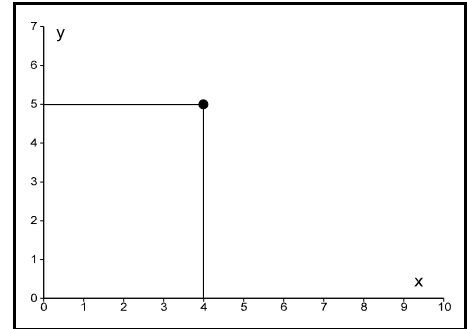
- Ved **nominalnivå**: Antall forskjellige verdier
- Ved **ordinalnivå**: Kvartilene

Disse angivelsene vil gi et inntrykk av spredningen.

9. Korrelasjon

Hvordan er sammenhengen mellom to variabler?

Hittil har vi sett på analyser med en variabel. Den fordelingen vi da får, kalles en *univariat fordeling*. Nå skal vi se på undersøkelser med to variabler, hvor vi har en såkalt *bivariat fordeling*. Dette er et spesialtilfelle av fordelinger med flere variabler, *multivariat fordeling*.



Figur 33 Plassering av punktet $x = 4, y = 5$

9.1. Punktdiagram (spredningsdiagram)

Når vi observerer to variabler, bruker vi ofte x og y for de to variablenes verdier. Sammenhengen mellom de to variablene kommer best fram i et *punktdiagram* (kalles også spredningsdiagram, scatterdiagram eller korrelasjonsdiagram). For hver enhet avsetter vi x -verdien på x -aksen, og trekker en linje parallelt med y -aksen. Så avsetter vi y -verdien på y -aksen, og trekker en linje parallelt med x -aksen. Punktet som viser enhetens verdier, er der de to linjene møtes. Se **Figur 33**.

Eksempel:

Tabell XX viser en datamatrix for temperatur og relativ fuktighet i et bokmagasin. De to variablene er målt 9 forskjellige dager. Hver dag er altså en undersøkelsesenhhet.

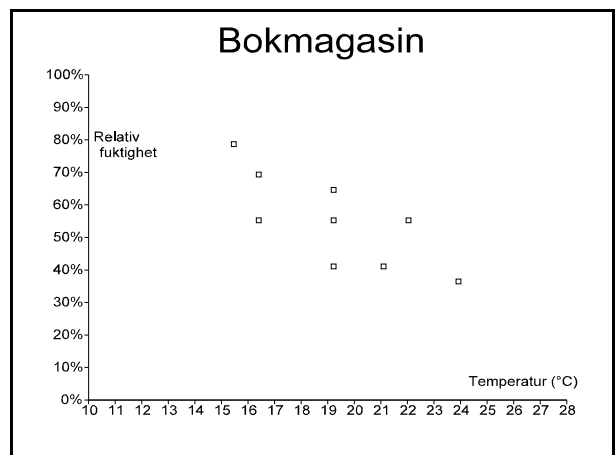
Tabell XX Observasjoner for to variabler

Enhet nr	Temperatur (x)	Relativ fuktighet (y)
1	20°	70%
2	22°	45%
3	20°	60%
4	16°	85%
5	17°	75%
6	23°	60%
7	20°	45%
8	25°	40%
9	17°	60%

Sammenhengen mellom de to variablene er vanskelig å se fra tabellen, mens det er en tydelig sammenheng når vi setter opp et diagram (se **Figur 34**). Her er temperatur markert på x -aksen, og fuktighet på y -aksen.

Hvis vi har grunn til å anta at den ene variabelen påvirker den andre, setter vi den som *påvirker* (den *uavhengige variabelen*) på x -aksen, mens den som *påvirkes* (den *avhengige variabelen*) avsettes på y -aksen.

I dette tilfellet har vi større grunn til å anta at temperaturen påvirker fuktigheten enn omvendt.



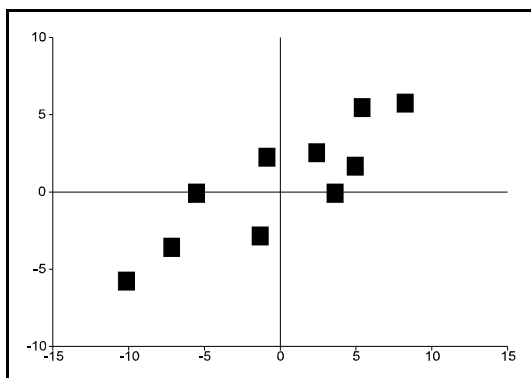
Figur 34 Spredningsdiagram

9.2. Produkt-moment-korrelasjon

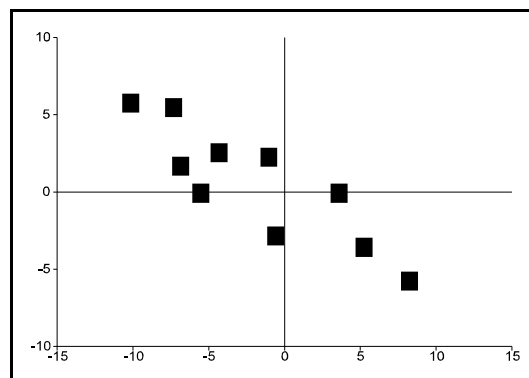
Vi kan ha forskjellig former for sammenheng mellom to variabler. Den enkleste formen for sammenheng, er den lineære. Den innebærer at alle punktene ligger på en rett linje.

Produkt-moment-korrelasjonen (kalles også Pearsons r eller bare korrelasjon) er et uttrykk for hvor godt data stemmer overens med en rett linje. Dette målet for korrelasjon egner seg ikke hvis sammenhengene følger en annen form for kurve.

Som vi skal se, kan det godt være sammenheng som ikke passer med en linje.



Figur 35 Positiv sammenheng



Figur 36 Negativ sammenheng

Korrelasjonen har to kjennetegn:

- **Retning.** Hvis y -verdiene viser tendens til å stige når x -verdien stiger, er det en **positiv** sammenheng. Hvis y stort sett er fallende med økende x , er det en **negativ** sammenheng. Retningen angis ved et fortegn, «+» eller «-».
- **Styrke.** Styrken viser hvor godt punktene passer med en rett linje. Hvis alle punktene ligger på en stigende rett linje, er korrelasjonen $+1$, mens den er -1 ved en fallende linje. I begge tilfellene er det en perfekt samvariasjon eller korrelasjon. Styrken er **absoluttverdien** til korrelasjonen, dvs. vi ser bort fra fortegn.

For produkt-moment-korrelasjon bruker vi symbolet r , som kalles **korrelasjonskoeffisienten**. Den regnes ut med formelen:

$$r = \frac{\sum (x - m_x) \cdot (y - m_y)}{s_x \cdot s_y \cdot (n - 1)}$$

Siden det er to variabler, må vi beregne to gjennomsnitt og to standardavvik. Symbolene m_x og s_x betyr henholdsvis gjennomsnitt og standardavviket for variabel x . Vi bruker tilsvarende symboler for gjennomsnitt og standardavvik for variabel y .

Uttrykket $\sum (x - m_x) \cdot (y - m_y)$ betyr at vi skal multiplisere $x - m_x$ med $y - m_y$ for hver enhet, og så legge sammen produktene. Antall enheter betegnes som vanlig med symbolet n .

9.3. Utregning av korrelasjonskoeffisienten

Vi går tilbake til eksemplet på s. 40, og **Tabell XXI** Mellomregning for å finne **r** skal regne ut:

$$r = \frac{\sum (x - m_x) \cdot (y - m_y)}{(n - 1) \cdot s_x \cdot s_y}$$

I **Tabell XXI** er de nødvendige mellomregninger satt opp. Her forutsettes at en på forhånd har regnet ut gjennomsnitt og standardavvik for hver variabel, jfr. kapittel 7 og 8. Disse målene er:

$$m_x = 20, s_x = 3$$

$$m_y = 60, s_y = 15.$$

Enh. nr.	Temp (x)	Fukt (y)	x-m _x	y-m _y	(x-m _x)·(y-m _y)
1	20	70	0	10	0
2	22	45	2	-15	-30
3	20	60	0	0	0
4	16	85	-4	25	-100
5	17	75	-3	15	-45
6	23	60	3	0	0
7	20	45	0	-15	0
8	25	40	5	-20	-100
9	17	60	-3	0	0
Sum	180	540	0	0	-275

Korrelasjonskoeffisienten blir:

$$r = \frac{\sum (x - m_x) \cdot (y - m_y)}{(n - 1) \cdot s_x \cdot s_y} = \frac{-275}{(9 - 1) \cdot 3 \cdot 15} = \underline{-0,76}$$

Det er opplagt en sammenheng mellom variablene. Hva denne sammenhengen skyldes, kan ikke statistikken gi svar på. Derimot kan faget fysikk gi en forklaring: Når temperaturen øker, minker den relative fuktigheten i et tett rom! (Og omvendt: Ved minkende temperatur øker den relative fuktigheten, jfr. ising på vindusruter om vinteren). Dette er også bakgrunnen for at temperatur blir satt på x-aksen (uavhengig variabel), mens fuktighet blir satt på y-aksen (avhengig variabel) i figuren.

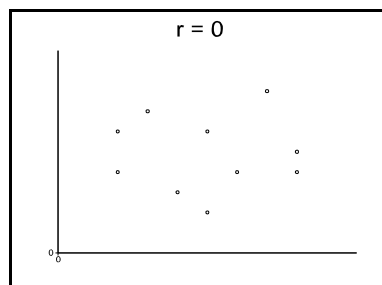
Om vi *ikke* kjenner gjennomsnittene og standardavvikene til de to variablene, trenger vi to kolonner til i **Tabell XXI**, nemlig $(x - m_x)^2$ og $(y - m_y)^2$. Fullstendig oppskrift til å finne korrelasjonskoeffisienten blir:

Utregning av korrelasjonskoeffisient

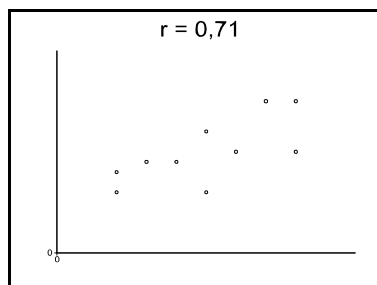
1. Addér verdiene for hver variabel, og finn gjennomsnittene
2. Trekk gjennomsnittet fra verdiene, for hver variabel
3. Kvadrer uttrykkene du finner, addér, og finn standardavvikene
4. Multipliser $x - m_x$ med $y - m_y$ for hver enhet, og addér
5. Finn korrelasjonskoeffisienten ut fra formelen

9.4. Tolkning av korrelasjonskoeffisienten

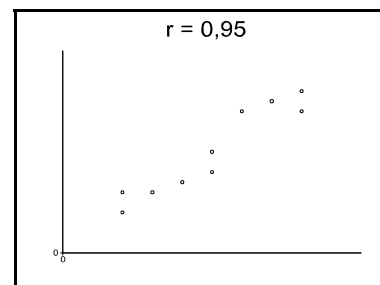
Figur 37 - Figur 39 viser resultatet av utregningen for 3 punktdiagrammer:



Figur 37



Figur 38



Figur 39

Hvis variabel y øker lineært med variabel x , vil korrelasjonen være nøyaktig 1.

Hvis det i grove trekk er slik at når x øker, så øker y , vil korrelasjonen være positiv.

Negativ korrelasjon betyr i likhet med positiv korrelasjon at det er samsvar. Negativ korrelasjon betyr at høyere verdi på den ene variabelen stort sett medfører lavere verdi på den andre.

En korrelasjon på 0, betyr derimot at det ikke er noen lineær sammenheng mellom variablene. Et spesialtilfelle er at y er konstant (rett linje, parallell med x -aksen).

En høy korrelasjonskoeffisient betyr at det er sammenheng. Dette betyr ikke nødvendigvis årsak-virkningsforhold mellom de to variablene. Det kan like gjerne være en *tredje* variabel som påvirker begge de to andre.

Eksempel:

I Danmark skal det visstnok være en liten, men positiv korrelasjon mellom antall storker pr km^2 i en kommune, og antall årlige fødsler pr. 1000 innbyggere i den samme kommunen.

Er det storke-tettheten som påvirker fødselstallet? Eller blir storkene tiltrukket av de mange spedbarna? Forklaringen, en felles faktor som både påvirker utbredelsen av storker og hyppigheten av fødsler er relativt enkelt å finne. Leseren oppfordres til å bruke sin fantasi!

Tolkningen av en korrelasjonskoeffisient kan være vanskelig. Selv om korrelasjonsmålet ligger mellom -1 og $+1$, kan korrelasjonen ikke tolkes som andel. Vi kan *ikke* si at en korrelasjon er dobbelt så stor som en annen, selv om tallverdien er dobbelt så stor.

Derimot kan en lettere tolke kvadratet av korrelasjonen, r^2 . Dette uttrykket forteller hvor stor del av *variansen* (kvadratet av standardavviket, dvs. s^2) i den ene variabelen som kan forklares ut fra variansen i den andre variabelen.

Ved en korrelasjon $r = 0,71$ er $r^2 = 0,5$. Her kan halvparten av variansen i den ene variabelen forklares ved variansen i den andre. Halvparten av variansen må forklares ved faktorer som ikke er felles.

9.5. Ikke-lineær sammenheng

Hvis det *er* en sammenheng, men denne sammenhengen ikke er lineær, bør en *ikke* bruke r som mål for sammenhengen. Produkt-moment-korrelasjonen vil i disse tilfellene gi en for *lav* verdi.

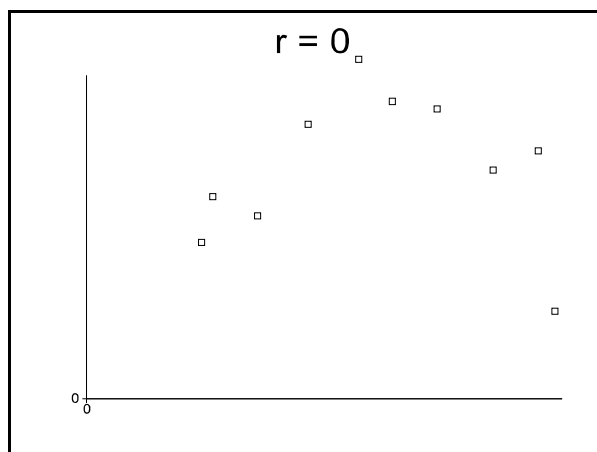
Eksempler:

I **Figur 40** ser det ut til å være en såkalt U-formet kurve som ligger til grunn for sammenhengen mellom x og y .

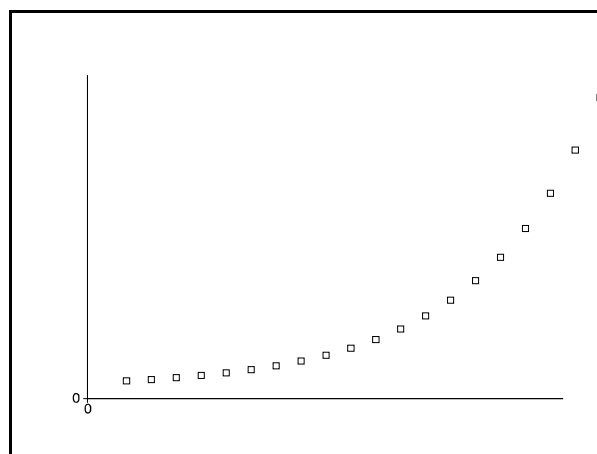
En slik kurve kan f.eks. beskrive sammenhengen mellom stress og prestasjon. Litt stress kan øke prestasjonen, men etter en grense minker prestasjonen ved større stress.

I **Figur 41** er det en klar matematisk sammenheng mellom x og y . Likevel vil korrelasjonskoeffisienten bare bli $r = 0,89$, fordi punktene ikke ligger langs en *rett* linje.

I denne figuren er sammenhengen mellom y og x uttrykt ved en såkalt eksponentialfunksjon. Eksempel på en slik sammenheng er økning i antall vitenskapelige tidsskrifter (y -aksen) som funksjon av tid (x -aksen).



Figur 40 En U-formet sammenheng



Figur 41 En eksponentiell sammenheng

9.6. Korrelasjon og målenivå

Variablene må vanligvis være målt på minst *intervallnivå* for at vi skal kunne regne ut r . Dette skyldes bl. a. at vi bruker standardavviket i utregningen, og utregning av standardavvik krever dette nivået.

Det er likevel et unntak: Hvis den ene eller begge variablene er *dikotome*, dvs. har 2 mulige verdier, kan vi likevel regne ut korrelasjonen. For dikotome variabler bruker vi da 0 og 1 som de to verdiene.

Likevel finnes det korrelasjonsmål som kan benyttes ved lavere målenivå. Hvis verdiene for begge variablene er rangert, fra laveste verdi til høyeste verdi, kan vi regne ut *rangkorrelasjonen*. Hvis forskjellen i rang mellom de to variablene er d , så er rangkorrelasjonen ρ (rho, gresk r). Dette korrelasjonsmålet bygger på Pearsons r .

$$\rho = 1 - \frac{6 \sum d^2}{n \cdot (n^2 - 1)}$$

10. Regresjon

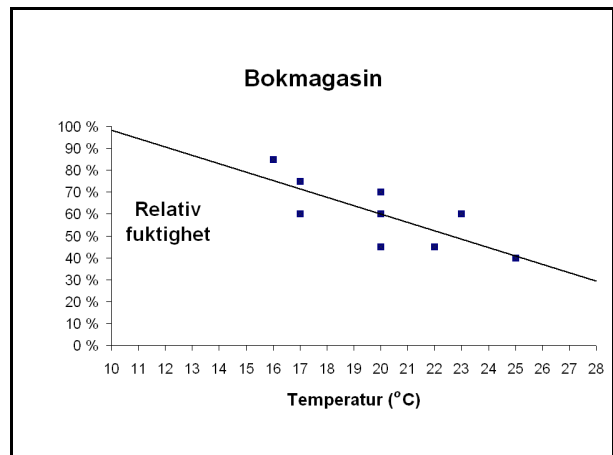
Hvordan kan vi forutsi en verdi i en bivariat fordeling?

Ved utregning av korrelasjonskoeffisienten finner vi hvor godt sammenhengen mellom to variabler passer med en rett linje. Det finnes en formel som beskriver denne linjen. Vi bruker linjen til å forutsi verdien til den ene variabelen (y) når vi kjenner den andre (x).

Eksempel:

I forrige kapittel så vi i **Figur 34** at det var en sammenheng mellom relativ fuktighet og temperatur i et bokmagasin. Utregningen av korrelasjonskoeffisienten ga et mål for denne sammenhengen. Stort sett vil den relative fuktigheten minke når temperaturen øker.

Hvis vi *ikke* vet noe om temperaturen, og skal anslå (gjette) hva fuktigheten er, vil det beste anslaget være gjennomsnittlig fuktighet, dvs. 60%. Hvis vi derimot kjenner temperaturen, må det gå an å gi et bedre anslag for fuktigheten. Hvis temperaturen er høy (over gjennomsnitt på 20°), kan vi anta at fuktigheten er lav. Hvis temperaturen er lav, kan vi tilsvarende regne med at fuktigheten er høy.



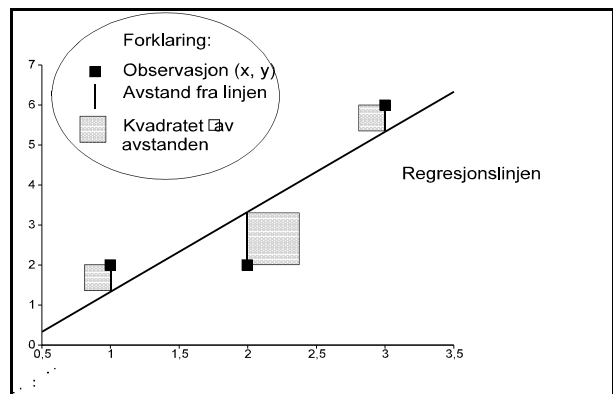
Figur 42 Spredningsdiagram med regresjonslinje

En rett linje vil gi oss en metode til å forutsi y hvis vi kjenner x. Hvordan kan vi gi et bedre tips for y når vi kjenner x?

Vi gjør dette ved å trekke en rett linje som ut fra visse kriterier ligger så nær *alle* punktene som mulig.

Matematisk utgangspunkt for «beste linje» er å se på avviket mellom y-verdiene for de gitte punktene i spredningsdiagrammet, og de tilsvarende y-verdiene som ligger på linja. Vi trekker en linje parallelt med y-aksen for å finne disse avvikene. Avvikene kvadreres (multipliseres med seg selv), og legges sammen. Den linja som da gir **minst** kvadrerte avvik, regnes som den beste.

Ut fra dette kalles metoden for «minste kvadraters metode»



Figur 43 Minste kvadraters metode

Linjen kalles for **regresjonslinje**, og metoden **lineær regresjon**.

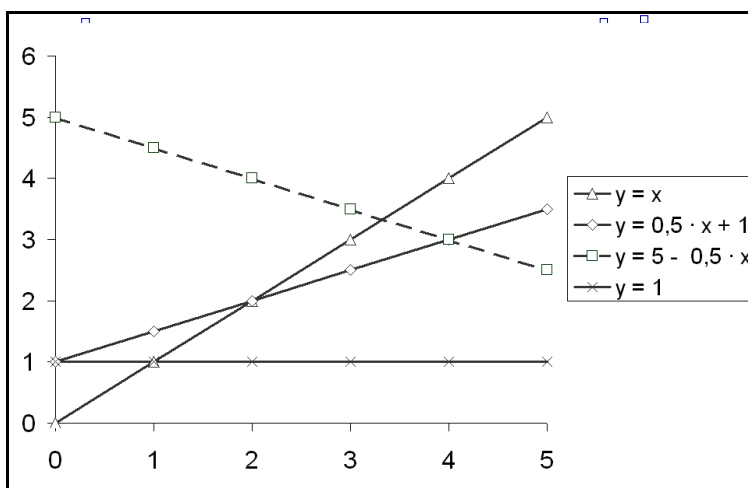
10.1. Formelen for en linje

Formelen for en rett linje i et aksekors kan helt generelt skrives på formen:

$$y = ax + b$$

Her er a og b to konstanter (tall). En gitt linje kan f.eks. uttrykkes som $y = 0,5 \cdot x + 1$.

Konstanten 0,5 (a) forteller hvor bratt linjen stiger. Den kalles for *stigningstallet* for linjen.



Figur 44 Forskjellige linjer i et koordinatsystem

Kjenner vi formelen for en linje, kan vi tegne den ved å velge to tilfeldige x-verdier, og finne de tilhørende y-verdiene ut fra formelen. Vi tegner inn de to punktene vi da finner, og trekker linja. Som kontroll på regning og tegning kan en godt regne ut y-verdien for en tredje x-verdi. Dette punktet bør ligge på samme linje!

10.2. Formelen for regresjonslinjen

Regresjonslinjen må også kunne uttrykkes ved $y = a \cdot x + b$. Konstantene a og b kalles *regresjonskoeffisientene*. De regnes ut slik:

$$a = \frac{\sum (x - m_x) \cdot (y - m_y)}{(n - 1) \cdot s_x^2} \quad \text{og} \quad b = m_y - a \cdot m_x$$

Oppsettet for å regne ut regresjonslinjen er svært likt utregning av korrelasjon.

Som for korrelasjon trenger vi: $\sum (x - m_x) \cdot (y - m_y)$

For å regne ut dette uttrykket, må vi kjenne m_x og m_y . Gjennomsnittene brukes også for å regne ut b.

Dessuten må vi finne s_x .

Har vi tidligere regnet ut korrelasjonskoeffisienten, er det meste av mellomregningen utført.

Eksempel:

Vi fortsetter eksemplet med bokmagasinet. Mellomregningen er gjort i **Tabell XXI** på s. 42. Her har vi:

$$m_x = 20, m_y = 60, s_x = 3, \sum (x - m_x) \cdot (y - m_y) = -275.$$

Vi regner først ut a:

$$a = \frac{\sum (x - m_x) \cdot (y - m_y)}{(n - 1) \cdot s_x^2} = \frac{-275}{8 \cdot 3^2} = -\frac{275}{72} = -3,82$$

Deretter kan vi finne b:

$$b = m_y - a \cdot m_x = 60 - (-3,82 \cdot 20) = 60 + 76,4 = 136,4$$

Vær oppmerksom på at det å trekke fra et negativt uttrykk er det samme som å legge til et positivt: $-(-3) = +3$!

Uttrykket for regresjonslinjen blir da $y = -3,82 \cdot x + 136,4$

10.3. Tegning av regresjonslinjen

Når linjen skal *tegnes*, velger vi to x-verdier. Vi regner ut tilhørende y-verdier. Det lønner seg å velge den ene x-verdien i ytterkant i figuren. Hvis $x = 0$ ligger i ytterkant av figuren, er det enkelt å regne ut y. Her blir: $y = a \cdot x + b = 0 + b = b$.

Den andre x-verdien kan gjerne være m_x . Tilhørende y-verdi blir da m_y !

Eksempel:

I vårt gjennomgangseksempel kan vi velge x-verdiene 10 og 20. Ønsker vi kontroll, regner vi ut y for $x = 25$ i tillegg. Oppsettet er gjengitt i **Tabell XXII**. Linjen tegnes så inn, slik det er gjort i **Figur 42**.

Tabell XXII

x	y
10	98,2
20	60,0
25	40,9

10.4. Prediksjon

Regresjonslinjen brukes til å gi det «best mulige» anslag for y, gitt at vi kjenner x. Forutsetningen er at vi har grunn til å anta at det er en lineær sammenheng, og at vi har regnet ut formelen for linjen ut fra et representativt utvalg.

Hvis vi f. eks. har data som viser at det er en sammenheng mellom avstand til biblioteket og bibliotekbruk, kan vi finne en regresjonslinje. Ut fra denne linjen kan vi gi et anslag for bibliotekbruken til en person med en kjent avstand til biblioteket. Den feilen vi da gjør, vil stort sett være *mindre* enn om vi ikke kjenner avstanden. Hvis vi ikke kjente avstanden, ville det beste anslaget for lånerens bibliotekbruk vært som for gjennomsnittet av lånerne.

En vanlig bruk av regresjonslinjen er forutsigelse av fremtidige verdier av en variabel, når en kjenner verdiene for tidligere år.

Tabell XXIII

Gjennomsnittspris på medisinske tidsskrifter

År	Pris i \$
1991	249
1992	276
1993	288

Eksempel:

I «Synopsis» finner vi fra år til år gjennomsnittsprisene på amerikanske tidsskrifter fordelt på fag. Se **Tabell XXIII**.

En kan ut fra de gitte tall finne et anslag for gjennomsnittlig pris i 1994. En kan da velge året som x , prisen som y . Dette gir imidlertid store tall ved regning, så en snarvei er å definere x som antall år etter 1990, slik at $x = 1$ for 1991 osv.

Mellomregningene er satt opp i **Tabell XXIV** Mellomregninger for å predikere pris på medisinske tidsskrifter

Under utregningen må vi finne gjennomsnittene:

$$m_x = \frac{\sum x}{n} = \frac{6}{3} = 2$$

$$m_y = \frac{\sum y}{n} = \frac{813}{3} = 271$$

År	x	Pris (y)	$x - m_x$	$(x - m_x)^2$	$y - m_y$	$(x - m_x) \cdot (y - m_y)$
1991	1	249	-1	1	-22	22
1992	2	276	0	0	5	0
1993	3	288	1	1	17	17
Σ	6	813	0	2	0	39

Deretter finner vi

$$s_x = \sqrt{\frac{\sum (x - m_x)^2}{n - 1}} = \sqrt{\frac{2}{2}} = 1, \text{ og vi kan finne a og b i regresjonsformelen:}$$

$$a = \frac{\sum (x - m_x) \cdot (y - m_y)}{(n - 1) \cdot s_x^2} = \frac{39}{2 \cdot 1} = 19,5$$

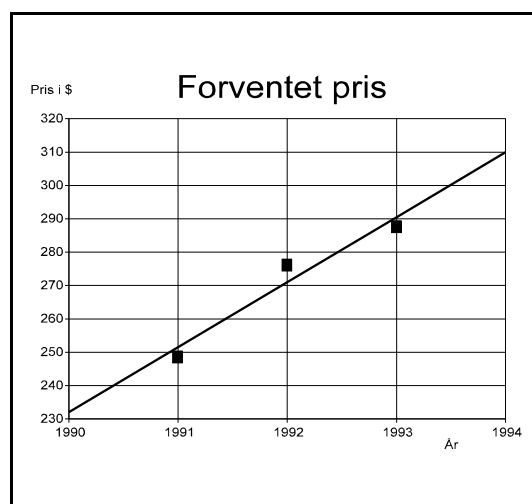
$$b = m_y - a \cdot m_x = 271 - 19,5 \cdot 2 = 271 - 39 = 232$$

Regresjonslinjen blir: $y = 19,5 \cdot x + 232$

Vårt anslag for tidsskriftpris i 1994 finner vi ved å sette inn $x = 4$ (4 år etter 1990):

$$y = 19,5 \cdot x + 232 = 19,5 \cdot 4 + 232 = \underline{310} (\$)$$

Figur 45 viser regresjonslinjen vi har funnet. Vi kan lese av prisen i denne figuren, ved å se hva y er for hvert år. I figuren er satt inn *årstallet*, som vi får ved å legge 1990 til x -verdiene.



Figur 45 Priser på medisinske tidsskrifter

Hele utregningen bygget på forutsetningen om linearitet, dvs. at prisstigningen stort sett er den samme fra år til år. Om vi har grunn til å tro at denne forutsetningen *ikke* er oppfylt, kan vi ikke bruke lineær regresjon. Vurderingen av om prisen stiger lineært på medisinske tidsskrifter overlates til leseren.

Her har vi antatt lik absolutt prisstigning. Mer korrekt ville det være å anta lik relativ (prosentvis) prisstigning.

En del feilaktige prediksjoner skyldes at forutsetningen om linearitet ikke holder. Mange personers privatøkonomi ble drastisk forverret p.g.a forventninger om fortsatt prisstigning i boligpriser, mens prisene tvert imot falt!

10.5. Regresjonslinje for x uttrykt ved y

Rent intuitivt skulle man tro at regresjonslinjen vi har funnet også kan brukes for å forutsi x med kjent y-verdi. Men dette er ikke tilfellet!

Grunnen er at minste kvadraters metode gir forskjellige resultater når vi skal se på avstanden mellom linjen og punktene i x-verdier og i y-verdier.

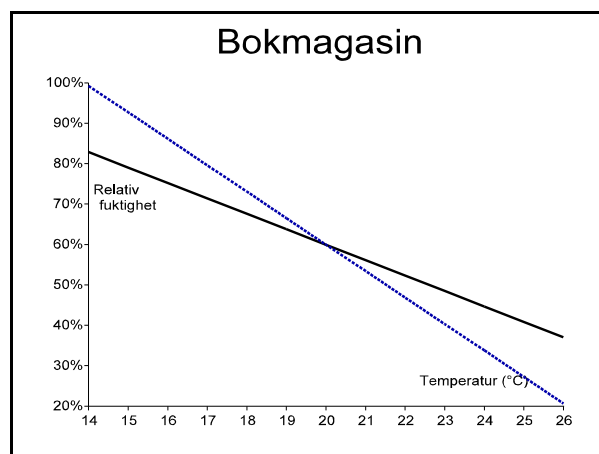
Regresjonslinjen $x = a_1 y + b_1$ regnes ut på samme måte som linjen $y = a x + b$. Her er symbolene a_1 og b_1 brukt for å ikke å forveksle dem med a og b. I formlene må en bytte x og y, slik at konstantene blir:

$$a_1 = \frac{\sum (y - m_y) \cdot (x - m_x)}{(n - 1) \cdot s_y^2} \quad \text{og} \quad b_1 = m_x - a_1 \cdot m_y$$

Eksempel:

Regresjonslinja som gir anslag for temperatur ut fra fuktighet blir: $x = -0,153 y + 29,2$.
Detaljene for utregningen overlates til leseren.

I **Figur 46** er *begge* regresjonslinjene for sammenhengen mellom temperatur og fuktighet tegnet inn. Begge linjene går gjennom punktet (m_x, m_y) . Jo større vinkelen er mellom regresjonslinjene, desto mindre er korrelasjonen.



Figur 46 Begge regresjonslinjer tegnet inn.
Den bratteste linjen er $x = a_1 \cdot y + b_1$

10.6. Korrelasjon, regresjon og EXCEL

Spredningsdiagram (punktdiagram)

Spredningsdiagrammet får vi fram ved å merke området med x- og y-verdiene. Som vanlig fortsetter vi gjennom diagramveviseren. Valget «punktdiagram» gir oss spredningsdiagrammet. Vi kan gjerne legge til tittelen til diagrammet og til aksene.

Når diagrammet er fullført kan det være hensiktsmessig å endre litt på skalaene. EXCEL velger gjerne verdier ned til 0 på begge akser. Hvis dette ikke er hensiktsmessig (f.eks. fordi verdiene ligger langt fra 0), kan aksene tilpasses. Aksen må merkes i diagrammet (med enkelt klikk på aksen). Med høyre musetast får du flere valg, bl.a. Formater akse. Her finnes valget skala, hvor maksimums- og minimumsverdiene kan innstilles.

Arket «punkt» i EXCEL-fila g:\felles\stat\karakter.xls viser resultat av gjennomgang i timen.

Regresjonslinje (trendlinje)

Når punktene i punktdiagrammet er merket (klikk en gang med venstre musetast på et av punktene), kan vi tegne inn regresjonslinja. Etter klikk med høyre musetast velger vi «Trendlinje».

Vi har visse valg ved trendlinja. Vi velger «Lineær trendlinje». Dessuten har vi visse alternativer, bl.a. kan EXCEL sette inn formelen for regresjonslinja, i formen $y = ax + b$. Vi kan også få skrevet inn kvadratet av korrelasjonskoeffisienten, dvs. r^2 .

Regresjonslinja blir bare tegnet inn i det området der x-verdiene forekommer. Ønsker vi at linja skal bli lengre, kan vi formatere trendlinja: Først et klikk med venstre tast for å merke linja, klikk på høyre tast for å få visse valg, deretter Formater trendlinje. Vi får da flere valgmuligheter, og under Alternativer kan vi be om «Prognose» forover eller bakover.

Arket «Punkt og linje» i EXCEL-fila g:\felles\stat\karakter.xls viser eksempel.

Spesielle funksjoner for korrelasjon og regresjon:

I formlene nedenfor betyr områdex det området der x-verdiene ligger, f. eks. b3:b14, mens områdey betyr området der y-verdiene ligger, f.eks. c3:c14. Områdene kan gjerne ha navn, f.eks. «Info», «LISA». Legg merke til rekkefølgen for uavhengig og avhengig variabel i regresjonsformlene!

$$=\text{korrelasjon}(\text{områdex};\text{områdey}) \dots\dots\dots \text{Tilsvare r} = \frac{\sum (x - m_x) \cdot (y - m_y)}{s_x \cdot s_y \cdot (n - 1)}$$

$$=\text{stigningstall}(\text{områdey};\text{områdex}) \dots\dots\dots \text{Tilsvare a} = \frac{\sum (x - m_x) \cdot (y - m_y)}{(n - 1) \cdot s_x^2}$$

$$=\text{skjæringspunkt}(\text{områdey};\text{områdex}) \dots\dots\dots \text{Tilsvare b} = m_y - a \cdot m_x$$

$$=\text{prognose}(x\text{-verdi};\text{områdey};\text{områdex}) \dots\dots\dots \text{Forventet y-verdi gitt en viss x-verdi}$$

11. Normalfordelingen

Normalfordelingen (også kalt Gaussfordelingen) er en symmetrisk, jevn, klokkeformet fordeling. Mange fordelinger som vi kan observere i virkeligheten ligner på normalfordelingen, f.eks. høyde, vekt, karakterer til eksamen.

Normalfordelingen er en matematisk fordeling. Den brukes som en *modell* for empiriske, klokkeformede fordelinger.

På samme måte som vi har en generell formel for rett linje, finnes det også en generell matematisk formel for normalfordelingen:

$$y = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{(x - \mu)^2}{2\sigma^2}}$$

der $\pi \approx 3,14$; $e \approx 2,72$

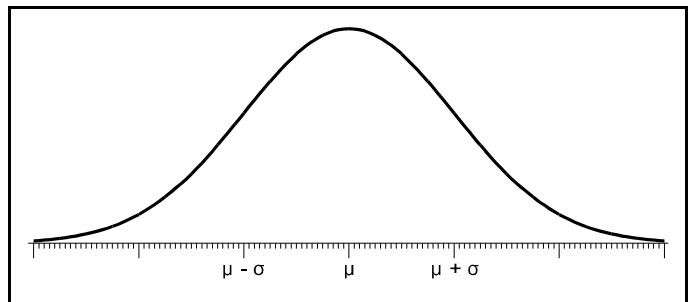
En gitt normalfordeling er entydig bestemt ved dens gjennomsnitt, μ , og standardavvik, σ .

Symbolene μ og σ brukes for gjennomsnitt og standardavvik i en *populasjon*. Det må være mange enheter for at fordelingen skal være så jevn som i normalfordelingen.

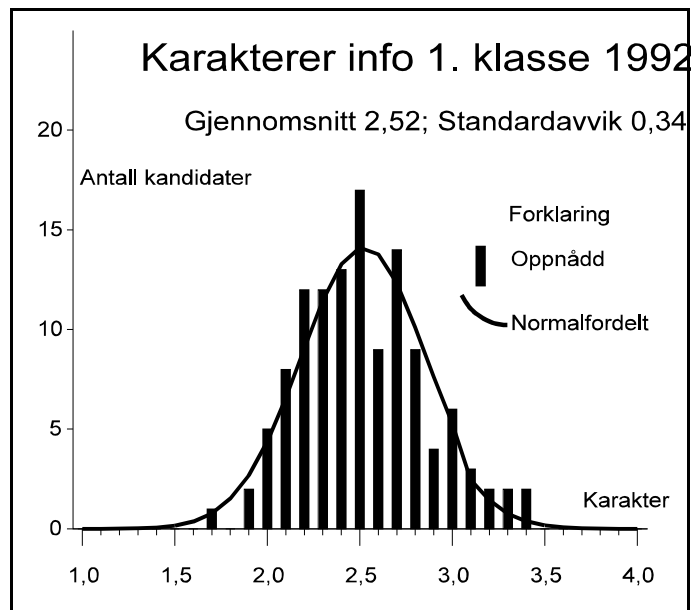
I **Figur 48** er karakterer til eksamen tegnet i et stolpediagram. Sammen med karakterfordelingen er tegnet en normalfordeling med samme gjennomsnitt og standardavvik. Antall enheter (n) er 121. Ved høyere antall kunne vi vente større likhet med normalfordelingen.

Det som skiller forskjellige normalfordelinger er deres gjennomsnitt og standardavvik. I **Figur 49** er det tegnet inn 4 forskjellige normalfordelinger.

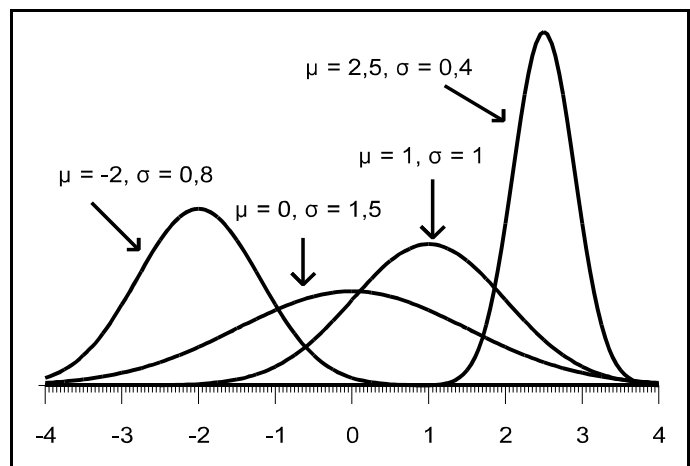
I kurven for normalfordelingen er det ikke *høyden* vi studerer, men *arealer*. Arealet under normalfordelingskurven settes vanligvis til 1, og vi leser av *andeler* under kurven. Under kurven finner vi andelen av enheter som ligger innenfor visse verdier.



Figur 47 Normalfordelingen



Figur 48 Karakterer og normalfordeling

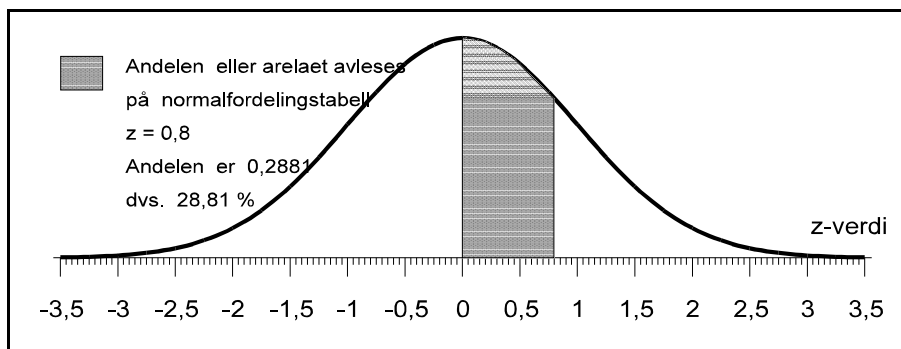


Figur 49 Forskjellige normalfordelinger

Normalfordeling kan sammenlignes med et frekvenspolygon. På x-aksen har vi verdier, på y-aksen frekvenser, men kurven er avrundet og jevn.

Siden kurven er symmetrisk, har gjennomsnitt, median og modus samme verdi. Kurven har et vendepunkt (der krumningen skifter fra konkav til konveks) som ligger ett standardavvik fra gjennomsnittet.

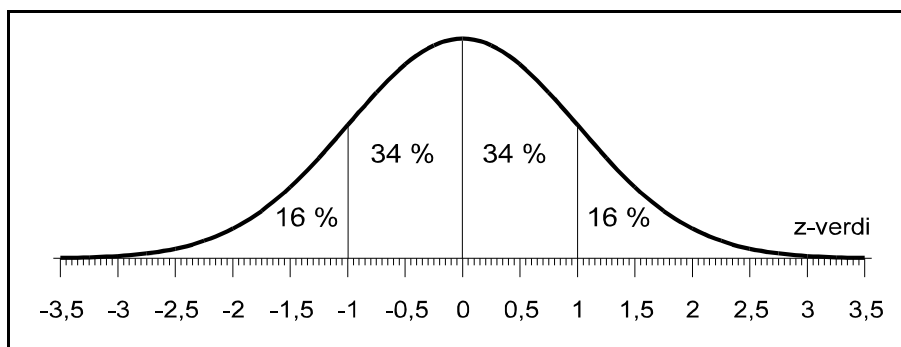
I normalfordelingen er det ikke de absolutte verdiene for frekvensene vi er ute etter, men arealene. Normalfordelingen er en kontinuerlig fordeling, så det har ikke noen hensikt å se etter frekvensen for en bestemt verdi.



Figur 50 Hvordan vi bruker normalfordelingstabellen

På side 72 er tabell over normalfordelingen med gjennomsnitt $\mu = 0$ og standardavvik $\sigma = 1$. For denne fordelingen bruker vi symbolet z for verdiene. I normalfordelingstabellen finner vi andelen enheter som har en z -verdi mellom 0 og en gitt annen verdi. I **Figur 50** gis et bilde av hva vi finner når vi slår opp 0,8 i normalfordelingstabellen. Vi finner en andel på 0,2881, og må multiplisere med 100 % for å finne prosentandelen.

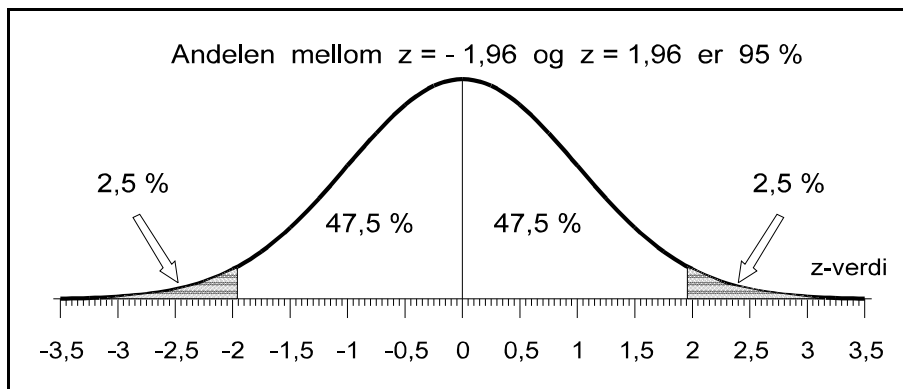
Av normalfordelingstabellen finner vi at ca. $2/3$ av enhetene ligger innenfor en avstand på ett standardavvik fra gjennomsnittet, mens ca. $1/3$ ligger utenfor. Jfr. **Figur 51**.



Figur 51 Andelene mellom $z = 0$ og $z = \pm 1$; og utenfor $z = \pm 1$

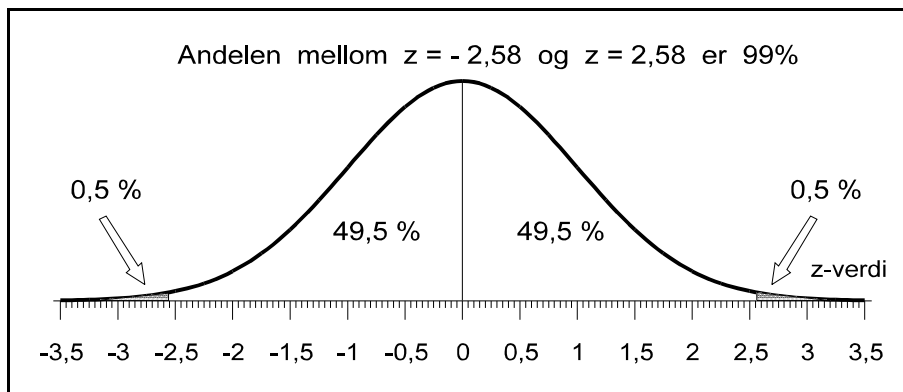
Ofte er vi interessert i å finne intervallet der de «fleste» enhetene ligger. Slår vi opp verdien 1,96 i normalfordelingstabellen, finner vi en andel på 0,475, dvs. 47,5 %. Dette er andelen enheter som har en verdi mellom 0 og 1,96. Siden kurven er symmetrisk, må det også være en andel på 47,5 % som har

en verdi mellom - 1,96 og 0. Totalt blir det 95 % mellom verdiene - 1,96 og + 1,96. Utenfor disse grensene faller de resterende 5 % av fordelingen.



Figur 52 $z = 1,96$ er et tall en gjerne bør merke seg!

Tilsvarende vil en andel på 99% av enhetene ha en z-verdi mellom - 2,58 og + 2,58:



Figur 53 $z = 2,58$ er også viktig å huske!

De to z-verdiene 1,96 og 2,58 brukes mye i forbindelse med konfidensintervall og hypotesetesting.

11.1. Standardkårer (z-verdier)

Siden alle normalfordelinger har samme form, trenger vi ikke tabeller for hver eneste fordeling. Hvis verdiene (x) er normalfordelt med gjennomsnitt μ og standardavvik σ , *transformerer* vi x til såkalte **standardkårer** eller **z-verdier**.

Disse verdiene regner vi ut slik: Først trekker vi gjennomsnittet fra råskåren, og deretter dividerer vi på standardavviket.

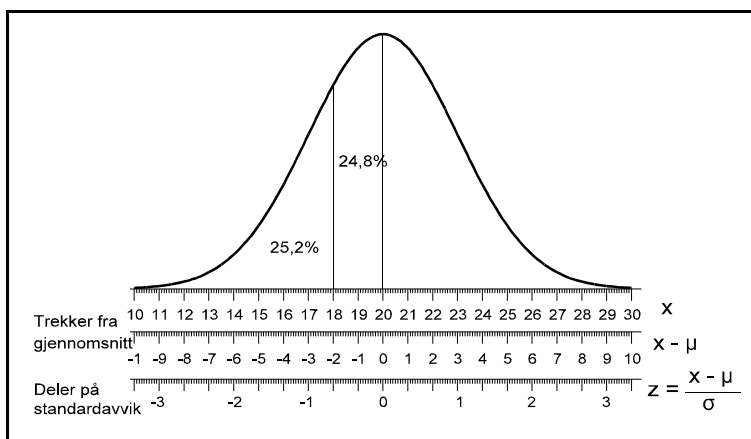
Formelen for dette er $z = \frac{x - \mu}{\sigma}$

Ved denne transformasjonen vil z-verdiene være normalfordelt med gjennomsnitt 0 og standardavvik 1. Standard-skårene forteller oss hvor mange standardavvik en verdi ligger fra fordelingsgjennomsnitt.

Når vi har transformert x-verdiene til z-verdier kan vi bruke en normalfordelingstabell. Her slår vi opp z-verdiene og finner andelen enheter med verdi mellom 0 og z.

Eksempel:

La oss anta at temperaturene i et bokmagasin er tilnærmet normalfordelt med gjennomsnitt $\mu = 20^\circ$ og standardavvik $\sigma = 3^\circ$. Hvor stor andel av dagene er det da en temperatur på under 18° ?



Figur 54 z-transformasjon for å finne andel av dager med temperatur på under 18°

Vi regner ut $\frac{x - \mu}{\sigma} = \frac{18^\circ - 20^\circ}{3^\circ} = \frac{-2^\circ}{3^\circ} \approx -0,67$. Deretter slår vi opp i normalfordelingstabellen på $z = 0,67$ og finner en andel på $0,2486 \cdot 100 \% = 24,86 \%$. Siden kurven er symmetrisk, må det være samme andel mellom $z = -0,67$ og $z = 0$ som mellom 0 og 0,67. Andelen dager med temperatur *under* 18° må da være:

$50\% - 24,86\% = 25,14 \% \approx 25\%$

At det er avrundet litt uvanlig på figuren skyldes at $z = 0,667$. Det er tatt hensyn til at avlesningen i tabellen på $z = 0,67$ gir litt for høyt resultat. I praksis har det liten betydning.

Det er lettest å forstå transformasjonen i to trinn. Først flytter vi gjennomsnittet til 0, ved å trekke μ fra verdiene. Deretter dividerer vi på standardavviket σ .

Skal vi tilbake fra z til x gjennomføres transformasjonen i motsatt rekkefølge: Først multipliserer vi med standardavviket σ , deretter legger vi til μ :

$$x = \mu + (z \cdot \sigma)$$

Dette kan vi også komme fram til ved å ta utgangspunkt i at z-verdien sier hvor mange standardavvik x ligger fra gjennomsnittet.

Eksempel (fortsett):

Vi vil gjerne vite grensene for temperatur slik at vi får med de midterste 95%. Hva blir disse grensene?

Den z-verdien som gir et areal mellom $-z$ og $+z$ på 0,95, er $z = 1,96$ (jfr. **Figur 52**). Vi regner oss da frem til disse grensene:

$$x_1 = \mu + (z \cdot \sigma) = 20^\circ + ((-1,96) \cdot 3^\circ) = 20^\circ - 5,88^\circ \approx 14,1^\circ$$

$$x_2 = \mu + (z \cdot \sigma) = 20^\circ + (1,96 \cdot 3^\circ) = 20^\circ + 5,88^\circ \approx 25,9^\circ$$

Siden normalfordelingstabellen gir oss andelen mellom 0 og forskjellige verdier av z , kan den også brukes til å finne andelen mellom to forskjellige verdier av z .

Eksempel (fortsett):

Vi vil finne andelen av dager som har en temperatur mellom $21,5^\circ$ og $24,5^\circ$. Vi finner de tilhørende z -verdier:

$$z_1 = \frac{x - \mu}{\sigma} = \frac{21,5 - 20}{3} = \frac{1,5}{3} = 0,5 \quad \text{og} \quad z_2 = \frac{x - \mu}{\sigma} = \frac{24,5 - 20}{3} = \frac{4,5}{3} = 1,5$$

Fra normalfordelingstabellen finner vi

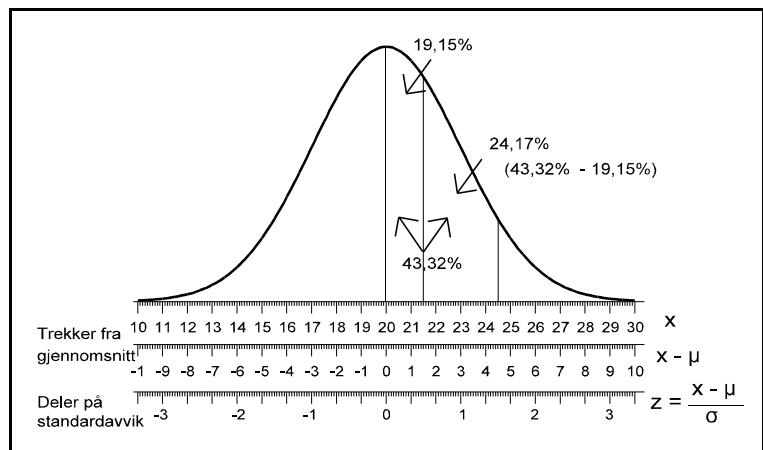
Andelen mellom $z = 0$ og $z = 0,5$ er $0,1915 = 19,15\%$

Andelen mellom $z = 0$ og $z = 1,5$ er $0,4332 = 43,32\%$

Andelen mellom $z = 0,5$ og $z = 1,5$ finner vi ved å trekke de to tallene fra hverandre. Andelen blir $43,32\% - 19,15\% = 24,17\% \approx 24\%$ (se **Figur 55**)

Hvor nøyaktig vi skal avrunde vil være et vurderings spørsmål. Hvis μ og σ er svært nøyaktige, kan vi også forsvare en større nøyaktighet i andelen. Vi bør også ta hensyn til hva vi skal bruke tallet til. Vanligvis brukes prosenter uten desimaler, eller med én desimal. Avrundingen gjennomføres først når hele regnestykket er utført.

Vi kan også tolke andelen som *sannsynlighet*. Vi vil vite sannsynligheten for at temperaturen er mellom $21,5^\circ$ og $24,5^\circ$ en tilfeldig dag.



Sannsynligheten blir $0,2417 = 24,17\% \approx 24\%$.

Figur 55 Finne andel av dager med temperatur $21,5^\circ$ - $24,5^\circ$

Hvis vi avrunder temperaturene til nærmeste grad, betyr en avlesning på 22° at den eksakte temperaturen ligger mellom $21,5^\circ$ og $22,49999\dots^\circ$, osv. Ved en slik avrunding vil vårt resultat bety at sannsynligheten for et måleresultat på 22° , 23° eller 24° er $24,17\% \approx 24\%$.

11.2. Fra andeler til antall

Hvis vi både kjenner det totale antall enheter i populasjonen og andelen enheter med verdier i et intervall, kan vi også finne antall enheter i intervallet.

Eksempel:

Magasintemperaturene er målt hver dag i et år (365 dager), med gjennomsnitt 20° og standardavvik 3° . Hvor mange dager har en temperatur mellom $21,5^\circ$ og $24,5^\circ$?

Vi kjenner andelen av dager som oppfyller kriteriet, nemlig 0,2417. Multipliserer vi dette med totalantallet, finner vi antallet, nemlig $0,2417 \cdot 365 \approx 88$.

Tar vi utgangspunkt i andelen uttrykt i prosenter, regner vi slik: $\frac{24,17\%}{100\%} \cdot 365 \approx 88$

Når det gjelder antall, bør vi selvsagt avrunde til nærmeste hele tall!

Trekker vi et tilfeldig utvalg, kan vi anslå antallet enheter med verdi innen et intervall på samme måte.

Eksempel:

Antallet dager med temperatur mellom $21,5^\circ$ og $24,5^\circ$ i et tilfeldig utvalg på 50 kan vi anslå til å være $0,2417 \cdot 50 \approx 12$.

11.3. Fra andeler til verdier

I visse tilfeller vil vi finne de verdiene som tilsvarer visse andeler. Særlig aktuelt er det å finne en z slik at en gitt andel av enhetene har verdi mellom $-z$ og $+z$. Hittil har vi f.eks. merket oss at 95% av alle enhetene har en z -verdi mellom $-1,96$ og $+1,96$. Tilsvarende vet vi at 99% av alle enhetene har en z -verdi mellom $-2,58$ og $+2,58$. Vi kan så regne videre fra z til x dersom vi kjenner fordelingsgjennomsnitt og standardavvik.

For å finne disse grensene må vi lete opp en *andel* i normalfordelingstabellen og lese av tilhørende z .

Eksempel:

Vi vil gjerne vite grensene for de midterste 90% av temperaturene i magasinet med gjennomsnitt $\mu = 20^\circ$ og standardavvik $\sigma = 3^\circ$. Halvparten av de midterste enhetene må da ligge på hver sin side av gjennomsnittet. Den z -verdien vi leter etter, må være slik at 45%, eller 0,45 av enhetene har verdi mellom 0 og z , mens 0,45 av enhetene har verdi mellom $-z$ og 0.

I normalfordelingstabellen finner vi at $z = 1,65$ svarer til en andel på 0,45. Grensene blir:

$$x_1 = \mu + (z \cdot \sigma) = 20^\circ + ((-1,65) \cdot 3^\circ) = 20^\circ - 4,95^\circ \approx 15^\circ$$

$$x_2 = \mu + (z \cdot \sigma) = 20^\circ + (1,65 \cdot 3^\circ) = 20^\circ + 4,95^\circ \approx 25^\circ$$

12. Samplingfordelingen for gjennomsnitt

Hvordan er sammenhengen mellom populasjonens og et utvalgs gjennomsnitt?

12.1. Sammenheng mellom utvalg og populasjon

Anta at vi har en normalfordelt populasjon med gjennomsnitt μ og standardavvik σ . Fra denne populasjonen trekker vi et utvalg på n enheter. Hva vil vi anslå utvalgets gjennomsnitt til å være?

Eksempel:

La oss tenke oss at vi måler temperaturen i magasinet på 4 forskjellige dager. Disse dagene er et utvalg fra en populasjon av dager, med gjennomsnittlig magasintemperatur på $\mu = 20^\circ$, og standardavvik $\sigma = 3^\circ$.

I **Tabell XXV** er det satt opp tenkte resultater i 4 slike utvalg. Vi ser at utvalgenes gjennomsnitt ligger rundt 20° , men ikke nødvendigvis er eksakt 20° . Standardavvikene ligger rundt 3° , men de varierer ganske mye.

Tabell XXV Magasintemperatur målt i 4 utvalg à 4 dager

Temperaturer i utvalgene	m	s
15°, 20°, 22°, 23°	20,0°	3,6°
17°, 19°, 19°, 21°	19,0°	1,6°
15°, 17°, 22°, 24°	19,5°	4,2°
19°, 20°, 24°, 25°	22,0°	2,9°

Hvis vi trekker mange slike utvalg, og deretter finner gjennomsnittet av gjennomsnittene i utvalgene, bør vi treffe svært nær populasjonens gjennomsnitt, μ .

Med matematisk statistikk kan en bevise at gjennomsnittet av alle tenkelige gjennomsnitt i utvalg av samme størrelse er lik populasjonens gjennomsnitt, μ . Vi kaller utvalgsgjennomsnittene for m , og gjennomsnittet av disse gjennomsnittene for μ_m . Sammenhengen er $\mu_m = \mu$.

Videre legger vi merke til at: Selv om utvalgenes gjennomsnitt ikke er helt lik $\mu = 20^\circ$, ligger disse gjennomsnittene nærmere μ enn det de enkelte observasjonene gjør. Sagt på en annen måte: Fordelingen av gjennomsnittene har mindre spredning enn populasjonens fordeling har.

Sammenhengen mellom standardavviket i fordelingen over gjennomsnitt, σ_m , og populasjonens

standardavvik er slik: $\sigma_m = \frac{\sigma}{\sqrt{n}}$

Dette uttrykket har et eget navn: **Standardfeilen** for utvalgenes gjennomsnitt. (Begrepet *utvalgsfeilen* blir også brukt om denne størrelsen).

Navnet kan vi forstå slik: Hvis vi bruker et utvalgs gjennomsnitt som anslag for populasjonens gjennomsnitt μ , er anslaget usikkert. Det blir vanligvis en liten feil. Standardavviket for denne feilen blir σ_m .

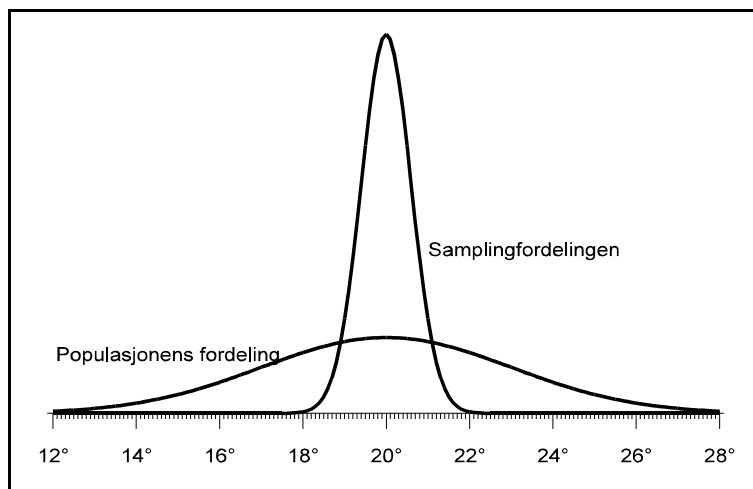
12.2. Samplingfordelingen

Fordelingen av gjennomsnitt i alle tenkelige utvalg på n enheter kalles *samplingfordelingen* for gjennomsnitt i utvalg på n .

Vi har sett at denne fordelingen har gjennomsnitt $\mu_m = \mu$ og standardavvik

$$\sigma_m = \frac{\sigma}{\sqrt{n}}. \text{ Hvis } n \text{ er minst } 25, \text{ vil}$$

samplingfordelingen være tilnærmet normalfordelt. Dette gjelder også selv om populasjonens fordeling ikke er det. Kjenner vi μ , σ , og n , kan vi tegne samplingfordelingen:



Figur 56 Sammenhengen mellom populasjonens fordeling og samplingfordelingen

Eksempel:

I **Figur 56** ser vi sammenhengen mellom de normalfordelte temperaturene og samplingfordelingen for gjennomsnitt i utvalg på 25.

Standardfeilen blir $\sigma_m = \frac{\sigma}{\sqrt{n}} = \frac{3^\circ}{\sqrt{25}} = \frac{3^\circ}{5} = 0,6^\circ$. Utregningen viser at standardavviket til samplingfordelingen er en femtepart av standardavviket til populasjonens fordeling.

Vi kan bruke samplingfordelingen når vi har gitt en populasjon med kjent gjennomsnitt. La oss tenke oss at vi trekker et utvalg, og vil vite hvilken forskjell vi kan vente å finne mellom utvalgets gjennomsnitt og populasjonens gjennomsnitt.

Eksempel:

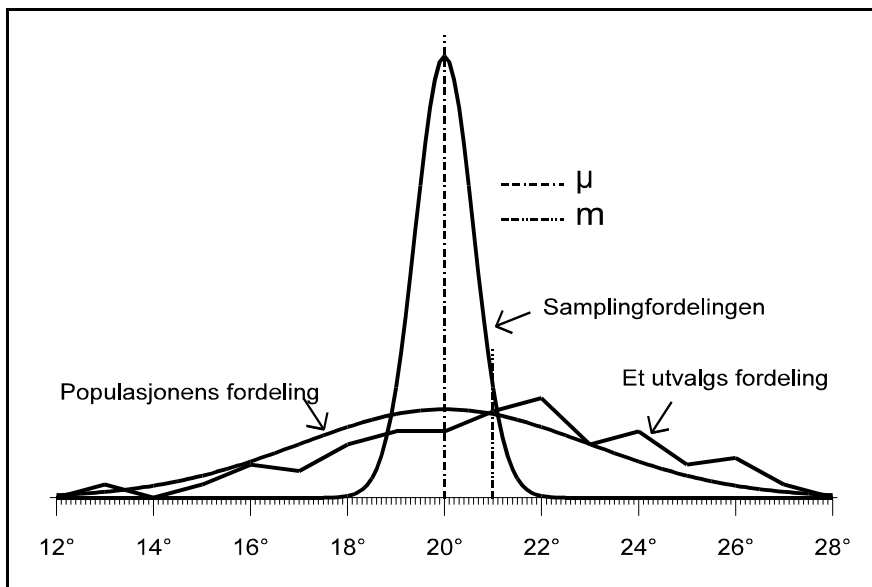
Vi måler temperaturen 25 dager, og finner et gjennomsnitt på $m = 21,0^\circ$. Virker dette rimelig, når populasjonens gjennomsnitt er $\mu = 20,0^\circ$ og standardavviket er $\sigma = 3^\circ$?

Vi sammenligner gjennomsnittet med populasjonens gjennomsnitt, $\mu = 20,0^\circ$. Utvalgets gjennomsnitt ligger 1° høyere, dvs. $m - \mu = 1^\circ$. Vårt spørsmål kan omformuleres slik: Er denne forskjellen liten eller stor, ved et utvalg på 25?

Vårt gjennomsnitt er ett av mange mulige. Fordelingen over alle tenkelige gjennomsnitt i utvalg på 25 kjenner vi som samplingfordelingen over gjennomsnitt.

Vi får et bilde av dette ved å tegne de tre fordelingene i samme diagram. Utvalgets gjennomsnitt, m , er en av alle mulige verdier i samplingfordelingen. Samplingfordelingen er nemlig ikke en fordeling av funne temperaturer, men en fordeling av *gjennomsnittstemperaturer*.

Vi tegner inn utvalgets gjennomsnitt, m . Når vi skal avgjøre om det ligger langt ut fra samplingsfordelingens gjennomsnitt, ser vi m som en verdi i *samplingfordelingen*.



Figur 57 Populasjonens fordeling, et utvalgs fordeling og samplingsfordelingen over gjennomsnitt.

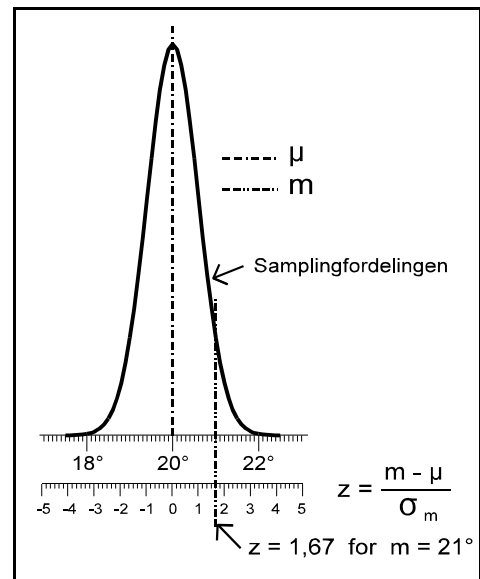
Når vi skal avgjøre om gjennomsnittet på 21° ligger langt fra populasjonens gjennomsnitt på 20° , regner vi ut z -verdien til 21° , men denne gangen i forhold til samplingsfordelingen:

$$z = \frac{m - \mu}{\sigma_m} = \frac{21^\circ - 20^\circ}{0,6^\circ} = \frac{1^\circ}{0,6^\circ} \approx 1,67$$

Slår vi opp $z = 1,67$ i normalfordelingstabellen, finner vi en andel på $0,4525 \approx 45\%$. Utenfor $z = 1,67$ er det da en andel under normalfordelingskurven på $50\% - 45\% = 5\%$. Vi kan kalle dette for samplingsfordelingens «hale».

Dette kan vi forstå slik: Av alle tenkelige gjennomsnitt i utvalg på 25, ligger 5% utenfor $m = 21^\circ$, dvs. på «halen», sett i samplingsfordelingen. Hvis vi vil undersøke hvor mange utvalgsgjennomsnitt som ligger lenger fra populasjonens gjennomsnitt enn $m = 21^\circ$, må vi regne andelen på begge «halene» under samplingsfordelingen. Den venstre halen er for m lavere enn 19° . Her blir $z = -1,67$; og andelen blir også her 5%. Andelen for begge halene blir $5\% + 5\% = 10\%$.

Konklusjonen blir at 10% av alle utvalgs gjennomsnitt ligger lenger fra populasjonens gjennomsnitt enn det vi har funnet.



Figur 58 Samplingsfordelingen

13. Konfidensintervall for gjennomsnitt

Hvordan kan vi slutte fra utvalgsgjennomsnitt til populasjongjennomsnitt?

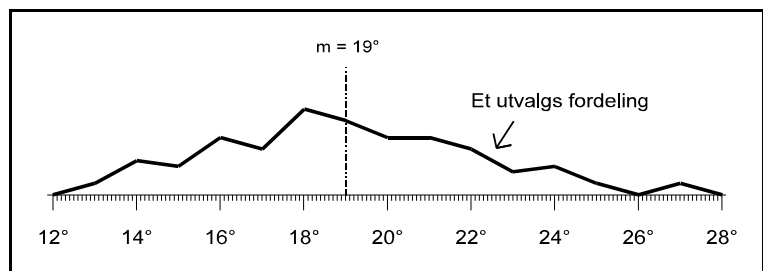
La oss anta at vi skal finne gjennomsnittlig verdi for en variabel. Populasjonen er stor, så det blir for dyrt, eller praktisk umulig å undersøke *alle* enhetene. Vi trekker et tilfeldig utvalg. I et slikt utvalg har alle enhetene like stor sannsynlighet for å bli trukket ut.

Det beste *estimatet* eller anslaget vi kan gi for populasjonens gjennomsnitt, er utvalgets gjennomsnitt. Spørsmålet er: Hvor trygt er det å bruke utvalgsgjennomsnittet? Hvilke grenser ligger feilen da innenfor?

Vi støter på begrepet *feilmargin* i meningsmålingene. Denne feilmarginen angir hvor stor forskjellen mellom populasjongjennomsnittet og utvalgsgjennomsnittet kan være.

Eksempel:

Vi har målt et magasins temperatur 25 forskjellige dager. Dagene er tilfeldig valgt, og vi har funnet et gjennomsnitt i utvalget på $m = 19,0^\circ$. Standardavviket i utvalget er $s = 3^\circ$.



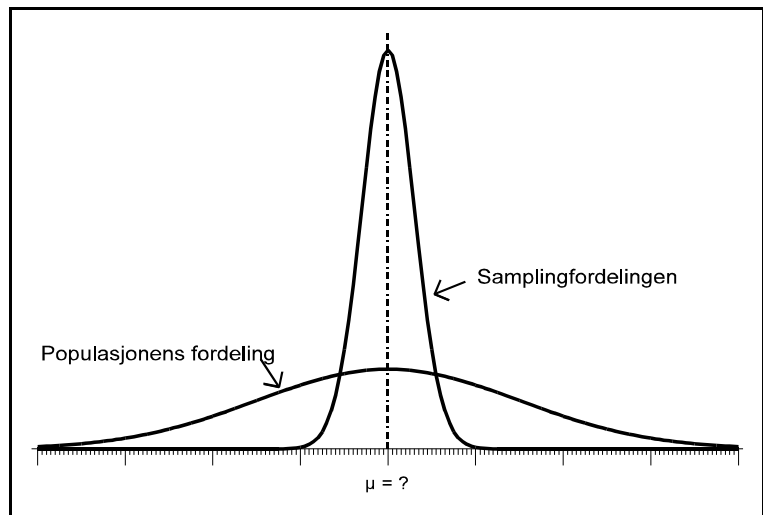
Figur 59 Utvalgets fordeling

Det beste anslaget vi kan gi for populasjonens gjennomsnitt, μ , er

det samme som utvalgets gjennomsnitt, m . Det betyr at vårt anslag blir $\mu \approx m = 19,0^\circ$. Tilsvarende er det beste anslaget vi kan gi for populasjonens standardavvik $\sigma \approx s = 3^\circ$.

Likevel er dette anslaget ikke nøyaktig. Vi vet at vi sannsynligvis gjør en feil, fordi det vanligvis er en viss forskjell mellom m og μ . Denne forskjellen, $m - \mu$, studerte vi i forrige kapittel. Der så vi også at alle mulige utvalgsgjennomsnitt var tilnærmet normalfordelt, med standardavvik σ_m .

Vi ser derfor på populasjonens fordeling og samplingfordelingen med ukjent μ . Vi vet fra forrige kapittel at sannsynligheten for at m ligger mer enn 1,96 standardfeil fra μ er 5%, mens det er 95% sannsynlig at m ligger mindre enn $1,96 \cdot \sigma_m$ fra μ .



Figur 60 Populasjonens gjennomsnitt er ukjent

Vi regner ut:

$$1,96 \cdot \sigma_m = 1,96 \cdot \frac{\sigma}{\sqrt{n}} = 1,96 \cdot \frac{3^\circ}{\sqrt{25}} = 1,96 \cdot \frac{3^\circ}{5} = 1,96 \cdot 0,6^\circ = 1,18^\circ$$

Vårt resonnement har gitt dette resultatet:

Det er 95% sannsynlig at forskjellen mellom μ og m er mindre enn $1,18^\circ$.

Ut fra dette kan vi regne ut et sannsynlig intervall for μ , nemlig mellom:

$$19,0^\circ - 1,18^\circ = 17,82^\circ \approx 17,8^\circ; \text{ og}$$

$$19,0^\circ + 1,18^\circ = 20,18^\circ \approx 20,2^\circ.$$

Vi kan da omformulere resultatet:

*Det er 95% sannsynlig at μ ligger mellom grensene $17,8^\circ$ og $20,2^\circ$. Vi sier at **konfidensintervallet** for μ er intervallet $[17,8^\circ, 20,2^\circ]$. Dette intervallet er imidlertid knyttet til den sannsynligheten vi har valgt, nemlig 95%.*

13.1. Sikkerhetsnivå

Når vi skal angi grensene for en populasjons gjennomsnitt, kan vi sjelden være helt sikre. Det er mulig vi har data fra et svært spesielt utvalg. Vi kan vanligvis ikke gi et 100% sikkert intervall, men f.eks. et intervall som er 95% sikkert. Sannsynligheten for at μ ligger innenfor konfidensintervallet, kaller vi **sikkerhetsnivået** til konfidensintervallet. De vanligste sikkerhetsnivåene er 95% og 99%.

Eksempel (fortsett):

Hvis vi i eksemplet skulle hatt sikkerhetsnivået 99%, er den eneste forskjellen at $z = 2,58$. Vi vet nemlig fra før (jfr. **Figur 53**) at 99% av enhetene ligger innenfor en z -verdi på 2,58 fra gjennomsnittet. Tilsvarende må 99% av alle tenkelige gjennomsnitt ligge innenfor 2,58 standardfeil fra gjennomsnittet.

Med 99 % sikkerhet må nå forskjellen mellom μ og m være høyst:

$$2,58 \cdot \sigma_m = 2,58 \cdot \frac{\sigma}{\sqrt{n}} = 2,58 \cdot \frac{3^\circ}{\sqrt{25}} = 2,58 \cdot \frac{3^\circ}{5} = 2,58 \cdot 0,6^\circ = 1,548^\circ$$

dvs. μ ligger mellom grensene $19^\circ - 1,548^\circ = 17,452^\circ \approx 17,5^\circ$ og $19^\circ + 1,548^\circ \approx 20,5^\circ$.
Konfidensintervallet blir $[17,5^\circ, 20,5^\circ]$.

13.2. Formelen for konfidensintervallet

Resultatene fra det gjennomgående eksemplet i dette kapitlet kan generaliseres slik:

Grensene for konfidensintervallet for populasjonens gjennomsnitt er:

$$\mu = m - z \cdot \sigma_m = m - z \frac{\sigma}{\sqrt{n}} \quad \text{og} \quad \mu = m + z \cdot \sigma_m = m + z \frac{\sigma}{\sqrt{n}}$$

Uttrykket z , som inngår i formelen, er knyttet til sikkerhetsnivået. Ved sikkerhetsnivå 95% er $z = 1,96$; mens vi bruker $z = 2,58$ ved et sikkerhetsnivå på 99%. Ved andre sikkerhetsnivåer må vi finne z i en normalfordelingstabell.

Vi kan sammenfatte dette slik:

Grensene for konfidensintervallet er at μ ligger innenfor $\mathbf{m} \pm z \cdot \sigma_m = \mathbf{m} \pm z \frac{\sigma}{\sqrt{n}}$

dvs. $m - z \cdot \sigma_m \leq \mu \leq m + z \cdot \sigma_m$

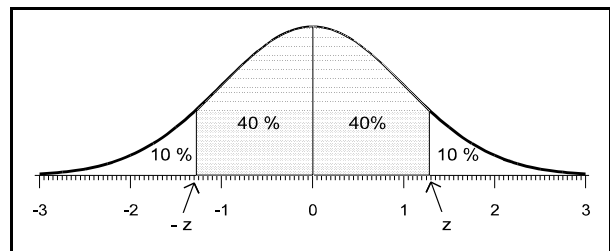
eller $\mathbf{m} - z \frac{\sigma}{\sqrt{n}} \leq \mu \leq \mathbf{m} + z \frac{\sigma}{\sqrt{n}}$

Hvis populasjonens standardavvik, σ , ikke er kjent, kan vi sette $\sigma \approx s$, og grensene bli da:

$$\mathbf{m} \pm z \frac{s}{\sqrt{n}}$$

Eksempel:

Vi vil finne z tilsvarende et sikkerhetsnivå på 80%. Sett i forhold til samplingfordelingen må halvparten av denne andelen, dvs. $40\% = 0,4$ ligge på hver side av gjennomsnittet. Vi leter opp 0,4 i normalfordelingstabellen, og finner $z = 1,28$ som den verdien som passer best. Formelen for konfidensintervallet blir da $\mu = m \pm z \cdot \sigma_m = m \pm 1,28 \cdot \sigma_m$.



Figur 61

13.3. Andre modeller og andre konfidensintervall

I dette kompendiet blir det bare gitt metoden for å finne konfidensintervall for gjennomsnitt. Det finnes metoder for å finne konfidensintervall for andre statistiske mål, f. eks. konfidensintervall for korrelasjonskoeffisienten.

I meningsmålinger over partipreferanser er det *andeler* som blir presentert. Også for andeler finnes det formler for konfidensintervall.

Den formelen som blir presentert her, bygger bl.a. på forutsetningen om at vi slutter fra et tilfeldig utvalg, hvor alle enhetene har samme sannsynlighet for å være med. For andre utvalgsmetoder, f.eks. *stratifiserte utvalg*, finnes metoder for å angi konfidensintervall.

Hvis n (utvalgets størrelse) er mindre enn 25, vil normalfordelingen være for unøyaktig for å beskrive samplingfordelingen. Vi har da en annen fordeling (t-fordelingen), som har store likhetstrekk med normalfordelingen. Det finnes egne tabeller for t-fordelingen, hvor vi bruker z på samme måte som i normalfordelingen.

14. Utvalgets størrelse

Hvor mange enheter må vi ha i utvalget?

14.1. Konfidensintervall og utvalgsstørrelse

Fra forrige kapittel vet vi at konfidensintervallet avhenger av antallet enheter i utvalget. Jo større utvalget er, desto mindre blir konfidensintervallet, fordi grensene for konfidensintervallet er:

$$m \pm z \frac{\sigma}{\sqrt{n}}$$

Her inngår uttrykket \sqrt{n} i nevneren i det siste leddet. Hvis n øker, vil også \sqrt{n} øke, og brøken vil bli mindre. Derfor blir også konfidensintervallet smalere.

Eksempel:

Vi har sett at grensene for konfidensintervallet med sikkerhetsnivå 95% er $m \pm 1,96 \cdot \sigma_m$. I vårt temperatur-eksempel ($m = 19^\circ$, $s = 3^\circ$), fant vi at ved utvalg på 25 ble:

$$\sigma_m = \frac{\sigma}{\sqrt{n}} = \frac{3^\circ}{\sqrt{25}} = \frac{3^\circ}{5} = 0,6^\circ \quad \text{Konfidensintervallet ble } [17,8^\circ, 20,2^\circ]$$

Hvis utvalget derimot er på 100, blir:

$$\sigma_m = \frac{\sigma}{\sqrt{n}} = \frac{3^\circ}{\sqrt{100}} = \frac{3^\circ}{10} = 0,3^\circ.$$

Her blir grensene for μ : $m \pm z \cdot \sigma_m = 19^\circ \pm 1,96 \cdot 0,3^\circ = 19^\circ \pm 0,588^\circ = 19^\circ \pm 0,6^\circ$

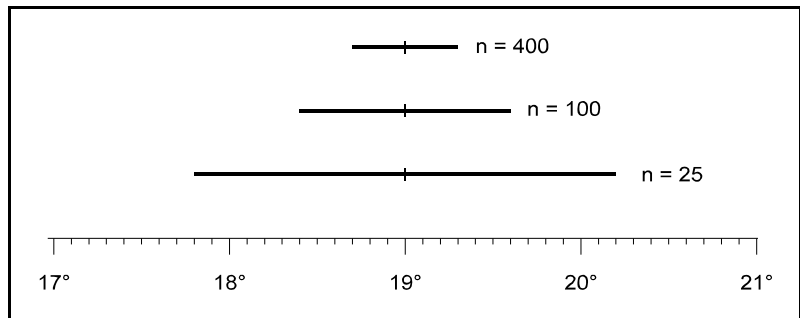
Med utvalg på 100 enheter blir konfidensintervallet $[18,4^\circ, 19,6^\circ]$, altså halvparten så bredt som når utvalget er på 25 enheter.

Vi ser at det som bestemmer konfidensintervallets bredde, er uttrykket:

$$z \cdot \sigma_m = z \cdot \frac{\sigma}{\sqrt{n}}$$

Videre ser vi at $|\mu - m| \leq z \cdot \sigma_m$, fordi $m - z \cdot \sigma_m \leq \mu \leq m + z \cdot \sigma_m$

Symbolet $|a|$ betyr *tallverdien* av a , dvs. tallet uten fortegn. $|\mu - m|$ er altså uttrykket $(\mu - m)$ uten fortegn.



Figur 62 Konfidensintervallet blir smalere jo flere enheter det er i utvalget

14.2. Feilmargin

Feilmarginen angir hvor stor forskjell vi må være forberedt på at det kan være mellom μ og m . Vi kan omformulere det slik:

Feilmarginen er den øvre grensen for uttrykket $|\mu - m|$ med et gitt sikkerhetsnivå.

Siden vi vet at $|\mu - m| \leq z \cdot \sigma_m$, kan feilmarginen angis som: $z \cdot \sigma_m$

I dette kompendiet brukes symbolet δ (delta, liten gresk d) for feilmargin.

Vi kan da angi feilmarginen som $\delta = z \cdot \sigma_m$.

14.3. Valg av utvalgsstørrelse

Feilmargin, sikkerhetsnivå og utvalgsstørrelse er matematisk knyttet til hverandre.

Ved utvalgsundersøkelser er ofte problemstillingen: Vi vil godta en viss gitt feilmargin. Vi har også bestemt et sikkerhetsnivå. Hvor stort utvalg bør vi da velge?

Vi har funnet et uttrykk for feilmarginen, nemlig $\delta = z \cdot \sigma_m$. Vi kan finne utvalgsstørrelsen hvis vi har valgt en feilmargin. Vi må finne n uttrykt ved de andre størrelsene, og løse en likning:

$$\delta = z \cdot \sigma_m = z \cdot \frac{\sigma}{\sqrt{n}} \quad \text{Vi opphøyer i 2. potens:}$$

$$\delta^2 = z^2 \cdot \frac{\sigma^2}{n} \quad \text{Vi multipliserer med } n:$$

$$n \cdot \delta^2 = z^2 \cdot \sigma^2 \quad \text{Vi dividerer på } \delta^2:$$

$$n = \frac{z^2 \cdot \sigma^2}{\delta^2} = \left(\frac{z \cdot \sigma}{\delta} \right)^2$$

Hvis vi ikke kjenner σ , kan vi bruke s som en tilnærming:

$$n = \left(\frac{z \cdot s}{\delta} \right)^2$$

Uttrykket vi har funnet, brukes altså til å velge utvalgets størrelse.

15. Hypotesetesting

Hypotesetesting går ofte ut på å bekrefte eller avkrefte antakelser om forhold ved *populasjoner* som er så store, at vi ikke kan undersøke *alle* enhetene. Ved å trekke *utvalg* fra populasjonene kan vi likevel, med en kjent grad av sikkerhet, si noe om forhold ved populasjonene selv.

Eksempel:

Vi har undersøkt et tilfeldig utvalg på 100 bøker av en bestemt type, og funnet at gjennomsnittsprisen (m) er på 214 kroner. Standardavviket (s) var på 60 kroner.

Forhandleren av denne typen bøker sier at gjennomsnittsprisen (μ) er på 200 kroner.

Har han rett, eller tar han feil? Vi har to muligheter. Han kan ha rett. Vi betegner denne påstanden H_0 . Dette kalles en *nullhypotese*. Den motsatte påstanden (han tar feil), betegner vi H_A , og kaller denne påstanden den *alternative hypotesen*.

Har forhandleren rett, er nullhypotesen riktig, dvs.

$H_0: \mu = 200$ kroner, eller skrevet slik: $H_0: \mu - 200 \text{ kroner} = 0$

Tar han feil, er derimot den alternative hypotesen riktig:

$H_A: \mu \neq 200$ kroner, eller skrevet slik: $H_A: \mu - 200 \text{ kroner} \neq 0$

15.1. Hypotese belyst ved konfidensintervall

Eksempel (fortsett):

Vi finner konfidensintervallet for μ . Grensene for dette er

$$\mu = m \pm z \cdot \frac{\sigma}{\sqrt{n}} \approx m \pm z \cdot \frac{s}{\sqrt{n}} = 214 \pm z \cdot \frac{60}{\sqrt{100}} = 214 \pm z \cdot 6$$

Grensene for konfidensintervallet blir:

Med sikkerhetsnivå 95 %: $(214 \pm 1,96 \cdot 6)$ kroner $\approx (214 \pm 11,8)$ kroner
dvs. kr. 202,20 og kr. 225,80

Med sikkerhetsnivå 99 %: $(214 \pm 2,58 \cdot 6)$ kroner $\approx (214 \pm 15,5)$ kroner
dvs. kr. 198,50 og kr. 229,50

Hvis vi nøyer oss med å være 95% sikre, kan vi si at $\mu > 200$ kr. Sagt på en annen måte: Siden H_0 er lite sannsynlig, er den andre muligheten, nemlig H_A , svært sannsynlig.

Hvis vi derimot vil være 99% sikre, er det mulig at $\mu = 200$ kr. Men kan vi si at det er sikkert, eller at det er sannsynlig? Nei, vi vet ikke noe mer enn at populasjonens gjennomsnitt med stor sannsynlighet ligger mellom kr. 198,50 og kr. 229,50. Populasjonens gjennomsnitt μ kan like godt være på kr. 228 som på kr. 200. Vi kan med andre ord verken sannsynliggjøre nullhypotesen eller den alternative hypotesen!

På sett og vis viser dette eksemplet at det å teste en hypotese om gitt gjennomsnitt er det motsatte av å finne et konfidensintervall.

15.2. Testing av $H_0: \mu = k$ (der k er en gitt størrelse)

Nullhypotesen har ikke nødvendigvis noe med tallet 0 å gjøre. Nullhypotesen er derimot en *entydig* hypotese. Ved omskrivingen $H_0: \mu - k = 0$ (her $\mu - 200 \text{ kr} = 0$) er det lettere å huske hva som er nullhypotese.

Vi starter med å anta at nullhypotesen er riktig, og ser hvor det fører hen. Vi antar at gjennomsnittlig innkjøpspris virkelig *er* 200 kroner. Hvor stor sannsynlighet er det da for at vi får et *så avvikende* resultat som det vi har fått?

Vårt utvalgs gjennomsnitt på 214 kroner avviker fra μ med 14 kroner. Er dette mye eller lite? Da må vi vurdere hvor stort avvik fra μ vi kan regne med ut fra tilfeldigheter alene. Fordelingen av alle tenkelige gjennomsnitt kjenner vi som samplingfordelingen, og vi må studere den.

Standardavviket for samplingfordelingen er standardfeilen, $\sigma_m = \frac{\sigma}{\sqrt{n}} = \frac{\text{kr. } 60}{\sqrt{100}} = \text{kr. } 6$

For å finne sannsynligheten for å få så pass avvikende gjennomsnitt må vi finne z-verdien for m (vårt oppnådde gjennomsnitt) på *samplingfordelingen*. Vi må finne z-verdien til $m = 214$ på samplingfordelingen, nemlig:

$$z = \frac{m - \mu}{\sigma_m} = \frac{214 - 200}{6} = \frac{14}{6} \approx 2,33$$

Slår vi opp i normalfordelingstabellen ser vi at sannsynligheten for et såpass stort avvik på *en «hale»* er $50\% - 49.01\% \approx 1\%$. På begge haler blir det dobbelt så mye: ca. 2%.

Det er med andre ord lite sannsynlig at vi kan få et så pass avvikende resultat ut fra slump alene. Nullhypotesen er derfor urimelig. Vi *forkaster* den, og godtar den alternative hypotesen.

Samplingfordelingen er et nødvendig mellomledd. Vi bruker den for å se om det er stor sannsynlighet for å få et gjennomsnitt i vårt utvalg som *avviker* så mye fra populasjonsgjennomsnittet som det vi faktisk har observert.

15.3. Signifikansnivå

Vi kunne gjerne sagt at det er 2% sannsynlighet for at vi får et resultat som er så avvikende som det vi har fått. Dette er bakgrunnen for at vi godtar den alternative hypotesen. Spørsmålet er likevel: Hva er *liten* sannsynlighet? Det er jo dette som hele resonnementet bygger på!

Som konvensjon opererer vi ofte med to forskjellige grenser for hva som er lite sannsynlig, 1% og 5%. Andre grenser kan også forekomme, f.eks. 10% (ti prosent) og 5‰ (fem promille). Grensen kalles *signifikansnivået*.

Signifikansnivå bør velges før vi gjennomfører hypotesetesten. La oss si at vi velger et signifikansnivå på 5%. Hvis det er *mindre* enn 5% sannsynlighet for at vi skal få et så avvikende resultat som det oppnådde, forkaster vi derfor nullhypotesen.

Hvis nullhypotesen forkastes, og vi godtar den alternative hypotesen, sier vi at resultatet er *signifikant på 5% nivået*.

I vårt eksempel kan vi si at resultatet $\mu \neq$ kr. 200 er signifikant på 5% nivået. Resultatet er derimot ikke signifikant på 1% nivået.

Signifikansnivå velger vi ut fra hvor sikre vi vil være på de konklusjonene vi trekker. Hvis vi f.eks. vil sjekke om folk med påståtte parapsykologiske evner kan tippe hvilket kort du ser på oftere enn hver 52. gang, vil vi antakelig ønske å være svært sikre. Her kan det være aktuelt å bruke et signifikansnivå på 1‰, noe som vi kan betrakte som et *strengt* nivå.

Hvis vi derimot skal undersøke om en medisin mot en antatt uhelbredelig sykdom har virkning (alternativ hypotese) eller om den ikke har det (nullhypotese), vil vi kanskje ikke være så strenge. Skal vi være 99% sikre? En 10% mulighet for at medisinen ikke er virksom kan kanskje tillates for å ta den i bruk for alvorlig syke. Da vil den jo med 90% sannsynlighet kunne redde liv.

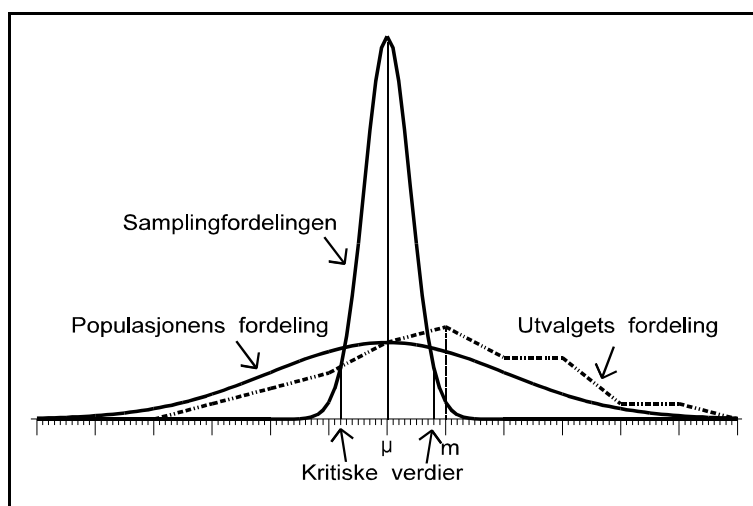
15.4. Kritisk verdi

Som vi så, forkastet vi nullhypotesen i vårt eksempel hvis sannsynligheten for å oppnå et så avvikende resultat som vi oppnådde var under det gitte signifikansnivå. Vi regnet ut en z-verdi i samplingfordelingen, ut fra formelen:

$$z = \frac{m - \mu}{\sigma_m} = \frac{m - \mu}{\frac{\sigma}{\sqrt{n}}}$$

I vårt tilfelle var resultatet signifikant på 5% nivået, fordi sannsynligheten for å oppnå så stor forskjell mellom utvalgets gjennomsnitt og populasjonens gjennomsnitt var på 2%, altså mindre enn 5%. Men ut fra det vi vet om normalfordelingen, vil sannsynligheten bli mindre enn 5% hvis z i tallverdi er større enn 1,96. I figuren overfor er det også tegnet en utvalgsfordeling hvor gjennomsnittet, m, ligger *utenfor* kritisk verdi.

Vi kaller $z = 1,96$ den *kritiske verdien* for 5% nivået.



Figur 63 Her ligger m utenfor kritisk verdi.
 H_0 forkastes

Tilsvarende er $z = 2,58$ den kritiske verdien for 1% nivået.

Hvis tallverdien av z ligger over den kritiske grensen, forkaster vi nullhypotesen, og godtar den alternative hypotesen. Hvis tallverdien ligger under den kritiske grensen, har vi ikke grunn til å forkaste H_0 .

15.5. Andre hypotesetester

En svært vanlig hypotesetest er å teste om to populasjoner har samme eller forskjellig gjennomsnitt. Da blir $H_0 : \mu_1 = \mu_2$. En kjenner gjennomsnittene (m_1 og m_2) i utvalg fra de to populasjonene.

Her forkastes nullhypotesen hvis uttrykket $z = \frac{m_1 - m_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ er større enn den kritiske verdien.

Kritisk verdi er også her 1,96 på 5% nivået og 2,58 på 1% nivået.

Har vi beregnet korrelasjonen mellom to variabler i et utvalg, kan vi teste om denne korrelasjonen er signifikant forskjellig fra 0. Dette blir ikke gjennomgått her.

Et spesialtilfelle for signifikanstesting av korrelasjon gjelder variabler på *nominalnivå*. Der bruker vi den såkalte χ^2 -testen (χ er den greske bokstaven «kji», testen kalles derfor kjikvadrat-testen).

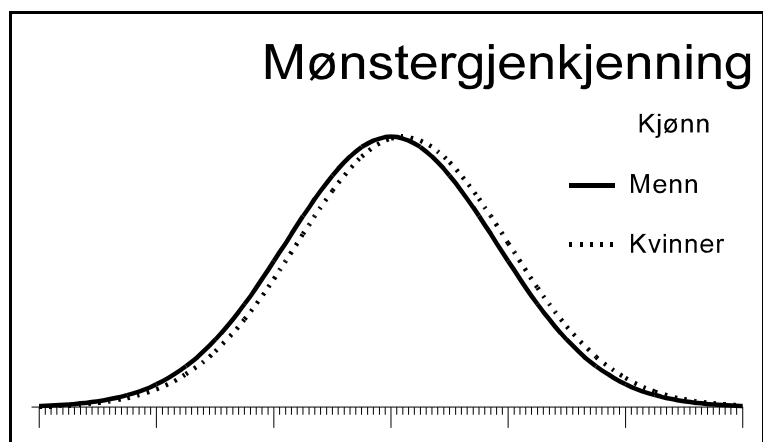
15.6. Begrensninger ved bruk av hypotesetesting

Innen visse tradisjoner i samfunnsfag har hypotesetesting fått svært stor plass. Ukritisk bruk av resultater som bygger på hypotesetesting kan skape store misforståelser.

1. Et signifikant resultat betyr at det er funnet en forskjell eller en sammenheng. Signifikans er ikke det samme som størrelse eller styrke.

La oss anta at vi har to populasjoner, f. eks. kvinner og menn. Vi er interessert i å se om det er forskjeller i en gitt egenskap mellom disse populasjonene.

Finner vi at det *er* en forskjell, behøver ikke dette bety stort. La oss anta at evnen til å gjenkjenne detaljer i mønsteret i en genser ser ut som i **Figur 64**.



Figur 64

Som vi ser, er det en viss forskjell på de to fordelingene, og en klar forskjell i gjennomsnittene. Men i et generelt perspektiv er fordelingene (kurvene) *nesten* sammenfallende.

Hvis vi hadde valgt store utvalg av menn og kvinner, f.eks. 10.000 av hver, ville vi fått et svært signifikant resultat, selv om den underliggende forskjellen er liten.

2. Et signifikant resultat sier ingen ting om årsakene til at resultatet er signifikant.

Igjen kan vi se på forskjeller mellom kjønn. Om vi finner kjønnsforskjeller (f.eks. når det gjelder språkbeherskelse eller tekniske ferdigheter), er det ikke gitt at disse er naturgitte (genetiske) forskjeller. Forskjellene kan godt skyldes miljø eller oppdragelse. Det må *andre* resonnementer til enn de rent statistiske for å bevise årsakssammenhenger.

3. Bruker vi mange signifikanstester samtidig eller etter hverandre, vil alltid noen av dem slå positivt ut.

La oss anta at vi har 100 forskjeller vi vil signifikant teste. Vi bruker et signifikansnivå på 5%. Vi forkaster da en nullhypotese hvis det er 5% sjanse eller mindre for at vi oppnår et så avvikende resultat ut fra slump.

Men i noen av tilfellene kan likevel et avvikende resultat inntreffe. Av de 100 testene kan vi vente at 5 stykker slår positivt ut *selv om* nullhypotesen gjelder! Ved å peke ut forskjeller eller sammenhenger ved disse 5 målingene, gjør forskeren en metode-feil. Hun burde korrigert for at flere signifikanstester blir utført samtidig.

4. Signifikanstestene bygger på forutsetningen om tilfeldig utvalg.

Dette er et sterkt krav i samfunnsforskning. Ved spørreundersøkelser vil bortfallet lett være systematisk. De som svarer, kan være de som er interessert i problemstillingen undersøkelsen skal belyse, eller de som har fordeler av undersøkelsen. Det kan også være andre systematiske skjevheter. For å kunne generalisere, må vi også korrigere for bortfall.

16. Formler

Gjennomsnitt og standardavvik:

$$m = \frac{\sum x}{n} \qquad m = \frac{\sum f \cdot x}{n} \text{ ved frekvensfordelinger}$$

$$s = \sqrt{\frac{\sum (x - m)^2}{n - 1}} \qquad s = \sqrt{\frac{\sum f \cdot (x - m)^2}{n - 1}} \text{ ved frekvensfordelinger}$$

$$n = \sum f$$

Median og kvartilavvik:

$$\text{md er verdien til enhet nr. } q_2 = \frac{n + 1}{2}$$

$$Q_1 \text{ er verdien til enhet nr. } q_1 = \frac{n}{4} + \frac{1}{2}$$

$$Q_3 \text{ er verdien til enhet nr. } q_3 = \frac{3 \cdot n}{4} + \frac{1}{2}$$

$$Q = \frac{Q_3 - Q_1}{2}$$

Korrelasjon og regresjon:

$$r = \frac{\sum (x - m_x) \cdot (y - m_y)}{(n - 1) \cdot s_x \cdot s_y}$$

$$y = (a \cdot x) + b, \text{ der } a = \frac{\sum (x - m_x) \cdot (y - m_y)}{(n - 1) \cdot s_x^2} \text{ og } b = m_y - (a \cdot m_x)$$

Normalfordeling og konfidensintervall:

$$z = \frac{x - \mu}{\sigma} \qquad x = \mu + (z \cdot \sigma)$$

$$\sigma_m = \frac{\sigma}{\sqrt{n}} \qquad z = \frac{m - \mu}{\sigma_m}$$

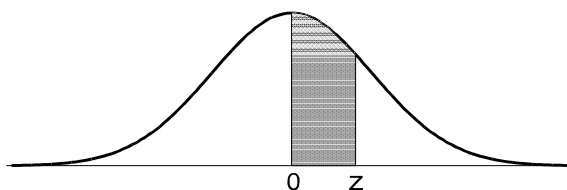
$$\text{Grenser for } \mu: m \pm z \frac{\sigma}{\sqrt{n}} \text{ dvs. } m - z \frac{\sigma}{\sqrt{n}} \leq \mu \leq m + z \frac{\sigma}{\sqrt{n}}$$

$$\text{Utvalgets størrelse: } n = \left(\frac{z \cdot \sigma}{\delta} \right)^2$$

17. Tabell over normalfordelingen

Tabellen viser andelen under normalfordelingskurven mellom 0 og z.

Linjene: z-verdi med 1. desimal
Kolonnene: 2. desimal i z-verdien



z	,00	,01	,02	,03	,04	,05	,06	,07	,08	,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,9	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000

18. Register

μ (my)	28	Gauss-fordelingen	51
σ (sigma)	35	Gjennomsnitt	28
Σ (stor sigma)	29	EXCEL	32
σ_m (standardfeil)	57	Grafiske fremstillinger	23
χ^2 -testen	69	Histogram	24
Alternativ hypotese	66	Hypotesetesting	66
Aritmetisk gjennomsnitt	28	Induktiv statistikk	5, 60
EXCEL	32	Intervallnivå	11
Avhengig variabel	40, 42	Intervallskala	10
Beskrivende statistikk	4	Kakediagram	24
Bivariat fordeling	40	kf (kumulert frekvens)	22
Celler	13	Kjikkvadrat-testen	69
Data	6	Klasser	8
Gruppering	18	Knapperader	
Datamatrikse	12	EXCEL	14
EXCEL	15	Konfidensintervall for gjennomsnitt	61
Deskriptiv statistikk	4, 5	Kontinuerlig skala	11
Diagrammer	23	Korrelasjon	40
EXCEL	26	EXCEL	50
Diagramveiviser		Korrelasjonskoeffisienten	41
EXCEL	26	Kritisk verdi	68
Dikotome variabler	11	Kumulative tabeller	21
Diskontinuerlig skala	11	Kumulativt diagram	25
Diskontinuerlige variabler	11	Kumulerte frekvenser	21
Diskrete variabler	11	EXCEL	22
Enheter	6	Kvartilavvik	38
Estimat	61	Kvartiler	36
EXCEL	13	EXCEL	39
f (frekvens)	9	Frekvensfordelinger	37
Feilmargin	61, 65	Linjediagram	23
Figurer	23	m (aritmetisk gjennomsnitt)	28
Forholdstalls-skala	11	md (median)	30
Forholdstallsnivå	11	Median	30
Formater		EXCEL	32
EXCEL	15	Frekvensfordelinger	37
Formellinje		Menylinje	
EXCEL	14	EXCEL	14
Formler	71	Modalverdi	28
Frekvens	9	Modus	28
Frekvenspolygon	23	EXCEL	32
Frekvenstabeller	17	Multivariat fordeling	40
EXCEL	20	my (μ)	28

Målenivå	10	Sigma	29, 35
Diagram	24	Signifikansnivå	67
Korrelasjon	44	Signifikanstesting	69
Sentraltendens	32	Sikkerhetsnivå	62
Spredning	39	Skjeve fordelinger	31
n (antall enheter)	7, 29	Slutningsstatistikk	5, 60
Navneboks		Spredning	34
EXCEL	14	EXCEL	39
Nedre kvartil	36	Spredningsdiagram	40
Nominalnivå	11	EXCEL	49
Nominalskala	10	Standardavvik	34
Normalfordelingen	51	EXCEL	39
Tabell	72	Standardfeil for gjennomsnitt	57
Nullhypotese	66	Standardskårer	53
Ordinalnivå	11	Statuslinje	
Ordinalskala	10	EXCEL	14
Pearsons r	41	Stigningstall	46
Populasjon	6, 51, 57	Stolpediagram	23
Produkt-moment-korrelasjon	41	Styrke	
Prosentvise tabeller	18	Korrelasjon	41
Punkttdiagram	40	Søylediagram	23
EXCEL	49	t-fordelingen	63
Q (kvartilavvik)	38	Tabeller	17
q_1 (enhetsnr. for nedre kvartil)	36	Tallverdi	64
Q_1 (nedre kvartil)	36	Typetall	28
q_2 (enhetsnr. for median)	30	Uavhengig variabel	40, 42
q_3 (enhetsnr. for øvre kvartil)	36	Undersøkelsesenheter	6
Q_3 (øvre kvartil)	36	Univariat fordeling	40
r (korrelasjonskoeffisient)	41	Utvalg	7, 57
Rangkorrelasjon	44	Utvalgsfeil	57
Ratio-skala	11	Variabler	7
Regneark	12	Variasjonsbredde	34
Regresjon	45	Verdier	7
EXCEL	49	Verktøylinjer	
Regresjonskoeffisienter	46	EXCEL	14
Relative tabeller	17	X-aksen	23
Råskårer	12	Y-aksen	23
s (standardavvik)	34	z-verdier	53
Samplingfordelingen for gjennomsnitt	57	Øvre kvartil	36
Sentraltendens	28		