Kvantitativ metode for lærere

Roar Bakken Stovner

Innhold

Forord

I denne læreboken dekker vi innholdet i et innføringskurs i kvantitative metoder. Boka er basert på et vanlig grunnkurs slik det vanligvis undervises for bachelorstudenter innen psykologi, helse eller samfunnsvitenskap, men er endret slik at det passer for studenter på femårige grunnskolelærerutdanninger.

Jeg hadde følgende krav til boka:

1. **Tilpasset lærerstudenter**: Jeg ville at stoffet og eksemplene skulle være fra utdanningsforskning og at de skulle være relevante for læreryrket.
2. **Uten matematikk**: Jeg vet at veldig mange lærerstudenter har så lite matematisk bakgrunn at å lære seg noe særlig av matematikken bak statistikken ville blitt rent pugg.
3. **Kort**: Boka er skrevet til et kurs på OsloMet som skal tilsvare 5 studiepoeng og tre forelesninger i kvantitativ metode. De fleste lærebøker for et innføringskurs i kvantitativ metode er tre ganger så lange og er skrevet slik at det er vanskelig å plukke bare noen deler.
4. **Viser analysene med en enkel og åpent tilgjengelig statistikkprogramvare**: Siden mange lærerstudenter skriver kvantitative masteroppgaver måtte boka ha med eksempler på hvordan man gjør statistisk analyse i en enkel programvare. Siden det finnes mye god statistikkprogramvare som er fritt tilgjengelig ønsket jeg å bruke noen av disse.
5. **Fremhever forskningsdesign og deskriptiv analyse**: Nesten alle statistikklærebøker benytter mye tid på hypotesetesting. Forutsetningene bak disse testene er nesten aldri oppfylt i praksis og de gir sjeldent gode svar på forskningsspørsmålene som stilles. Det er derfor bedre å bruke tiden til å lære mer om forskningsdesign og deskriptiv statistikk.

Jeg fant ingen eksisterende bok som dekket disse punktene, så jeg valgte å tilpasse en lærebok for psykologi-studenter som var gitt ut med en åpen lisens.

Boka er en oversatt og omarbeidet versjon av Navarro og Foxcroft (2025) sin bok «Learning statistics with jamovi: A Tutorial for Beginners in Statistical Analysis» https://www.learnstatswithjamovi.com/, som igjen er en omarbeidet versjon av Navarro (2011) sin bok «Learning statistics with R: A tutorial for psychology students and other beginners. (Version 0.6)» https://learningstatisticswithr.com/.

Hvis du finner noen feil eller har forslag til forbedringer kan du melde inn et «issue» her https://github.com/roarstovner/kvant_for_laerere/issues eller ta kontakt med meg via ansattsiden min.

Jeg håper dere setter pris på å ha en lærebok i kvantitativ metode som dekker de fem punktene i tillegg til å være gratis tilgjengelig på nett.

Roar Stovner, 4. august, 2025

Om 2026-versjonen

Boka fikk gode tilbakemeldinger i 2025, men årets versjon er nok bedre på svært mange områder, blant annet takket være gode forslag fra studenter. Jeg har

- forkortet boka ved å stryke delene om hypotesetesting og skrevet en grundigere og mer forståelig forklaring på estimering og konfidensintervall.
- endret strukturen på delen om deskriptiv statistikk; heller enn å være inndelt i et kapittel om utregninger og et om diagrammer, er den delt inn i deskriptiv statistikk på én variabel og flere variabler.
- lagt til en del om etikk ved kvantitativ måling i Kapittel 2..
- gjort mindre forbedringer på alle delkapitlene, spesielt ved at eksemplene nå er enda tettere knyttet til skole og lærergjerningen.

En spesiell takk til Anne Kristine Øgreid som tok på seg å være en slags redaktør for boka. Jeg står ansvarlig for innholdet, men hun har forbedret språket i hele boka, fått kapitlene raskere til poengene, og forbedret forklaringene. Takk også til fakultetet, LUI ved OsloMet, som har gitt økonomisk støtte (såkalte «strategimidler») til å utvikle boka videre.

Roar Stovner, 17. august, 2026

(Hvis noen ønsker å benytte boka i egen undervisning er det fritt frem, men ta gjerne kontakt via ansattsiden min, for det er gøy å vite at boka blir brukt! Å gjøre endringer i boka er tillatt, siden boka er utgitt under lisensen CC-BY-SA. Det eneste kravet er at de tidligere versjonene av boka (inkludert denne) krediteres, og at den nye versjonen gjøres tilgjengelig med samme lisens. Bokas kildekode ligger på github.)

Sitering: Roar B. Stovner. (2026). *Kvantitativ metode for lærere*. https://kvant.roarstovner.no, https://doi.org/10.5281/zenodo.17068063.

DOI-en over er bokas hoved-DOI, og den peker alltid til nyeste utgave. Hver utgave har i tillegg sin egen DOI. Siterer du et konkret sidetall eller en formulering, bør du bruke DOI-en til utgaven du faktisk leser:

- Roar B. Stovner. (2026). *Kvantitativ metode for lærere* (2026-utgaven). https://doi.org/10.5281/zenodo.22002981
- Roar B. Stovner. (2025). *Kvantitativ metode for lærere* (2025-utgaven). https://doi.org/10.5281/zenodo.17068064

1. Hvorfor lære kvantitative metoder?

Thou shalt not answer questionnaires
Or quizzes upon World Affairs,
Nor with compliance
Take any test. Thou shalt not sit
With statisticians nor commit
A social science
– W.H. Auden[1]

Mange blir overrasket over at de skal lære seg kvantitativ metode i lærerutdanninga. Hvis du virkelig digget kvantitative metoder eller statistikk, ville du sannsynligvis vært påmeldt et statistikkurs nå, ikke et obligatorisk kurs i vitenskapsteori og metode for lærere. Likevel er studentmassen på OsloMet, hvor jeg jobber, delt på midten i synet på kvantitativ metode; mange ser ikke vitsen, mens mange setter pris på å ha blitt tvunget gjennom stoffet, fordi det gir et mer solid grunnlag for å forstå og bruke utdanningsforskning. Dessuten er mange nysgjerrige på hvordan man har utledet all kunnskapen som de er blitt undervist i på lærerutdanninga, og for å vite det må man vite noe om forskningsmetode, blant annet kvantitativ metode.

Jeg mener det er fem grunner til at lærere bør lære seg noe om kvantitativ metode, og de fortjener hvert sitt delkapittel.

Kvantitativ metode hjelper deg å vurdere påstander du møter i læreryrket

Som lærer møter du mange påstander som hevdes å være forskningsbaserte. Læreverk og IKT-verktøy markedsføres som evidensbaserte, pedagogiske tilnærminger presenteres som «dokumentert effektive», og entusiastiske rektorer kommer springende med nye undervisningsmetoder fra forskningsfronten. Det er en oppfatning blant både forskere og praktikere at kvantitativ forskning er mer pålitelig enn kvalitativ forskning, muligens fordi studier med store utvalg og statistiske beregninger ofte assosieres med høyere grad av generaliserbarhet og dermed større troverdighet, vi kan kalle det «*Bigger is better*»-resonnementet. Med grunnleggende kunnskaper i kvantitativ metode er du bedre rustet til å gjennomskue slike feilslutninger og kan heller foreta en kritisk vurdering av kvantitativ forskning.

[1] Sitatet kommer fra Audens dikt fra 1946, *Under Which Lyre: A Reactionary Tract for the Times*, som var en advarsel om samfunnsutviklingen etter andre verdenskrig.

Vi tar et eksempel: Da Nicolai Tangen, lederen for oljefondet, intervjuet psykologen og utdanningsforskeren Angela Duckworth til podcasten sin, spurte han om det var en sammenheng mellom kjønn og *grit*, altså pågangsmot (Tangen, 2022, ca. 40:30). Duckworth svarte at menn og kvinner hadde like mye pågangsmot, i hvert fall var utvalget hennes så stort at kjønnsforskjellen måtte være liten. Tangen svarte at han var en «big believer in sample size» og virket fornøyd med Duckworth sitt svar. Tangen betraktet altså utvalgsstørrelsen nærmest som en garanti for kvalitet, men, som du kommer til å lære i løpet av dette kurset, så er utvalgsstørrelse bare én av mange faktorer som er avgjørende for å vurdere kvaliteten og holdbarheten i slike påstander, og ofte langt fra den viktigste.

Du kommer for eksempel til å lære at du må undersøke hvordan målingene er gjort. Hvordan målte Duckworth hvor mye pågangsmot forskningsdeltagerne hadde? Jo, ved å la forskningsdeltagerne svare på noen påstander om seg selv. For hver påstand skulle de krysse av på en gradert skala fra «Very much like me» til «Not like me at all»:

- New ideas and projects sometimes distract me from previous ones.
- Setbacks don't discourage me.
- I have been obsessed with a certain idea or project for a short time but later lost interest.
- I am a hard worker.
- I often set a goal but later choose to pursue a different one.
- I have difficulty maintaining my focus on projects that take more than a few months to complete.
- I finish whatever I begin.
- I am diligent.

Når du nå vet hvordan deltagernes pågangsmot ble målt, hva tenker du da om konklusjonen om at menn og kvinner har like stort pågangsmot? Tenker du for eksempel at menn har en tendens til å rangere seg selv høyere enn kvinner på skalaen når de tar stilling til påstandene, selv om de faktisk ikke har så mye mer pågangsmot sammenlignet med kvinner? I så fall er Duckworths konklusjon feil, samme hvor mange deltagere hun har spurt! Du lærer om dette i Kapittel 2..

Det overordnede formålet med boka er å gi kunnskapen som trengs for å vurdere slike påstander som Duckworth framsatte. Lærere blir jevnlig møtt med påstander som bygger på kvantitative data, for eksempel påstander om elevers lesekompetanse fra PISA-undersøkelsen, eller påstander om læringseffekten av et nytt IKT-verktøy fra de som ønsker å selge det. Hvis man kan vurdere kvantitativ forskning står man bedre rustet til å forholde seg kritisk til denne typen utsagn.

Kvantitativ metode er premissleverandør for utdanningspolitikk og utdanningsforskning

Kvantitative undersøkelser og statistikk danner ofte grunnlaget for beslutninger i utdanningssektoren. Det fremste eksempelet på dette er betydningen den såkalte PISA-undersøkelsen (*Programme for International Student Assessment*) har hatt for utformingen av norsk utdanningspolitikk, ettersom

resultatene derfra har lagt grunnlaget for en rekke tiltak for å styrke elevenes kompetanse i lesing, matematikk og naturfag. Da de første norske resultatene kom i år 2000 ble det ramaskrik – de norske elevene var helt gjennomsnittlige! Vi som trodde vi hadde blant verdens beste skolesystemer! Dette ble fort omdøpt «PISA-sjokket» (Bergesen, 2006). Her måtte det ryddes opp, og skolesystemet måtte forbedres. Resultatene fra PISA-undersøkelsen medvirket slik til en politisk prosess som endte med etableringen av nasjonale prøver i lesing, regning og engelsk i 2004 og læreplanen Kunnskapsløftet, LK06, i 2006.

Kvantitative undersøkelser kan altså være så potente at de har en betydelig påvirkning på hvilke mål vi har for skolen og hvordan vi styrer den. Her er en liten smørbrødliste over undersøkelser som har hatt tydelig påvirkning på norsk utdanningspolitikk:

- **PISA-undersøkelser** og andre **store internasjonale undersøkelser** brukes som argument for reformer og endringer i skolesystemet. I denne boka ser vi nærmere på noen variabler fra den store internasjonale ICCS-undersøkelsen i kapittel 4 og 5.
- **Nasjonale prøver** påvirker prioriteringer og ressursfordelinger i skolen, og det er politisk omstridte spørsmål om skolers skår på nasjonale prøver skal være offentlig kjent og hvordan ressurser skal fordeles basert på resultatene.
- **Randomiserte kontrollerte eksperimenter** blir benyttet til å argumentere for at noen undervisningsmetoder er bedre enn andre.
- Kvantitative **elevundersøkelser** om trivsel, mobbing og læringsmiljø danner grunnlag for tiltak for å bedre elevenes skolemiljø.
- Statistikk over karakterer og over hvor mange elever som gjennomfører videregående skole, brukes til å vurdere kvaliteten til den enkelte skole.

Som lærer vil du måtte forholde deg til premissene for en utdanningspolitikk som blant annet baserer seg på slike undersøkelser. Å forstå grunnlaget disse undersøkelsene hviler på, er derfor ikke bare viktig for profesjonsutøvelsen, men også for å kunne delta informert i debatten om hvordan skolen skal og bør utvikle seg.

Kvantitativ metode hjelper deg med *assessment literacy*, altså å utvikle din vurderingskompetanse

Som lærer møter du vurderingssituasjoner hele tiden. Du mottar resultater om klassen din på nasjonale prøver, du sitter i evalueringsmøte med skoleledere som lurer på hvorfor klassen din skåret dårlig på tentamen, eller du lager vurderingssituasjoner for å finne ut hvor mye elevene har lært. Kvantitative metoder kan gi deg bedre forståelse for hva slike vurderinger betyr; du kan få en grunnleggende *assessment literacy* som hjelper deg å forberede slike vurderingssituasjoner og tolke resultatene fra dem.

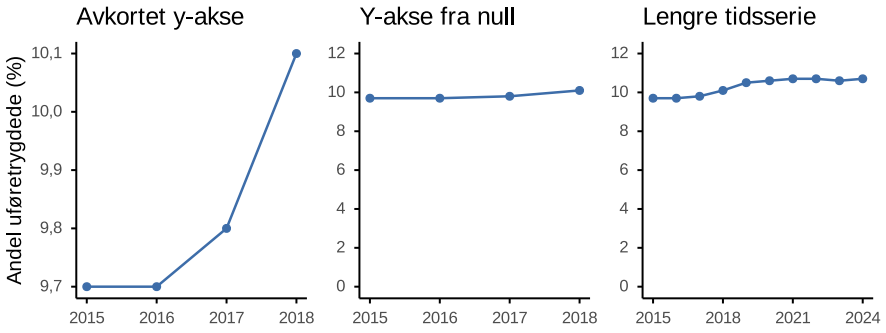
Kvantitativ metode hjelper deg som lærerstudent

Videre i studiet ditt kommer du til å møte forskningslitteratur som benytter kvantitative metoder. Hvis du har innsikt i kvantitative metoder, vil du forstå denne forskningen bedre og lettere vurdere hva kunnskapen bygger på.

Kvantitativ metode hjelper deg i hverdagen

Statistikk dukker opp overalt i nyhetsbildet, og norske medier er ikke immune mot å presentere tall på misvisende måter. I februar 2020 publiserte Aftenpostens A-magasinet en sak om at stadig flere nordmenn er uføretrygdede. Artikkelen inkluderte en graf over andelen uføretrygdede i befolkningen fra 2015 til 2018. I nettversjonen av artikkelen gikk y-aksen fra 9,7 % til 10,1 %, noe som fikk en økning på 0,4 prosentpoeng til å se dramatisk ut. Sjefredaktør Trine Eilertsen beklaget i etterkant og kalte det «misbruk av statistikk» (Eilertsen, 2020).

Figur 1.1 illustrerer to grep som kan gjøre en graf misvisende. Det venstre panelet viser grafen omtrent slik A-magasinet presenterte den på nett, med en avkortet y-akse som skaper et inntrykk av kraftig økning. Det midterste panelet viser de samme dataene med y-aksen fra null; nå er økningen knapt synlig. Men begge disse grafene viser bare fire år. Det høyre panelet viser hele tidsserien fra 2015 til 2024, og her ser vi at andelen uføretrygdede flatet ut etter 2019. A-magasinet viste altså bare den bratteste delen av kurven.



Figur 1.1: Andel uføretrygdede i befolkningen. Venstre panel viser 2015–2018 med avkortet y-akse, liknende slik A-magasinet presenterte det. Midtre panel viser det samme med y-akse fra null. Høyre panel viser hele tidsserien 2015–2024. Kilde: SSB, tabell 11714.

Poenget er ikke at journalister er dårlige i statistikk, men at en grunnleggende kunnskap om kvantitativ metode er nyttig for å oppdage når tall blir presentert på en misvisende måte, enten det gjelder helsestatistikk, kriminalitet, økonomi eller andre temaer du møter som privatperson. En bivirkning ved å kunne noe om kvantitativ metode er at du oftere blir irritert på aviser eller internett.

1.1 En advarsel mot å «tro på dataene»

Kvantitativ forskning har en snedig makt over sinnene våre. Når man har tallfestet noe, fremstår det som autoritativt, sikkert, objektivt og krystallklart. Jeg tror det er bakgrunnen for at man hører uttalelser som «vi må stole på dataene» eller «vi må høre på det evidensen sier». Dette er det reneste vrøvl, og jeg skal nå forklare hvorfor. Jeg forklarer med en halvlang anekdote for å illustrere at enhver statistisk analyse er forankret i en teori om verden. Det er ikke sant at «dataene snakker for seg selv», dataene snakker alltid ut fra et teoretisk ståsted. Et bedre utsagn er «hva dataene betyr avhenger av teorien du har», eller kanskje enda bedre, «at teorier stemmer overens med data er bare et av flere kriterier man kan bruke for å bedømme hvilke teorier som er mest sanne».

Anekdoten er jeg temmelig sikker på at er sann. I 1973 hadde University of California, Berkeley, noen bekymringer rundt hvilke studenter som slapp inn på utdanningene deres. De syntes kjønnsfordelingen var problematisk (Tabell 1.1).

Tabell 1.1: Berkeley-studenter etter kjønn

	Antall søkere	Prosent tilbudt plass
Menn	8442	44%
Kvinner	4321	35%

Berkeley var rett og slett bekymret for å bli saksøkt fordi en mindre andel kvinner ble tilbudt plass på studiene deres. I USA er det forbudt å forskjellsbehandle søkere etter kjønn, og når kvinner systematisk blir tatt opp i lavere grad enn menn, kan det se ut som ulovlig diskriminering. Av mennene fikk 44 % plass, mot 35 % av kvinnene, en forskjell på 9 prosentpoeng i opptaksraten. Med nesten 13 000 søkere er en så stor forskjell altfor stor til å være tilfeldig, og det er nok med litt sunn fornuft for å anta at kvinner ble diskriminert. *Man må jo lytte til dataene.*

Men da andre så nærmere på opptaksdataene, viste de en litt annen historie (Bickel et al., 1975). I stedet for å undersøke opptaksraten til universitetet som helhet, underøkte de opptaksraten på hvert fakultet. da viste det seg at de fleste fakultetene hadde *høyere* opptaksrate for kvinnelige søkere. Tabell 1.2 viser opptaksdataene for de seks største fakultetene.

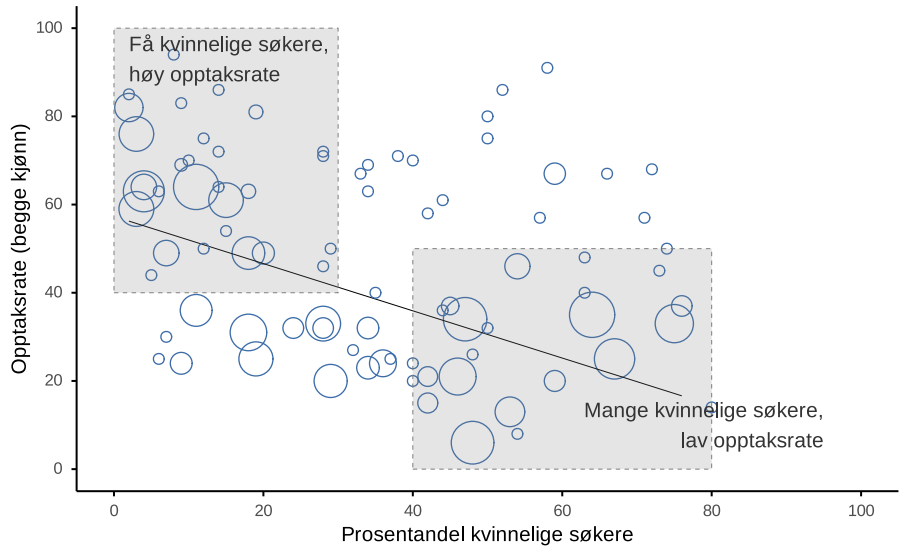
Tabell 1.2: Berkeley-studenter etter kjønn for de seks største fakultetene

	Menn		Kvinner	
fakultet	Søkere	Prosent tilbudt plass	Søkere	Prosent tilbudt plass
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

De fleste fakultetene hadde altså en *høyere* opptaksrate for kvinner enn for menn! Likevel var den totale opptaksraten ved universitetet lavere for kvinner enn for menn. Hvordan kan begge disse påstandene være sanne samtidig?

For det første, merk at fakultetene har ulike opptaksrater: fakultet A og B tok opp mange søkere, mens fakultet C, D, E og F avslo de fleste. For det andre, menn og kvinner søkte til ulike fakulteter. Menn søkte i størst grad til fakultet A og B, altså de som var lettest å komme inn på. Kvinner søkte i størst grad til fakultet C, D, E og F, altså de som var vanskeligst å komme inn på. Dermed fikk kvinner en lavere opptaksrate totalt, selv om de hadde en høyere opptaksrate på de fleste fakultetene. Hvis vi ser på Figur 1.2, ser vi at denne trenden er systematisk, faktisk helt slående.

Denne statistiske effekten er kjent som **Simpsons paradoks**: Sammenhenger som gjelder hele utvalget kan være snudd på hodet i undergruppene. Simpsons paradoks illustrerer et viktig poeng: Hvordan kan man vite om man skal stole på dataene når resultatet avhenger av hvordan man analyserer dem?



Figur 1.2: Berkeley-opptaksdata for 1973. Denne figuren viser opptaksraten for de 85 fakultetene som hadde kvinnelige søkere. De fakultetene som hadde størst andel kvinnelige søkere er til høyre, og disse har lav opptaksrate. Fakultetene som hadde minst andel kvinnelige søkere (til venstre) hadde høy opptaksrate. Figuren er en nytegning av Figur 1 fra Bickel et al. (1975). Større sirkler betyr flere søkere; de minste fakultetene (færre enn 40 søkere) vises som små sirkler.

Men selv denne mer detaljerte analysen hviler på et valg. Simpsons paradoks handler egentlig om hvilke variabler vi tar hensyn til. Den første analysen så bare på *kjønn* og *opptak*. Den andre tok i tillegg hensyn til *fakultet*. Men beslutningen om å inkludere fakultet er ikke noe dataene selv kan fortelle deg, den kommer fra en antakelse om at diskriminering skjer ved *fakultetenes* opptaksbeslutning, ikke universitetets.

Andre kan tenke at kjønnsdiskriminering skjer på et annet sted i prosessen: Hvorfor var det plass til så mange flere studenter på utdanningene med mange mannlige søkere enn på utdanningene med mange kvinnelige søkere? Hva om administrasjonen prioriterte å finansiere mange studieplasser på studier som tiltrakk seg menn? Det ville også vært kjønnsdiskriminerende, men det fanges ikke opp av noen av disse analysene. Denne alternative tolkningen er i seg selv en hypotese som trenger egne data for å undersøkes, for eksempel referater fra møter der det ble besluttet hvor mange plasser hvert studieprogram skal ha. Poenget er at valget av hva man analyserer, gjenspeiler teorien og hva den sier er viktig.

De tre måtene å lese Berkeley-tallene på hviler hver på sin antakelse om *hvor* en eventuell diskriminering skjer:

1. Den første analysen (som bare ser på kjønn og opptak) antar at diskriminering viser seg i universitetets samlede opptaksrate. Universitetet behandles da som én beslutningstaker.
2. Den andre analysen (som tar hensyn til fakultet) antar at diskriminering skjer i det enkelte fakultetets opptaksbeslutning, slik at kjønnene derfor

må sammenlignes innenfor hvert fakultet. Det ser på fakultetet som beslutningstagerne.
3. En tredje analyse (som ser på hvordan studieplassene er fordelt mellom fakulteter) antar at diskrimineringen ligger i finansieringen av antall studieplasser, ikke i selve opptaket. Da er det finansieringsbeslutningene, ikke opptaksbeslutningene, man må analysere.

Det anekdoten viser, er at slutningene du kan trekke fra data eller statistiske undersøkelser er teoriavhengige. Basert på et teoretiske ståsted velger man hva slags data og metode som er relevant for å besvare forskningsspørsmålene. Den kjente statistikeren Sander Greenland har en minneverdig formulering:

> Den største illusjonen om forskning: «La dataene tale for seg selv» Men DATAENE SIER IKKE NOE SOM HELST! De er bare merker på papir eller bits [på en datamaskins harddisk] som bare ligger der og gjør ingenting. Hvis du hører dem snakke bør du umiddelbart oppsøke hjelp! (Greenland, 2022, s. 7, min oversettelse)

Men dette betyr ikke at alle tolkninger er like gode, eller at kvantitativ forskning bare er «meninger med tall». Tvert imot: Nettopp fordi tolkninger bygger på antakelser, kan antakelsene gjøres eksplisitte, og da kan andre forskere kritisere dem. Forskere kan påpeke at analysen ikke er gjort i tråd med teorien, eller utfordre selve teorien. Hvis forskerne derimot *ikke* er bevisste sine antakelser, risikerer de å forsterke egne eksisterende oppfatninger og gå ut i verden med en ny klubbe og bare hamre løs: «Dere må jo lytte til dataene!»

Kvantitative metoder er altså ikke verktøy som gir deg ferdige svar. De er verktøy som undersøker data med et teoretisk blikk. Ulike antakelser gir ulike analyser, men det betyr ikke at alle analyser er like gode. I eksempelet fra Berkeley vil jeg argumentere for at analysen som tar hensyn til fakulteter er bedre, men merk at dette er et argument som baserer seg på kunnskap om hvordan opptaksbeslutninger i USA fungerer, ikke noe dataene alene kan fortelle deg. Teorier kan vurderes etter hvor godt de henger sammen internt, hvor mye de klarer å forklare, og om de tåler å bli testet mot nye data. Når du leser en kvantitativ studie, kan du stille deg spørsmål som: Hvilke antakelser er gjort? Hva er inkludert og hva er utelatt? Hvilke andre forklaringer er mulige?

1.2 Oppsummering

Ettersom årene går, og flere og flere lærere i skolen har mastergrad og noe kjennskap til kvantitativ forskning, får vi forhåpentligvis en lærerstand som kan benytte seg kritisk av kvantitativ forskning, en lærerstand som både kan avvise forskning som ikke er god nok og som kan samle seg om og implementere lærdom fra god forskning.

I dette kapittelet har du lært:

- Fem grunner til at lærere bør kunne litt om kvantitativ metode

- Hvorfor data og statistikk ikke i seg selv gir noen objektiv sannhet, men at de er viktige deler av hvordan teorier blir underbygget.

I Studiedesign

2. Måling i utdanningsforskning

> *When you can measure what you are speaking about, and express it in numbers, you know something about it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts advanced to the stage of science.*
> – Lord Kelvin[2]

Sitatet fra Lord Kelvin, en av attenhundretallets store vitenskapsmenn, viser en stor tro på verdien av å måle ting. Hvis man ikke kan måle noe, vet man ikke noe vitenskapelig om det. Nuvel, i samfunnsvitenskapene, og dermed i utdanningsforskning, gjelder nok ikke dette. Kanskje gjelder heller det motsatte, at hvis noe er målt, så bør vi være ekstra skeptiske til det:

> *We're seeing that the published literature–you know, in some of our best journals–features measures that have little or no validity evidence. Measurement, schmeasurement. Applied researchers are, as a norm, not engaged in this process of construct validation.*
> – Professor Jessica Kay Flake[3]

I dette kapittelet skal vi lære hvorfor vi bør være skeptiske til kvantitative målinger i utdanningsforskning og hva som skal til for å gjøre gode målinger. Men først må vi lære de vanligste begrepene man benytter for å beskrive målinger. Dette kapittelet bygger i stor grad på Campbell & Stanley (1963) og på Stevens (1946) for diskusjonen om måleskalaer.

2.1 Begreper om målinger

Veien fra idé til data går i tre steg. Vi starter med et *teoretisk konstrukt*, altså den teoretiske beskrivelsen av det vi ønsker å undersøke. Så lager vi en *målemetode*, som er en presis oppskrift på hvordan konstruktet skal måles. Til slutt gjennomfører vi målingene, og resultatet er *variabler* med *verdier* i et datasett. Vi tar de tre stegene i tur.

[2] Foredrag til *Institution of Civil Engineers*, 3. mai, 1883. Kilde: https://en.wikiquote.org/wiki/William_Thomson

[3] Foredrag til *RIOT Science Club*, 23. november, 2020. Kilde: https://youtu.be/Cq6n7AS_r8w?si=PSLqDe2VGExXp2qV&t=637

2.1.1 Teoretiske konstrukter og operasjonalisering

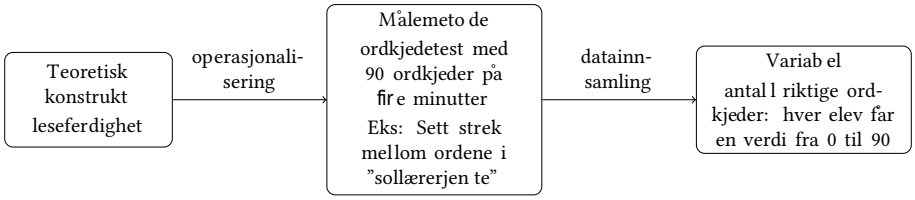
Fenomenene vi ønsker å måle i utdanningsforskning, slik som leseferdighet, motivasjon, læring og undervisning, er beskrevet teoretisk. Den teoretiske beskrivelsen kalles et *teoretisk konstrukt*, ofte omtalt som bare *konstruktet*.

La oss bruke leseferdighet som eksempel. Hva er egentlig leseferdighet? Leseforskere er ikke helt enige, men de fleste beskrivelser inneholder minst disse delene:

- *avkoding*, altså å omsette bokstavene på arket til ord, raskt og uanstrengt
- *leseforståelse*, altså å skape mening av det som står i teksten
- *refleksjon*, altså å vurdere teksten, hva forfatteren vil med den og hva den er verdt

Noen leseforskere regner også *leselyst* eller *leseengasjement* som en del av leseferdigheten, mens andre holder det utenfor og heller ser på det som noe som påvirker leseferdigheten. Man kan altså helt fint være uenige om hva teoretiske konstrukter er.

Et konstrukt kan man ikke måle direkte. Man må først bestemme seg for en helt presis oppskrift på hvordan målingen skal foregå. Denne oppskriften kalles en *målemetode*, og arbeidet med å komme fra det vage konstruktet til den presise målemetoden kalles å **operasjonalisere** konstruktet. Selve målemetoden omtales som en *operasjonalisering* av konstruktet. Begrepene er oppsummert i Figur 2.1.



Figur 2.1: Fra teori til data. Konstruktet *leseferdighet* operasjonaliseres til en målemetode, og datainnsamlingen gir en variabel med verdier.

Kort fortalt handler operasjonalisering om å omforme et konstrukt, som ofte er ganske vagt, til en helt presist definert måling. Operasjonalisering omfatter blant annet:

- **å avgrense hva som skal måles.** Skal vi måle hele leseferdigheten, eller bare avkodingen? Skal leselyst regnes med? Svaret avhenger av hva vi trenger målingen til. En logoped som utreder mistanke om dysleksi er interessert i avkodingen, mens en norsklærer som planlegger undervisning i sakprosa er mer interessert i forståelse og refleksjon.
- **å bestemme målemetode.** Avkoding måles ofte med en *ordkjedetest*, der eleven setter strek mellom ordene i kjeder som «busshundmelktyv». Leseforståelse måles gjerne ved at eleven leser en tekst og svarer på spørsmål til den, slik som i nasjonale prøver i lesing. Leselyst måles kanskje ved å svare på spørsmål i et spørreskjema?
- **å definere hvilke verdier målingene kan ha.** Skal ordkjedetesten telle antall riktige, eller angi kategoriene *trenger ekstra oppfølging* eller *trenger ikke ekstra oppfølging*? Skal leselyst måles med tall (1 til 7 eller 1 til 4) eller

tekst (fra *ikke interessant* til *veldig interessant*)? Skal elevene kunne skrive hva de vil? Hvordan skal vi i så fall lage verdier av svarene etterpå?

Legg merke til at det samme konstruktet kan operasjonaliseres på mange måter, og at operasjonaliseringene ikke er likeverdige. En ordkjedetest og en nasjonal prøve i lesing måler begge noe vi kaller «leseferdighet», men de gir forskjellige tall og rangerer ikke elevene helt likt. En elev som avkoder langsomt, men forstår godt det hun leser, kommer dårlig ut på ordkjedetesten og godt ut på den nasjonale prøven. Når du leser en forskningsartikkel er det derfor ikke nok å vite at forskerne har målt «leseferdighet»; du må vite *hvordan* de har målt den.

Operasjonalisering er komplisert, og det finnes ikke én riktig måte å gjøre det på. Hvordan du velger å operasjonalisere et konstrukt til en formell måling, avhenger av hva du trenger målingen til. Ofte vil du oppdage at forskerne som jobber innen ditt område har veletablerte praksiser for hvordan man skal måle konstruktet, men andre ganger må du utvikle en operasjonalisering av konstruktet som passer til din forskning.

2.1.2 Variabler og verdier

Når konstruktet er operasjonalisert, kan man gjøre målingene. Kvantitative analyser baserer seg på at man har gjort den samme målingen av mange forskjellige fenomener. I utdanningsforskning er fenomenene gjerne skoler, lærere, klasser, elever, rektorer, læreplaner, lærebøker, elevtekster, undervisningstimer eller liknende. Måling er å tildele disse fenomenene tall, merkelapper eller andre veldefinerte beskrivelser. Alle følgende eksempler vil regnes som målinger:

- Petters **antall riktige ordkjeder** er *48*.
- Petters **leselyst** er *3 av 4*.
- Petters **alder** er *11* år.
- Petters **selvidentifiserte kjønn** er *gutt*.
- Timens **arbeidsro** var *middels*.
- Rektors **holdning til lærernes autonomi** var *lærerautonomi er svært viktig*.

I denne listen viser den fete skriften det fenomenet som måles, som kalles **variabelen**, og den kursiverte skriften viser resultatet av målingen, som kalles **verdien**. En variabel representerer altså fenomenet som måles, mens verdien er det variabelen får i et bestemt tilfelle. Skillet mellom variabelen og de ulike verdiene variabelen kan anta, er sentralt:

- **Antall riktige ordkjeder** kan ha verdiene *0, 1, 2, 3* og så videre opp til *90*, siden testen har 90 ordkjeder. Ingen elev kan få 91 riktige, og ingen kan få 48,5.
- **Leselyst**, målt med spørsmålet «jeg leser bare hvis jeg må», kan ha verdiene *helt enig*, *litt enig*, *litt uenig* og *helt uenig*. Merk at eleven med *høyest* leselyst her svarer *helt uenig*.
- **Alder** kan ha verdiene *0, 1, 2, 3* og så videre. Den øvre grensen er litt uklar, men i praksis kan vi si at den høyeste mulige alderen er *150*, siden ingen mennesker noensinne har levd så lenge.

- **Selvidentifisert kjønn** er for de fleste elever enten *gutt* eller *jente*, men eleven kan også velge å identifisere seg med *ingen av delene*, eller eksplisitt kalle seg selv *transperson*.

Som eksemplene over viser, virker verdiene for noen variabler ganske opplagte, slik som for antall riktige ordkjeder, men for andre variabler er det vanskeligere å bestemme verdiene. Måten du spesifiserer de tillatte måleverdiene på, er med andre ord en viktig del av å planlegge en studie.

2.1.3 Forskjellige målemetoder

Å bestemme målemetode er kanskje det viktigste valget i en operasjonalisering, og utvalget er større enn man skulle tro. Disse seks er de vanligste i utdanningsforskning:

- *prøver og tester*, der man gir folk oppgaver som skal løses
- *registerdata*, som fødselsdato og karakterer, hentet fra offisielle registre
- *selvrapportering*, der folk oppgir verdien selv, som regel i et spørreskjema
- *tredjepartsrapportering*, der noen andre oppgir den, for eksempel læreren eller foreldrene
- *observasjon*, der forskeren registrerer direkte det som skjer
- *fysiologiske mål*, også kalt biomarkører, for eksempel pulsmåling eller registrering av øyebevegelser

Hvilken av dem som passer, avhenger først og fremst av konstruktet du skal måle. Vi skal diskutere dem nøyere i Seksjon 2.4, for vi må først få på plass begrepene som skal til for å vurdere en målemetode.

2.2 Målenivåer

Resultatet av rekke målinger av samme fenomen kalles altså en variabel. Vi skal lære om fire ulike **målenivåer** som variabler kan ha. Disse målenivåene, eller **måleskalaene**, beskriver hvilke typer verdier en variabel kan anta, og hvilke statistiske operasjoner som er gyldige for den.

2.2.1 Nominell variabel

En **nominell variabel** (også kjent som en **kategorisk variabel**) er en variabel der de mulige verdiene er navn på kategorier. Ordet *nominell* kommer av *nomen* som er latin for *navn* – en nominell variabel har altså verdier som er navnet på kategorier. Et eksempel er øyenfarge; de mulige verdiene er «grønn», «blå», «grå», «brun», og så videre. Kjønn er en nominell variabel som ofte har bare to mulige verdier, «mann» og «kvinne», men som kan ha flere. I utdanningsforskning er ofte variabelen *skole* av denne typen, der de mulige verdiene er navn på skoler («Dalsiden Skole», «Vangenhaug Skole», osv.). For denne typen variabler gir det ikke mening å si at en av dem er «større» eller «bedre» enn noen annen, man kan altså ikke sortere dem i noen naturlig rekkefølge. Og det gir absolutt ikke mening å snakke om gjennomsnittet av målingene – hva skulle «elevene i studien sin gjennomsnittlige øyenfarge» ha vært? For å regne gjennomsnitt måtte man summert de 10 verdiene og delt på 10, men hva blir blå + grønn + blå + brun? Kort sagt, verdiene til nominelle variabler er kun navnet på en kategori. Verdiene har

ikke noen naturlig rekkefølge, de kan ikke adderes, multipliseres eller regnes med på noen måte.

Men man kan telle antallet innenfor hver kategori. Anta at jeg forsker på hva som kjennetegner lærernes pendling til jobb.[4] En variabel jeg burde målt er hva slags transportmiddel lærerne bruker for å komme seg på jobb. Denne «transporttype»-variabelen kan ha ganske mange mulige verdier, slik som «tog», «buss», «bil» og «sykkel». La oss anta at disse fire er de eneste mulighetene og forestill deg at jeg spør 100 lærere om hvordan de kom seg på jobb i dag, med dette resultatet (Tabell 2.1).

Tabell 2.1: Hvordan 100 lærere pendlet til jobb i dag

Transportmiddel	Antall lærere
(1) Tog	12
(2) Buss	30
(3) Bil	48
(4) Sykkel	10

Så, hva var den gjennomsnittlige transporttypen? Åpenbart er svaret her at det ikke finnes noen. Det er et tåpelig spørsmål å stille. Du kan si at å reise med bil er den mest populære metoden, og at å reise med tog er den minst populære metoden, men det er stort sett alt. Legg også merke til at rekkefølgen jeg viser alternativene i, ikke er veldig interessant. Jeg kunne ha valgt å vise dataene som i Tabell 2.2.

Tabell 2.2: Hvordan 100 lærere pendlet til jobb i dag, en annen visning

Transportmiddel	Antall lærere
(3) Bil	48
(1) Tog	12
(4) Sykkel	10
(2) Buss	30

... og ingenting endrer seg egentlig.

2.2.2 Ordinal variabel

En **ordinal variabel** er en nominell variabel der de ulike kategoriene har en meningsfull rekkefølge. Ordet *ordinal* kommer av samme ord som «orden» eller «ordning». Her er et typisk eksempel: Tenk deg at jeg er interessert i elevers holdninger til klimaendringer. Jeg ber noen elever om å velge det utsagnet som best samsvarer med deres holdning ut fra disse fire utsagnene:

[4] Her er «lærernes pendling» konstruktet, og transportmiddel er én av variablene det kan operasjonaliseres i. Andre variabler kunne vært reisetid og reiselengde. Vær oppmerksom på at forskere til daglig benytter begrepene konstrukt og variabel litt om hverandre.

1. Temperaturene stiger på grunn av menneskelig aktivitet.
2. Temperaturene stiger, men vi vet ikke hvorfor.
3. Temperaturene stiger, men ikke på grunn av mennesker.
4. Temperaturene stiger ikke.

Legg merke til at disse fire utsagnene faktisk har en naturlig rekkefølge i hvordan de måler enighet sett i sammenheng med nåværende vitenskapelig konsensus. Utsagn 1 ligger tett opp til den nåværende konsensus, utsagn 2 ligger noe mindre tett opp til den, utsagn 3 avviker i betydelig grad, og utsagn 4 står i klar motsetning til den etablerte konsensus. Derfor kan jeg ordne elementene som $1 > 2 > 3 > 4$, som viser at «holdning til klimaendring», operasjonalisert på denne måten, er en ordinal variabel.

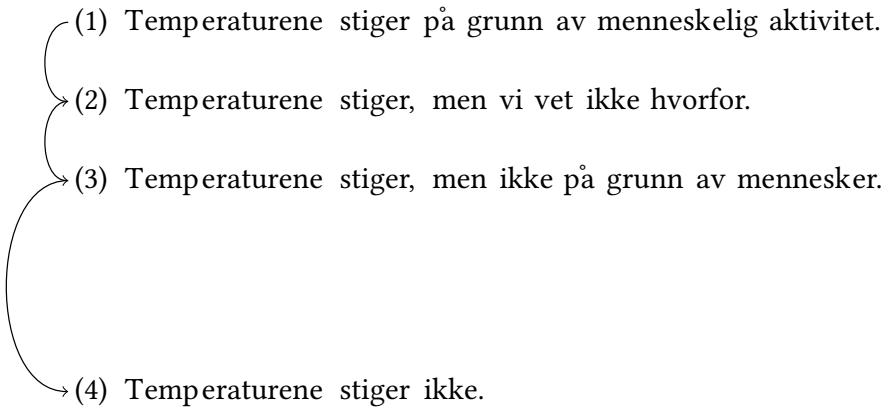
La oss anta at jeg stilte 100 elever disse spørsmålene og fikk svarene vist i Tabell 2.3.

Tabell 2.3: Holdninger til klimaendringer

Respons	Antall
(1) Temperaturene stiger på grunn av menneskelig aktivitet.	51
(2) Temperaturene stiger, men vi vet ikke hvorfor.	20
(3) Temperaturene stiger, men ikke på grunn av mennesker.	10
(4) Temperaturene stiger ikke.	19

Det rimelig å gruppere, for eksempel, (1), (2) og (3) sammen, og si at 81 av 100 personer var i det minste *delvis enig* med vitenskapen. Og det er også ganske rimelig å gruppere (2), (3) og (4) sammen og si at 49 av 100 personer registrerte i det minste *noe uenighet*. Men det ville være fullstendig merkelig å gruppere (1), (2) og (4) sammen og si at 90 av 100 personer sa… hva da? Det er ingenting fornuftig som tillater deg å gruppere disse svarene sammen.

Det er fristende å regne med svaralternativene (1) til (4) som om de var tall. For eksempel er det fristende å si at gjennomsnittet av svar (2), (3) og (4) er $\frac{2+3+4}{3} = 3$. Det ville vært galt, siden det regnestykket krever at det er like stor forskjell mellom svaralternativ (2) og (3) som mellom (3) og (4). Dette er ikke garantert, og man kan argumentere for at konstruktet «enighet med vitenskapelig konsensus» er mer som i Figur 2.2: De tre første utsagnene godtar alle at temperaturene stiger, mens det fjerde avviser hele premisset, trass i at termometre verden over måler høyere og høyere gjennomsnittstemperatur. Da kan forskjellen i *holdning til klimaendringer* være mye større fra (3) til (4) enn mellom de tre første.



Figur 2.2: I en ordinal variabel er ikke avstanden mellom kategoriene nødvendigvis like stor. Her markerer pilene avstanden mellom kategoriene, og avstanden mellom kategori (3) og (4) er større enn de andre.

Når en ordinal variabel markeres med tall, som er helt vanlig, skal man altså ikke tolke tallstørrelsene bokstavelig og anta at man kan regne med dem – tallene er der kun for å markere kategorienes rekkefølge.

Et vanlig eksempel på en ordinal variabel i utdanningsforskning er «høyeste fullførte utdanningssnivå». Du *kan* si at en person som kun har fullført ungdomsskolen har mindre utdanning enn en person som har fullført videregående, og så videre. Med ordinale variabler vet man ikke *hvor mye* større en verdi er i forhold til en annen. For eksempel gir det ikke mening å spørre om det er større forskjell mellom «ungdomsskole» og «videregående» enn mellom «videregående» og «bachelor». Er utdanningsnivået på en bachelor mye høyere eller litt høyere enn fullført videregående? Hvis variabelen er ordinal kan man ikke besvare det spørsmålet.

2.2.3 Intervallvariabel

I motsetning til variabler på nominelt og ordinalt målenivå er **intervallvariabler** og **forholdstallsvariabler** variabler der tallverdien er meningsfull. For en variabel som er målt på intervallskala, er *forskjellene* mellom tallverdiene tolkbare, men variabelen har ikke en «naturlig» nullverdi. Et godt eksempel på en variabel målt på intervallskala, er temperaturmåling i grader celsius. Hvis det for eksempel var 15° i går og 18° i dag, så er differansen mellom dem, 3°, meningsfull og måler en forskjell i temperatur. Dessuten er disse tre gradene i forskjell *nøyaktig den samme forskjellen* som mellom 7° og 10°. Kort sagt er regneartene addisjon og subtraksjon meningsfulle for variabler på intervallskala. Merk at dette *ikke* var tilfelle for ordinale variabler.

Men legg merke til at 0° ikke betyr «ingen temperatur i det hele tatt». Det betyr faktisk «temperaturen der vann fryser». Nullpunktet er valgt av praktiske grunner og dermed ganske vilkårlig. Mangelen på et naturlig nullpunkt gjør det meningsløst å multiplisere og dividere temperaturer. Det er feil å si at 20° er dobbelt så varmt som 10°, akkurat som det er rart og meningsløst å hevde at 20° er negativ to ganger så varmt som −10°.

Intervallvariabler er vanlige i utdanningsforskning. Anta at jeg er interessert i å undersøke bruken av nettbrett over tid. Da er det naturlig å

sammenlikne bruken av nettbrett i forskjellige år, og en variabel i datasettet vil være årstall. Dette er en variabel på intervallskala. En elev som begynte på skolen i 2018, begynte 5 år før en student som begynte i 2023; differanser gir altså mening. Det ville derimot vært helt meningsløst å dele 2023 på 2018 og si at den andre eleven begynte 0,24 % senere enn den første eleven; ganging og deling gir altså ikke mening. Derfor er årstall på en intervallskala.

Tid på døgnet er en annen variabel på intervallskala. Det gir mening å sammenlikne klokka 14 og klokka 7 ved å si «sju timer senere», men det gir ikke mening å sammenlikne dem ved å si «dobbelt så mye tid på døgnet.»

2.2.4 Forholdstallsvariabel

Den fjerde og siste typen variabel er variabler på *forholdstallsnivå*. Dette er variabler med et nullpunkt der det er greit å multiplisere og dividere. Et godt eksempel på en variabel på forholdstallsnivå er tid. Når man måler elevers ferdigheter, er det vanlig å registrere hvor lang tid elevene bruker på å løse en oppgave, fordi det er en indikator på hvor vanskelig oppgaven er. Anta at Roar bruker 2,3 sekunder på å svare på et spørsmål, mens Bård bruker 3,1 sekunder. Som med en intervallskalavariabel er både addisjon og subtraksjon meningsfulle her. Bård brukte virkelig $3,1 - 2,3 = 0,8$ sekunder lenger tid enn Roar. Men legg merke til at multiplikasjon og divisjon også gir mening her: Bård brukte $3,1/2,3 = 1,35$ ganger så lang tid som Roar på å svare på spørsmålet. Og grunnen til at du kan gjøre dette for en forholdstallsvariabel er at «null sekunder» virkelig betyr «ingen tid i det hele tatt».

Her er noen flere variabler på forholdstallsnivå:

- inntekt (0 kr betyr virkelig ingen inntekt)
- antall barn (0 barn betyr virkelig ingen barn)
- antall år avlagt utdanning (0 år betyr ingen utdanning)
- antall riktige på matematikkprøven (0 riktig betyr ingen riktig)

Det siste eksempelet krever en analyse. Ofte måles «kompetanse», «ferdighet» eller liknende ut fra antall riktige svar på en prøve eller annen ferdighetstest. La oss ta matematikkompetanse som et eksempel. Hvis Bård besvarer 20 spørsmål riktig på en matematikkprøve og Roar besvarer 10 riktig, kan man si at Bård fikk dobbelt så mange riktige. Man kan også si at han fikk 10 flere riktig. Addisjon, subtraksjon, multiplikasjon og divisjon gir mening, altså er det en forholdstallsvariabel. Men er det riktig å si at Bård har dobbelt så høy *matematikkompetanse* som Roar? Det kommer kanskje an på prøven og hvilke spørsmål man svarte riktig på? Hva hvis Roar svarte riktig på 10 vanskelige spørsmål mens Bård raste gjennom de 20 enkleste spørsmålene på prøven uten å få til noe mer? Eller hva hvis Roar ikke hadde besvart noen spørsmål riktig, hadde det vært riktig å si at han har null matematikkompetanse? Dette viser at hva som er riktig målenivå avhenger av hvordan konstruktet er operasjonalisert. Hvis forskeren operasjonaliserer konstruktet matematikkompetanse som *antall riktige svar på prøven*, er variabelen på forholdstallsnivå, men hvis forskeren operasjonaliserer matematikkompetanse som *en lærers vurdering av elevens kompetanse på en skala fra 1 til 5*, er variabelen på ordinalnivå.

2.2.5 Kontinuerlige og diskrete variabler

Man kan også karakterisere variabler ut fra om de er kontinuerlige eller diskrete variabler (Tabell 2.4). Kort sagt handler det om dette:

- En **kontinuerlig variabel** er en variabel hvor det for alle par av verdier du kan tenke deg alltid er mulig å ha en annen verdi imellom.
- En **diskret variabel** er en variabel som ikke er kontinuerlig, altså vil det for en diskret variabel være to verdier uten noen annen verdi imellom.

Disse definisjonene virker sannsynligvis litt abstrakte, men de er ganske enkle når du ser noen eksempler. For eksempel er svartid kontinuerlig. Hvis Roar bruker 3,1 sekunder og Bård bruker 2,3 sekunder på å svare på et spørsmål, vil Louises responstid ligge imellom hvis hun bruker 3,0 sekunder. Og selvfølgelig vil det også være mulig for Ingrid å bruke 3,031 sekunder på å svare, noe som betyr at hennes svartid vil ligge mellom Louises og Roars. Fordi vi i prinsippet alltid kan finne en ny verdi for svartiden mellom to andre, er svartid en kontinuerlig variabel.

Tabell 2.4: Forholdet mellom målenivåene og skillet mellom diskret/kontinuerlig. Celler med en x tilsvarer en kombinasjon som er mulig

Målenivå	kontinuerlig	diskret
nominal		x
ordinal		x
intervall	x	x
forholdstall	x	x

Diskrete variabler er det motsatte. For eksempel er nominelle variabler alltid diskrete. Det finnes ikke en transporttype som faller «mellom» tog og fly, i hvert fall ikke på den samme måte som 2,65 faller mellom 2 og 3. Derfor er transporttype en diskret variabel. På samme måte er ordinale variabler alltid diskrete. Har en skala svaralternativene «Helt enig», «Ganske enig» og «Hverken enig eller uenig», finnes det ikke noe svaralternativ mellom «Helt enig» og «Ganske enig» på den skalaen. Man kan så klart lage en ny skala med flere alternativer, for eksempel ved å føye til «Litt enig», men da har man en ny skala som også er diskret. Intervallskala- og forholdstallskalavariabler kan gå begge veier. Som vi så ovenfor, er forholdstallsvariabelen *svartid* kontinuerlig. Intervallnivåvariabelen *temperatur i grader celsius* er også kontinuerlig. Imidlertid er intervallnivåvariabelen *året du gikk i første klasse* diskret. Det er ikke noe år mellom 2002 og 2003. Forholdstallsvariabelen *antall spørsmål du får riktig på en flervalgsprøve* er også diskret. Siden et flervalgsspørsmål ikke kan være «delvis korrekt», er det ingenting mellom 5 av 10 riktige og 6 av 10 riktige. Tabell 2.4 oppsummerer forholdet mellom målenivåene og skillet mellom diskret/kontinuerlig. Celler med en x er mulige.

2.2.6 Noen komplikasjoner: Likert-skalaer

I den virkelige verden er det mange variabler som ikke passer så godt inn i disse målenivåene. Det er ikke problematisk, for målenivåene gir forskere likevel en mer presis måte å snakke om variablene på. Et vanlig svarformat på spørreundersøkelser som ikke helt passer inn i målenivåene, er **Likert-spørsmål**. Likert-spørsmål er det vanligste verktøyet i spørreundersøkelser. Du har selv fylt ut mange slike. Anta at du får følgende påstand med svaralternativer i en undersøkelse:

> Jeg har vanskeligheter med å avslutte det jeg begynner på.

1. Sterkt uenig
2. Uenig
3. Verken enig eller uenig
4. Enig
5. Sterkt enig

Disse svaralternativene er et eksempel på en 5-punkts Likert-skala, der folk blir bedt om å velge mellom én av flere (i dette tilfellet 5) tydelig ordnede muligheter, vanligvis med en verbal beskrivelse gitt i hvert tilfelle. Det er imidlertid ikke nødvendig at alle elementene er beskrevet. Dette er også et eksempel på en 5-punkts Likert-skala:

1. Sterkt uenig
2.
3.
4.
5. Sterkt enig

Likert-skalaer er veldig praktiske, om enn noe begrensede, verktøy. Men hva slags målenivå er en Likert-skala på? De er åpenbart diskrete, siden du ikke kan gi et svar på 2,5. De er åpenbart ikke nominelle skalaer, siden elementene er ordnet; og de er heller ikke forholdstallsskalaer, siden det ikke er noe naturlig nullpunkt.

Men er de ordinalskala eller intervallskala? Da må vi bestemme om forskjellene mellom nabosvaralternativene er like. Ett argument sier at vi egentlig ikke kan bevise at forskjellen mellom «sterkt enig» og «enig» er like stor som forskjellen mellom «enig» og «verken enig eller uenig». Det virker egentlig ganske åpenbart at de ikke er like i det hele tatt, så dette antyder at vi burde behandle Likert-skalaer som ordinale variabler. På den annen side ser det i praksis ut til at de fleste deltakerne tolker skalaen fra «sterkt uenig» til «sterkt enig» som «på en skala fra 1 til 5». For eksempel, hvis man spør folk om hvor fornøyde de er med inntekten sin på en 7-punkts Likert-skala og sammenlikner dette med deres faktiske inntekt, så ser man at hvert nivå på Likert-skalaen svarer til omtrent like mye penger. Som en konsekvens behandler mange forskere variabler med Likert-skalaer som en

intervallskala.[5] Det er ikke nødvendigvis en intervallskala, men i praksis er det nært nok til at vi tenker på det som en.

2.2.7 Samlevariabler

Mange teoretiske konstrukter er komplekse og fanges ikke opp av bare én måling. Hvis man skal måle kvaliteten på en lærers praksis i «vurdering for læring» må man kanskje måle både hvordan læreren vurderer kompetansen til elever ut fra prøvesvar i tillegg til hvordan læreren gir tilbakemeldinger på prøven. Hvis man utelater en av delene, har man ikke målt hele konstruktet «kvalitet på ‹vurdering for læring›-praksis», så derfor må man bruke flere enkeltvariabler for å måle ett og samme fenomen. Når man kombinerer flere ulike målinger for å lage én variabel kalles det en **samlevariabel**.

I spørreundersøkelser måles gjerne konstruktene med samlevariabler, og et typisk eksempel er fra den store internasjonale undersøkelsen om elevers matematikk- og naturfags-kompetanse, TIMSS. For å måle elevenes indre motivasjon i matematikk fikk elevene ni Likert-påstander som de skulle ta stilling til på en firepunkts skala fra «Svært enig» til «Svært uenig». Her er påstandene[6]:

1. Jeg liker å lære matematikk
2. Jeg skulle ønske at jeg ikke var nødt til å lære matematikk
3. Matematikk er kjedelig
4. Jeg lærer mye interessant i matematikk
5. Jeg liker matematikk
6. Jeg liker alt skolearbeid som har med tall å gjøre
7. Jeg liker å løse oppgaver i matematikk
8. Jeg gleder meg til timene i matematikk
9. Matematikk er et av de fagene jeg liker best

Svarene på disse ni spørsmålene slås sammen til en samlevariabel for indre motivasjon. Man kan tenke seg at man regner gjennomsnittet til de ni svarene og lar det være verdien på elevens indre motivasjon, men i virkeligheten er det en mer avansert utregning for å lage samlevariabelen. Merk at spørsmål 2 og 3 måler indre motivasjon motsatt av de andre spørsmålene, altså at «svært uenig» betyr at man har høy indre motivasjon. Før slike spørsmål analyseres, reverserer man ofte svaralternativene slik at man kan tolke svarene likt som i de andre spørsmålene.

Hvilket målenivå har slike samlevariabler? Det kommer an på. De kan være nominelle, for eksempel hvis en spesialpedagog lar elever ta et testbatteri, altså en samling av flere tester, og ut fra svarene lager en samlevariabel *dysleksi* med to mulige verdier, «har dysleksi» og «har ikke dysleksi». Dette er en samlevariabel fordi den er regnet ut basert på mange andre. Men de aller vanligste samlevariablene er de som er laget på bakgrunn av mange Likert-spørsmål, slik som indre motivasjon ble målt i TIMSS-eksempelet over, og de behandles som om de er på en kontinuerlig intervallskala. Disse variablene er ikke helt kontinuerlige, for hvis man lager en samlevariabel

[5] Kanskje er den egentlige grunnen at de statistiske utregningene med intervallskalaer er mye enklere enn de tilsvarende utregningene med ordinale skalaer.

[6] Hvis du er interessert i å se hele spørreskjemaet er det tilgjengelig her: https://www.uv.uio.no/ils/forskning/prosjekter/timss/2019/elevskjema_9trinn.pdf

ved å ta gjennomsnittet av ni spørsmål, som hver har et heltall som svar, så kan man ikke få alle mulige verdier. Siden en nidel er 0,11, kan man få svarene 3,11 og 3,22, men ikke 3,15. Likevel er det nært nok til at man benytter variablene som om de var kontinuerlige. Strengt tatt er de kanskje heller ikke på intervallnivå, av samme grunner som ble nevnt om Likert-skalaer over, men i praksis behandles de som om de var det.

2.3 Å vurdere en målemetodes kvalitet: Reliabilitet og validitet

Vi har nå lært begreper for å beskrive hvordan man operasjonaliserer et teoretisk konstrukt og dermed skaper en målemetode. Det vi alltid må vurdere er om målingene er av god kvalitet, eller om de rett og slett er dårlige. Forskere diskuterer målemetoders kvalitet med begrepene *reliabilitet* og *validitet*. Enkelt sagt forteller **reliabiliteten** til en målemetode hvor stabil målingen er, mens **validiteten** til en målemetode forteller hvor gyldig den er.

2.3.1 En målings reliabilitet

Reliabilitet betyr altså hvor stabil målemetoden er. Baderomsvekta mi gir en svært reliabel måling av vekta mi, fordi den gir meg det samme svaret når jeg går av og på vekta flere ganger etter hverandre. Legg merke til at dette ikke betyr at baderomsvekta viser riktig vekt; det kan være at springfjæra i vekta har blitt litt løs slik at den viser noen kilo for mye. I så fall stemmer ikke vekta sin måling overens med min sanne vekt i det hele tatt, noe som betyr at målingen har lav validitet. Den har likevel høy reliabilitet, fordi den viser samme resultat fra gang til gang.

Reliabilitet betyr altså at målingen gir samme verdi på tvers av situasjoner. Dette gir opphav til mange forskjellige undertyper av reliabilitet, for eksempel:

- **Test-retest-reliabilitet**. Hvis vi gjentar målingen på et senere tidspunkt, får vi det samme svaret?

  En prøve har høy test-retest-reliabilitet hvis elever får samme karakter når de tar samme prøve på ulike tidspunkt, noe som betyr at elevens dagsform ikke påvirker prøveresultatet.

  En forskers vurdering av kvaliteten på lærerens undervisning har høy test-retest-reliabilitet hvis det ikke spiller noen rolle hvilken time forskeren observerer.

- **Inter-rater-reliabilitet**. Hvis en annen person gjennomfører målingen, får vi da det samme svaret?

  En måling av elevers adferd i en musikktime har høy inter-rater-reliabilitet hvis resultatet er uavhengig av hvem som observerer timen.

  Hvis karakteren på skriftlig eksamen i engelsk ikke avhenger av hvilken sensor som vurderer besvarelsen, har den høy inter-rater-reliabilitet.

Hvordan man bør vurdere en variabels reliabilitet, avhenger av det teoretiske konstruktet. Anta at man skal måle konstruktet «elevers utagerende adferd» ved å observere én undervisningstime. Hvis formålet er å gi en generell

vurdering av elevenes utagerende adferd, vil målingen ha lav test-retest-reliabilitet, fordi den sannsynligvis vil gi svært forskjellige svar avhengig av om du utfører målingen siste time fredag eller første time tirsdag. Hvis formålet derimot er å finne ut av hvordan elevenes adferd varierer i en undervisningsuke, er hele poenget med målingen å fange opp disse variasjonene. Derfor må man tenke over det teoretiske konstruktet og forskningsformålet for å gi en vurdering av en målemetodes reliabilitet.

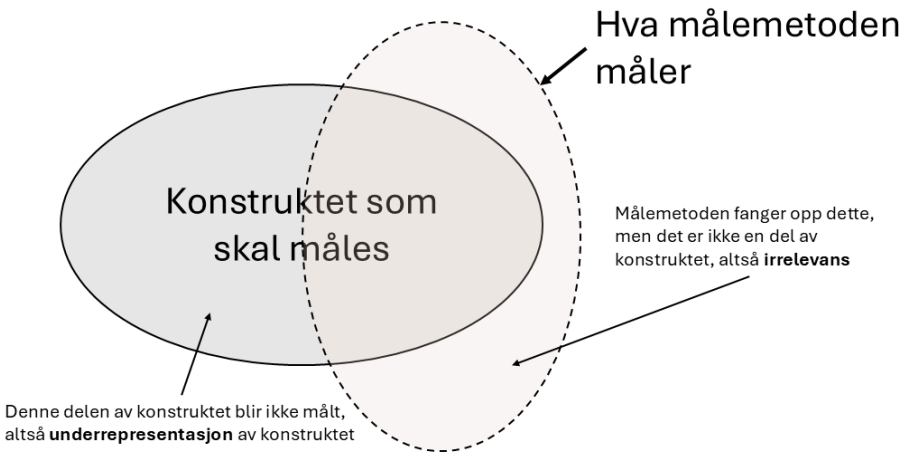
2.3.2 En målings validitet

En målemetode har høy **validitet** hvis den måler det man tror den måler. Validitet kalles også «gyldighet», som er et godt ord for validitet. Man skiller ofte mellom mange typer validitet, men den viktigste, som kanskje innebefatter alle andre måter å snakke om en målings validitet på, er konstruktvaliditet.

2.3.3 Konstruktvaliditet

Konstruktvaliditet, også kalt begrepsvaliditet, er i bunn og grunn et spørsmål om hvorvidt du måler det du ønsker å måle. En måling har god konstruktvaliditet hvis den faktisk måler det teoretiske konstruktet, og dårlig konstruktvaliditet hvis den ikke gjør det. For å gi et veldig enkelt, om enn latterlig, eksempel, anta at jeg prøver å undersøke hyppigheten av juks blant ungdomsskoleelever på norsktentamen. Måten jeg forsøker å måle det på, er å be de elevene som jukser om å reise seg i klasserommet slik at jeg kan telle dem. Når jeg gjør dette med en klasse på 30 elever, er det 0 elever som reiser seg. Så jeg konkluderer derfor med at andelen juksere i klassen min er 0 %. Det er åpenbart en urimelig slutning. Poenget her er ikke å presentere et spesielt avansert metodologisk eksempel, men å illustrere hva konstruktvaliditet innebærer. Problemet ligger i operasjonaliseringen: mens jeg prøver å måle «andelen personer som jukser», måler jeg faktisk «andelen personer som er ærlige nok til å innrømme juks, eller selvskadende nok til å late som om de juksa». Åpenbart er ikke disse det samme! Så målemetoden min har mislyktes, den måler ikke konstruktet jeg prøvde å måle, så den har dårlig konstruktvaliditet.

Det finnes flere måter en måling kan ha svak konstruktvaliditet på. For eksempel kan målingen bare måle en liten del av konstruktet, kalt (konstrukt-)*underrepresentasjon*, eller konstruktet kan måle andre ting enn det det skal, kalt (konstrukt-)*irrelevans*, se Figur 2.3.



Figur 2.3: At målemetoder fanger for lite eller for mye kalles underrepresentasjon og irrelevans

For å ha høy konstruktvaliditet kan det være nødvendig å ha høy **prediktiv validitet**. Dette betyr at målingene dine samsvarer med andre variabler – altså predikerer andre variabler – slik man ville forvente. Hvis man måler en lærers motivasjon for yrket, kan man kontrollere den prediktive validiteten ved å sjekke om de motiverte lærerne velger å bli i yrket. Hvis dette ikke er tilfelle, altså hvis høyt motiverte lærere ikke blir lenger i yrket enn lite motiverte, har målingen lav prediktiv validitet. Da har du kanskje ikke målt lærernes motivasjon.

2.4 Målemetoder

Nå har vi begrepene vi trenger for å forklare målemetodene fra Seksjon 2.1.3 én for én. Et viktig skille for å finne ut hva slags målemetode man skal bruke er om man måler noe eleven (eller personen) *gjør*, eller noe eleven *føler* og *tenker*? Ferdigheter og kompetanse er noe man gjør, og da måler man gjerne med prøver eller observasjon. Sinnstilstander og følelser kan man ikke be noen om å demonstrere, så da må du enten be folk fortelle deg om dem, eller så må du slutte deg til dem fra det du ser.

Derfor følger to gjennomgangseksempler dette delkapittelet. **Leseferdighet** er en ferdighet, og illustrerer de to første målemetodene. **Elevers indre motivasjon for matematikk** er en sinnstilstand, og illustrerer de fire siste. Motivasjonseksempelet er forholdsvis langt, men jeg håper du bruker tid på det og får en viss ærefrykt over komplikasjonene man støter på når man forsøker å måle noe i utdanningsforskning.

2.4.1 Prøver og tester

En prøve gir eleven oppgaver som skal løses. Dette er den vanligste målemetoden når konstruktet er en ferdighet eller en kompetanse, rett og slett fordi ferdigheter er noe man kan be folk om å vise fram. Eksempler er ordkjedetesten fra innledningen, nasjonale prøver, PISA og andre internasjonale undersøkelser eller en prøve lærere lager til elevene sine.

Vi så i avsnittet om operasjonalisering at nasjonale prøver og ordkjedetester ikke nødvendigvis rangerer elevene likt, og nå skjønner vi at det er på grunn av konstrukt-underrepresentasjon. En prøve rekker over et utvalg av konstruktet, sjeldent hele. Ordkjedetesten måler avkoding og sier ingenting om refleksjon over tekst, og en nasjonal prøve på halvannen time sier lite om hvordan eleven leser en roman over fjorten dager.

Til gjengjeld er reliabiliteten ofte god på prøver og tester. Poengsummen på en flervalgsprøve blir den samme uansett hvem som retter den, og en prøve med mange oppgaver gir stort sett samme resultat om eleven tar den på nytt en uke senere. Men den gode reliabiliteten faller så snart oppgavene krever skjønn i vurderingen. En norskeksamen der elevene skriver lange tekster, fanger opp mer av leseferdigheten enn en ordkjedetest gjør, men to sensorer kan lande på ulik karakter for samme besvarelse. Prøver som skal måle store kompliserte konstrukter, er ofte mindre reliable enn prøvene som måler kun en liten del av dem.

2.4.2 Registerdata

Noen verdier trenger du ikke måle selv, fordi andre allerede har registrert dem. Fødselsdato, foreldrenes utdanningsnivå og elevenes karakterer ligger i offisielle registre. Slike data er billige og dekker gjerne alle elevene i landet.

Prisen er at du må godta målingen slik den er. Bruker du standpunktkarakteren i norsk som mål på leseferdighet, har du en variabel som også fanger opp skriving, muntlig aktivitet, kunnskap om litteraturhistorie, sjangerkunnskap, og mye mer, altså konstrukt-irrelevans i rikt monn. I tillegg settes karakteren av elevens egen lærer, og lærere er ikke enige med hverandre om hva som skiller en firer fra en femmer. Registerdata er derfor sjelden den mest valide operasjonaliseringen av et konstrukt, men ofte den eneste som lar seg gjennomføre når du trenger data om mange tusen elever.

2.4.3 Selvrapportering

Selvrapportering er at folk oppgir verdien på variabelen selv, som regel ved å svare på spørsmål i et spørreskjema. Det er raskt, billig og enkelt, men det fungerer bare med folk som er ærlige og kan lese godt på det språket spørsmålene er stilt på. Selvrapportering benyttes gjerne til å måle indre sinnstilstander som tanker, holdninger og følelser.

I et eksempel over viste jeg hvordan TIMSS målte indre motivasjon ved hjelp av ni spørsmål i et spørreskjema. Da ber man elevene om å selvrapportere om sin indre motivasjon for matematikk. Dette virker godt; elevene vet selv hvilke følelser de har knyttet til matematikk, om de liker å jobbe med matematikk, gleder seg til matematikktimen og så videre. Trolig kan ingen andre, slik som foreldre eller lærere, rapportere like godt om elevens indre motivasjon, for det er tross alt elevens egen sinnstilstand man ønsker å måle. Kun eleven har tilgang til den.

Selvrapportering av indre motivasjon har høy reliabilitet. Hvis man setter elevene til å svare på spørreskjemaet en annen dag, vil de sannsynligvis svare mer eller mindre likt (test-retest-reliabilitet). Det spiller heller ikke så stor rolle hvordan påstandene som elevene skal ta stilling til, er utformet. For eksempel vil de to påstandene «jeg er glad i matematikk» eller «jeg liker å

jobbe med matematikk» sannsynligvis besvares nokså likt av en og samme elev. Og inter-rater-reliabiliteten er perfekt, for det er helt entydige svar på Likert-påstander (se Seksjon 2.2.6 for en oppfriskning) i et spørreskjema.

Reliabiliteten er altså trolig god. Validiteten er verre. Selvrapportering kan påvirkes av en rekke faktorer, for eksempel hvor sosialt akseptabelt det er å være interessert i matematikk i elevenes eget miljø. Hvis elever befinner seg i en omgangskrets der matematikk oppfattes som «ukult», kan de, bevisst eller ubevisst, nedtone sin indre motivasjon i svarene. Det motsatte kan også skje: Dersom elever går i en klasse der nesten alle misliker matematikk, men de selv synes faget er helt greit, kan de overvurdere sin egen motivasjon fordi de sammenligner seg med et miljø der interessen er enda lavere. Respondenter i intervjuer og spørreundersøkelser har også tendens til å svare det de tror at forskeren vil ha. Elever kan svare at de er interessert i matematikk fordi de tror det er det svaret forskeren ønsker. Alt dette er konstrukt-irrelevans som selvrapporteringen fanger opp.

2.4.4 Tredjepartsrapportering

Du kan også spørre andre om å oppgi verdien på variabelen, for eksempel foreldrene eller læreren til eleven. Det er utveien når den man vil måle er for ung til å svare selv, eller man måler noe som det er vanskelig å selvrapportere om.

Sett fra andre sitt ståsted kan nemlig se ut om elevens selvrapportering ikke er så pålitelig. Lærere kan observere at elever som svarer at de liker å jobbe med matematikk, ikke jobber med matematikk i timen, men heller foretrekker å tulle med sidemannen eller tegne i skriveboka si. Og omvendt kan læreren observere at elever som svarer at de er enige i at «matematikk er kjedelig» jobber iherdig og tilsynelatende motivert med individuell oppgaveløsing og kaster seg ut i helklassediskusjoner. Hva hvis elevenes selvrapportering om motivasjon ikke samsvarer med hvordan andre observerer motivasjonen?

Da er det fristende å konkludere med at læreren har rett og eleven tar feil. Men læreren ser bare eleven i matematikktimen, ser ikke inn i hodet på henne, og kan lett forveksle det å jobbe stille med det å være motivert. Vi vet ikke om den ene målingen er riktig og den andre er gal, og vi har ingen fasit å sammenlikne med.

2.4.5 Observasjon

Ved observasjon registrerer forskeren direkte det som skjer. For mange konstrukter er dette den mest valide tilnærmingen. Skal man måle undervisningskvalitet, er det åpenbart bedre å se på undervisningen enn å be læreren vurdere den selv. Prisen er tid og penger, og derfor ender de fleste likevel opp med selvrapportering, der lærere beskriver egen undervisning, eller tredjepartsrapportering, der elever rapporterer om hvordan læreren underviser.

Siden selvrapportering av indre motivasjon er utsatt for målefeilene vi så over, kan det være fristende å heller prøve å observere motivasjon. Hvis en elev er indre motivert for matematikk, innebærer det sannsynligvis at eleven jobber villig med matematikk i timen, noe som kan observeres. Da kan man ganske enkelt sette en forsker til å observere elevenes motivasjon. Men merk

at en elev som er *ytre* motivert også vil jobbe villig med matematikk i timen, så å observere kan gi en alvorlig konstrukt-irrelevans ved at det ikke er mulig å skille mellom indre og ytre motivasjon når en observerer en elev i arbeid.

Derfor må man gi matematikkoppgaver eleven i mindre grad blir ytre motivert av, for eksempel hvis læreren ikke er til stede og oppgaven ikke gir øving på stoff som skal vurderes. Vil observasjon da være et godt mål på en elevs indre motivasjon? Man kunne simpelthen målt hvor lenge eleven jobbet med en slik oppgave. Jeg liker tanken! «Hvor motivert var Petter?» «Sju minutter!» Selv om denne målemetoden unngår problemer med selvrapportering, introduserer den mange nye. For eksempel kan elevene bli påvirket til å jobbe fordi det sitter en forsker bak i klasserommet. Dessuten kan reliabiliteten bli lav, for forskere kan være uenige i akkurat når eleven slutter å jobbe med oppgaven, eller målingen kan være veldig avhengig av hvilken oppgave man bruker, slik at test-retest-reliabiliteten blir lav. Men kanskje viktigst av alt: Siden indre motivasjon er en sinnstilstand, er det unektelig rart å måle indre motivasjon ved observasjon av ytre adferd.

2.4.6 Fysiologiske mål

Fysiologiske mål, også kalt biomarkører, er lite benyttet i utdanningsforskning. Man kan følge øyebevegelsene til en leser for å se hvor hun stopper opp, du kan anslå hvor stresset en elev er under en prøve ved å måle hjerteslagene, eller du kan måle hjerneaktivitet i et elektroencefalogram ved å koble elektroder til hodet. Dette høres veldig forlokkende ut. Kroppen svarer jo ikke strategisk; pulsen bryr seg ikke om hva som er kult i klassen; elektroencefalogrammet kan ikke fakes.

Det er derfor fristende å droppe både selvrapportering og observasjon av indre motivasjon og heller finne en biomarkør. Kan det finnes et hormon eller en nevrotransmitter som slår ut når man jobber med en oppgave man er indre motivert for? Kanskje dopamin er et sånt stoff? Eller kan man måle puls eller hjernebølger? Man kunne målt en biomarkør i elevenes kropper mens de jobbet med matematikk og slik fått en objektiv måling av elevenes motivasjon for matematikk. Det er lite som tilsier at slike biomarkører faktisk ville fungere i praksis, og det er sannsynlig at man fortsatt ville undret seg over hvorfor målingene fra biomarkøren ikke samsvarer med selvrapportert eller observert indre motivasjon.

2.4.7 Når målemetodene ikke er enige

Dette er til å bli gal av. Jeg har presentert fire målemetoder for å måle indre motivasjon, og ingen av dem stemmer overens med de andre. Dette er ikke kun et tankeeksperiment; det er svært reelt. Manglende samsvar mellom selvrapportering og observasjon går igjen i måling av mange konstrukter som handler om elevers og læreres indre sinnstilstander, slik som tanker, holdninger, motivasjon, kunnskap, læring og følelser (Leatham, 2006).

Men det er også ofte manglende samsvar når konstruktet ikke er en indre sinnstilstand, men handler om helt konkrete hendelser i et klasserom. For eksempel sier mange lærere at de synes helklassediskusjoner er viktige, at de setter av mye tid til det i timene og at de gir elevene rom til å diskutere heller enn å kun komme med korte svar. Observasjon av helklasseundervisning

har imidlertid vist det motsatte: Det er få helklassediskusjoner, og de er ofte lærerstyrte der elevene kun kommer med korte svar. Altså har målinger av helklassediskusjoner via selvrapportering lav validitet. Er elever og lærere løgnere, som ikke velger å svare sannferdig; hyklere, som gjør stikk motsatt av det de sier er viktig; eller er det noe annet som foregår? En nærliggende forklaring er rett og slett at selvrapportering er vanskelig, slik at vi ikke kan vente annet enn avvik mellom observasjon og selvrapportering.

Huff. Hvis du led under forvirringen at å måle noe var enkelt, håper jeg du nå er kurert.

2.5 Etiske sider ved kvantitativ måling av folk

Det er ikke bare vanskelig å lage reliable og valide målinger av mennesker; målingene kan også ha dype etiske konsekvenser.[7] Jeg viser deg bare to av de mest nærliggende etiske problemene.

Det første etiske problemet kan oppsummeres med at «målinger skaper folk». Alex starter i første klasse som en vanlig gutt. På femte trinn tar han nasjonale prøver i lesing, og havner på «mestringsnivå 1». Underveis forsvinner usikkerheten prøven er beheftet med (kanskje Alex skåret rett under grensen for nivå 2), og at prøven var urettferdig for ham (Alex sleit med teksten som handlet om skøyter, for han har hverken sett eller prøvd skøyter). Igjen står én opplysning, mestringsnivå 1.[8]

I de voksnes (og kanskje Alex sitt) hode blir «mestringsnivå 1» **tingliggjort**, at man behandler noe som virkelig bare fordi det er satt et begrep på det. «Mestringsnivå 1» er strengt tatt ikke annet enn en betegnelse på at skåren havnet på den ene siden av en grense noen har trukket, men på lærerværelset og rundt kjøkkenbordet blir Alex fort til «en svak leser» (Mosvold & Ohnstad, 2016). Kategorien har sluttet å være en prestasjon på en prøve og har blitt til en egenskap ved Alex. Ingenting har skjedd med Alex, men «mestringsnivå 1» har liksom blitt noe virkelig ved ham i de voksnes øyne.

Derfor kan merkelappen «mestringsnivå 1» bli en selvoppfyllende profeti for Alex. En stein bryr seg ikke om hvilken kategori en geolog sier den tilhører, men mennesker merker at de blir satt i en kategori, og de blir påvirket. Læreren senker forventningene og gir Alex lettere tekster, foreldrene slutter å foreslå bøker til jul, og Alex, som nå vet hva han er, slutter å prøve seg på tykke bøker. Etter et års tid leser han dårligere enn han ville gjort uten merkelappen. Kategorien har ikke bare beskrevet Alex, den har vært med på å lage ham. Målinger skaper folk, eller, i hvert fall kan de være delaktige i det.

Man kan kanskje innvende «men han leser jo litt dårlig, da, og det må det være fint å finne ut av slik at han kan få hjelp», og det kan være riktig! Noen ganger utløser målingene ekstra ressurser. Kanskje blir Alex kalt inn til spesialpedagoger som tar en annen kvantitativ måling, en som skal kartlegge

[7] Mange forskere mener at konsekvensene som målingen har for folk er en del av dens validitet, såkalt konsekvensvaliditet. Dette er omstridt, fordi det plumper ut i debatten om vitenskapelig tenkning bør være fri for verdispørsmål, det *verdifrie ideal*.

[8] En pikant detalj er at en feil i det statistiske analyseverktøyet som ble benyttet av nasjonale prøver gjorde at målingene ble gale fra 2015 til 2021. Det er anslått at 200 000 elever ble gitt feil mestringsnivå (https://www.utdanningsnytt.no/nasjonale-prover-utdanningsdirektoratet-utdanningsforbundet/over-200000-nasjonale-prover-ble-feilvurdert-av-udir/470804). På grunn av feilen oppdaget man heller ikke at norske elevers regne- og leseferdigheter gikk ned i perioden, mens elevenes engelskferdigheter gikk opp.

dysleksi. Sier prøven at Alex har dysleksi, får han mange rettigheter; sier den nei, får han ingenting. La oss inderlig håpe grenseverdien er satt på bakgrunn av en svært grundig vurdering av både etikk og validitet.

Det andre etiske problemet kan oppsummeres med at «målinger forvrenger målsetninger og målingen selv». Kvantitative målinger benyttes til å fordele knapphetsgoder som ressurser og anseelse. Skoler med gode resultater får anseelse; politikere som sitter med makten når PISA-resultatene går opp får smigrende omtale; lærere med lav strykprosent blir godt likt av rektor og elever, og så videre. Derfor er det en fare for at man, bevisst eller ubevisst, legger om målene sine til å skåre bedre på målingen.

Alle skoleeiere har som mål at elevene skal lære å lese bedre. Da kan det være de måler skolene på hvor stor andel elever som havner på mestringsnivå 1 på nasjonale prøver i lesing. Da er det stor fare for at skolene endrer fokus; for eksempel at de bare fokuserer på de elevene som er gode nok til å kunne nå mestringsnivå 2, og ikke på de som virkelig sliter. Da er målsetingen forvrengt fra å lære elevene å lese til å flytte dem over en litt vilkårlig grense. Dette er ikke bare en teoretisk fare. I en amerikansk barneskole under et testbasert ansvarsregime fant Booher-Jennings (2005) at lærerne systematisk kanaliserte ressursene mot elevene som lå like under bestått-grensen, kalt «bubble kids», mens de svakeste elevene ble nedprioritert eller henvist til spesialundervisning slik at resultatene deres ikke skulle telle med.

Men også målingen selv forvrenges. I USA er det godt kjent at skoler og hele skoledistrikter legger om undervisningen til å dreie seg om flervalgsprøver, og at noen har holdt svake elever borte fra statlige prøver for å heve skolens resultater (Koretz, 2017). Da stiger skårene uten at elevene leser bedre, så prøven har mistet validiteten som et mål på lesekompetanse.

Dette er kjent som *Campbells lov*: Jo mer en kvantitativ indikator brukes til å ta beslutninger, jo mer utsatt blir den for press som ødelegger både indikatoren selv og det den skulle måle (Campbell, 1979). Campbell selv brukte nettopp skoleprøver som eksempel!

Du får kanskje lyst til å droppe kvantitativ måling helt, men det er ikke noe jeg vil anbefale; kvantitative målinger kan like gjerne ha gode konsekvenser. Og alternativet er ofte beslutninger bli tatt av folks profesjonelle (og uprofesjonelle) skjønn, som inneholder trynefaktorer, fordommer og andre uspesifiserte preferanser. En profesjonell lærer møter derfor kvantitative målinger verken med uforbeholden tillit eller avvisning.

2.6 Oppsummering

Målinger er en viktig del av kvantitativ forskning, og trolig er vi for lite bevisste på kvaliteten på målingene når vi evaluerer forskning. I første kapittel refererte jeg til en samtale mellom oljefondets leder Nicolai Tangen og forskeren Angela Duckworth. Da Duckworth sa at hun hadde forsket på folks pågangsmot og at hun hadde veldig store utvalg, burde ikke Tangen blitt imponert over utvalgsstørrelsen. Han burde spurt: «Hvordan målte du pågangsmotet til så mange folk?» Svaret er jo at alle har svart på et spørreskjema med ti spørsmål. Tror du dette spørreskjemaet måler pågangsmot på en god måte?

I dette kapittelet har du lest om:

- Teoretiske konstrukter og operasjonalisering. Hva betyr det å operasjonalisere et teoretisk konstrukt? Hva betyr verdier, variabler, konstrukter og målemetoder?
- Variabler og verdier.
- Målenivåer og variabeltyper. Husk at det er to forskjellige distinksjoner her. Det er forskjellen mellom diskrete og kontinuerlige variabler, og det er forskjellen mellom fire ulike målenivå (nominal, ordinal, intervall og forholdstall).
- Vurdering av målingers reliabilitet. To typer reliabilitet, og begge handler om man får samme resultat om man måler den samme tingen to ganger.
- Vurdering av målingers validitet. Måler målemetoden din det du ønsker? Hvis ikke, har målingen konstrukt-underrepresentasjon eller konstrukt-irrelevans?
- Målemetoder. Prøver og tester, registerdata, selvrapportering, tredjepartsrapportering, observasjon og fysiologiske mål. Hvilken som passer, kommer an på konstruktet. Merk komplikasjonene med å måle sinnstilstander.
- Etiske sider ved kvantitativ måling av folk. «Målinger skaper folk» (tingliggjøring) og «målinger forvrenger målsetninger og målingen selv» (Campbells lov), men likevel er det ikke sikkert du bør droppe kvantitative målinger.

3. Forskningsdesign i kvantitativ metode

3.1 Formålet med forskningen

Man skiller ofte mellom tre ulike formål for kvantitative undersøkelser, å beskrive noe, å predikere noe og å finne årsaksforhold. Vi går gjennom hver i tur.

3.1.1 Beskrive

Beskrivende forskning er en av de mest grunnleggende formene for kvantitativ forskning i utdanningsfeltet. Beskrivende kvantitativ forskning har som hovedformål å gi et systematisk og nøyaktig bilde av hvordan en eller flere variabler fordeler seg i en bestemt populasjon. Denne typen forskning søker å svare på spørsmål som «hva», «hvor mange» og «hvor mye», men ikke nødvendigvis «hvorfor». I utdanningsforskning kan beskrivende studier for eksempel kartlegge:

- Gjennomsnittlig karakternivå på videregående skoler i ulike fylker
- Ungdomsskolelæreres utdanningsbakgrunn
- Leseferdighetene til femteklassinger i Norge

Disse eksemplene beskriver bare én variabel av gangen. Her ønsker forskeren å få oversikt over hvordan denne variabelen ser ut i populasjonen. Dette kan innebære å beregne sentralmål som gjennomsnitt, median og typetall, samt spredningsmål som standardavvik og variasjonsbredde, eller å vise dataene med en passende figur.

Ofte er forskere interessert i å beskrive sammenhenger mellom to eller flere variabler. Et klassisk eksempel fra internasjonal utdanningsforskning er å undersøke sammenhengen mellom elevers leseferdigheter og hvilket land de kommer fra. Gjennom studier som PISA (Programme for International Student Assessment) kan forskere beskrive hvordan leseferdighetene varierer systematisk mellom land. De kan rapportere at elever i Norge i gjennomsnitt skårer lavere enn elever i Finland, uten nødvendigvis å forklare hvorfor denne forskjellen eksisterer.

Beskrivende forskning er viktig selv om den ikke gir oss forklaringer på hvorfor noe skjer. Beskrivende studier gir oss:

- **Oversikt over status quo**: Hvor står vi i dag innenfor et bestemt område?
- **Grunnlag for sammenligning**: Hvordan ser situasjonen ut i ulike grupper eller kontekster?
- **Identifikasjon av mønstre**: Finnes det systematiske forskjeller eller likheter som fortjener nærmere oppmerksomhet?
- **Utgangspunkt for årsaksforklaringer**: Hvilke sammenhenger observerer vi som kan undersøkes videre?

Beskrivende forskning er derfor ikke bare et første steg, men en verdifull forskningsaktivitet i seg selv som bidrar til å bygge kunnskapsgrunnlaget i utdanningsfeltet.

3.1.2 Predikere

Prediksjon er et viktig formål innen utdanningsforskning som handler om å forutsi fremtidige utfall basert på tilgjengelig informasjon. I motsetning til beskrivende studier som fokuserer på å kartlegge eksisterende forhold, har prediksjonsbasert forskning som mål å kunne si noe om hva som vil skje i fremtiden.

En av de mest utbredte anvendelsene av prediksjon i utdanningsfeltet er kartleggingsprøver. Disse prøvene administreres tidlig i elevenes skoleløp, ofte på 1.-3. trinn, med det formål å identifisere elever som kanskje ikke vil klare å tilegne seg lærestoffet uten ekstra tilrettelegging. Målet er å kunne sette inn tiltak på et tidlig tidspunkt, før problemene blir for store. Kartleggingsprøver i lesing kan for eksempel teste elevers fonologiske bevissthet, bokstavkunnskap og avkodingsferdigheter. Basert på resultatene fra disse prøvene ønsker man å predikere hvilke elever som vil ha vanskeligheter med å følge ordinære undervisning. Elever som scorer lavt på slike prøver, kan dermed få tilbud om intensivert opplæring eller spesialundervisning.

Et sentralt problem med prediksjonsmodeller er at de forutsigelsene man ønsker å gjøre, ofte baserer seg på data som er generert under andre forhold enn de dataene modellen skal benyttes på. For eksempel kan en kartleggingsprøve være utviklet og testet på elever i nærheten av NTNU i Trondheim i 2010. Da er det ikke sikkert den fungerer godt i Nordland eller på Nordstrand. Når en prøve tas i bruk på elever med andre kjennetegn eller under andre samfunnsforhold vil prediksjonene ikke bli like gode.

3.1.3 Finne årsaksforhold

Det tredje formålet med kvantitativ utdanningsforskning er å finne årsaksforhold, altså ikke bare beskrive eller predikere noe, men å forklare årsaken til at det er slik. Dette kalles å stadfeste **kausalitet**. Når vi kan etablere kausalitet, kan vi si at en variabel faktisk forårsaker endringer i en annen variabel. Man uttrykker ofte årsaker med piler. Hvis årsaken er A og virkningen er B kan man uttrykke «A forårsaker B» som $A \rightarrow B$. Man kaller A for **årsaken** og B for **virkningen** eller **effekten**.

La oss se på et eksempel på kausalitet. På en skole sliter de med at elevene på femtetrinn har for dårlig engelsk ordforråd. De setter i gang en klassekonkurranse, der den klassen som får høyest gjennomsnitt på en stor gloseprøve på slutten av året vinner. På slutten av året finner de at femtetrinn skårer skyhøyt på ordforråd. Da har klassekonkurransen **forårsaket** en forbedring i ordforråd, altså

$$\text{klassekonkurranse} \rightarrow \text{økt ordforråd} \tag{3.1}$$

Vi kan si det er en **kausal sammenheng** mellom klassekonkurransen og ordforrådet; økt ordforråd var en **virkning** av klassekonkurransen.

Årsaksforklaringer er viktige i utdanningsforskning fordi hvis vi vet hva som forårsaker noe, så har vi ofte kontroll over det og kan endre det. Hvis

vi hadde visst hva som forårsaket det dårlige engelsk-ordforrådet kunne vi fjernet årsaken, og fordi vi vet at klassekonkurransen har en kausal effekt på ordforrådet kan vi utbedre det. Hvis vi hadde visst hva som forårsaket at gutter oftere ikke fullfører videregående enn jenter kunne vi kanskje gjort noe med det. Kunnskap om årsaker er viktige for lærere, rektorer, kommuner, politikere – rett og slett for alle som vil endre skolefeltet til det bedre. Hvis vi hadde visst hva som hadde fått norske elever til å trives enda bedre på skolen og skåre høyere på internasjonale undersøkelser er jeg temmelig sikker på at politikere kunne satt av penger til det. Problemet er at å finne ut at noe forårsaker noe annet, altså å stadfeste kausalitet, er vanskelig.

3.2 Variablenes roller: avhengige og uavhengige variable

Forskningsformålene (beskrive, predikere og årsaksforklare) gir oss to forskjellige typer Variabler. Hvis man vil beskrive kjønnsforskjeller i motivasjon for kroppsøving, tenker man gjerne at «kjønn» forklarer motivasjon for kroppsøving og ikke motsatt, at motivasjonen for kroppsøving forklarer hvilket kjønn du er. Når vi predikerer noe, er det én variabel vi ønsker å predikere og mange andre variable vi ønsker å bruke til å gjøre prediksjonen. I en årsaksforklaring er én variabel virkningen og potensielt flere variabler er årsakene. Det er viktig å holde de to rollene - «ting som forklarer» og «ting som blir forklart» - adskilt fra hverandre.

La oss være tydelige på dette ved å introdusere matematiske symboler for å beskrive variabler (noe du vil møte igjen og igjen). Vi kan betegne variabelen som «skal forklares» som Y, og variablene som «gjør forklaringen» som X_1, X_2, og så videre. Når vi gjør en analyse, har vi forskjellige navn for X og Y siden de spiller ulike roller. De klassiske navnene er **uavhengig variabel** og **avhengig variabel**. Uavhengig variabel er variabelen du bruker til å gjøre forklaringen (altså Xene), mens avhengig variabel er variabelen som blir forklart (altså Y). Logikken bak disse navnene er at hvis det virkelig er et årsaksforhold mellom X og Y, kan vi si at Y avhenger av X, mens X er uavhengig av Y.

Jeg må innrømme at jeg synes disse navnene er problematiske. De er vanskelige å huske og ganske misvisende, fordi (1) uavhengig variabel er stort sett avhengig av en hel masse variable, og (b) hvis det ikke er noe årsaksforhold, avhenger ikke den avhengige variabelen faktisk av den uavhengige variabelen. Siden jeg ikke er alene om å synes «uavhengig variabel» og «avhengig variabel» er uklare navn, finnes det heldigvis flere alternativer som er mer intuitive. Begrepene jeg vil bruke for Y i denne boka er **utfall** eller **utfallsvariabel**. Tanken her er at samme om du beskriver en sammenheng, predikerer en variabel eller årsaksforklarer en variabel, så er «utfall» et vanlig ord å bentytte som er mindre forvirrende enn avhengig variabel. Begrepene jeg vil bruke for variablene X_1, X_2, og så videre, er **forklaringsvariabel**, **prediktor** og **årsaksvariabel**. Tanken her er at ingen av ordene er egentlig dekkende på tvers av de tre forskningsformålene, så da er flere ord i vanlig bruk. Det er ikke en regel at man må bruke ordet kun til det

ene forskningsformålet, og du kommer til å se at ordene blir benyttet om hverandre, sikkert i denne boken også.[9] Dette er oppsummert i Tabell 3.1.

Tabell 3.1: Navn på variablenes roller

navn	skal forklares	skal forklare
matematisk symbol	Y	X
klassisk navn	avhengig variabel	uavhengig variabel
beskrivende navn	utfall	forklaringsvariabel
predikerende navn	utfall	prediktor
årsaksforklarende navn	utfall	årsaksvariabel

3.3 Eksperimentelle studier og observasjonsstudier

Et av de viktigste skillene du bør kjenne til er forskjellen mellom forskning som benytter **eksperimenter** og forskning som ikke gjør det. Dette skillet handler om hvor mye kontroll forskeren har over deltakerne og situasjonen i studien.

Et **eksperiment** er en studie der forskeren prøver å kontrollere en eller flere sider ved situasjonen i studien. For eksempel kan det være forskeren ønsker å undersøke hvordan elevene responderer på et nytt undervisningsopplegg hen har laget. Da må forskeren forsikre seg om at undervisningsopplegget blir benyttet. Det kan gjøres ved at forskeren eller en assistent underviser opplegget, eller at de får en lærer til å utføre opplegget. Det opplegget forskeren har planlagt for å påvirke undervisningen kalles en **intervensjon**. Forskeren manipulerer eller varierer forklaringsvariablene (altså undervisningsopplegget) med intervensjonen, mens utfallsvariabelen (elevenes respons) får variere naturlig. Da kan forskeren studere verden slik den ser ut når forklaringsvariabelen er akkurat slik forskeren ønsker.

En sentral trussel mot gyldigheten til eksperimentelle studier er at intervensjonen ikke virker. I utdanningsforskning er det vanlig å ville teste forskjellige undervisningsopplegg, men det er vanskelig for lærerne å følge undervisningsopplegget til punkt å prikke, kanskje fordi de har for dårlig tid til å sette seg inn i opplegget eller fordi opplegget er så annerledes fra slik de vanligvis underviser at de ikke får det til. I så fall kan det være forskerne tror de forsker på *undervisningsopplegget slik det er planlagt*, men i virkeligheten forsker de på *undervisningsopplegget slik det ble seende ut i eksperimentet*. Dette kan også være interessant, men da er det viktig å være klar over forskjellen og at man ikke tror intervensjonen har kontrollert forklaringsvariabelen. Derfor trengs ofte kvalitativ følgeforskning for å forsikre seg om at intervensjonen var vellykket.

[9] Det finnes dessverre mange forskjellige navn som brukes i litteraturen. Jeg vil ikke liste alle – det ville ikke tjene noen hensikt – bortsett fra å nevne at «responsvariabel» noen ganger brukes der jeg har brukt «utfall». Denne typen terminologisk forvirring er dessverre svært vanlig i forskningsfeltet.

Studier uten eksperimenter kalles **observasjonsstudier**. Da prøver forskeren å unngå å påvirke situasjonen hen forsker på. Man kan for eksempel sette opp små kameraer og mikrofoner i mange klasserom og gjøre opptak av undervisningen – etter å ha søkt etisk godkjenning og innhentet samtykke, så klart. En sentral trussel mot gyldigheten til observasjonsstudier er at forskeren påvirker situasjonen, samme hvor mye hen prøver å la være. Det kan jo være læreren forbereder seg ekstra til undervisningtimene som skal forskes, eller at elevene oppfører seg annerledes når kameraene gjør opptak.

3.4 Korrelasjon er ikke kausalitet

En sentral innsikt i kvantitativ forskning er at korrelasjon ikke nødvendigvis er kausalitet. Vi skal lære mer presist senere hva korrelasjon er, men for nå holder det å vite at to variable er korrelert hvis den ene variabelen har en høy verdi når den andre variabelen har det, og omvendt, at de har lave verdier når den andre har det. Forskere vet at det er en sterk positiv korrelasjon mellom variabelen *antall bøker hjemme* og variabelen *elevers leseferdigheter*, altså at at elever med flere bøker hjemme har en tendens til å lese bedre.[10] Man kunne lett tenke at det å ha mange bøker hjemme forårsaker bedre leseferdigheter. Tenk over dette – tror du at det stemmer? I så fall sier du deg enig i at en elevs leseferdighet hadde gått opp hvis du hadde kjøpt en større bokhylle til familien og fylt den med masse bøker.

La oss analysere korrelasjonen i detalj. La oss kalle antall bøker for A og leseferdighet for B. Hvis det hadde vært en årsakssammenheng mellom A og B, altså $A \to B$, så hadde A og B vært korrelerte, og korrelasjonen hadde oppstått på grunn av årsakssammenhengen mellom A og B. Vi kan oppsummere det i en regel

En årsakssammenheng $A \to B$ skaper en korrelasjon mellom A og B.

Dette er et nyttig resultat som jeg tror er enkelt å skjønne.

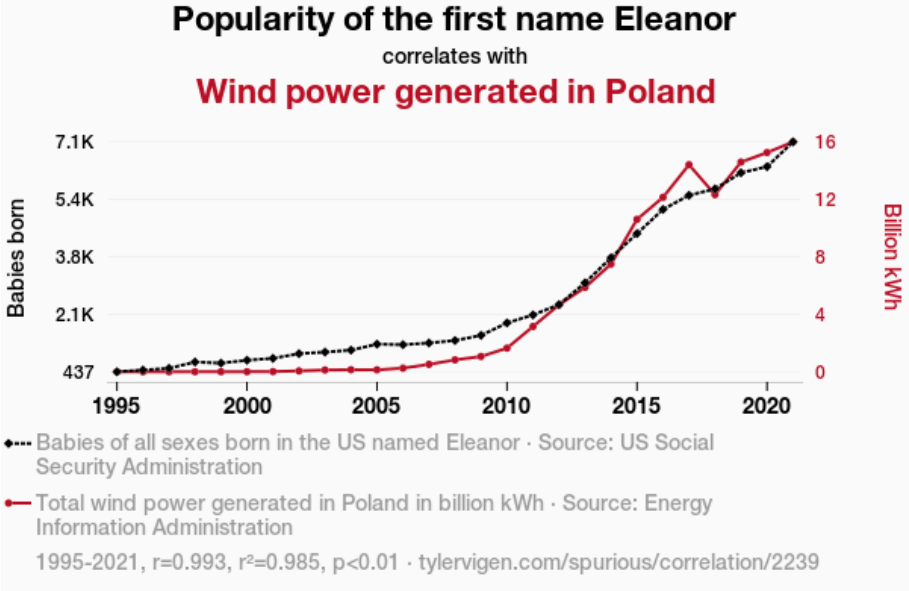
> ⚠ Advarsel
>
> Vel, regelen er sann i de fleste tilfeller. Men det *kan* være at A har en årsakspåvirkning på B, men så er det noe som hindrer B fra å inntreffe. Det å utføre et vellykket ran, A, kan gjør deg mer tilbøyelig til å utføre flere ran, B, men B vil ikke inntreffe hvis man blir arrestert først. I så tilfelle vil ikke A og B bli korrelert. Du kan sammenlikne det med å påføre kraft på en boks; det vil forårsake at boksen flytter seg (korrelasjon mellom *å påføre kraft* og *bevegelse*), men det gjelder ikke hvis boksen står inntil en vegg.

Men den «motsatte» regelen er ikke sann – det er ikke slik at en korrelasjon mellom A og B betyr at man kan konkludere med $A \to B$. Dette er ikke like enkelt å skjønne. Hvorfor i huleste heiteste er det korrelasjon hvis den ene

[10] OK, jeg vet ikke om dette gjelder lenger nå som alle har digitale dingser. Men før var det i hvert fall sterk korrelasjon mellom antall bøker foreldrene hadde i bokhylla hjemme og alt mulig av positive utfall for elevene.

ikke forårsaker den andre? Hvis de var helt urelaterte burde det jo være ingen korrelasjon? Er korrelasjonen bare en tilfeldighet?

Tilfeldige korrelasjoner har blitt manges forklaring på hvorfor korrelasjon ikke er kausalitet. En tilfeldig korrelasjon er en korrelasjon som oppstår av, nettopp, tilfeldigheter. Hvis man hadde gjort samme studie om igjen hadde ikke korrelasjonen vært der, fordi den oppsto på grunn av tilfeldigheter med utvalget eller noe annet. Ideen om tilfeldige korrelasjoner har blitt popularisert gjennom det meget vittige nettstedet https://www.tylervigen.com/spurious-correlations som generer tusenvis av tilfeldige korrelasjoner som Figur 3.1.



Figur 3.1: En tilfeldig korrelasjon. Figuren er fra www.tylervigen.com og er utgitt under en CC-BY-lisens.

Korrelasjonen i Figur 3.1 er mellom antall babyer i USA som heter Eleanor og antall kilowattimer vindkraft produsert i Polen. Det er ingen årsakssammenheng mellom variablene, for hvis alle i USA kalte jentebarna sine Eleanor ville ikke vindkraftutbyggingen i Polen skutt fart. Denne korrelasjonen er nok helt tilfeldig.

Men de fleste korrelasjoner er ikke tilfeldige, man kan finne dem igjen i studie etter studie. Det kan være helt tydelig at korrelasjonen alltid vil være der, men *likevel* trenger det ikke være slik at den ene variabelen forårsaker den andre. Vi skal se på to problemer som gjør at man ikke kan konkludere med en årsakssammenheng mellom to variable selv om de er korrelerte, retningsproblemet og konfundere.

3.4.1 Retningsproblemet

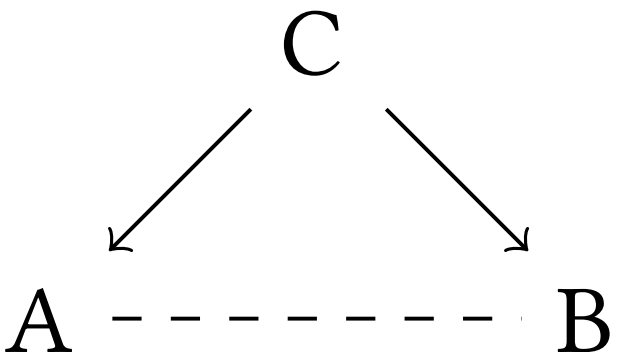
Retningsproblemet er ganske enkelt at hvis man har en korrelasjon mellom A og B, så kan det like gjerne være at A påvirker B som at B påvirker A. Man kan rett og slett ikke finne ut av hvilken vei kausaliteten går – hvilken av følgende piler som stemmer

I eksempelet vårt betyr retningsproblemet at vi ikke vet om antall bøker hjemme fører til økt leseferdighet eller om elevens leseferdighet fører til mange bøker hjemme. Hvis et barn er veldig flink og interessert i å lese, så høres det jo sannsynlig ut at foreldrene kjøper flere bøker for å imøtekomme barnets interesse. Det trenger altså ikke være at å ha mange bøker hjemme gjør at eleven får mye god lesetrening.

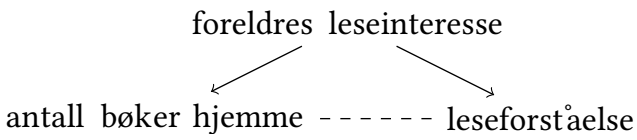
3.4.2 Konfundere

En **konfunder** til variablene A og B er en variabel C som er slik at den påvirker både A og B. Altså $C \to A$ og $C \to B$. Ofte tegner man dette i et diagram slik som i Figur 3.2a.

> **i** Note
>
> «Konfunder» uttales med samme trykk som «vidunder». *Konfunder* er en norsk oversettelse av det engelske *confounder*. Det finnes dessverre ingen vanlige oversettelser av *confounder*, og både tredjevariabel, forvirringsvariabel, forstyrrende variabel, og bakenforliggende faktor er i bruk, i tillegg til konfunder. Konfunder er like vanlige som disse, og jeg bruker konfunder slik at man lettere forstår ordet hvis man leser det på engelsk. Å konfundere betyr egentlig å forvirre.



(a) Her er C en konfunder for A og B.



(b) Her er «foreldres leseinteresse» en konfunder for korrelasjonen mellom «antall bøker hjemme» og «leseforståelse».

Figur 3.2:

Konfundere kan meget gjerne være grunnen til at det er en korrelasjon mellom antall bøker hjemme og en elevs leseferdighet. I Figur 3.2b har vi foreslått en mulig konfunder, nemlig foreldrenes leseinteresse. Foreldrene med høy

leseinteresse har trolig mange bøker hjemme. Foreldrenes leseinteresse kan også påvirke barnas leseinteresse, kanskje fordi de har lest mye for barna eller fordi de har gener for leseinteresse som er brakt videre til barna, noe som gjør at barna også er interesserte i å lese. Da oppstår det en korrelasjon mellom antall bøker hjemme og elevens leseinteresse, men det er ikke fordi antall bøker hjemme påvirker leseinteressen, det er fordi foreldrenes leseinteresse påvirker *både* antall bøker hjemme og elevens leseinteresse.

Det finnes statistiske teknikker for å fjerne påvirkningen til en konfunder. Dette kalles å «kontrollere for» eller å «ta hensyn til» variabelen, og er en av hovedgrunnene til å bruke, for eksempel, regresjonsanalyse, som ikke dekkes i denne boka. Slike statistiske teknikker kan fungere fint, men er ofte utilstrekkelige. Et hovedproblem er at man bare kan kontrollere for variable man har målt. I en stor studie kan det være man har målt 8-10 variable, og disse variablene kan man kontrollere for, men det finnes uendelig antall variable man ikke har målt. Derfor kan det finnes flere viktige konfundere man ikke har målt. Dette kalles **ikke-målt konfundering** og er et problem det ikke finnes statistiske teknikker for å unnslippe. Hvis du har tviholdt rundt håpet om å lære en statistisk teknikk som lar deg finne kausalitet i korrelasjoner bør du slippe taket nå – dette håpet vil kun føre til skuffelse.

At korrelasjon ikke betyr kausalitet høres enkelt ut, nesten banalt, men all erfaring viser at det ikke er enkelt – folk tråkker rett og slett feil for ofte. I avisoverskrift etter avisoverskrift hører man om «sammenhengen mellom tomater og kreft» og liknende. Uten unntak er dette aviser som hinter til kausalitet (tomater fører til kreft) ut fra korrelasjoner (folk som spiser tomater får oftere kreft). Også i utdanningsforskning er det mange eksempler. I mange (hver eneste?) av årgangene av PISA-studiene er variabelen «utforskende arbeidsformer i naturfag» korrelert med lavere skår på variabelen «naturfagskompetanse» (Jensen et al., 2024). Hver eneste gang benytter grupper som ønsker mer lærerstyrt undervisning dette resultatet til å skrike høyt om at utforskende arbeidsformer i naturfag gir dårligere resultater, altså en kausal slutning. Man benytter ganske enkelt kvantitativ forskning til å bekrefte eksisterende oppfatninger. (En god øving kan være å forklare hvorfor denne kausale slutningen ikke nødvendigvis holder, altså komme opp med noen konfundere eller vise at retningsproblemet kan gjelde? Se i informasjonsboksen for fasit.)

i Note

Retningsproblemet gjelder i hvert fall: Det kan være at hvis lærere har en elevgruppe som har lav naturfagskompetanse, så gjør man mye utforskende undervisning.

En mulig konfunder kan være variabelen «læreren har spesialisering i naturfag». Lærere med spesialisering i naturfag vil kanskje gi elevene høyere naturfagskompetanse. Hvis lærerne med spesialisering i naturfag gjør mindre utforskende undervisning, vil det oppstå en korrelasjon mellom høy naturfagskompetanse hos elevene og lite utforskende undervisning.

Foreldrepress er en annen mulig konfunder. Kanskje ønsker foreldrene til de akademisk sterkeste barna tradisjonell undervisning, ikke noe sånn «nymotens utforskende nånnsens», og legger press på læreren. I så fall vil «foreldrepress» konfundere korrelasjonen.

Jeg må legge inn et tilbakeblikk til Kapittel 2. også, fordi det er så nærliggende for meg å tro at korrelasjonen oppstår på grunn av svak måling av variabelen «utforskende undervisning». Variabelen «naturfagskompetanse» tror jeg er glitrende målt av PISA, for PISAs hovedformål er å måle elevenes kompetanse. Jeg tror derimot at «utforskende undervisning» er målt dårligere. For det første er det målt med spørreskjema til elevene, og jeg tror ikke elevene er så presise informanter om hvorvidt de har hatt utforskende undervisning. Variabelen er målt med Likert-spørsmål som *Students are given opportunities to explain their ideas*, og jeg tror sterke og svake elever svarer forskjellig på dette spørsmålet. Det kan jo være at svaktpresterende elever synes de blir spurt om å forklare ting alt for ofte, mens de høytpresterende gjerne skulle forklart ting oftere. Da vil svakt- og høytpresterende elever svare forskjellig på spørsmålet, selv om de har blitt spurt om å forklare ideene sine like ofte. Da vil korrelasjonen oppstå på grunn av en målefeil!

Måleproblemene er uendelige i dette tilfellet. Tror du at en norsk elev, en kinesisk elev og en estisk elev vil svare sammenliknbart på disse spørsmålene? Det *må* de gjøre hvis det skal være noen vits i å sammenlikne på tvers av land! Man kan kose seg en hel helg med å tenke over alle måleproblemene som kan få denne korrelasjonen til å oppstå, noe jeg vil anbefale på det sterkeste.

Man kan også gå i motsatt felle, at man blankt avviser alle årsaksforklaringer basert på at to variable er korrelert. Men husk, en kausal sammenheng mellom to variable skaper en korrelasjon. En korrelasjon er derfor et sterkt hint om at det kan være en kausal sammenheng mellom variablene og bør ikke avvises blankt. Om det er lurt å konkludere med kausalitet basert på en korrelasjon bør vurderes fra tilfelle til tilfelle, og vurderingen kommer til å avhenge av det teoretiske ståstedet ditt.

3.5 Randomiserte kontrollstudier

Forrige delkapittel forklarte hvorfor det er vanskelig å finne ut at noe forårsaker noe annet. Heldigvis finnes det et studiedesign man kan bruke for å

utelukke alle tenkelige konfundere, nemlig **randomiserte kontrollstudier**. Randomiserte kontrollstudier kalles også *randomiserte kontrollerte forsøk* og *randomiserte kontrollerte eksperimenter*, og du møter ofte på begrepet på engelsk, der det heter *randomized controlled trials*. Først forklarer vi hva slike studier er, og så forklarer vi hvorfor det selv med randomiserte kontrollstudier ikke er enkelt å finne ut av årsaksforklaringer.

En randomisert kontrollstudie er et eksperimentelt forskningsdesign hvor deltakere tilfeldig tildeles ulike grupper for å teste effektiviteten av en intervensjon. I utdanning betyr dette å tilfeldig dele inn elever, lærere eller skoler i en **intervensjonsgruppe** og en **kontrollgruppe**. Intervensjonsgruppa mottar intervensjonen, som kan være en ny undervisningsmetode, en digital dings, et antimobbeopplegg eller hva som helst annet. Kontrollgruppa får ikke intervensjonen.

Det som gjør randomiserte kontrollstudier så gode er **tilfeldig gruppeinndeling**. Når vi fordeler deltakerne til ulike grupper basert på ren tilfeldighet, oppnår vi noe ganske elegant: de som naturlig ville reagert positivt eller negativt på intervensjon blir jevnt spredt utover alle gruppene. Resultatet? Eventuelle forskjeller vi ser i resultatene vil skyldes intervensjonen vi tester, og ikke at de som responderer best eller dårligst på intervensjonen tilfeldigvis havnet i samme gruppe.

Forklaringen i avsnittet over, at randomiseringen gjør at verdien på utfallsvariabelen blir jevnt fordelt på tvers av gruppene, er tilstrekkelig, men mange lurer likevel på hvorfor ikke konfundere er et problem. Vil ikke konfundere (for eksempel underliggende forskjeller mellom deltagerne) kunne påvirke resultatet? Nei, for alle de underliggende forskjellene mellom deltakerne – som evner, motivasjon, bakgrunn og andre personlige egenskaper — blir også balansert mellom intervensjons- og kontrollgruppen.

Et eksempel på en randomisert kontrollert studie fra Norge er om lærende og låste tankesett (Bettinger et al., 2018). Forskerne utviklet en intervensjon som ønsket å påvirke elevenes tankesett til å være mer lærende, altså slik at elevene får tro på at de kan lære. Intervensjon besto av en 45 minutters læringsmodul som elevene skulle fullføre individuelt på sin egen digitale enhet. Forskerne rekrutterte én videregående skole i Rogaland som tilbydde modulen til alle elevene og der elevene kunne samtykke til å delta i forskningen hvis de ville. Det var 458 elever den dagen modulen skulle gjennomføres, og 385 samtykket til å være med i forskningen. Da elevene samtykket ble de randomiserte til intervensjons- og kontrollgruppa. Intervensjonsgruppa ble vist en læringsmodul som forklarte med nevrovitenskap hvordan alle kan lære hvis de gjør en innsats. Kontrollgruppas læringsmodul forklarte andre ting om hjernen.

Denne intervensjonen skulle påvirke forklaringsvariabelen *tankesett* til å være mer lærende. De hadde to utfallsvariable, *iherdighet* (perseverance) og *algebraferdighet*. Iherdighet ble målt ved om elevene valgte vanskeligere oppgaver på en algebratest, og algebraferdighet ble målt ved om de fikk til flere av oppgavene.

Noen uker etterpå var andre økt. Da målte forskerne først forklaringsvariabelen (tankesett) for å finne ut om intervensjonen hadde virket. Tankesettet var langt mer lærende i intervensjonsgruppa enn i kontrollgruppa (faktisk 56% av et standardavvik høyere), så forskerne konkluderte med at interven-

sjonen hadde lyktes. Deretter ga de en algebratest for å måle *iherdighet* og *algebraferdighet*. Siden mange var fraværende fra den andre økta var det bare 289 elever som var til stede i begge øktene, altså 63% av de 458 elevene som ble spurt om å delta.

Datafilen deres må da ha sett ut omtrent som i Tabell 3.2:

Tabell 3.2: De første radene på en tenkt datafil tilhørende Bettinger et al. Bettinger et al. (2018) sin randomiserte kontrollstudie. Hver rad tilhører én student. Én nominell variabel angir om studenten er i intervensjons- eller kontrollgruppa, *tankesett* er forklaringsvariabelen og *iherdighet* og *algebraferdighet* er utfallsvariable.

gruppe	tankesett	iherdighet	algebraferdighet
kontroll	1,4	8	2,8
intervensjon	3,1	8	2,1
intervensjon	2,8	5	3,8
kontroll	2,1	5	1,9
intervensjon	3,7	7	3,6

Resultatene viste at intervensjonsgruppa skåret høyere på iherdighet, de valgte altså vanskelige algebraspørsmål oftere enn kontrollgruppa. (Forskjellen var 24% av et standardavvik.) De skåret også litt høyere på algebraferdighet (rundt 12 % av et standardavvik). (Mer om usikkerhet i slike anslag i Kapittel 7..) Siden dette var en randomisert kontrollstudie er vi temmelig sikre på at forskjellene ikke skyldes at intervensjons- eller kontrollgruppa hadde annerledes gjennomsnittsverdi på iherdighet og algebraferdighet fra før av.

Randomiserte kontrollstudier er ikke perfekte, og det er flere utfordringer vi bør være oppmerksomme på:

- **Effekten på individnivå er ukjent** i randomiserte kontrollstudier. De gir kun et estimat av intervensjonens gjennomsnittlige effekt. Man kan samle inn mer data om deltagerne for å beskrive effekten i undergrupper, for eksempel effekten for forskjellige kjønn, men da finner man bare gjennomsnittseffekten for hvert kjønn. Du kan aldri finne ut hva effekten vil være for deg, barna dine, eller noen andre. Kanskje vil en intervensjon som ser ut til å ha positive effekter ha negative effekter for akkurat de du bryr deg om.
- **Frafall** kan skape problemer. Av de 458 elevene som opprinnelig ble invitert til å delta, var det kun 289 som gjennomførte hele studien. Dette frafallet skjer sjelden tilfeldig – kanskje var det nettopp de minst motiverte elevene som hoppet av underveis? Så lenge elevene i begge gruppene har like stor sannsynlighet for å falle fra, påvirker ikke dette resultatene våre. Men hvis for eksempel elevene i kontrollgruppen systematisk forsvinner fordi læringsmodulen deres var for kjedelig, kan dette påvirke konklusjonene selv om vi har randomisert.

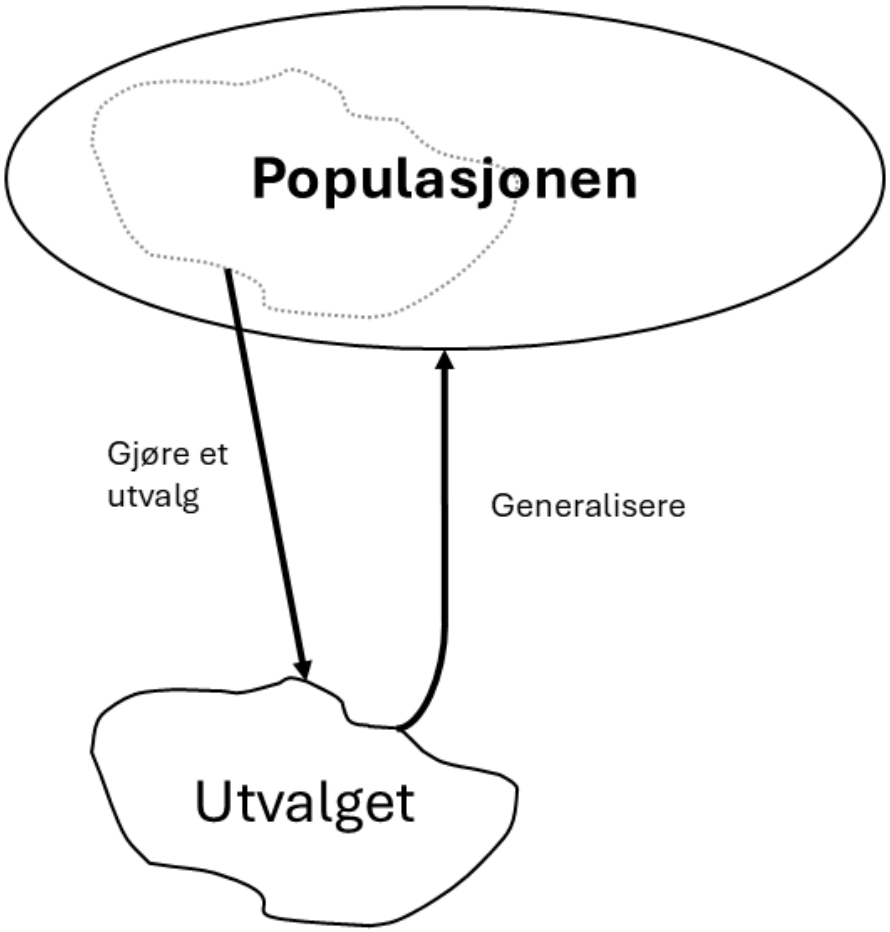
- **Generalisering** er en annen utfordring. I vårt eksempel deltok bare én skole i Rogaland, og randomiseringen forteller oss ikke hvordan samme tiltak vil fungere på andre skoler, i andre deler av landet, eller for andre aldersgrupper.
- **Etiske hensyn** kan gjøre randomiserte kontrollstudier kompliserte. Noen ganger fordi vi ikke kan gjennomføre intervensjonen av etiske årsaker – tenk for eksempel på hvor umulig det ville være å studere hvordan bakrus påvirker konsentrasjonen til videregåendeelever. Andre ganger fordi det kan være uetisk å *ikke* gi intervensjonen til kontrollgruppen, særlig hvis de foreløbige resultatene er svært lovende. I medisinsk forskning har faktisk flere studier måttet stoppes fordi behandlingen viste seg å være så effektiv at det ble ansett som uetisk å holde den tilbake fra kontrollgruppen. Det kan også være vanskelig å rekruttere deltakere fra visse befolkningsgrupper, for eksempel hvis de er skeptiske til forskning eller myndigheter. Dette fører til at sårbare grupper ofte er underrepresentert i studier, og da vet vi ikke om funnene våre gjelder for dem også.
- **Tilfeldige avvik** forekommer. Randomisering er, vel, tilfeldig, og noen ganger er de som responderer godt på intervensjonen samlet i den ene gruppa. Da får du et dårlig estimat på effekten. Men randomisering kan analyseres matematisk, så i tillegg til intervensjonens gjennomsnittseffekt kan man regne ut en usikkerhet i estimatet, se Seksjon 7.8.

3.6 Utvalg fra en populasjon

3.6.1 Populasjon, utvalg og analyseenhet

Mange forskningsspørsmål i kvantitativ metode dreier seg om å finne ut av noe for en viss **populasjon**. Populasjonen er den mengden av folk eller ting man ønsker at forskningen sin skal gjelde for. For eksempel kan det være jeg vil finne ut hvordan motivasjonen for skole endrer seg mellom barneskolen og ungdomsskolen. I så fall vil populasjonen kanskje være noe slikt som «alle norske elever som starter på ungdomsskolen i 2026». Merk at populasjonen er det jeg *ønsker* at forskningen skal gjelde for, og at vi sjeldent kan undersøke hele populasjonen.

I praksis må man gjøre et **utvalg** fra populasjonen. Siden man ikke velger hele populasjonen, men kun et utvalg, er en sentral del av kvantitativ forskning å **generalisere** det som gjelder for utvalget til hele populasjonen. I kvantitativ metode trekker vi slutninger *fra* utvalget *om* populasjonen. Dette er illustrert i Figur 3.3.



Figur 3.3: Utvelgelse fra, og generalisering til, populasjonen.

I stedet for å forske på alle norske elever som skal starte på ungdomsskolen må jeg forske på et mindre utvalg – alt annet er umulig med mindre man har uendelig med penger. Hvis jeg skulle forsket på alle elevene, kunne jeg analysert data fra «Elevundersøkelsen», som gis hvert år og er obligatorisk for alle norges elever på 7. trinn. Men for å finne ut av motivasjonen på ungdomsskolen trenger data fra 8. trinn også, og der er Elevundersøkelsen frivillig å gjennomføre, så det blir bare et utvalg av populasjonen. Hvordan skal utvalget ellers være? Skal jeg rekruttere 100 elever fra hvert fylke? Skal jeg legge ut en reklame for studien på TikTok og nøye meg med de svarene jeg får? Slike valg er helt sentrale for å få gode svar på kvantitative undersøkelser.

Et sentralt spørsmål er hvilken **analyseenhet** man skal benytte. Analyseenheten er de enhetene (tingene) du samler data fra. Hvis jeg undersøker elevenes motivasjon ved å spørre elevene selv, er det *eleven* som er analyseenheten. Hvis jeg spør læreren om å vurdere motivasjonen til klassen, er det *læreren* eller *klassen* som er analyseenheten. Hvis jeg sender ut observatører i mange klasserom for å forsøke å observere klassens motivasjon i forskjellige undervisningsøkter, er det kanskje *undervisningsøkter* som er analyseenheten.

I det følgende presenterer vi et lite knippe vanlige utvalgsmetoder og deres fordeler og ulemper.

3.6.2 Enkle tilfeldige utvalg

For å generalisere til populasjonen vil i de fleste tilfeller det beste utvalget er et utvalg som er helt likt populasjonen, bare mye mindre. I stedet for å velge ut alle de ca. 60 000 syvendeklassingene i Norge (populasjonen) og målt motivasjonen deres, hadde det vært ideelt å valgt ut 600 syvendeklassinger slik at disse svarer likt som populasjonen. Dette vil si at hvis 10 % av elevene i populasjonen er topp motiverte, så er 10 % av utvalgets elever det også; hvis 20 % av populasjonen er ganske motiverte, så er 20 % av utvalgets elever det også, og så videre. Denne egenskapen, at utvalget er likt populasjonen, kalles at utvalget er **representativt med hensyn til variabelen**.

Aller helst bør ikke utvalget bare være representativt med hensyn til én variabel, det bør være likt på alle andre mulige måter også. Hvis man ønsker å finne ut om elever med lave eller høye karakterer har forskjellig motivasjon, er det viktig at de elevene i utvalget med høye (eller lave) karakterer representerer elevene i populasjonen med høye (eller lave) karakterer. Hvis utvalget er representativt med hensyn på alle mulige variabler, kaller vi utvalget **representativt**. Hvis utvalget ikke er representativt er det **skjevt**.

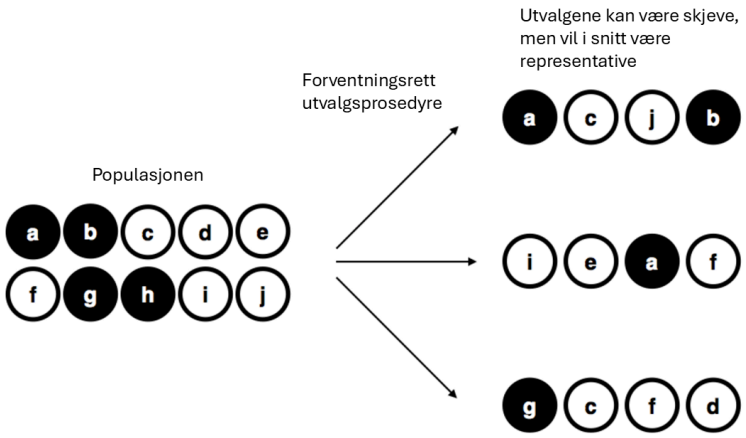
Det virker helt uoppnåelig å skaffe et representativt utvalg. Tenk på det, det finnes uendelig mange variable, og man må forsikre seg om at utvalget er likt populasjonen på alle sammen. Men det finnes et triks som fikser et representativt utvalg for deg, **randomisering**.

For å holde ting enkle kan vi forestille oss at vi har en pose med 10 brikker. Hver brikke har sin egen unike bokstav, så vi kan skille dem fra hverandre. Brikkene kommer i to farger - svart og hvit. Dette settet med brikker er populasjonen vi er interessert i, og den er vist grafisk til venstre i Figur 3.4. Som du kan se på bildet, er det 4 svarte brikker og 6 hvite brikker. Men i virkeligheten ville vi selvfølgelig ikke vite dette uten å kikke i posen først!

La oss nå forestille oss at du kjører følgende «eksperiment»: Du rister posen godt, lukker øynene og trekker ut 4 brikker uten å legge noen av dem tilbake. Først kommer a-brikken (svart), så c-brikken (hvit), deretter j (hvit) og til slutt b (svart). Hvis du har lyst, kan du legge alle brikkene tilbake i posen og gjenta hele eksperimentet - akkurat som vist på høyre side av Figur 3.4. Hver gang vil du få forskjellige resultater, selv om du følger nøyaktig samme prosedyre. Dette at samme utvalgsmetode kan gi forskjellige utvalg hver gang, kaller vi et **tilfeldig utvalg**.

Siden vi ristet posen før vi trakk ut brikkene, virker det rimelig å anta at hver brikke har like stor sjanse for å bli valgt. En slik prosedyre – hvor hvert medlem av populasjonen har samme sjanse for å bli valgt – kalles et **enkelt tilfeldig utvalg**. Det at vi ikke legger brikkene tilbake etter å ha trukket dem, betyr at du ikke kan observere den samme brikken to ganger.[11]

[11] Det er også verdt å nevne en tredje prosedyre. Denne gangen lukker vi øynene, rister posen og trekker ut en brikke. Men nå registrerer vi observasjonen og *legger så brikken tilbake i posen* før vi trekker neste brikke. Vi gjentar denne prosessen til vi har 4 brikker. Datasett som genereres på denne måten er fortsatt enkle tilfeldige utvalg, men siden vi legger brikkene tilbake umiddelbart etter at vi har trukket dem, kalles det et utvalg **med tilbakelegging**. Forskjellen mellom denne situasjonen og den første er at det nå er mulig å observere samme brikke flere ganger.

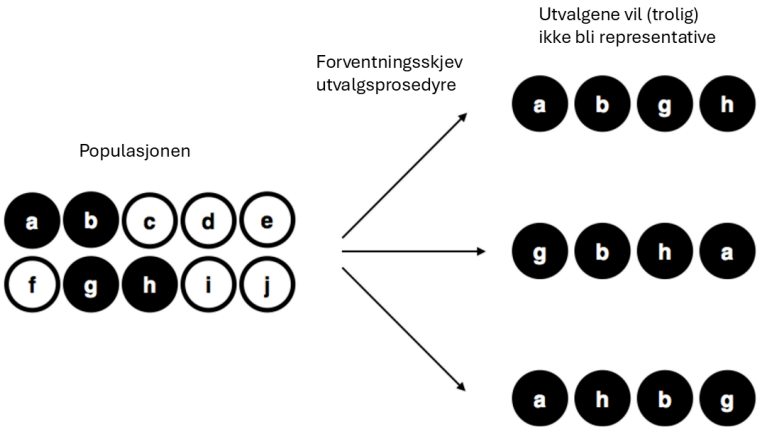


Figur 3.4: Enkelt tilfeldig utvalg uten tilbakelegging fra en endelig populasjon

Hvis du studerer utvalgene i Figur 3.4 nøye ser du at ingen av dem gjengir populasjonen riktig. I populasjonen er 40% av brikkene svarte, og i utvalgene er 50%, 25% og 25% svarte. Så utvalgene er skjeve, altså ikke representative, men hvis man hadde repetert utvalgsmetoden mange, mange ganger, hadde man hatt 40% svarte brikker i utvalgene sine i gjennomsnitt. Derfor sier man at enkelt tilfeldig utvalg er en **forventingsrett** utvalgsmetode.[12] Hvert enkelt utvalg er ikke nødvendigvis representativt, men hvis man repeterer utvalget mange ganger kan man forvente at gjennomsnittet er representativt, og hvis man gjør utvalget veldig veldig stort vil avviket fra et representativt utvalg være veldig lite, se [^sec-law-of-large-numbers].

For å sikre at du forstår hvor viktig utvalgsprosedyren er, kan vi tenke på en alternativ måte å gjennomføre eksperimentet på. La oss si at min fem år gamle sønn hadde åpnet posen og bestemt seg for å bare trekke ut fire svarte brikker, uten å legge noen tilbake. Dette skjeve utvalget er illustrert i Figur 3.5. Tenk nå på hvor mye vi kan lære av å se 4 svarte brikker og 0 hvite brikker. Det avhenger helt klart av hvordan utvalget ble gjort, ikke sant? Hvis du vet at prosedyren var skjev og bare valgte svarte brikker, forteller et utvalg med bare svarte brikker deg ikke særlig mye om populasjonen som helhet! Derfor liker statistikere virkelig godt når et datasett kan betraktes som et enkelt tilfeldig utvalg – det gjør dataanalysen *mye* enklere.

[12]På engelsk heter forventningsrett «unbiased». Mange har begynt å bruke biased og unbiased på norsk også, men jeg vil utfordre deg til å si at «utvalget er forventningsskjevt» eller «dommeren er partisk» i stedet for å bare bruke «biased».



Figur 3.5: Skjevt utvalg uten tilbakelegging fra en endelig populasjon.

I utdanningsvitenskap er nesten alle utvalg uten tilbakelegging, siden samme person vanligvis ikke får lov til å delta i undersøkelsen to ganger. Men det meste av statistisk teori bygger på antakelsen om at dataene kommer fra et enkelt tilfeldig utvalg **med tilbakelegging**. I praksis spiller dette sjelden noen rolle. Hvis populasjonen vi er interessert i er stor (for eksempel har mer enn 10 enheter!), er forskjellen mellom utvalg med og uten tilbakelegging så liten at vi ikke trenger å bekymre oss for den. Forskjellen mellom enkle tilfeldige utvalg og skjeve utvalg derimot - det er ikke noe vi kan se bort fra så lett!

Og så for å hamre inn et siste poeng: tilfeldige utvalg er forventningsrette med hensyn på alle mulige variable. I eksemplene tok vi frem variabelen svart/hvit, men brikkene kan variere i størrelse, materiale, eller alder også. Enkle tilfeldige utvalg er forventningsrette med hensyn på disse variablene og alle andre variable du kan tenke deg, også. Du bør dvele ved dette, for det er det nærmeste man kommer magi i forskningsdesign. Ved et enkelt triks kan velge et utvalg som representerer populasjonen på alle mulige måter – wow! Men det er med enkle tilfeldige utvalg som med magi – de finnes ikke på ordentlig.

Vi ramser opp noen grunner til at enkle tilfeldige utvalg ikke finnes:

- **Manglende liste over enheter i populasjonen**: For å få et enkelt tilfeldig utvalg trenger man først en liste over alle enhetene i populasjonen. Stort sett finnes ingen slik liste. Hvis jeg vil forske på alle musikklærere på ungdomstrinnet i Norge skulle jeg gjerne hatt en liste over alle disse musikklærerne, men en slik liste finnes ikke.
- **Selektiv deltagelse**: Hvis man er så heldig at det finnes lister over alle i populasjonen, for eksempel hvis populasjonen din er «ungdomsskoler i Oslo», så kan du banne på at ikke alle ønsker å være med i studien din.
- **Selektiv dropout**: Og mens studien pågår kommer garantert noen til å droppe ut av studien.

Du har ikke et enkelt tilfeldig utvalg hvis 40% av de tilfeldig utvalgte skolene sier de ikke ønsker å være med og hvis 10 % av skolene som ville delta slutter

å svare på epostene dine halvveis ut i studien. Og de skolene du mister fra utvalget ditt skiller seg sannsynligvis fra de andre skolene – for eksempel ved at de har mange egne problemer de må løse og derfor ikke har tid til å delta – og derfor heter det *selektiv* deltagelse og dropout. Dessverre er selektiv deltagelse og dropout helt vanlig, og derfor er enkle tilfeldige utvalg noe man bare kan tilnærme seg, men sjeldent eller aldri oppnå.

3.6.3 Tilfeldige klyngeutvalg

Et tilfeldig klyngeutvalg brukes der populasjonen er samlet i naturlige klynger som er fysisk nær hverandre. I utdanningsforskning er de desidert vanligste klyngene skoler eller klasser. Et **tilfeldig klyngeutvalg** er et utvalg der man i steg 1 velger klynger tilfeldig og i steg 2 velger deltagere innad i hver klynge. Dette betyr at man først velger klasser tilfeldig, og deretter bestemmer seg for hvilke elever man skal velge. Man kan enten velge alle deltagerne i klyngen eller velge tilfeldig innenfor hver klynge.

Tilfeldige klyngeutvalg gir også representative utvalg, slik som enkle tilfeldige utvalg, men man må bruke mer avanserte statistiske teknikker i utregningene og man får mer variasjon i svarene. Fordelene med klyngeutvalg er enklere å utføre, og i mange tilfeller der enkle tilfeldige utvalg er umulige å gjennomføre er klyngeutvalg mulige! Vi nevnte tidligere at det ville være umulig å gjøre et enkelt tilfeldig utvalg med norske musikklærere på ungdomstrinnet, fordi det ikke finnes noen liste over norske musikklærere på ungdomstrinnet. Derfor er et enkelt tilfeldig utvalg umulig. Men det finnes lister over alle ungdomsskoler i landet, så et klyngeutvalg er mulig! Man velger først ungdomsskoler tilfeldig, og deretter spør man rektorene om hvem musikklærerne på skolen er.

Store internasjonale undersøkelser, som for eksempel PISA, benytter gjerne klyngeutvelgelse for å få representative utvalg.

3.6.4 Ikke-tilfeldige utvalg

Som du ser fra listen over mulige populasjoner jeg viste tidligere, er det nesten umulig å få et enkelt tilfeldig utvalg fra de fleste populasjoner vi er interessert i. Vi rekker ikke å diskutere andre utvalgsmetoder grundig i denne boken, men for å gi deg en følelse av hva som finnes der ute, vil jeg liste opp noen av de viktigste.

Bekvemmelighetsutvalg er mer eller mindre det det høres ut som. Utvalgene velges på en måte som er praktisk for forskeren. Et vanlig eksempel i utdanningsforskning er at man velger ut lærere og elever som befinner seg på skoler i nærheten av forskerne. Dette høres kanskje lite prinsipielt ut, men det kan være gode grunner for å gjøre det slik, for da sparer man reisepenger i forskningsbudsjettet og de besparelsene kan man bruke til å heve kvaliteten på andre sider ved studien.

Stratifisert utvalg er å gjøre utvalget fra flere forskjellige underpopulasjoner, også kalt strataer. I stratifiserte utvalg velger man deltagere innad i hvert av strataene. Hvis du ønsker å inkludere lærere med ulik mengde erfaring, kan du velge 50 lærere fra gruppene «0 til 5 års erfaring», «5 til 10 års erfaring», «10 til 15 års erfaring» og «mer enn 15 års erfaring», for eksempel. Stratifisert utvalg kan være tilfeldig og ikke-tilfeldig.

Stratifisert utvelgelse kan være mer effektivt, særlig når noen av underpopulasjonene er sjeldne. For eksempel, hvis du skal sammenlikne hvordan det går med elever med og uten ADHD-diagnose på ungdomsskolen, bør du gjøre et stratifisert utvalg. Siden rundt 6% av elevmassen har en ADHD-diagnose ville et enkelt tilfeldig utvalg med 100 elever inneholdt kun seks med diagnose. Da måtte utvalget vært veldig stort for å få med mange diagnostiserte elever. Med et stratifisert utvalg kunne man valgt 100 elever med diagnose og 100 elever uten diagnose, noe som ville vært mer effektivt. (Merk at det ikke er lett å få et utvalg med spesifikke diagnoser, for du må først få tilgang til en liste over hvem som har diagnosen.) Stratifiserte utvalg kan være tilfeldige og ikke-tilfeldige.

Snøball-utvalg er en teknikk som er spesielt nyttig når man skal ta utvalg fra en «skjult» eller vanskelig tilgjengelig populasjon, og brukes ofte i samfunnsvitenskapene. La oss si at forskere ønsker å gjennomføre en meningsmåling blant transpersoner. Forskerteamet har kanskje bare kontaktopplysninger til noen få transpersoner, så undersøkelsen starter med å spørre dem om å delta (fase 1). På slutten av undersøkelsen blir deltakerne bedt om å oppgi kontaktopplysninger til andre personer som kanskje vil delta. I fase 2 blir disse nye kontaktene intervjuet. Prosessen fortsetter til forskerne har tilstrekkelig data. Den store fordelen med snøball-utvalg er at det gir deg data i situasjoner som ellers kunne vært umulige å få tak i. Hovedulempen at utvalget kommer til å avhenge sterkt av informantene i fase 1 og derfor ikke nødvendigvis representerer populasjonen «transpersoner» godt. En annen ulempe er at utvalgsprosessen kan føre til store ulemper for deltagerne og kanskje bli uetisk. Skjulte populasjoner er ofte skjulte av en grunn. Jeg valgte transpersoner som eksempel her for å fremheve dette problemet. Hvis du ikke var forsiktig, kunne du ende opp med å «oute» personer som ikke ønsker å bli outet. Det kan også være invaderende å bruke folks sosiale nettverk til å studere dem. I mange tilfeller kan den enkle handlingen å kontakte dem og si «hei, vi vil studere deg fordi du er trans» være sårt hvis man allerede blir objektivisert fordi man er trans. Da burde man innhentet informert samtykke på forhånd, men hvordan innhenter man samtykke før man har kontaktet noen? Sosiale nettverk er komplekse ting, og selv om du *kan* bruke dem til å få data, betyr ikke det at du *bør*.

3.6.4.1 Hvor stor rolle spiller det at du ikke har et tilfeldig utvalg?

I virkeligheten, og i hvert fall i utdanningsforskning, ser vi sjeldent enkle tilfeldige utvalg. Men hvor stort problem er egentlig dette? Det kan virke som et kritisk problem, bare sammenlikn Figur 3.4 og Figur 3.5, men ofte er situasjonen ikke like håpløs.

Noen typer skjeve utvalg er faktisk helt uproblematiske, for eksempel i tilfeldige stratifiserte utvalg. Da vet du nøyaktig hva skjevheten består i siden du har skapt den med vilje – ofte for å gjøre studien mer effektiv. Og det finnes statistiske teknikker som kan justere for skjevhetene, selv om vi ikke dekker dem i denne boken. I slike tilfeller er det altså ingen grunn til bekymring.

Andre ganger trenger man ikke enkle tilfeldige utvalg for å generalisere, fordi man kan generalisere med teoretiske argumenter. Hvis jeg lurer på om multiplikasjons-spillet jeg har utviklet er passe vanskelig for norske elever, holder det at jeg gjør et bekvemmelighetsutvalg og tester det på noen

nærskoler der jeg har et godt forhold til rektorene, eller må jeg teste det ut i Finnmark og Agder også? Det teoretiske argumentet mitt for at det holder å teste det ut på nærskoler kan være at kun ferdighetsnivå og språkbakgrunn teller, og jeg finner ulike ferdighetsnivåer og språkbakgrunner på nærskolene.

Det er også viktig å huske at tilfeldig utvalg er et verktøy for å nå et mål, ikke målet i seg selv. La oss si du har brukt et bekvemmelighetsutvalg som sannsynligvis er skjevt. Dette blir bare et problem hvis det fører til at du trekker feil konklusjoner. Sett fra dette perspektivet trenger vi ikke at utvalget skal være representativt med hensyn på alle mulige konstrukter – det holder at det er representativt med hensyn på konstruktene vi studerer.

To viktige poenger skjuler seg i denne diskusjonen:

- Når du designer egne studier, tenk nøye over hvilken populasjon du faktisk bryr deg om, og prøv å gjøre et passende utvalg. I praksis blir du ofte tvunget til å bruke bekvemmelighetsutvalg, og da bør du bruke tid på å tenke gjennom hvilke problemer utvalget skaper.
- Hvis du skal kritisere andres forskning for å ha brukt bekvemmelighetsutvalg i stedet for tilfeldige utvalg, bør du gi teoretiske argumenter for hvordan dette kan ha påvirket resultatene.

3.7 En studies validitet

Som forsker er det viktigste målet ditt at forskningen skal være «gyldig», noe som kalles **validitet**. Vi drøftet målemetoders validitet i Kapittel 2., men selv hvis målingene er valide betyr ikke det at studiens slutninger er valide. Ideen bak validitet er egentlig ganske enkel: Kan du stole på resultatene fra studien din? Hvis svaret er nei, er studien ugyldig.

Dette høres kanskje enkelt ut, men i praksis er det utfordrende å vurdere en studies gyldighet fordi det er så mange ting som spiller inn på om studiens slutninger er gyldige. Bare i løpet av dette kapittelet har vi møtt mange trusler mot en studies validitet, konfundere, retningsproblemet, skjeve utvalg, selektiv deltagelse og dropout, og flere. Derfor er det nyttig å bryte validitet ned i forskjellige undertyper. De viktigste er indre- og ytre validitet.

3.7.1 Indre validitet

Indre validitet er om slutningene dine er gyldige for utvalget ditt. Det inkluderer målevaliditet, se Seksjon 2.3.2, og i forskning med formål om å finne årsaksforklaringer, så inkluderer det om årsaksforklaringen er godt underbygget. Det kalles «indre» fordi det dreier seg om sammenhengene mellom ting som skjer innenfor selve studien.

Et enkelt eksempel er en studie som ønsker å finne ut om universitetsutdanning gjør folk til bedre skribenter. Du samler en gruppe førsteårsstudenter, ber dem skrive et essay på 1000 ord, og teller opp stavefeil og grammatiske feil. Så gjør du det samme med tredjeårsstudenter, som jo har mer universitetsutdanning bak seg. Hvis tredjeårsstudentene gjør færre feil, kan du da konkludere med at universitetsutdanning forbedrer skriveferdighetene?

Her ligger problemet: Tredjeårsstudentene er jo også eldre og har helt naturlig fått mer skriveerfaring over tid, så alder er en konfunder. Så hva er egentlig årsaken til at de skriver bedre? Er det alderen? Når man blir eldre

får man jo skrevet mer, uavhengig av om man går på universitetet eller ikke. Eller er det universitetsutdanningen? Du kan rett og slett ikke vite årsken basert på denne studien. Dette er et klassisk eksempel på svak indre validitet — studien klarer ikke å utelukke konfundere på en god nok måte.

3.7.2 Ytre validitet

Ytre validitet handler om hvor godt funnene dine kan overføres til andre situasjoner og utvalg. Spørsmålet er: Vil du se det samme mønsteret i «den virkelige verden» som du så i studien din? Tenk på det slik: Studien din vil alltid ha ganske spesifikke rammer — bestemte oppgaver, miljøer, deltakere, og så videre. Hvis resultatene ikke holder seg når du går utenfor disse rammene, har du et problem med ytre validitet. Det er to hovedproblemer med ytre validitet i utdanningsforskning, å generalisere til populasjonen, som er beskrevet i detalj over, og økologisk validitet.

Ideen bak **økologisk validitet** er at resultatene fra studien ikke kan fungere i den virkelige verden fordi studiesituasjonen skiller seg fra den virkelige situasjonen. Mange randomiserte kontrollstudier i utdanningsforskning sliter med økologisk validitet, fordi eksperimentet ikke vil fungere likedan utenfor studiesituasjonen. Hvis Oslo Kommune bestemmer seg for å gjennomføre en randomisert kontrollstudie der eksperimentet er å ta ut elever til smågruppeundervisning med kvalifiserte lærere, kan man ikke forvente å få samme resultat hvis kommunen etterpå vedtar å utvide tilbudet til alle skoler. Grunnen er at det ikke finnes nok kvalifiserte lærere, så eksperimentet vil se annerledes ut i virkeligheten enn i studien. Et annen eksempel er studier som undersøker effekten av å skrive på skjerm heller enn papir. For å gjøre skjerm-situasjonen så sammenliknbar som mulig med papir-situasjonen slår forskerne gjerne av stavekontrollen – men de fleste har jo på stavekontrollen i den virkelige verden! Da finner man effekten av et eksperiment som ikke likner på den virkelige verden.

3.8 Sammendrag

Denne korte introduksjonen til forskningsdesign har dekket noen av de viktigste temaene man må tenke på hvis man skal forske selv eller vurdere eksisterende forskning. Måling er en så sentral del av forskningsdesignet at det fikk et eget kapittel, Kapittel 2., mens resten fikk plass i dette kapittelet. Vi har dekket:

- **Formål med kvantitativ forskning**: å beskrive, predikere og å finne årsaksforhold.
- **Variables «roller»**: avhengige og uavhengige variable, i tillegg til en bråte andre navn. Jeg foretrekker prediktor og utfall.
- **Eksperimentelle studier og observasjonsstudier**.
- **Korrelasjon og kausalitet**: spesielt tilfeldige korrelasjoner, retningsproblemet og konfundere.
- **Utvalg**: to typer tilfeldige utvalg og tre typer ikke-tilfeldige utvalg.
- **Vurdering av en studies validitet**: Indre og ytre validitet, men husk også målevaliditet fra Kapittel 2..

II Deskriptiv statistikk

4. Komme i gang med jamovi

Robots are nice to work with.
– Roger Zelazny[13]

Dette kapittelet er kort, rett og slett fordi jeg ikke rakk å skrive det ferdig før 2026-utgaven. Her får du bare det aller nødvendigste: hvordan du skaffer deg jamovi, hvordan du får bokas datafiler inn i programmet, og hvordan resten av boka hjelper deg videre. Du trenger ikke jamovi for å forstå resten av boka, men jeg tror de fleste får mer ut av kapitlene om de klikker seg gjennom analysene i jamovi mens de leser.

Er du en av dem som ser for deg å skrive en kvantitativ masteroppgave, er dette kapittelet ekstra verdt tiden din. Da er jamovi godt alternativ til store og dyre statistikkprogrammer jamovi gjør de aller fleste analysene en masteroppgave har bruk for. Og skulle du trenge noe det ikke har fra før, finnes det et enkelt system med tilleggsmoduler som du installerer med noen klikk inne i programmet, så du slipper å lære deg et nytt program for å komme videre.

4.1 Å skaffe seg jamovi

jamovi er gratis og åpent tilgjengelig, og du finner det på jamovi.org. Der har du to muligheter. Enten laster du ned programmet og installerer det på maskinen din, eller så bruker du «Cloud»-versjonen, som kjører i nettleseren uten at du installerer noe som helst. Cloud-versjonen er den raskeste veien i gang, særlig hvis du sitter på en maskin der du ikke får lov til å installere programmer. Men vær klar over at den krever at du oppretter en brukerkonto og logger inn, og at datafilene dine da blir liggende på en server hos noen andre. Hvis du skal laste opp egne data om personer blir det et personvernspørsmål du må avklare. Den installerte versjonen krever ingen innlogging, holder dataene på din egen maskin, og er dessuten raskere når datasettene blir store. Til bokas datasett fungerer begge helt likt, så velg det som passer deg; skal du analysere dine egne innsamlede data, vil jeg anbefale den installerte versjonen.

Grensesnittet er ikke stort å sette seg inn i. Øverst ligger fire faner: **Variables**, som viser en oversikt over variablene i datasettet, **Data**, som viser selve dataene i et regneark, **Analyses**, der du finner alle analysene, og **Edit**. Nesten alt vi gjør i denne boka skjer under Analyses. Der klikker du deg fram til en analyse, og resultatene dukker opp i panelet til høyre, side om side med dataene.

[13] Kilde: *Dismal Light* (1968).

4.2 Bokas datafiler

Analysene i boka bruker to datasett, og begge kan du laste ned og åpne selv:

- *ICCS.omv* er hoveddatasettet, og brukes i Kapittel 5. og Kapittel 6.. Det er et utdrag fra den store internasjonale undersøkelsen ICCS (Fraillon et al., 2024), med 16473 elever fra Norge, Spania, Polen og Brasil, og 15 variabler om blant annet kjønn, forventet utdanning, kunnskap om demokrati og holdninger til likestilling. Ønsker du dataene i et format andre programmer kan lese, ligger de samme dataene i *ICCS.csv*.
- *anscombe.csv* er et lite datasett som brukes ett sted i Kapittel 6., der poenget er at fire datasett kan ha helt like tall og likevel se helt forskjellige ut.

Forskjellen på de to filtypene er verdt å merke seg. En **.omv-fil** er jamovis eget format. I tillegg til tallene husker den hvilket målenivå hver variabel har, hva verdiene heter, og hvilke analyser du har kjørt. Det er derfor *Forventet utdanning* allerede er ordinal og sortert i riktig rekkefølge når du åpner ICCS.omv; noen (jeg) har satt det opp på forhånd. En **.csv-fil** er derimot bare en tekstfil med tall og komma. Den kan leses av så å si alle programmer, men den husker ingenting om målenivåer, så det må du sette selv etter importen. Bruk .omv-filen når du kan.

> I jamovi: åpne en datafil
>
> Last først ned filen til maskinen din, for eksempel til Nedlastinger-mappen. Høyreklikk gjerne på lenken og velg «Lagre lenke som», så er du sikker på at filen havner et sted du finner igjen.
>
> Klikk deretter på de tre strekene (≡) øverst til venstre i jamovi og velg **Open**. Velg **Browse** under «This Device» (bruker du Cloud-versjonen, heter valget «Upload» eller liknende), finn filen du lastet ned, og åpne den.
>
> Dataene dukker nå opp i Data-fanen, med én rad per elev og én kolonne per variabel. Gå til Variables-fanen for å se hvilket målenivå hver variabel har. Åpnet du en .csv-fil, bør du sjekke målenivåene ekstra nøye, siden .csv-formatet ikke lagrer dem.

Når du har åpnet ICCS.omv, bør du se noe som likner på skjermbildet i Figur 5.1 i neste kapittel: en tabell med elever nedover og variabler bortover. Da er du klar til å følge analysene i resten av boka.

4.3 Slik hjelper resten av boka deg

Fra og med neste kapittel er hovedteksten opptatt av *hva* en analyse er og *hvorfor* vi gjør den, ikke av hvor du skal klikke. Klikkingen har fått sin egen plass. Hver gang boka viser fram en analyse du kan gjøre selv, ligger det en liten boks like ved, med en tittel som starter med «I jamovi:». Boksen over er et eksempel på en slik boks.

Hver boks er skrevet slik at den skal kunne leses for seg selv, så du trenger ikke lete i teksten rundt for å skjønne hva du skal gjøre.

Vil du ha en grundigere innføring i jamovi enn dette kapittelet gir, kan du lese kapittelet «Getting started with jamovi» i boka denne er basert på

(Navarro & Foxcroft, 2025). Det er på engelsk, men det er godt illustrert og dekker programmet langt mer utførlig enn jeg gjør her.

5. Å beskrive én variabel

Når du skal analysere et nytt datasett, må du finne måter å beskrive dataene på som er både konsise og lett forståelige. Dette kalles **deskriptiv statistikk**, eller beskrivende statistikk. I denne boka er det to kapitler om deskriptiv statistikk, dette kapittelet og det neste. I dette kapittelet beskriver vi én variabel, og i neste kapittel beskriver vi sammenhengen mellom flere variabler.

Man kan dele deskriptiv statistikk inn i to hovedformer:

- Man kan bruke tall for å beskrive dataene, for eksempel utregning av gjennomsnittet av en variabel.
- Man kan lage et diagram som viser dataene, for eksempel et søylediagram. Diagrammer kalles også figurer og plott.

I de neste to kapitlene lærer du om vanlige tall og diagrammer som man bruker for å beskrive data.

> ⚠ Ikke tenk at du beskriver populasjonen
>
> Merk deg at deskriptiv statistikk kun beskriver de dataene som er samlet inn, altså utvalget. Ofte ønsker man å generalisere til en populasjon, men da må man bruke inferensiell statistikk, noe du lærer i Kapittel 7..

Rå data er sjelden informative i seg selv. Jeg har forberedt et lite datasett fra den store internasjonale undersøkelsen ICCS, International Civic and Citizenship Education Study (Fraillon et al., 2024). Dette datasettet inneholder variabler for kjønn, skole, land, forventet høyeste fullførte utdanning og en samlevariabel for hvor enig man er i at kjønnene bør ha like rettigheter. Datasettet har 16473 rader, én rad for hver elev, men vi viser bare 50 rader. Se i Tabell 5.1 for å vurdere hvor informative disse dataene er.

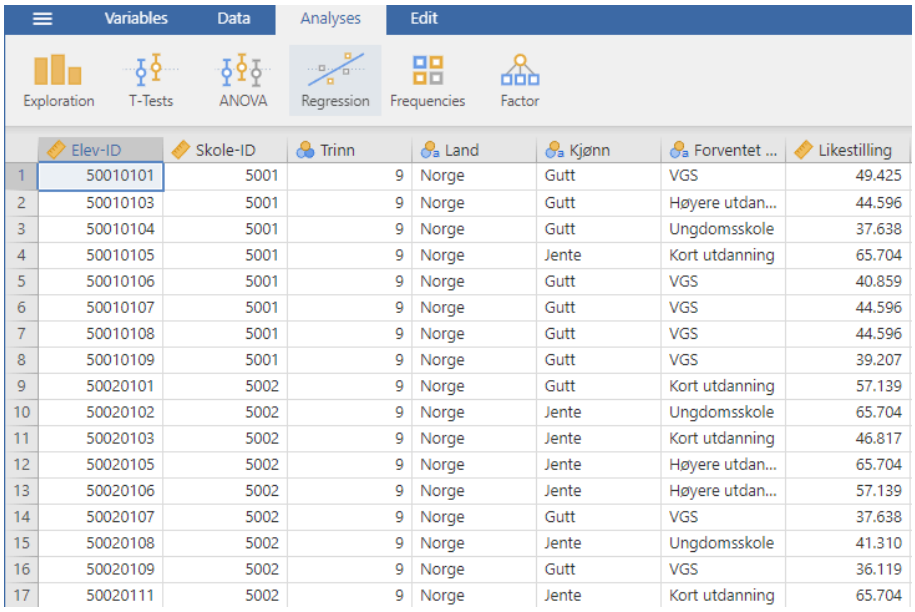
Tabell 5.1: Utvalgte variabler fra ICCS-studien.

Skole-ID	Land	Kjønn	Forventet utdanning	Likestilling
5150	Spania	Gutt	VGS	46,8
5016	Polen	Jente	Ungdomsskole	40,9
5134	Spania	Gutt	Høyere utdanning	65,7
5018	Polen	Jente	Høyere utdanning	65,7
5062	Polen	Jente	Høyere utdanning	65,7
5025	Norge	Gutt	VGS	65,7
5173	Brasil	Jente	Høyere utdanning	52,7
5126	Polen	Gutt	Høyere utdanning	42,6
5037	Norge	Jente	Kort utdanning	65,7
5041	Brasil	Gutt	VGS	37,6
5123	Norge	Jente	Høyere utdanning	65,7
5019	Brasil	Gutt	Høyere utdanning	52,7
5169	Brasil	Gutt	Høyere utdanning	65,7
5106	Polen	Jente	Høyere utdanning	57,1
5107	Norge	Gutt	Ungdomsskole	39,2
5120	Brasil	Gutt	Kort utdanning	52,7
5071	Norge	Jente	VGS	37,6
5126	Norge	Jente	Ungdomsskole	57,1
5020	Norge	Gutt	VGS	57,1
5170	Polen	Jente	Kort utdanning	52,7
5064	Brasil	Jente	Høyere utdanning	52,7
5149	Norge	Jente	Høyere utdanning	49,4
5076	Polen	Gutt	VGS	39,2
5013	Polen	Jente	Høyere utdanning	49,4
5006	Norge	Jente	VGS	65,7
5158	Polen	Gutt	Kort utdanning	65,7
5201	Brasil	Jente	Høyere utdanning	49,4
5010	Spania	Gutt	Kort utdanning	40,9
5114	Brasil	Gutt	Kort utdanning	49,4
5013	Norge	Gutt	Høyere utdanning	46,8
5092	Polen	Jente	Høyere utdanning	57,1
5012	Norge	Gutt	Høyere utdanning	65,7
5151	Polen	Jente	Høyere utdanning	46,8
5112	Norge	Jente	Høyere utdanning	65,7
5036	Norge	Gutt	Høyere utdanning	65,7
5[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Gir denne tabellen deg egentlig noen klar forståelse av mønstre i dataene, for eksempel forskjeller mellom grupper, typiske verdier eller sammenhenger mellom variabler? Sannsynligvis ikke. De fleste vil oppleve at det er vanskelig å få øye på slike trekk direkte i rådataene, og at man trenger at dataene blir oppsummert og fremstilt på en mer oversiktlig måte.

Det er dette som er deskriptiv statistikk, å gjøre rå data om til tall og figurer som gir innsikt.

I dag gjør man all statistikk på en datamaskin, så la oss vise dataene i statistikkprogrammet, jamovi. Etter å ha åpnet filen *ICCS.omv* ser det ut som i Figur 5.1. Trenger du hjelp til å få filen inn i jamovi, står det forklart i Seksjon 4.2.



	Elev-ID	Skole-ID	Trinn	Land	Kjønn	Forventet ...	Likestilling
1	50010101	5001	9	Norge	Gutt	VGS	49.425
2	50010103	5001	9	Norge	Gutt	Høyere utdan...	44.596
3	50010104	5001	9	Norge	Gutt	Ungdomsskole	37.638
4	50010105	5001	9	Norge	Jente	Kort utdanning	65.704
5	50010106	5001	9	Norge	Gutt	VGS	40.859
6	50010107	5001	9	Norge	Gutt	VGS	44.596
7	50010108	5001	9	Norge	Gutt	VGS	44.596
8	50010109	5001	9	Norge	Gutt	VGS	39.207
9	50020101	5002	9	Norge	Gutt	Kort utdanning	57.139
10	50020102	5002	9	Norge	Jente	Ungdomsskole	65.704
11	50020103	5002	9	Norge	Jente	Kort utdanning	46.817
12	50020105	5002	9	Norge	Jente	Høyere utdan...	65.704
13	50020106	5002	9	Norge	Jente	Høyere utdan...	57.139
14	50020107	5002	9	Norge	Gutt	VGS	37.638
15	50020108	5002	9	Norge	Jente	Ungdomsskole	41.310
16	50020109	5002	9	Norge	Gutt	VGS	36.119
17	50020111	5002	9	Norge	Jente	Kort utdanning	65.704

Figur 5.1: Et skjermbilde av jamovi som viser variablene lagret i filen *ICCS.omv*

For å få en nærmere forståelse av dataene må vi beskrive dem med tall og figurer. Vi organiserer framstillingen etter typen variabel vi analyserer. Først presenteres kategoriske variabler (nominelle og ordinale), og deretter kontinuerlige variabler (intervall- og forholdstallsnivå). De kategoriske variablene er markert med et symbol med tre fargede sirkler i jamovi, og er altså *Land*, *Trinn*, *Kjønn* og *Forventet utdanning*.

5.1 Kategoriske variabler

Hvis du ser på variabelen *Forventet utdanning* i Figur 5.1, vil du se at den har verdiene «Høyere utdanning», «Kort utdanning», «VGS» og «Ungdomsskole». Denne variabelen viser hva elevene ser for seg at de har som høyeste utdanning når de er ferdige med studiene. Hvordan ville du analysert denne variabelen? Sannsynligvis på de to måtene du skal lære nå.

5.1.1 Frekvenstabeller

En av de mest grunnleggende oppgavene innen dataanalyse er å telle antall observasjoner for de ulike verdiene til nominelle og ordinale variabler. For variabelen *Forventet utdanning* vil det si å telle hvor mange som ser for seg å ta høyere utdanning, en kort utdanning, kun VGS eller kanskje stoppe etter ungdomsskolen. Siden et slikt antall kalles en *frekvens*, kalles tabellen der man viser opptellingen for en **frekvenstabell**.

Frekvenstabellen vises i Tabell 5.2.

> I jamovi: lage en frekvenstabell
>
> Analyses ⟶ Exploration ⟶ Descriptives. Legg *Forventet utdanning* i «Variables» og kryss av for «Frequency tables».

Tabell 5.2: Frekvenstabell for *Forventet utdanning*

Forventet utdanning	Antall	Prosent
Ungdomsskole	923	5,6
VGS	3786	23,1
Kort utdanning	2362	14,4
Høyere utdanning	9342	56,9

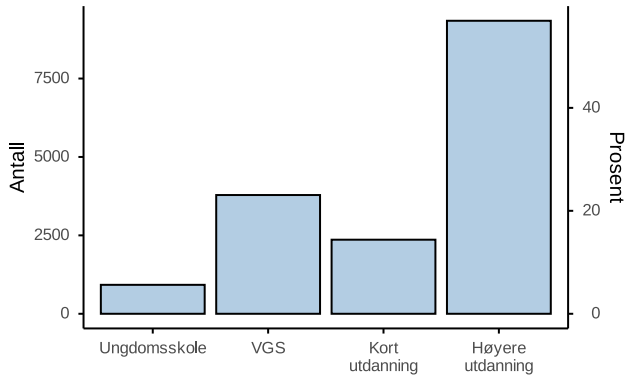
Tabellen viser en opptelling av *Forventet utdanning*. Radene svarer til de ulike verdiene, og i *Antall*-kolonnen finner du hvor mange det var av hver verdi, altså frekvensen. I tillegg viser kolonnen *Prosent* hvor stor andel av totalen hvert antall utgjør. Legger vi sammen prosentene for de to laveste verdiene, ser vi at 28,7 % av elevene ser for seg «Ungdomsskole» eller «VGS» som høyeste fullførte utdanning.

Stort mer er det ikke å si om frekvenstabeller, annet enn hvordan man lager et diagram for å vise det samme.

5.1.2 Stolpediagram

Frekvenstabeller er enkle å lage og lese, men ofte er en visualisering enda mer slående. For nominelle og ordinale data er **stolpediagrammet** den vanligste visualiseringen. Et stolpediagram for *Forventet utdanning* vises i Figur 5.2. Siden *Forventet utdanning* er en ordinal variabel, er stolpene sortert i riktig rekkefølge.

> I jamovi: lage et stolpediagram
>
> Analyses ⟶ Exploration ⟶ Descriptives. Legg variabelen i «Variables», åpne «Plots» og kryss av for «Bar plot».



Figur 5.2: Et stolpediagram over *Forventet utdanning*. Venstre y-akse viser antall elever, høyre y-akse viser samme stolper avlest som andel i prosent.

Man kan velge om stolpediagrammet skal vise antall eller prosent. I Figur 5.2 har jeg laget to y-akser for å vise begge. Venstre viser antall elever, altså frekvensen fra *Antall*-kolonnen i frekvenstabellen. Høyre akse viser nøyaktig de samme stolpene, men avlest som andel i prosent, de samme tallene som står i *Prosent*-kolonnen i Tabell 5.2.

Hvorfor vise begge? For én variabel er det egentlig samme bilde, bare lest på to måter. Disse to lesemåtene blir imidlertid viktige når vi senere skal sammenligne grupper av ulik størrelse. Antallet kan da være misvisende: Gruppen med flest elever vil få høyest stolpe nesten uansett hva andelene er. Andelene gjør stolpene sammenliknbare på tvers av grupper og datasett av ulik størrelse, og det er som regel det vi er ute etter. Vi viser slike analyser i neste kapittel, i Seksjon 6.1.3.

I jamovi: stolpediagram med y-aksen i prosent

Descriptives gir kun y-aksen i antall. For prosent-aksen: Analyses → Frequencies → N Outcomes – χ^2 goodness of fit. Legg variabelen i «Variable», åpne «Plots», kryss av for «Bar plot» og velg «Y-axis: Percentages». Jamovi viser én akse av gangen, ikke begge samtidig.

5.1.3 Typetallet

Typetallet til en variabel er den verdien som forekommer hyppigst. Typetallet er det eneste sentraltendensmålet som fungerer for nominelle variabler; hverken median eller gjennomsnitt gir mening når verdiene ikke har noen naturlig rekkefølge eller tallverdi. En nominell variabel i ICCS-dataene er *Kjønn*, som har de tre mulige verdiene «Jente», «Gutt» og «Annet». Typetallet til *Kjønn* vil si oss hvilket kjønn det er flest av i dataene. De syv første elevene i datasettet er:

Gutt, Jente, Gutt, Jente, Jente, Gutt, Jente.

Typetallet til *Kjønn* for disse syv elevene er «Jente» fordi det var flest av den verdien.

For å finne typetallet i hele datasettet lager vi en **frekvenstabell** for *Kjønn* på samme måte som vi gjorde for *Forventet utdanning* tidligere i kapittelet. Frekvenstabellen, Tabell 5.3, viser at «Jente» er den hyppigste verdien, dog med knapp margin, så typetallet til *Kjønn* er «Jente».

Tabell 5.3: Frekvenstabell for *Kjønn*

Kjønn	Antall	Prosent
Jente	8317	50,5
Gutt	7992	48,5
Annet	164	1

Selv om typetallet er det eneste sentraltendensmålet som fungerer for nominelle variabler, kan det også være nyttig å vite typetallet til en ordinal-, intervall- eller forholdstallsvariabel. Hvis vi ser tilbake på frekvenstabellen for *Forventet utdanning* i Tabell 5.2, så var «Høyere utdanning» den hyppigste verdien. Typetallet til *Forventet utdanning* er altså «Høyere utdanning».

5.1.4 Medianen for ordinale data

For ordinale data kan vi i tillegg bruke **medianen** som mål på sentraltendens. Medianen er den midterste verdien når observasjonene er sortert i rekkefølge. Den fungerer når verdiene har en naturlig rekkefølge, selv uten tallverdier.

La oss se på *Forventet utdanning* for de 7 første elevene i utvalget:

> VGS, Ungdomsskole, Høyere utdanning, Høyere utdanning, Høyere utdanning, VGS, Høyere utdanning.

Sorterer vi dem i rekkefølge får vi:

> Ungdomsskole, VGS, VGS, Høyere utdanning, Høyere utdanning, Høyere utdanning, Høyere utdanning.

Middelverdien er «Høyere utdanning», og det er altså medianen til *Forventet utdanning* for disse 7 elevene.

For tallvariabler regner vi medianen på akkurat samme måte: Sorter, og finn middelverdien. Av og til er det ingen verdier helt i midten, slik som her:

> 1, 3, 5, 6, 10, 15

Her er både 5 og 6 like nære midten. Da er medianen definert til å være gjennomsnittet av disse to verdiene, altså 5,5. For ordinale variabler går ikke dette, siden verdiene ikke kan legges sammen; er de to midterste verdiene forskjellige, oppgir man dem derfor gjerne begge som median.

Vi kommer tilbake til medianen i Seksjon 5.2.2.2, der vi ser hvordan den står seg mot gjennomsnittet.

5.2 Tallvariabler

Nå beveger vi oss over til tallvariabler, altså data på intervall- eller forholdstallsnivå. I resten av kapittelet skal vi fokusere på variabelen *Likestilling* fra ICCS-dataene. Som du kanskje oppfattet fra Kapittel 2. bør man først spørre seg hva variabelen egentlig måler. Dette er viktig nok til at vi bruker litt tid på det.

Jeg fant frem i dokumentasjonen til ICCS 2022 (Fraillon et al., 2024) at *Likestilling* er en samlevariabel satt sammen av følgende Likert-påstander (se Kapittel 2.):

- Men and women should have equal opportunities to take part in government.
- Men and women should have the same rights in every way.
- Women should stay out of politics.
- When there are not many jobs available, men should have more right to a job than women.
- Men and women should get equal pay when they are doing the same jobs.
- Men are better qualified to be political leaders than women.

Alle påstandene ser jo ut til å måle likestilling på en relevant måte, men i Norge kan operasjonaliseringen kanskje virke litt tam, siden nesten alle norske elever støtter likestilling mellom kjønnene. For norske elever ville man kanskje valgt påstander som det faktisk er uenighet om. En mer relevant vinkling kunne være holdninger til likestilling i idrett: er det rettferdig at mannlige fotballspillere tjener mye mer og har bedre betingelser enn kvinnelige spillere, eller er forskjeller mellom kvinner og menn i idretter som skihopp akseptable? Slike påstander ligger nærmere ungdomsskoleelevenes egen erfaring og er lettere å være uenige i. Men nasjonalt tilpassede påstander som dette ville neppe passet like godt i alle de andre landene der ICCS gjennomføres.

Den verdien eleven får på variabelen *Likestilling* gir altså bare mening i lys av disse påstandene. Hadde påstandene vært annerledes, hadde verdien vært en annen og kanskje fortalt en annen historie. I tillegg regner ICCS om elevenes svar på disse påstandene til en samlevariabel. Den går fra ca. 15 (svært imot likestilling mellom kjønnene) til ca. 67 (svært for). Hvis du er interessert i å vite hvordan ICCS gjør denne omregningen, er det bare å slå opp i dokumentasjonen, men det trenger vi ikke for denne boka.

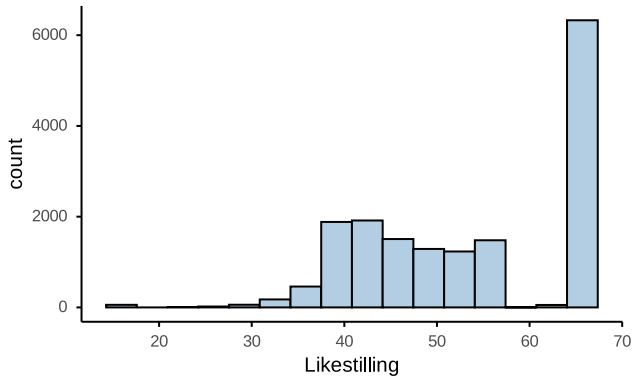
Nå skal vi beskrive denne variabelen ved hjelp av deskriptiv statistikk, og vi starter med histogrammer.

5.2.1 Histogrammer

Histogrammer er blant de mest grunnleggende og nyttige måtene å visualisere data på. De fungerer for tallvariabler, altså variabler på intervall- eller forholdstallsnivå, og gir et tydelig inntrykk av hvordan tallene fordeler seg. De fleste kjenner nok til histogrammer, siden de brukes så ofte, men jeg vil likevel forklare kort hvordan de fungerer.

Prinsippet er enkelt: Du deler opp de mulige verdiene i intervaller og teller hvor mange observasjoner som faller innenfor hvert intervall. Denne tellingen kalles frekvensen til intervallet og vises som en vertikal stolpe. Jeg tror

noe av grunnen til at de oppfattes som så enkle, er at de minner om stolpediagrammer. Forskjellen på histogram og stolpediagram er at histogrammet brukes for tallvariabler og viser fordelingen med sammenhengende stolper, mens stolpediagrammet brukes for kategoriske variabler og viser kategorier med adskilte stolper. Figur 5.3 viser et histogram av *Likestilling* for alle elevene i datasettet, altså elever fra Norge, Spania, Polen og Brasil. Vi ser at de fleste elevene skårer høyt på variabelen, men at en betydelig andel svarer mye lavere. Stolpen helt til høyre representerer de elevene som fikk høyest skår, og høyden på stolpen viser at det er rundt 6000 av dem. Antallet elever leser du av på y-aksen, mens selve skåren står på x-aksen.

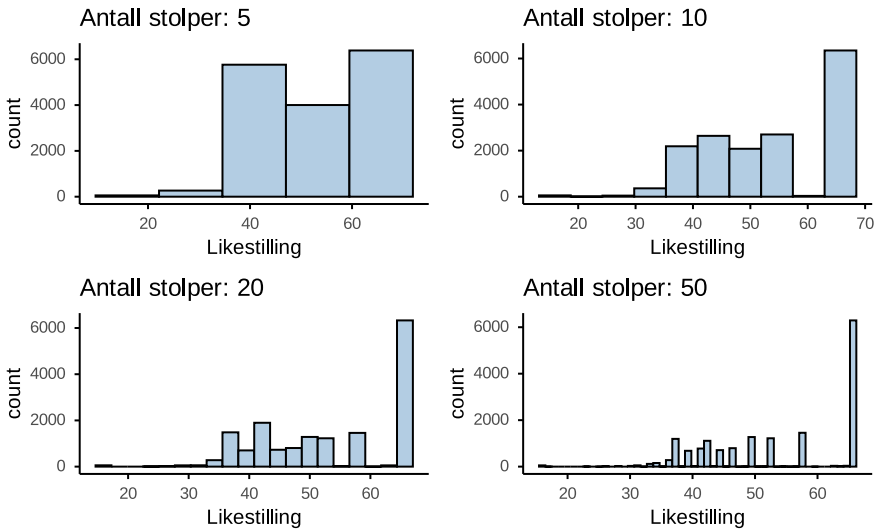


Figur 5.3: Et histogram av variabelen *Likestilling* fra ICCS 2022.

I jamovi: lage et histogram

Analyses ⟶ Exploration ⟶ Descriptives. Legg variabelen i «Variables», åpne «Plots» og kryss av for «Histogram».

Problemet med histogrammer er at utseende deres er veldig avhengig av hvor mange stolper man lager. I Figur 5.4 ser vi den samme variabelen med litt forskjellig valg av antall stolper. De blir seende helt forskjellige ut.



Figur 5.4: Variabelen *Likestilling* framstilt med ulikt antall stolper.

Det finnes ingen fasit for hvor mange stolper man skal lage i et histogram. Antall stolper bør velges slik at histogrammet gir et tydelig og informativt bilde av fordelingen, uten at viktige mønstre forsvinner eller drukner i detaljer.

5.2.2 Sentraltendensmål for tallvariabler

Å tegne histogrammer er en effektiv måte å vise fram hovedbudskapet i dataene dine, men det kan også være nyttig å sammenfatte dataene med noen få enkle tall. Det første du som regel vil vite noe om, er **sentraltendensen**, altså hva som er «midten», «gjennomsnittet», «det typiske» for dataene. Du har allerede møtt to sentraltendensmål, typetallet i Seksjon 5.1.3 og medianen i Seksjon 5.1.4. Nå legger vi til **gjennomsnittet**.

5.2.2.1 Gjennomsnittet

Gjennomsnittet finner vi ved å legge sammen alle verdiene og dele på antall verdier. De første 7 verdiene for variabelen *Likestilling* i Tabell 5.1 var

46,8, 40,9, 65,7, 65,7, 65,7, 65,7, 52,7,

så gjennomsnittet blir:

$$\begin{aligned} &\frac{46,8+40,9+65,7+65,7+65,7+65,7+52,7}{7} \\ &= \frac{403,2}{7} \\ &= 57,6 \end{aligned} \tag{5.1}$$

Dette er sannsynligvis godt kjent for de fleste.

Det som kanskje er nytt, er at man nesten aldri regner dette ut for hånd. Når du har mange observasjoner, slik som 16473 elever i ICCS-dataene, gjør

vi heller beregningene digitalt. Tabell 5.4 viser standard deskriptiv statistikk for *Likestilling*.

> I jamovi: regne ut gjennomsnitt og annen deskriptiv statistikk
>
> Analyses ⟶ Exploration ⟶ Descriptives. Legg *Likestilling* i «Variables». En tabell med standard deskriptiv statistikk dukker opp på høyre side av skjermen.

Tabell 5.4: Standard deskriptiv statistikk for variabelen *Likestilling*.

Mål	**Likestilling**
N	16 473
Gjennomsnitt	53,4
Median	52,7
Standardavvik	11,4
Minimumsverdi	15,9
Maksimumsverdi	65,7

Tabellen viser at gjennomsnittsverdien for *Likestilling* er 53,4. I tillegg får du annen nyttig informasjon som det totale antallet observasjoner (N = 16473)[14], samt median-, minimumsverdi- og maksimumsverdi og standardavviket for variabelen.

Det er viktig å huske at disse sentraltendensene og spredningsmålene er en oppsummering av hele gruppa, ikke en beskrivelse av noen enkelt elev. At gjennomsnittsskåren på *Likestilling* er 53,4, betyr ikke at den typiske eleven skåret 53,4; elevene fordeler seg over et stort spenn rundt det tallet.

Variabelen *Likestilling* gir tallverdier, men hvordan gjør vi det hvis vi vil regne ut gjennomsnittet av variabelen *Forventet utdanning*? De 7 første verdiene av denne variabelen er

> VGS, Ungdomsskole, Høyere utdanning, Høyere utdanning, Høyere utdanning, VGS, Høyere utdanning.

Hvordan skal man regne gjennomsnittet av dette? Det går ikke, for man kan ikke legge sammen verdiene. Derfor kan man bare regne gjennomsnitt for variabler med tall, altså variabler på intervall- og forholdstallskala.

5.2.2.2 Gjennomsnitt eller median?

Det er ikke nok å vite hvordan man regner ut gjennomsnitt og median; du må også forstå hva de forteller deg om dataene dine. Dette er illustrert i Figur 5.5. Tenk på gjennomsnittet som «tyngdepunktet» til datasettet ditt, mens medianen er «middelverdien» i dataene. Her er noen praktiske retningslinjer:

[14] Bokstaven *n*, enten den er stor eller liten, beskriver alltid utvalgsstørrelsen.

- **For nominelle variabler** kan du hverken bruke gjennomsnitt eller median, typetall er den eneste sentraltendensen som fungerer.
- **For ordinale variabler** vil medianen stort sett være det beste valget. Medianen bruker kun rekkefølgeinformasjonen i dataene dine (hvilke verdier som er større eller mindre enn andre), noe som passer perfekt for ordinale data. Gjennomsnittet derimot, er avhengig av presise tallverdier. Det er likevel ganske vanlig å regne gjennomsnitt av variabler på ordinalnivå. Da setter man for eksempel det laveste nivået til tallet 1, det neste nivået til tallet 2 osv.
- **For tallvariabler** er både median og gjennomsnitt relevante valg.

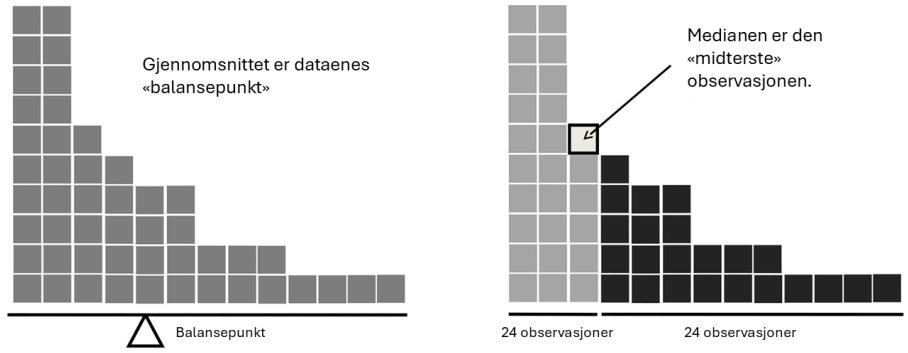
Hva du velger av median og gjennomsnitt for å analysere tallvariabler, avhenger av hva du ønsker å fremheve.

Jeg må forklare med et eksempel. Tenk deg at Arne (månedsinntekt 50 000 kr), Berit (60 000 kr) og Cato (65 000 kr) sitter ved et bord. Gjennomsnittsinntekten ved bordet er 58 333 kr og medianinntekten er 60 000 kr. Begge disse er gode sentralverdier for dataene.

Så kommer Erling Braut Haaland og setter seg ved bordet (månedsinntekt 30 000 000 kr). Plutselig gjør gjennomsnittsinntekten et byks til 7 543 750 kr, mens medianen kun stiger til 62 500 kr! Gjennomsnittsverdien representerer plutselig ikke de andre i det hele tatt, fordi den ekstreme verdien til Erling Braut Haaland trekker gjennomsnittet opp for mye. Medianen, derimot, ser ikke den ekstreme verdien i det hele tatt.

Medianens styrke er altså at den er robust mot ekstreme verdier. Gjennomsnittet kan likevel være relevant hvis du er interessert i den totale inntekten ved bordet, men hvis du vil vite hva som er en typisk inntekt ved Haaland-bordet, ville medianen gi deg et mye mer representativt bilde.

Virkeligheten er vanligvis ikke like ekstrem som Haaland-eksempelet, men samme tankemåte spiller inn når fordelingen er skjev. Figur 5.5 illustrerer at i en skjev fordeling, med hovedtyngden til venstre og en hale av ekstreme verdier til høyre, ligger medianen nærmere hovedtyngden, mens gjennomsnittet trekkes mot halen.



Figur 5.5: En illustrasjon av forskjellen mellom hvordan gjennomsnittet og medianen bør tolkes. Gjennomsnittet er "balansepunktet" til datasettet. Hvis du forestiller deg at histogrammet av dataene er et klosser på et brett, så er gjennomsnittet balansepunktet. Medianen er derimot den midterste observasjonen, slik at halvparten av klossene ligger til venstre og halvparten ligger til høyre for medianen.

Valget mellom gjennomsnitt og median er viktig når dataene har en hale som trekker opp eller ned gjennomsnittet. I slike tilfeller bør du velge median for at sentraltendensmålet skal samsvare best med hovedtyngden av dataene. Slike fordelinger med en lang hale på én side kalles **skjeve**, og du vil møte begrepet i forskningsartikler. Et eksempel der skjevhet er viktig er sammenlikninger av gjennomsnittsinntekt. USA har høyere gjennomsnittsinntekt enn Norge, men lavere medianinntekt. Årsaken er at USA har flere superrikinger som trekker opp gjennomsnittet på samme måte som Erling Braut Haaland gjorde i vårt eksempel. Men den vanlige lønnstaker sin inntekt er mer lik medianinntekten, og en vanlig lønnstaker tjener mer i Norge enn i USA.

5.2.3 Spredningsmål

Alt vi har sett på så langt handler om sentraltendens, altså hvilke verdier som ligger «i midten» eller som er mest «typiske» i dataene våre. I tillegg trenger vi ofte å forstå hvor spredt dataene er. Ligger de fleste observasjonene tett rundt gjennomsnittet, eller er de spredt jevnt utover? Dette kaller vi **spredning**.

Spredning er også en viktig egenskap ved dataene. Tenk deg to klasser som begge har 4,0 i snitt på en prøve. I den ene ligger nesten alle elevene rundt fireren. I den andre er halve klassen på topp og halve på bunn, slik at snittet på 4,0 skjuler at nesten ingen elever egentlig ligger der. Som lærer vet du at dette er to helt forskjellige klasser å undervise, men gjennomsnittet alene klarer ikke å skille dem. Det er nettopp dette et spredningsmål fanger opp: om elevene ligger tett rundt gjennomsnittet, eller spredt utover.

La oss utforske fire ulike måter å måle spredning på ved hjelp av ICCS-dataene. Hvert spredningsmål har sine egne fordeler og ulemper, så det er nyttig å kjenne til flere.

5.2.3.1 Variasjonsbredde

Variasjonsbredden er det aller enkleste spredningsmålet. Den regnes ut ved å ta den største verdien minus den minste verdien. Tabell 5.5 viser variasjonsbredden og andre spredningsmål for *Likestilling*.

> I jamovi: regne ut spredningsmål
>
> Analyses ⟶ Exploration ⟶ Descriptives. Under «Statistics» finner du seksjonen «Dispersion» med avkryssingsbokser for «Std. deviation», «Variance», «Range», «Minimum» og «Maximum». Variasjonsbredde heter «Range» på engelsk, så kryss av for den.

Tabell 5.5: Spredningsmål for variabelen *Likestilling*.

Mål	Likestilling
Variasjonsbredde	49,8
Kvartilbredde	23,1
Standardavvik	11,4
Varians	129,9
Minimumsverdi	15,9
Maksimumsverdi	65,7

Variasjonsbredden er lett å forstå, men har en egenskap som ofte er en ulempe: Den påvirkes veldig av ekstreme verdier.

Se for deg denne lille datamengden: −100, 2, 3, 4, 5, 6, 7, 8, 9, 10

Her får vi en variasjonsbredde på hele 110 på grunn av den ene ekstreme verdien (-100). Men hvis vi fjerner denne ekstreme verdien, blir variasjonsbredden bare 8. Det er en enorm forskjell! Dette må man være obs på når man benytter variasjonsbredden.

5.2.3.2 Kvartilbredde

Ordet «kvartil» i **kvartilbredden** er satt sammen av «kvart» og «persentil». «Kvart» betyr en firedel, og tanken er at vi deler det sorterte datasettet i fire like store deler. En **persentil** er en verdi som en viss andel av observasjonene ligger under: den 25. persentilen er verdien som 25 % av observasjonene er mindre enn. Kvartilene er navn på tre bestemte persentiler, 25., 50. og 75., og kalles henholdsvis første, andre og tredje kvartil.

Kvartilbredden er litt som variasjonsbredden, men i stedet for forskjellen mellom den største og minste verdien ser vi på forskjellen mellom første og tredje kvartil, altså mellom 25. og 75. persentil.

I jamovi: regne ut persentiler

Analyses → Exploration → Descriptives → «Percentile Values» og kryss av for «Percentiles».

Kvartilbredden (*interquartile range, IQR*) er altså differansen mellom 25. og 75. persentil, og for variabelen *Likestilling* er

- første kvartil lik 42,6,
- andre kvartil lik 52,7,
- tredje kvartil lik 65,7.

Vi kan finne kvartilbredden ved å regne avstanden mellom første og tredje kvartil: 65,7 − 42,6 = 23,1. Siden 25% av dataene er mindre enn 42,6 og 75% av dataene er mindre enn 65,7, ligger 50% av dataene i et intervall som er 23,1 bredt. Legg også merke til at 50. persentil og medianen har samme verdi, 52,7, akkurat slik det skal være. I forskningsartikler oppsummeres alt dette slik:

> The median value of *gender equality* was 52.7 (IQR = 42.6–65.7).

Dette er en effektiv måte å vise sentraltendens og spredning. Medianen viser hva midtpunktet i dataene er og kvartilbredden viser hvor spredt halvparten av verdiene ligger rundt medianen. Senere skal du lese om boksplott, som er bygd rundt median og interkvartilbredde, Seksjon 5.2.4.

Kvartilbredden har den samme gode egenskapen som medianen: den påvirkes lite av ekstreme verdier. Har du mange observasjoner, vil én enkelt ekstremverdi som Haalands lønn knapt flytte 25.- og 75.-persentilen, mens gjennomsnittet spretter opp slik vi så i Seksjon 5.2.2.2. Derfor fungerer kvartilbredden dårlig til å vise spredningen rundt et gjennomsnitt, siden gjennomsnittet kan ligge utenfor 25. og 75. persentil. Det ville den gjort i Haaland-eksempelet.

Kvartilbredden fungerer teknisk sett også for ordinale data, men i praksis brukes den nesten utelukkende på tallvariabler.

5.2.3.3 Standardavvik og varians

De to spredningsmålene vi har sett på så langt er differanser, enten mellom største og minste verdi eller mellom tjuefemte og syttifemte persentil. En annen tilnærming er å regne ut hvor stort et «typisk» avvik er fra gjennomsnittet. Det mest brukte spredningsmålet som gjør dette er **standardavviket**.

La oss regne det ut for hånd på de samme 7 verdiene av *Likestilling* som vi brukte for gjennomsnittet, se *Data* i Tabell 5.6. Gjennomsnittet av dem var 57,6. **Avviket** til en observasjon er hvor langt den ligger fra gjennomsnittet, altså verdien minus gjennomsnittet; det første avviket er for eksempel 46,8 − 57,6 = −10,8. Avvikene vises som *Avvik* i samme tabell.

Tabell 5.6: Utregning av standardavviket for de syv første *Likestilling*-verdiene, én verdi per elev. Avvikene har gjennomsnitt null, mens gjennomsnittet av de kvadrerte avvikene er variansen.

	1	2	3	4	5	6	7	Gjennomsnitt
Data	46,8	40,9	65,7	65,7	65,7	65,7	52,7	57,6
Avvik	−10,8	−16,7	8,1	8,1	8,1	8,1	−4,9	0,0
Kvadrert avvik	116,6	278,9	65,6	65,6	65,6	65,6	24,0	97,4

Vi vil sammenfatte avvikene til ett tall som forteller hvor stort et typisk avvik er. Det nærliggende er å ta gjennomsnittet av avvikene, men det går ikke: de positive og negative avvikene opphever hverandre, så gjennomsnittet blir alltid nøyaktig null, slik *Gjennomsnitt* i Tabell 5.6 viser. Trikset er å **kvadrere** hvert avvik først, slik at alt blir positivt; det første avviket, $-10,8$, blir da $(-10,8)^2 = 116,6$, se raden *Kvadrert avvik*. Nå kan vi ta gjennomsnittet: snittet av de kvadrerte avvikene er 97,4, og standardavviket er kvadratroten av det, slik at vi kommer tilbake til samme størrelsesorden som dataene:

$$s = \sqrt{97,4} = 9,9 \tag{5.2}$$

Et typisk avvik fra gjennomsnittet er altså omtrent 9,9 likestillingspoeng for disse elevene. Det virker jo ganske fornuftig ut fra *Avvik*-verdiene i tabellen.

Tallet under rottegnet, 97,4, har et eget navn: det er **variansen**, altså det gjennomsnittlige kvadrerte avviket nederst til høyre i tabellen. Variansen er viktig, og vi møter det igjen i Kapittel 7..

Teknisk om variansen

Det er vanskelig å tolke variansen direkte fordi enheten også er kvadrert. Akkurat som «meter» og «kvadratmeter» er helt forskjellige ting, så er «likestillingspoeng» og «kvadratlikestillingspoeng» helt forskjellige ting. Hvis du hører fra gymlæreren at elevene i klassen din sprang i gjennomsnitt 2300 meter på Cooper-testen, så vil du synes det var rart om han sa at variasjonen var ca. 160 000 kvadratmeter. Det gir mye mening å ta kvadratroten og oppgi at variasjonen var 400 m. Standardavviket er intuitivt fordi vi tar kvadratroten til slutt. I Kapittel 7. skal vi se hvorfor variansen også er viktig.

> ⚠ Teknisk advarsel om jamovi
>
> En liten teknisk detalj: vi delte summen av de kvadrerte avvikene på antallet observasjoner, $n = 7$. Statistikkprogrammer som jamovi deler i stedet på $n - 1$, og ville derfor fått et litt annet tall enn vårt. Forskjellen er ubetydelig for store datasett, men *hvorfor* det gjøres slik henger sammen med skillet mellom å beskrive utvalget og å estimere populasjonen. Det kommer vi tilbake til i Kapittel 7..

Husk likevel at standardavviket sammenfatter spredningen i ett enkelt tall og sier ingenting om *formen* på fordelingen. To variabler kan ha samme standardavvik selv om histogrammene deres ser helt forskjellige ut. Dette gjelder for alle sentraltendensene og spredningsmålene: Figurer gir ofte et bedre bilde enn tallene alene.

Når man rapporterer standardavviket i en forskningsartikkel gjøres det gjerne slik:

> The mean of the variable *Gender equality* was 53.4 (SD = 11.4).

Her står SD for *standard deviation*, som er engelsk for standardavvik.

5.2.3.4 Hvilket spredningsmål skal du bruke?

Jeg har gitt deg fire ulike spredningsmål: variasjonsbredde, kvartilbredde, varians og standardavvik. Jeg oppsummerer egenskapene deres kort:

- **Variasjonsbredde**: Viser avstanden mellom den største og minste verdien i dataene. Variasjonsbredden påvirkes lett av ekstremverdier, så den brukes vanligvis kun når du har spesielle grunner til å fokusere på ytterpunktene.
- **Kvartilbredde**: Beskriver hvor «den midterste halvdelen» av dataene befinner seg. Den er robust mot ekstremverdier og passer perfekt sammen med medianen. Et godt valg!
- **Standardavvik**: Er kvadratroten av variansen. Den kombinerer matematisk nytteverdi og praktisk tolkbarhet siden den bruker samme enheter som dataene. Den klare favoritten når gjennomsnittet er din foretrukne sentraltendens. Standardavviket er definitivt det mest brukte spredningsmålet!
- **Varians**: Grunnlaget standardavviket er utledet fra. Variansen bruker kvadrerte enheter og er derfor vanskelig å tolke direkte, men den er nyttig i videre statistikk.

Kort fortalt: Kvartilbredde og standardavvik er de to klart mest populære valgene for å beskrive variabilitet, men variansen er den nyttigste i senere utregninger. Variasjonsbredden bruker mange studenter som spredningsmål i mastergradene sine, sikkert fordi den er så intuitiv, så det er viktig å vite om dens styrker og svakheter også.

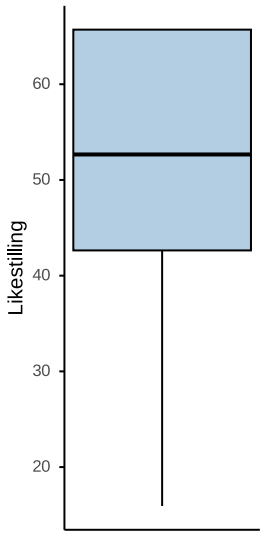
5.2.4 Boksplott

Et godt alternativ til histogrammer er **boksplott**. Ideen bak et boksplott er å gi deg en rask og tydelig visuell oversikt over medianen og kvartilbredden.

Siden denne informasjonen var enkel å regne ut også før datamaskiner ble vanlig, og fordi boksplott presenterer all denne informasjonen på en kompakt måte, er boksplott populære. Vi tar en titt på variabelen *Likestilling*, se i Figur 5.6.

I jamovi: lage et boksplott

Analyses → Exploration → Descriptives. Legg variabelen i «Variables», åpne «Plots» og kryss av for «Box plot».



Figur 5.6: Et boksplott av variabelen *Likestilling*.

Den blågrå boksen starter og slutter ved første og tredje kvartil, og viser derfor kvartilbredden. Den tykke linjen midt i boksen viser medianen. De tynne linjene viser variasjonsbredden ved at de strekker seg ut til de mest ekstreme datapunktene i hver retning, i hvert fall nesten. I dette plottet går det ingen strek oppover, fordi ingen verdier er høyere enn 75. persentil; så mange elever fikk toppskår på *Likestilling*. Som jeg skrev i Seksjon 5.2.3.1, så er variasjonsbredden veldig følsom for hvor ekstreme de mest ekstreme verdiene er. Derfor viser de fleste boksplott ikke variasjonsbredden, men kutter strekene ved en eller annen grense. Som standard er denne grensen satt til 1,5 ganger kvartilbredden (IQR). Hvis noen observasjoner faller utenfor dette området, vises de som prikker eller sirkler i stedet for. Disse kalles **avviksverdier** (*outliers* på engelsk).[15]

I våre *Likestilling*-data er det ingen observasjoner som er avvikende, så det er ingen avviksverdier og derfor ingen prikker.

Boksplottet kommer virkelig til sin rett når man sammenlikner undergrupper, som kommer i Seksjon 6.2.1.

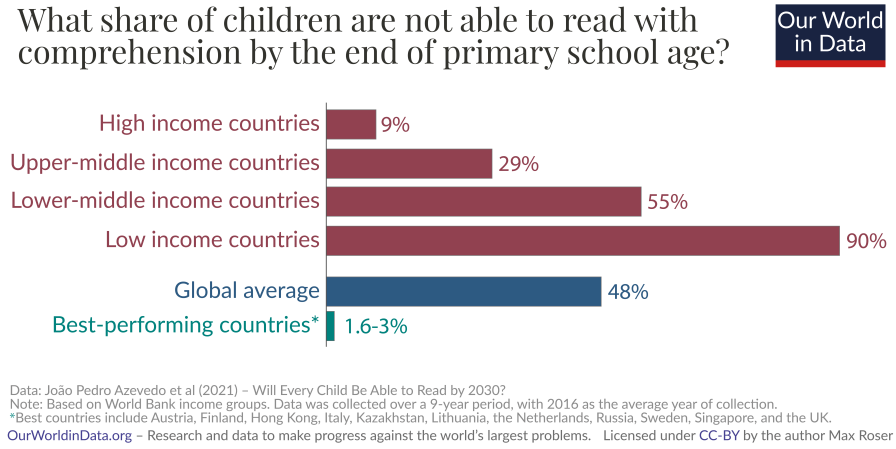
[15] Av naturlige årsaker blir dette ofte oversatt til «uteliggere» på norsk. Av like naturlige årsaker velger jeg å *ikke* benytte ordet «uteligger» for avviksverdiene.

5.3 Profesjonelle grafer

Alle grafene i dette kapittelet er laget i jamovi, og jamovi tegner standard-grafer som er tydelige og minimalistiske. Hva skiller dem fra mer forseggjorte grafer? Stort sett fargevalg og tekst-annoteringer som må skreddersys til hver graf. Se for eksempel på stolpediagrammet i Figur 5.7. Det er et vanlig stolpediagram, men

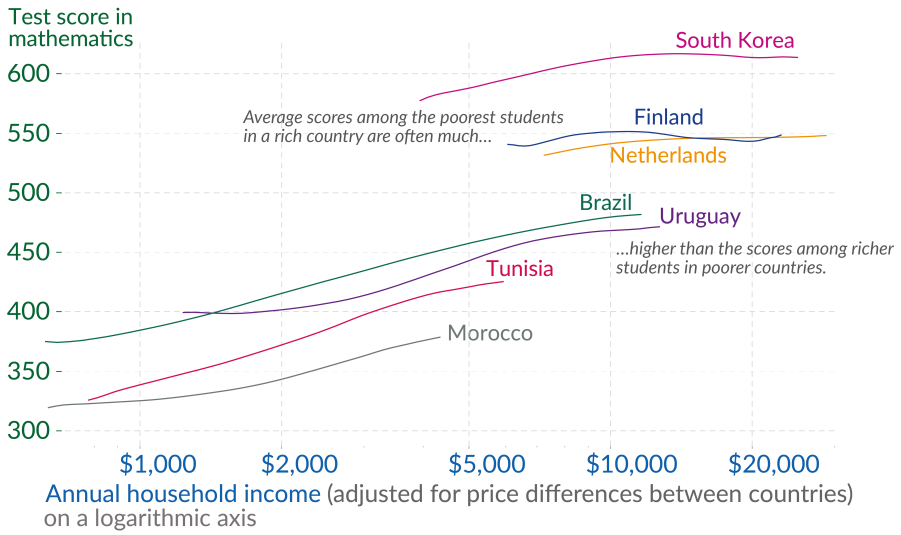
- det er lagt horisontalt slik at teksten skal bli lettere å lese,
- hver søyle er markert med prosenter,
- aksene er fjernet,
- farger er brukt for å vise grupper av søyler,
- tittel og figurnoter forteller hele historien.

Alt i alt blir øyet ledet rundt i figuren slik at den samlet sett forteller en interessant historie om globale utdanningsforskjeller.



Figur 5.7: Et profesjonelt stolpediagram. Legg merke til forklarende tekst og farger. Grafen er laget av Roser (2022). Utgitt under lisensen CC-BY.

Eksempelet i Figur 5.7 er et standard stolpediagram; andre diagrammer er spesiallaget. Se for eksempel på Figur 5.8, som på en svært illustrerende måte viser at ungene lærer mer i Sør-Korea, Finland og Nederland enn i Brasil og Uruguay, selv hvis man sammenlikner familier med samme kjøpekraft.



Figur 5.8: Et profesjonelt linjediagram. Legg merke til forklarende tekst og farger. Grafen er laget av Roser (2022). Utgitt under lisensen CC-BY.

Andre visualiseringer er enda mer spesialiserte, som for eksempel for å vise hvordan amerikanere tilbringer dagene sine eller hele denne saken fra NRKs team med data-journalister som visualiserer naturnedbygging i Norge. Det er lov å la seg inspirere!

5.4 Lagre bildefiler med jamovi

Du lurer kanskje på hvordan du lagrer alle grafene vi har laget. Bare høyreklikk på grafen og eksporter det til en fil. Du kan velge mellom flere formater som «png», «eps», «svg» eller «pdf». Alle disse formatene gir deg høy kvalitet på bildene, perfekt for å dele med kolleger, inkludere i oppgaver eller bruke i artikler og presentasjoner, men «.png» er trolig det som vil samarbeide best med Microsoft Word.

5.5 Oppsummering

Å beregne grunnleggende deskriptiv statistikk og lage gode figurer er noe av det første du gjør når du analyserer data. Deskriptiv statistikk beskriver utvalget og forteller altså ingenting om verdiene for populasjonen.

Jeg har dekket disse temaene:

- **Kategoriske variabler**: Jeg viste hvordan man lager Frekvenstabeller og Stolpediagram for nominelle og ordinale data, og hvordan man finner Typetallet.
- **Histogrammer**: Histogrammer viser fordelingen av tallvariabler og er ofte det mest informative verktøyet.
- **Mål for sentraltendens**: Typetallet fungerer for alle variabeltyper, medianen krever ordinal- eller tallvariabler, og gjennomsnittet krever tallvariabler. Bruk medianen hvis dataene er skjeve; bruk gjennomsnittet ellers.

- **Mål for spredning**: Spredningsmål viser deg hvor «spredt» dataene dine er. De viktigste målene er kvartilbredde og standardavvik, mens varians spiller en nøkkelrolle i videre statistikk.
- **Boksplott**: Boksplott er en kompakt måte å visualisere spredning og sentraltendens, spesielt nyttig for å sammenligne undergrupper.

6. Å beskrive sammenhenger mellom variabler

I forrige kapittel lærte du å beskrive én variabel om gangen, med blant annet frekvenstabeller, histogram, gjennomsnitt og standardavvik. Det er nyttig for å forstå *hva* dataene inneholder, men i de fleste studier er vi mest interessert i *hvordan variabler henger sammen*. Er det forskjell mellom elever fra ulike land? Henger sosioøkonomisk status sammen med holdninger til likestilling? Slike spørsmål krever at vi ser på to variabler samtidig.

Framgangsmåten avhenger av hva slags variabler vi har. Derfor er kapittelet organisert etter variabeltype: først undersøker vi sammenhengen mellom to kategoriske variabler, deretter mellom en kategorisk og en kontinuerlig variabel, og til slutt mellom to kontinuerlige variabler.

Vi bruker syv variabler fra ICCS-datasettet gjennom kapittelet. Tabell 6.1 viser målenivået for hver variabel og hva verdiene representerer. Inndelingen i *Type* (kategorisk eller kontinuerlig) er den grove grupperingen kapittelet er organisert etter, mens *Målenivå* er det mer finmaskede vokabularet (nominal, ordinal, intervall, forhold) fra Kapittel 5.. Det kan være lurt å gå tilbake til Tabell 6.1 underveis i kapittelet hvis du er usikker på variablene.

Tabell 6.1: Variablene fra ICCS-datasettet som vi bruker i dette kapittelet.

Variabel	Type	Målenivå	Verdier
Land	Kategorisk	Nominal	Norge, Spania, Polen, Brasil
Kjønn	Kategorisk	Nominal	Jente, Gutt, Annet
Forventet utdanning	Kategorisk	Ordinal	Ungdomsskole < VGS < Kort utdanning < Høyere utdanning
Likestilling	Kontinuerlig	Intervall	Høyere skår betyr mer positiv til likestilling
Demokratikunnskap	Kontinuerlig	Intervall	Høyere skår betyr mer kunnskap om demokrati
Sosioøkonomisk status	Kontinuerlig	Intervall	Standardisert indeks; 0 ≈ gjennomsnittet
Forventet politisk deltakelse	Kontinuerlig	Intervall	Høyere skår betyr mer forventet politisk engasjement

6.1 Kategorisk × kategorisk

Når begge variablene er kategoriske (nominelle eller ordinale) bruker vi krysstabeller og stablede stolpediagram for å beskrive sammenhengen mellom variablene.

6.1.1 Krysstabeller

I Kapittel 5. lærte du å lage frekvenstabeller for enkeltvariabler. Hvis du ønsker en tabell med to variabler (for eksempel for å kombinere *Forventet utdanning* og *Land* for å se om det er forskjell i hvor lang utdanning elevene ser for seg i de ulike landene), trenger du en **krysstabell** (*contingency table* på engelsk). Resultatet for vårt eksempel ser du i Tabell 6.2. Tallene er antall elever, og prosentene i parentes er andelen innenfor hvert land (kolonneprosent).

Tabell 6.2: Krysstabell over *Forventet utdanning* og *Land*. Kolonnene viser antall elever og kolonneprosent (andel innenfor hvert land).

Forventet utdanning	Norge	Spania	Polen	Brasil
Ungdomsskole	568 (12,2 %)	225 (7,5 %)	130 (3,1 %)	NA
VGS	820 (17,7 %)	831 (27,6 %)	1541 (36,5 %)	594 (12,9 %)
Kort utdanning	770 (16,6 %)	492 (16,3 %)	202 (4,8 %)	898 (19,5 %)
Høyere utdanning	2465 (53,1 %)	1450 (48,2 %)	2319 (55,0 %)	3108 (67,5 %)
NA	17 (0,4 %)	13 (0,4 %)	26 (0,6 %)	4 (0,1 %)

I jamovi: lage en krysstabell

Frequencies → Contingency Tables → Independent Samples. Legg *Forventet utdanning* i «Rows» og *Land* i «Columns». Under «Cells» kryss av for «Counts» og «Percentages: Columns».

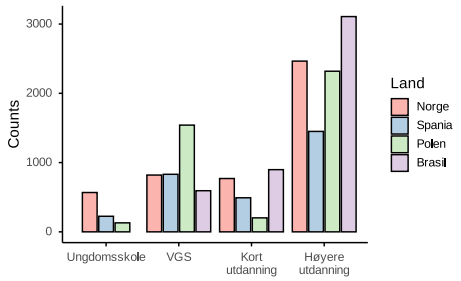
Når du tolker krysstabellen, kan du for eksempel se på cellen for spanske elever som ser for seg en kort utdanning (som kan være ett- til to-årig fagskole). Tallet utenfor parentesen er frekvensen, altså hvor mange spanske elever i utvalget dette gjelder. Tallet i parentesen er kolonneprosenten, og siden *Land* ligger på kolonnene betyr det «andel innenfor Spania som ser for seg kort utdanning». Hvis jeg i stedet hadde vist radprosenter, ville prosentverdien vært en annen og betydd «andelen av elevene som ser for seg kort utdanning, som er fra Spania». Det er to ulike spørsmål, så det er viktig å være bevisst på hvilken type prosent som brukes.

Krysstabeller er nyttige! Se for eksempel at norske elever skiller seg ut ved at en betydelig andel ser for seg å avslutte utdanningen etter ungdomsskolen, mens ingen brasilianske elever gjør det.

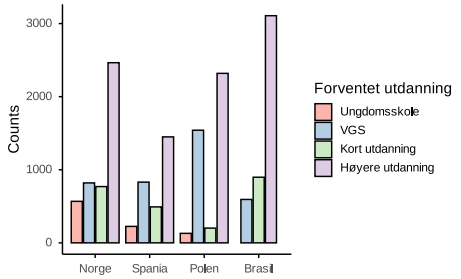
6.1.2 Grupperte stolpediagram

Du lærte å lage stolpediagram i Kapittel 5.. Et **gruppert stolpediagram** lar deg sammenligne undergrupper ved å dele opp dataene etter en ekstra kategorisk variabel. I Figur 6.1a er *Forventet utdanning* plassert på x-aksen, og hvert utdanningsnivå har én stolpe per *Land*. I Figur 6.1b er det byttet om, slik at *Land* er plassert på x-aksen og hvert land har én stolpe per utdanningsnivå. Hvis du lurer på hvilken du skal velge, må du tenke over hvilken visualisering som best kommuniserer ditt poeng eller som best besvarer forskningsspørsmålet ditt.

> I jamovi: stolpediagram for undergrupper
>
> Analyses ⟶ Exploration ⟶ Descriptives. Legg variabelen i «Variables», åpne «Plots» og kryss av for «Bar plot». Legg gruppevariabelen i «Split by».



(a) *Forventet utdanning* på x-aksen, gruppert etter *Land*.

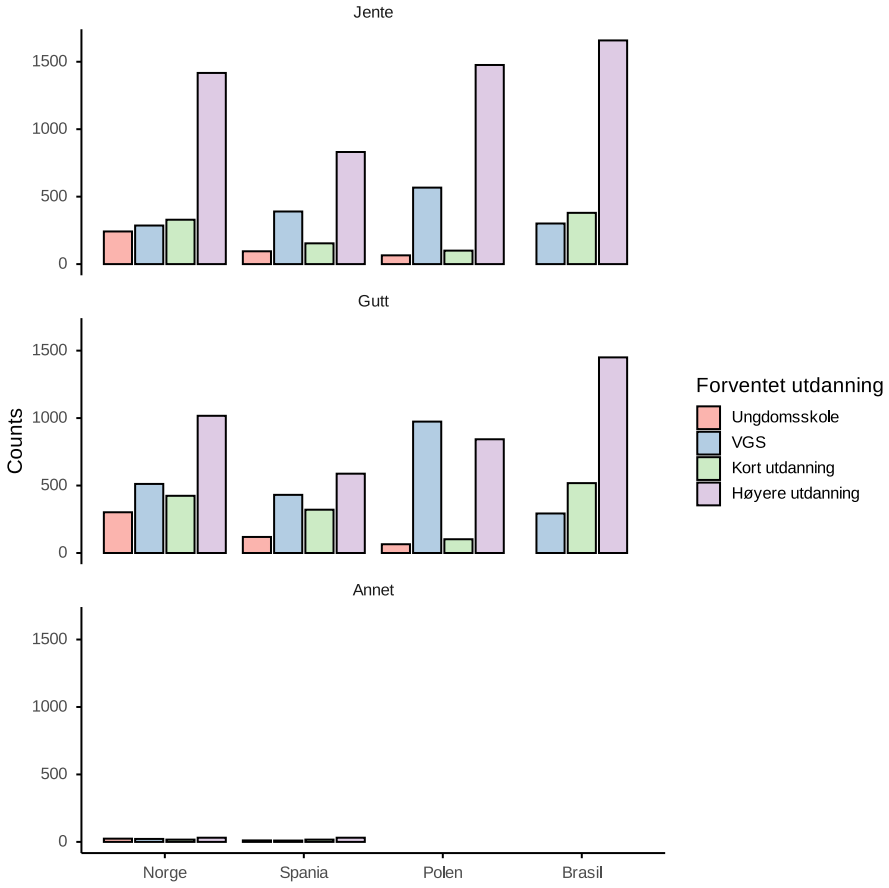


(b) *Land* på x-aksen, gruppert etter *Forventet utdanning*.

Figur 6.1: Gruppert stolpediagram av *Forventet utdanning* og *Land*, vist på to måter.

Vi kan også legge til *enda en variabel*, slik at vi sammenligner undergrupper langs to dimensjoner samtidig. I Figur 6.2 ser du resultatet for *Land*, *Forventet utdanning* og *Kjønn*, der hvert kjønn vises for seg. Figuren blir imidlertid omfattende og noe uoversiktlig. Dette gjelder særlig diagrammet for «Annet»

(nederst), der det lave antallet observasjoner gjør mønstrene vanskelige å lese. I neste delkapittel viser jeg hvordan slike fremstillinger kan forbedres ved hjelp av stablede stolpediagrammer.



Figur 6.2: Antall elever per *Forventet utdanning* og *Land*, der hvert kjønn vises for seg.

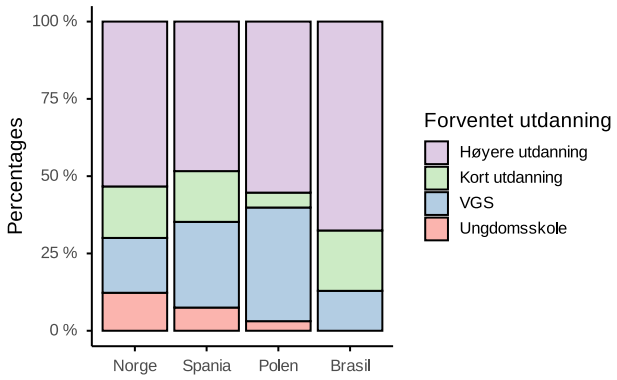
6.1.3 Stablede stolpediagram

I Seksjon 5.1.2 så du at vi ofte ikke ønsker å vise antallet i hver stolpe, men andelen, fordi antallet blir misvisende når gruppene har ulik størrelse. **Stablede stolpediagram** er en god måte å vise andelene på. Du stabler stolpene oppå hverandre slik at de summerer til 100 %, og hver stolpe viser sammensetningen innenfor den aktuelle stolpen. Det blir både kompakt og direkte sammenliknbart mellom grupper, se Figur 6.3.

I jamovi: lage et stablet stolpediagram

Frequencies → Independent samples (χ^2 test of association). Legg *Forventet utdanning* på «Rows» og *Land* på «Columns». Under «Plots» kryss av for «Bar plot» og velg:

- Bar Type: «Stacked»
- Y-axis: «Percentages»
- «Percentages within column» (regner prosenter innenfor hvert land)



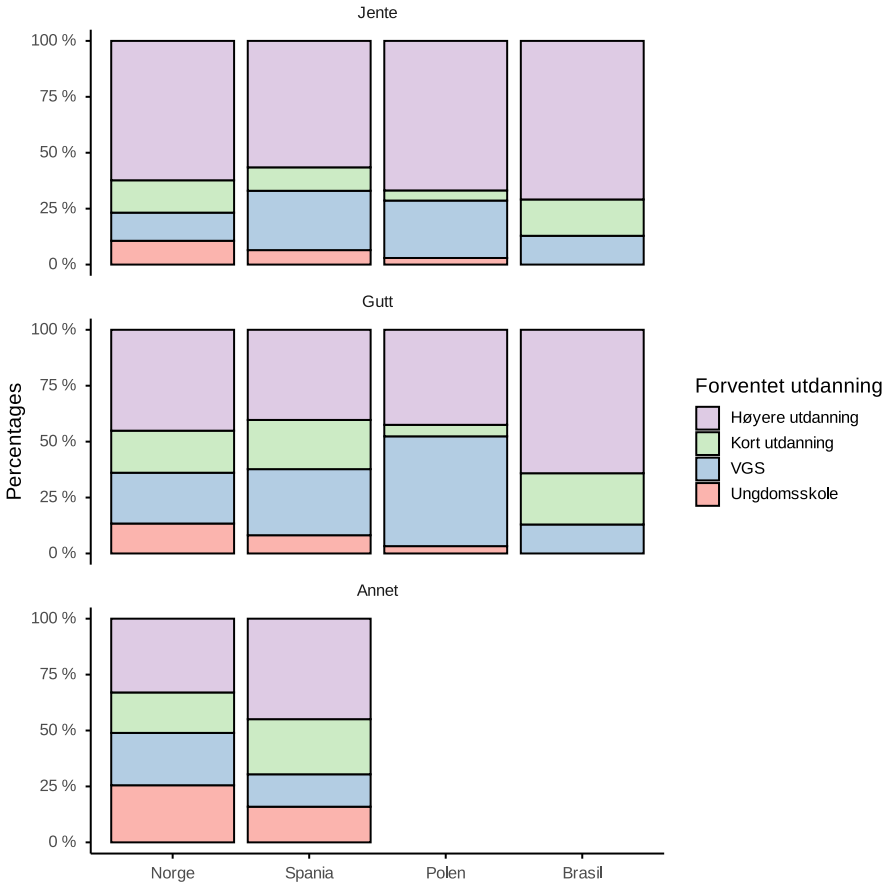
Figur 6.3: Stablet stolpediagram av *Forventet utdanning* for hvert *Land*. Hver stolpe summerer til 100 %, og segmentene viser andelen innenfor landet.

I Figur 6.3 ser vi et stablet stolpediagram som viser *Forventet utdanning* for hvert land. Vi ser tydelig at en større andel elever i Brasil planlegger å ta høyere utdanning enn i Norge. Forskjellene er forholdsvis små, og i stolpediagrammene Figur 6.1a og Figur 6.1b, er det vanskelig å se at brasilianere oftere ser for seg å ta høyere utdanning.

Med vanlige stolpediagram var det vanskelig å gjøre en analyse på tvers av to variabler, siden Figur 6.2 ble så stor. Vi kan prøve å lage en liknende figur med stablede stolper. Resultatet blir som i Figur 6.4, og det stablede stolpediagrammet er mye mer kompakt og lesbart enn samme informasjon vist som et vanlig stolpediagram (Figur 6.2). Den største forskjellen mellom de to diagrammene er at «Annet»-gruppen var nesten umulig å lese med det vanlige stolpediagrammet fordi stolpene ble så små. Med stablede stolper står hver stolpe like høyt (100 %), og vi ser sammensetningen innenfor hver liten gruppe. (Men husk at disse stolpene er basert på veldig få elever, så andelene er usikre.)

I jamovi: stablet stolpediagram med tre variabler

Behold oppsettet for stablet stolpediagram, men legg den tredje variabelen i «Layers».



Figur 6.4: Stablet stolpediagram av *Forventet utdanning* for hvert *Land* og *Kjønn*. Hver stolpe summerer til 100 %.

For å oppsummere: Analyse av sammenhengen mellom to kategoriske variabler betyr å gjøre en undergruppeanalyse der den ene variablen bestemmer undergruppene. Resultatene kan framstilles enten i en krysstabell eller som flere stolpediagram ved siden av hverandre. Stablede stolpediagram gir en kompakt framstilling av andeler, som ofte er det mest relevante i slike analyser.

6.2 Kategorisk × kontinuerlig

Hva om du vil analysere sammenhengen mellom en *kontinuerlig* variabel og en kategorisk variabel? For eksempel vil en analyse av *Land* og *Likestilling* vise om elever fra ulike land svarer forskjellig på variabelen *Likestilling*. I likhet med forrige delkapittel er det bare å lage deskriptiv statistikk *separat for hver gruppe*, der gruppen er definert av den kategoriske variabelen, akkurat som i krysstabellen og stolpediagrammene i forrige delkapittel.

6.2.1 Sentraltendens og spredning per gruppe

Egentlig er dette å regne sentraltendens og spredning per gruppe det samme grepet som vi allerede gjorde med krysstabellen (Tabell 6.2) og de grupperte stolpediagrammene (Figur 6.1): vi splittet dataene etter en kategorisk variabel

og tok for oss hvert utsnitt for seg. Forskjellen nå er at det vi regner ut innenfor hver gruppe er gjennomsnitt, median og standardavvik, ikke antall eller andel. Det hadde for eksempel vært interessant å vite om de spanske og norske elevene svarer forskjellig på variabelen *Likestilling*, eller om de som ser for seg høyere utdanning svarer forskjellig fra de som ser for seg å avslutte utdanningen etter videregående. Slike spørsmål er svært interessante, og derfor er undergruppeanalyser ofte det masterstudenter og andre forskere benytter til å besvare forskningsspørsmålene sine.

Undergruppene vi kan undersøke er nettopp de forskjellige verdiene til de tre kategoriske variablene fra Tabell 6.1: *Land*, *Forventet utdanning* og *Kjønn*. En måte å tenke på verdiene til nominelle og ordinale variabler er altså som undergrupper.

> Å gjøre deskriptiv statistikk for undergrupper er å gjøre utregningene separat for hver verdi av en kategorisk variabel.

For å regne ut deskriptiv statistikk separat for «Gutt», «Jente» og «Annet», må vi dele opp dataene etter verdiene i variabelen *Kjønn*. Resultatet ser du i Tabell 6.3.

I jamovi: deskriptiv statistikk per undergruppe

Analyses ⟶ Exploration ⟶ Descriptives. Legg variabelen i «Variables» og gruppevariabelen i «Split by».

Tabell 6.3: Deskriptiv statistikk for *Likestilling* separat for hver verdi av *Kjønn*.

Kjønn	N	Gjennomsnitt	Median	SD
Jente	8 317	58,2	65,7	9,6
Gutt	7 992	48,4	46,8	10,9
Annet	164	50,3	49,4	13,5

I Tabell 6.3 ser vi at jentene har mye høyere gjennomsnittskår på *Likestilling* enn guttene, 58,2 mot 48,4. De som oppga «Annet», ligger nærmere guttene, med et gjennomsnitt på 50,3. Jentene har dessuten litt mindre standardavvik enn guttene, 9,6 mot 10,9, mens de som oppga «Annet» har mer sprikende svar (SD = 13,5). Bildet er det samme for alle tre gruppene om vi heller ser på medianen og kvartilbredden enn gjennomsnittet og standardavviket.

Vi kan også undersøke om det er forskjell mellom landene for hvert kjønn. Da gjør vi undergruppeanalysen inndelt etter både *Kjønn* og *Land*, se Tabell 6.4.

Tabell 6.4: Deskriptiv statistikk for *Likestilling* separat for hver kombinasjon av *Land* og *Kjønn*. Polen og Brasil har ingen elever som har oppgitt *Annet* kjønn.

Land	Kjønn	N	Gjennomsnitt	Median	SD
Norge	Jente	2 281	61,6	65,7	8,0
Norge	Gutt	2 264	49,5	46,8	11,5
Norge	Annet	95	50,0	49,4	13,7
Spania	Jente	1 475	59,3	65,7	9,2
Spania	Gutt	1 467	51,0	49,4	10,7
Spania	Annet	69	50,7	52,7	13,4
Polen	Jente	2 220	57,5	57,1	8,7
Polen	Gutt	1 998	44,6	42,6	10,0
Polen	Annet	0	—	—	—
Brasil	Jente	2 341	54,8	57,1	10,7
Brasil	Gutt	2 263	49,1	46,8	10,5
Brasil	Annet	0	—	—	—

Mønsteret fra Tabell 6.3 går igjen i alle land: jentene har høyere gjennomsnittskår på *Likestilling* enn guttene. Men det er noen interessante variasjoner. De norske jentene er mer for likestilling enn de spanske (61,6 mot 59,3), mens de norske guttene er *mindre* for likestilling enn de spanske guttene (49,5 mot 51,0). Polske gutter ligger lavest av alle med et gjennomsnitt på 44,6. Men vær forsiktig med å tolke resultatet: dette er forskjeller mellom gjennomsnitt i ett utvalg, og noen av forskjellene kan være tilfeldige. Mer om dette i Kapittel 7..

Når man deler inn i undergrupper på denne måten, kan det være at enkelte undergrupper har svært få eller ingen respondenter. ICCS-utvalget fra Polen og Brasil inneholder ingen elever som har oppgitt «Annet» kjønn, og selv i Norge og Spania er gruppen liten (henholdsvis 95 og 69 respondenter). Det er fremdeles mange nok til at det gir mening å regne gjennomsnitt der vi har dem, men andre ganger vil det være like greit å si at man ikke har mange nok respondenter til å gjøre en undergruppeanalyse for akkurat den gruppen. Merk at dette er et etisk dilemma for forskere: hvis forskning skal gagne mennesker, er det et problem om noen grupper mennesker systematisk blir utelatt fordi de er for små til å analyseres.

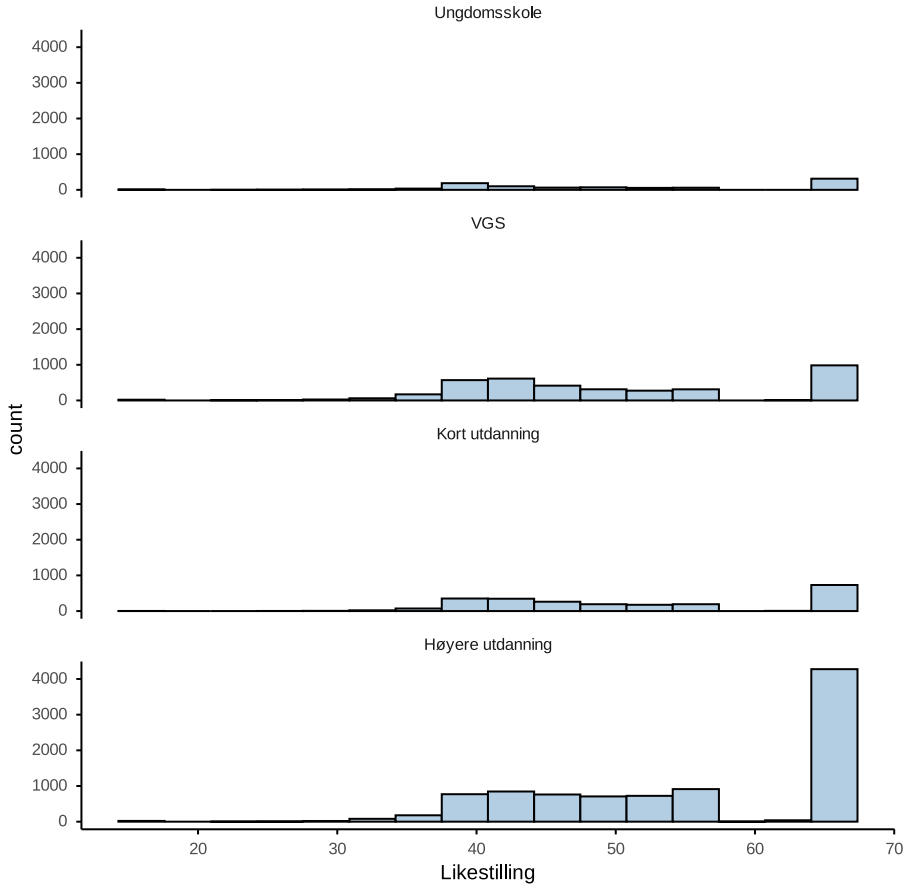
6.2.2 Boksplott og histogrammer per gruppe

Undergruppeanalyse med tall er nyttig, men en visualisering gjør sammenhengen mye tydeligere. Du lærte i Kapittel 5. at boksplott og histogrammer egner seg godt for kontinuerlige variabler, og ved å dele dataene inn i undergrupper får vi automatisk ett plott per gruppe.

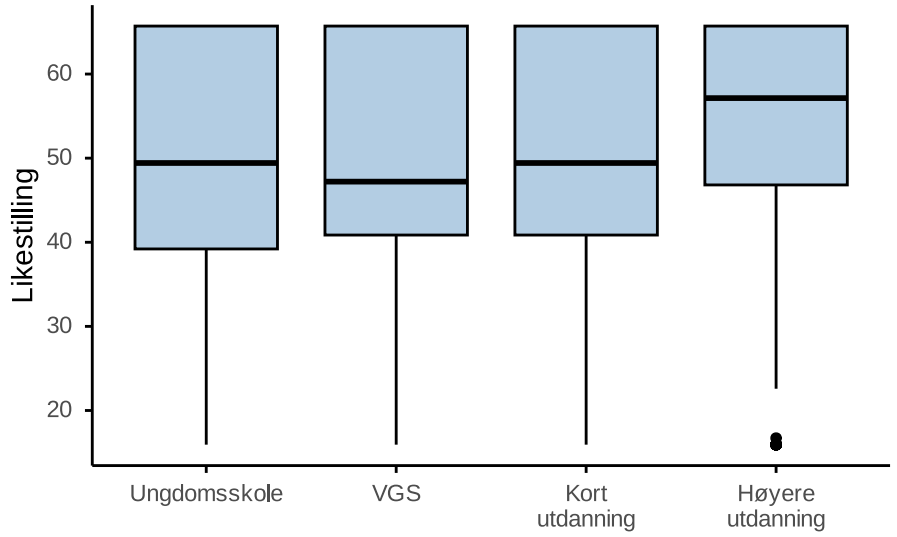
I jamovi: boksplott eller histogram per undergruppe

Analyses → Exploration → Descriptives. Legg variabelen i «Variables», åpne «Plots» og kryss av for «Box plot» eller «Histogram». Legg gruppevariabelen i «Split by».

Se Figur 6.5 for en sammenlikning av visualisering med henholdsvis histogram og boksplott. Det er vanskelig å sammenlikne undergruppene med histogrammet fordi det er forskjellig antall observasjoner i hver søyle, men i boksplottet ser vi klart at elever som ser for seg *Høyere utdanning* er mer for likestilling enn elever med lavere utdanningsambisjoner. Forskjellen kommer kanskje tydeligst frem ved å titte på medianen, den svarte streken i midten av boksen.



(a) Med histogrammet er det vanskelig å se forskjell på gruppene.



(b) Med boksplott kommer forskjellen tydelig frem; særlig at *Høyere utdanning* ligger høyere enn de andre gruppene.

Figur 6.5: Sammenlikning av histogram og boksplott for å vise variabelen *Likestilling* for hvert forventede utdanningsnivå.

For å oppsummere: Å analysere sammenhengen mellom en kategorisk og en kontinuerlig variabel er en undergruppeanalyse, men det vi regner ut innenfor hver gruppe er sentraltendens og spredning i stedet for antall. Resultatet kan vises som tabell eller som ett boksplott per gruppe. Boksplottet er som regel et bedre visuelt valg enn histogrammet, fordi det lar oss sammenligne medianer direkte og gir en tydelig framstilling som ikke påvirkes like mye av ulik gruppestørrelse.

6.3 Kontinuerlig × kontinuerlig

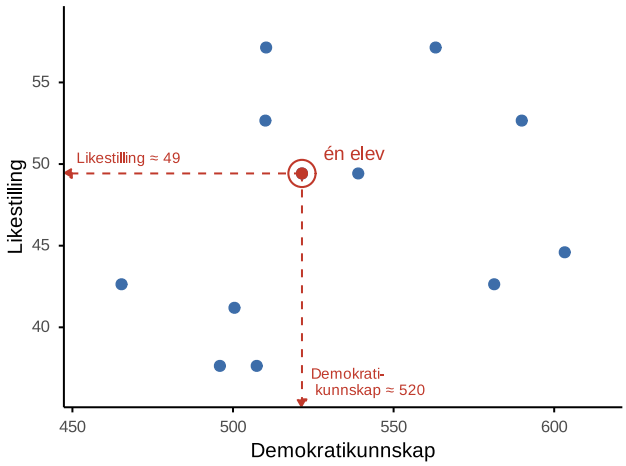
Hva om begge variablene er kontinuerlige? Da kan vi ikke dele inn i grupper, siden det ville bli altfor mange. I stedet bruker vi spredningsplott og korrelasjon.

6.3.1 Spredningsplott

Spredningsplott viser sammenhengen mellom to tallvariabler. ICCS-datasettet vårt har fire kontinuerlige variabler (se Tabell 6.1); to av dem er *Demokratikunnskap* og *Likestilling*. *Demokratikunnskap* er hovedvariabelen i ICCS, og måler elevenes kunnskap om demokrati og samfunn, for eksempel hvordan lover blir til, hva som kjennetegner frie valg, og hvilke rettigheter og plikter innbyggere har. Et naturlig spørsmål å stille er om elever som kan mer om demokrati og samfunn også svarer annerledes på spørsmål om likestilling mellom kjønnene.

Hvis man skulle analysere dette på samme måte som for de kategoriske variablene, måtte man gjort en undergruppeanalyse av *Likestilling* der gruppene er definert av de ulike verdiene av *Demokratikunnskap*. Men fordi *Demokratikunnskap* er satt sammen av mange enkeltspørsmål, har den 15817 ulike verdier, og da ville man støtt på problemer. Man kan ikke vise så mange undergrupper!

Løsningen er å legge hver sin variabel på hver sin akse og plotte hver elev som ett punkt: elevens *Demokratikunnskap* bestemmer hvor langt til høyre punktet ligger, og elevens *Likestilling* hvor langt opp. Figur 6.6 viser prinsippet for noen få elever.

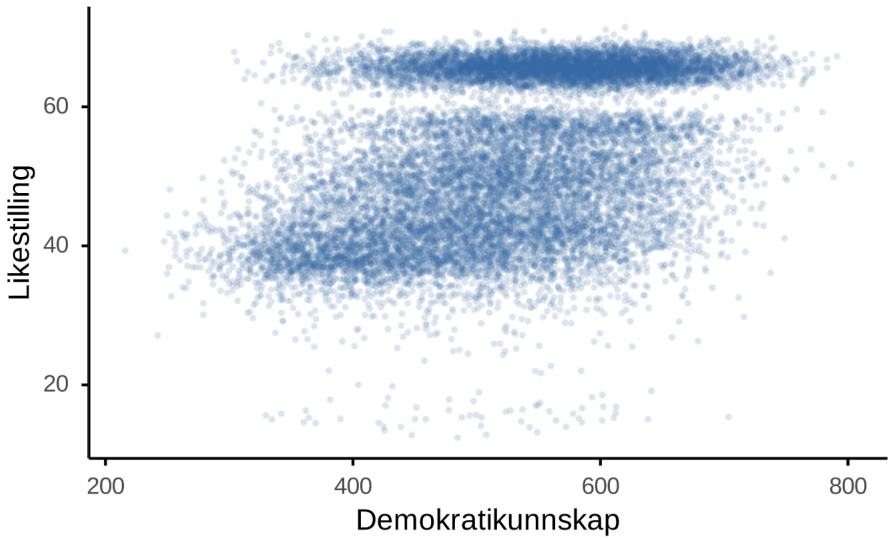


Figur 6.6: Slik leser du et spredningsplott: hvert punkt er én elev, og elevens verdi på hver variabel leser vi av ved å følge punktet loddrett ned til x-aksen (*Demokratikunnskap*) og vannrett bort til y-aksen (*Likestilling*). Her er bare tolv elever tegnet inn, slik at hvert punkt er lett å se.

Gjør vi dette for alle de rundt 16 000 elevene, får vi Figur 6.7. Vi ser en tendens til at punkter som ligger lengre til høyre også ligger litt lenger opp. Dette betyr at elever med høyere *Demokratikunnskap* i større grad mener det skal være *Likestilling* mellom kjønnene.

I jamovi: lage et spredningsplott

Analyses ⟶ Exploration ⟶ Scatterplot. Legg én variabel i «X-Axis» og én i «Y-Axis».



Figur 6.7: Spredningsplott over *Demokratikunnskap* og *Likestilling*. Hvert punkt er en elev; den mørke randen langs toppen viser at svært mange elever har høyest mulig skår på *Likestilling*.

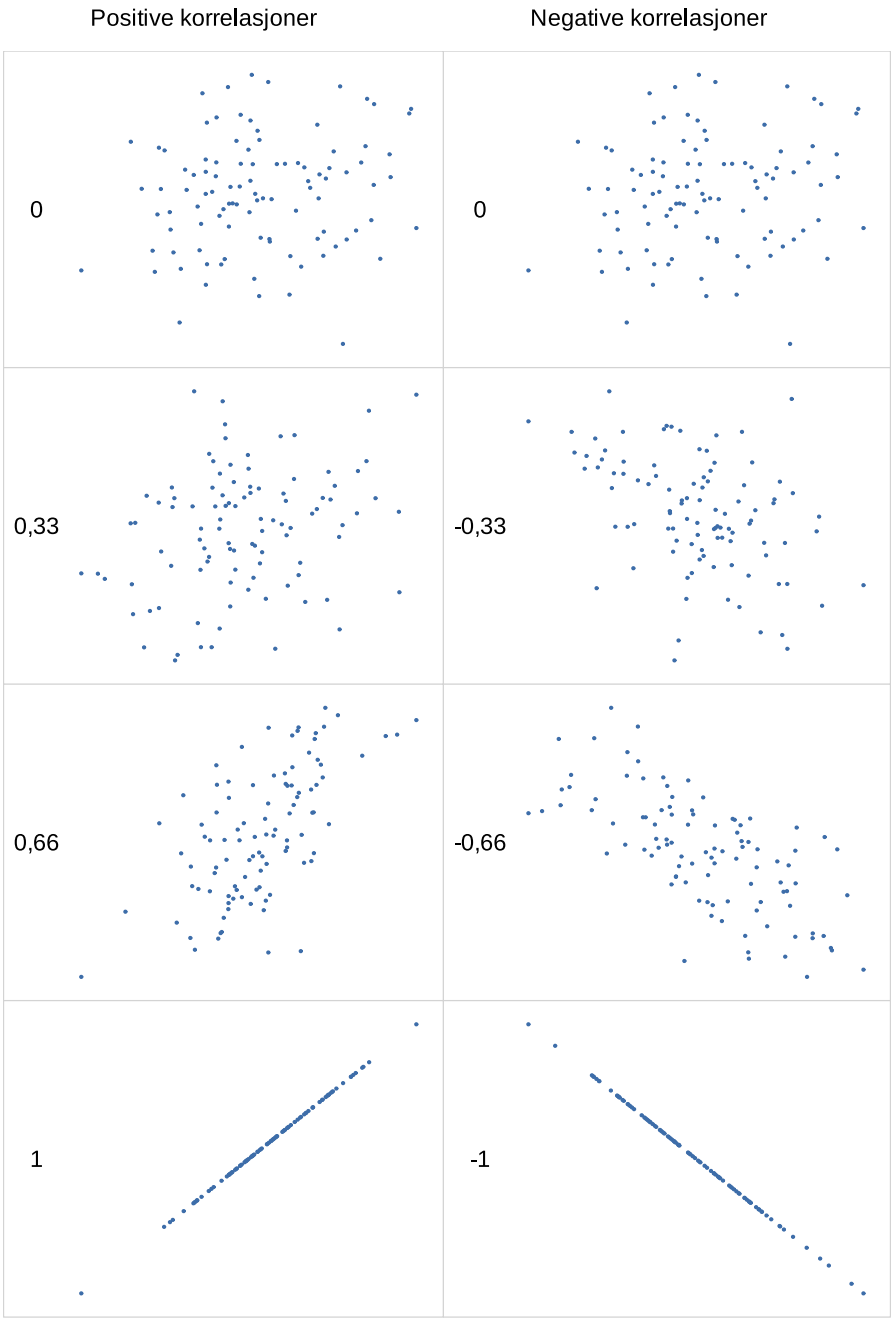
Spredningsplottet antyder en sammenheng ved at punktskyen trekker svakt oppover mot høyre, men det er vanskelig å si *hvor sterk* sammenhengen er. Er den sterk, moderat eller svak? For å avgjøre det trenger vi et tall som oppsummerer styrken. Det tallet er korrelasjonskoeffisienten.

6.3.2 Korrelasjonskoeffisienten

Korrelasjonskoeffisienten kalles mer presist Pearsons korrelasjonskoeffisient og skrives som r. Det er et tall mellom -1 og 1 som måler styrken og retningen på sammenhengen mellom to variabler, X og Y.

- En $r = 1$ betyr en perfekt positiv lineær sammenheng.
- En $r = -1$ betyr en perfekt negativ lineær sammenheng.
- En $r = 0$ betyr at det ikke er noen lineær sammenheng i det hele tatt.

Hvordan ser ulike korrelasjoner ut? Figur 6.8 viser eksempler. Disse punktskyene er ikke hentet fra virkelige observasjoner, de er *simulerte* slik at vi kan vise nøyaktige verdier av r i pene spredningsplott.



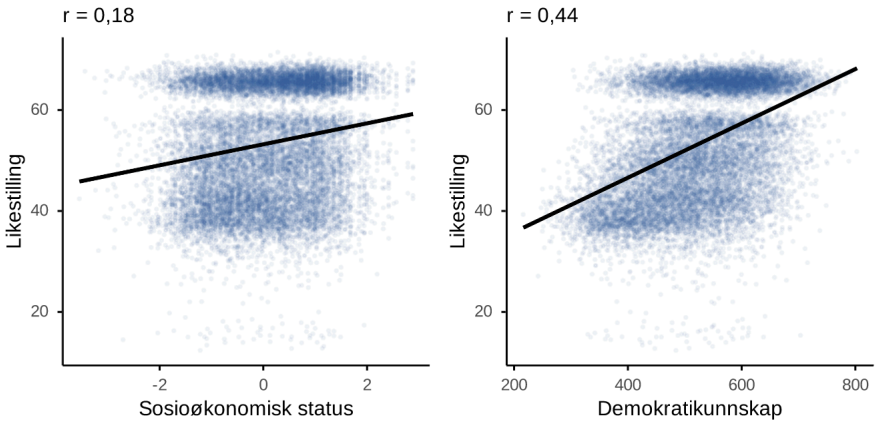
Figur 6.8: Illustrasjon av effekten av å variere styrken og retningen på en korrelasjon. I venstre kolonne er korrelasjonene $0; 0,33; 0,66$ og 1. I høyre kolonne er korrelasjonene $0; -0,33; -0,66$ og -1.

Korrelasjonen måler altså i hvor stor grad punktene faller på en rett linje. Jo nærmere r ligger ± 1, jo tettere samler punktene seg rundt linja. Jo nærmere r er 0, jo mer ser plottet ut som en sky uten retning. Merk at r ikke viser hvor bratt linja er, den viser bare hvor tett punktene ligger rundt linja.

6.3.3 Styrke og retning i ekte data

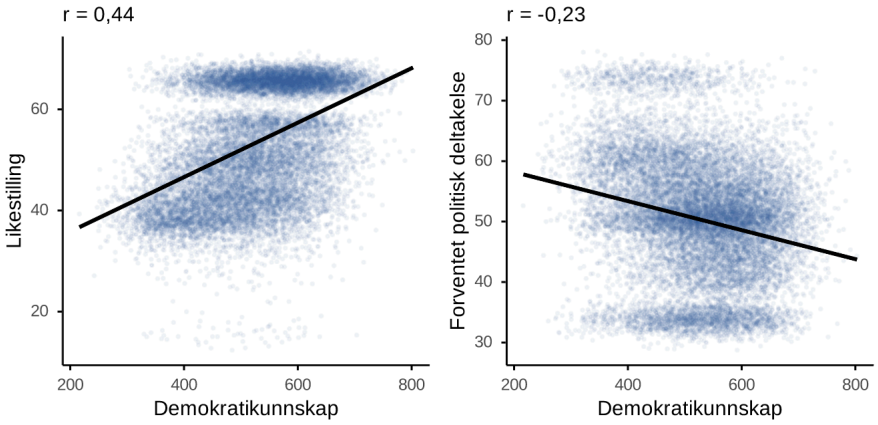
Stiliserte plott er rene og tydelige, men ekte data ser sjelden så pene ut. Nå vender vi tilbake til ICCS og bruker korrelasjonskoeffisienten r til å sammenligne ekte korrelasjoner. I figurene som følger, har vi regnet ut den rette linja som best viser sammenhengen mellom variablene fordi det gjør det lettere å se hellingen i en rotete punktsky. For å vise ulike sammenhenger henter vi inn enda en variabel fra ICCS, nemlig *Sosioøkonomisk status* som er en samlevariabel for utdanningen, inntekten og yrket til elevenes foresatte.

Styrken ser vi best ved å sammenligne to plott der retningen er den samme. I Figur 6.9 står *Sosioøkonomisk status* × *Likestilling* ($r = 0,18$) til venstre og *Demokratikunnskap* × *Likestilling* ($r = 0,44$) til høyre. Begge er positive, men korrelasjonen til høyre er omtrent dobbelt så sterk. Det vises ved at punktene samler seg tettere rundt linja, ikke ved at linja er brattere. Sammenligner du med de stiliserte plottene i Figur 6.8, ser du raskt at de ekte dataene ser helt annerledes ut og at det ikke er enkelt å se i hvilket spredningsplott sammenhengen mellom variablene er sterkest. Slik er det med ekte data, så da er det godt å ha et tall som kan måle det for oss.



Figur 6.9: Sammenligning av to positive sammenhenger med ulik styrke. Til venstre *Sosioøkonomisk status* × *Likestilling*; til høyre *Demokratikunnskap* × *Likestilling*. Over hvert plott vises korrelasjonskoeffisienten.

Retningen ser vi best ved å bytte ut én av variablene. I Figur 6.10 beholder vi *Demokratikunnskap* på x-aksen, men sammenligner *Likestilling* ($r = 0,44$) til venstre med *Forventet politisk deltakelse* ($r = -0,23$) til høyre. Det viktige er fortegnet. Til venstre heller linja oppover (positiv sammenheng), til høyre heller den nedover (negativ sammenheng). Den negative korrelasjonen er overraskende, for den betyr jo at elever som kan mer om demokrati forventer å delta *mindre* politisk! Det er et resultat vi kommer tilbake til i Seksjon 6.4.4.



Figur 6.10: Sammenligning av en positiv og en negativ sammenheng, med *Demokratikunnskap* på x-aksen i begge. Til venstre *Likestilling* på y-aksen; til høyre *Forventet politisk deltakelse*. Over hvert plott vises korrelasjonskoeffisienten.

6.3.4 Korrelasjonsmatrise

Vi har nå sett på tre par av variabler og lest av r for hvert par. Når vi vil sammenligne korrelasjoner for mange variabler samtidig, er det vanlig å samle alle korrelasjonene i én tabell, en **korrelasjonsmatrise**. Tabell 6.5 viser matrisen for de fire ICCS-variablene vi har jobbet med: *Demokratikunnskap*, *Likestilling*, *Sosioøkonomisk status* og *Forventet politisk deltakelse*.

> I jamovi: lage en korrelasjonsmatrise
>
> Regression → Correlation Matrix. Legg de kontinuerlige variablene du vil korrelere i variabelboksen. Pearsons r er valgt som standard.

Tabell 6.5: Korrelasjonsmatrise for fire kontinuerlige ICCS-variabler. Hver celle er Pearsons r mellom variabelen i raden og variabelen i kolonnen.

	Demokratikunnskap	**Likestilling**	**Sosioøkonomisk status**	**Forventet politisk deltakelse**
Demokratikunnskap	1,00			
Likestilling	0,44	1,00		
Sosioøkonomisk status	0,39	0,18	1,00	
Forventet politisk deltakelse	−0,23	−0,14	−0,03	1,00

Hver celle er korrelasjonen mellom variabelen i raden og variabelen i kolonnen. Diagonalen er $1,00$ fordi en variabel henger perfekt sammen med seg selv, og den øvre halvdelen er tom fordi den ville gjentatt den nedre: korrelasjonen mellom *Likestilling* og *Demokratikunnskap* er den samme som mellom *Demokratikunnskap* og *Likestilling*.

Matrisen bekrefter mønsteret fra Figur 6.9 og Figur 6.10 og legger til to par vi ikke har sett. *Demokratikunnskap* henger positivt sammen med både *Likestilling* ($r = 0,44$) og *Sosioøkonomisk status* ($r = 0,39$); det er de to sterkeste sammenhengene i tabellen. *Forventet politisk deltakelse* skiller seg ut ved å henge negativt sammen med alle de tre andre variablene, sterkest med *Demokratikunnskap* ($r = -0,23$) og svakest med *Sosioøkonomisk status* ($r = -0,03$, som er så nær null at sammenhengen i praksis er fraværende).

6.4 Tolkning av deskriptiv statistikk

Vi har i disse to kapitlene sett mange måter å beskrive data på, som grovt sett kan deles opp i oppsummerende tall og diagrammer. Resten av kapittelet handler om noen forhold du bør være bevisst på når du leser eller lager slike tall og diagrammer.

6.4.1 Hvordan tolke størrelsen på oppsummerende tall?

De fleste oppsummerende tall sier lite i seg selv. Tenk for eksempel over at 53,1 % av norske elever i ICCS så for seg å ta høyere utdanning. Er dette tallet høyt eller lite? For å besvare dette må man tolke tallet i kontekst, for eksempel ved å sammenlikne med naboland eller opp mot en eller annen politisk målsetting om hvor mange vi ønsker at skal ta høyere utdanning.

Størrelsen på korrelasjoner kan i hvert fall være vanskelige å tolke. I det virkelige liv ser man sjelden korrelasjoner på 1. Hvordan skal man tolke en korrelasjon på, si, $r = 0,4$? Det ærlige svaret er at det avhenger av hva du skal bruke dataene til, og hvor sterke sammenhengene i fagfeltet ditt pleier å være. Ser man for eksempel etter korrelasjoner mellom undervisning og læring, gjør man det svært bra hvis man får en korrelasjon på $0,3$, fordi læring påvirkes av så mye annet enn undervisningen at slike sammenhenger er vanskelige å påvise. Men korrelerer man hvor godt en elev gjør det på den samme rettskrivningprøven på mandag og tirsdag, bør korrelasjonen være svært høy. Tolkningen avhenger i stor grad av konteksten.

Egenskaper ved målingen er en del av denne konteksten. Hvis variablene er lite reliable, blir sammenhengen mellom dem svakere enn den «egentlig» er. Det er en av grunnene til at korrelasjoner i utdanningsforskning sjelden er store; se Kapittel 2. for mer om hva som påvirker målekvalitet.

6.4.2 Visualiser alltid

Jeg kan ikke få sagt det nok: Visualiser dataene før du tolker oppsummerende tall. Grunnen er enkel: Tallene betyr ikke alltid det du tror.

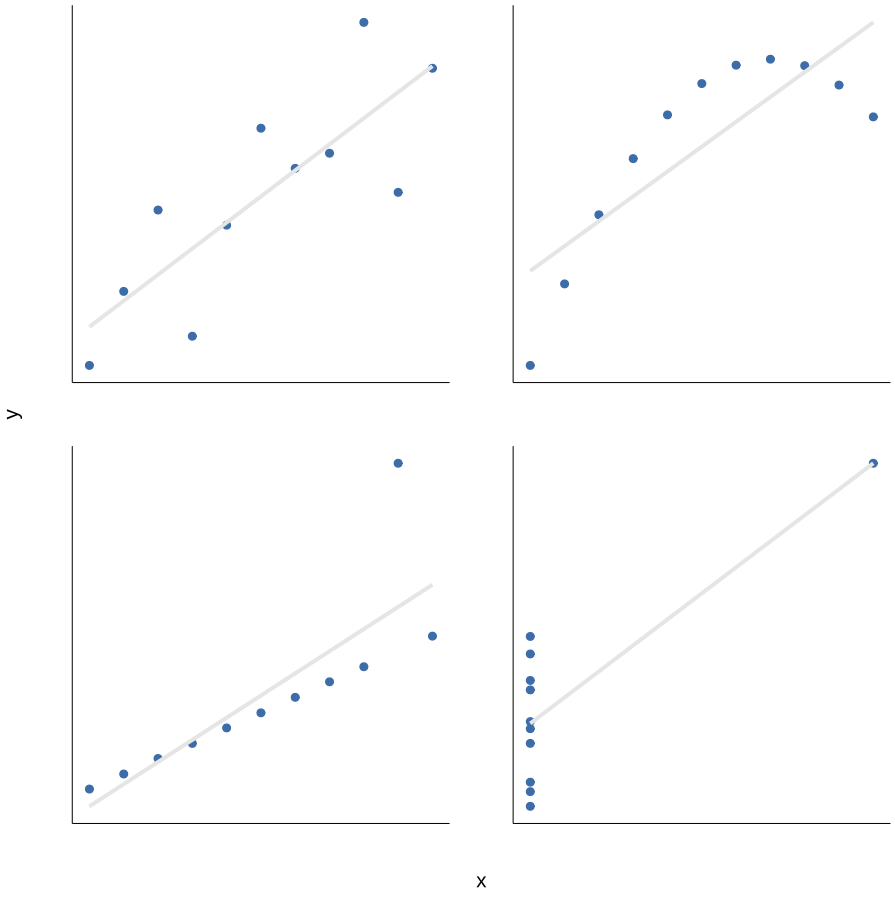
Et klassisk eksempel som viser dette perfekt, er «Anscombes kvartett» (Anscombe, 1973). Kvartetten består av fire datasett, og hvert av dem har en X- og en Y-variabel. Kvartetten er laget slik at de deskriptive statistikkene er så å si identiske for alle fire:

- Gjennomsnittet for alle X-variablene er $9,0$ og for alle Y-variablene er $7,5$.
- Standardavvikene er også praktisk talt like for både X og Y.
- Og for å toppe det hele, er korrelasjonen mellom X og Y i *alle* de fire tilfellene $r = 0,816$.

Dette kan du forresten enkelt sjekke selv, for dataene ligger klare i filen *anscombe.csv*.

Basert på kun disse tallene skulle man tro at datasettene er nokså like, men det er de absolutt ikke. Først når vi visualiserer dataene som spredningsplott, slik du ser i Figur 6.11, avsløres sannheten: De fire datasettene er helt ulike. Da skjønner vi hvorfor mange statistikere har følgende leveregel, som mange dessverre glemmer i praksis:

Visualiser alltid dataene dine!



Figur 6.11: Spredningsplott for Anscombes kvartett. Alle fire datasettene har en Pearson-korrelasjon på $r = 0,816$, men de er likevel svært forskjellige fra hverandre.

6.4.3 Gruppestatistikk er ikke individstatistikk

Når vi rapporterer et gjennomsnitt for en gruppe, er det fristende å lese tallet som om det beskriver den typiske personen i gruppa. Men gjennomsnittet

er en oppsummering av mange ulike individer, ikke en beskrivelse av noen enkelt av dem. Å forveksle de to nivåene kalles en **økologisk feilslutning**.[16]

Tenk på en setning som «Norge skårer rundt 490 på PISA i lesing». Det betyr ikke at «den typiske norske 15-åringen skårer rundt 490». Det betyr at *gjennomsnittet* over alle norske 15-åringer er omtrent 490, mens enkeltelevene fordeler seg over et stort spenn rundt det tallet. Standardavviket er nettopp et mål på hvor lite «typisk» gjennomsnittet er for et tilfeldig enkeltindivid.

Samme logikk gjelder mange steder i utdanningssystemet. Klassegjennomsnittet på en prøve forteller deg ikke karakteren hvert individ fikk. Skolegjennomsnittet forteller deg ikke hvilken klasse som drar snittet opp eller ned. Det er greit å sammenligne grupper med grupper, men ikke å hente ut en påstand om enkeltindivider fra et grupperesultat.

Vi har sett dette i ICCS-tallene allerede. Det landsgjennomsnittet på variabelen *Likestilling* skjuler at jenter og gutter svarer ganske forskjellig (Tabell 6.4). Stopper man ved landgjennomsnittet, mister man at norske elever er en heterogen gruppe.

Og noen ganger blir det enda verre, for sammenhengen mellom to variabler kan endre seg, eller til og med snu, når vi undersøker undergrupper. Det skal vi se på nå.

6.4.4 Sammenhenger kan endre seg i undergrupper

I Seksjon 6.3.3 så vi at korrelasjonen mellom *Demokratikunnskap* og *Forventet politisk deltakelse* var negativ ($r = -0,23$): elever som vet mer om demokrati, forventer å delta *mindre* politisk. Hmm, snodig. Men se hva som skjer hvis vi regner korrelasjonen separat for hvert land, Tabell 6.6.

Tabell 6.6: Korrelasjon mellom *Demokratikunnskap* og *Forventet politisk deltakelse* totalt og innenfor hvert land.

Land	N	r
Alle land	16 473	−0,23
Norge	4 640	−0,10
Spania	3 011	−0,13
Polen	4 218	−0,11
Brasil	4 604	−0,31

Korrelasjonen er fortsatt svakt negativ, men den er mye svakere innenfor hvert land enn det er totalt. Hvordan kan det henge sammen?

Forklaringen ligger i Brasil. Brasilianske elever har lavest gjennomsnittlig *Demokratikunnskap* av de fire landene, men høyest *Forventet politisk deltakelse*. Når vi regner korrelasjonen med alle elevene sammen, drar denne forskjellen *mellom* land korrelasjonen tydelig nedover. Sammenhengen *in-*

[16] Ingen gir dårligere navn enn statistikere. Hvorfor kalle det en *økologisk* feilslutning når det ikke har noe med økologi å gjøre? Hvor er økosystemene, liksom? Jeg velger å tenke at informasjon om økologien (økosystemet) ikke gir presis informasjon om individene, for da passer navnet litt.

nenfor hvert land (altså om en elev som vet mer enn klassekameraten sin også tenker hun skal delta mer politisk) er langt svakere.

Dette er et eksempel på et mer generelt poeng: en samlet korrelasjon kan være drevet av forskjeller mellom undergrupper, ikke av sammenhengen innenfor undergruppene. I ekstreme tilfeller snur korrelasjonen fortegn når man studerer undergruppene hver for seg; det kalles **Simpson-paradokset**, og du så et eksempel på dette allerede i Kapittel 1. med opptaksdata fra Berkeley-universitetet.

6.4.5 Sammenheng er ikke kausalitet

Jeg må minne om lærdommen fra Seksjon 3.4, nemlig at korrelasjon ikke er kausalitet. Selv om variabelen *Demokratikunnskap* hang positivt sammen med *Likestilling* ($r = 0,44$), betyr det ikke at vi kan endre elevers syn på likestilling ved å heve demokratikunnskapene deres. Faktisk bør du kunne navngi grunner til at vi ikke kan trekke denne kausale slutningen.

6.4.6 Teorien velger hva vi ser etter

Det er ikke dataene som forteller oss hvilke par av variabler det er interessant å sammenligne, det gjør forskningsspørsmålet vårt og teorien vår. Med teori mener vi her forestillingene våre om hvilke begreper som kan henge sammen og hvorfor; den gir føringer for hva vi ser etter og hvordan vi analyserer dataene. Vi regnet korrelasjonen mellom *Forventet utdanning* og *Land* fordi vi hadde en forestilling om at landenes skoletradisjoner påvirket om elevene ser for seg å ta høyere utdanning, ikke fordi datasettet ba oss om det. Tester man derimot «alt mot alt» i et stort datasett, finner man nesten alltid sammenhenger ved ren tilfeldighet, og slike funn betyr lite. Selv om noe finnes i dataene, betyr ikke det at det er sant; du må ikke lytte utelukkende til dataene, slik vi advarte i Kapittel 1., for forskningsspørsmål og teori spiller en avgjørende rolle.

6.5 Oppsummering

Deskriptiv statistikk kan fortelle deg om enkeltvariabler i dataene (Kapittel 5.), men også om sammenhenger mellom variabler (dette kapittelet). I dette kapittelet har vi sett at man undersøker sammenhenger forskjellig avhengig av variabeltypen:

- **Kategorisk × kategorisk**: Bruk krysstabeller eller stablede stolpediagram.
- **Kategorisk × kontinuerlig**: Bruk sentraltendens og spredningsmål per gruppe, eller boksplott og histogram per gruppe for å vise forskjellene visuelt.
- **Kontinuerlig × kontinuerlig**: Bruk spredningsplott for å visualisere sammenhengen, og korrelasjonskoeffisienten for å beskrive styrken og retningen med et tall.

Tallene og diagrammene er bare oppsummeringer. Tolkning av deskriptiv statistikk minner deg om noen forhold som er lette å glemme: Tolkningen avhenger av konteksten. Du må alltid visualisere dataene, et gruppegjennom-

snitt sier lite om enkeltindivider, en sammenheng mellom variabler kan endre seg i undergrupper, en sammenheng er ikke det samme som en årsakssammenheng, og det er forskningsspørsmålet og teorien (ikke dataene) som bør styre hvilke variabler du undersøker.

III Inferensiell statistikk

7. Usikkerhet og konfidensintervall

[God] has afforded us only the twilight … of Probability.
– John Locke

I de foregående kapitlene har vi beskrevet dataene våre ved å regne ut gjennomsnitt og andre oppsummerende tall og laget figurer. Men forskere ønsker nesten alltid å si noe som går *utover* dataene vi faktisk har samlet inn. Da trenger vi **inferensiell statistikk**, og det møter du faktisk oftere enn du tror.

Samme dag som jeg skriver dette, våren 2025, er hovedoppslagene i media en ny meningsmåling der 28% av 1000 respondenter svarer at, om det hadde vært valg i dag, så hadde de stemt på Arbeiderpartiet. Dette viser at Arbeiderpartiet ligger an til å gjøre et godt valg, fortelles det. Ser du at dette går forbi deskriptiv statistikk? At 280 av 1000 personer svarte at de ville stemme Arbeiderpartiet betyr ikke at de vil gjøre et godt valg, for Arbeiderpartiet trenger mange flere stemmer enn 280 for å gjøre det godt. Media antar at disse 1000 personene sier noe om de 4 350 000 andre stemmeberettigede i Norge også. Dette krever inferensiell statistikk: Ved å spørre 0,02% av populasjonen kan man finne ut noe om de resterende 99,98%.

Dette kapittelet handler kun om et spørsmål fra inferensiell statistikk: Hvor godt kan man estimere gjennomsnittsverdien i en populasjon fra et utvalg? Å *estimere* betyr å anslå størrelsen på noe. Når vi estimerer gjennomsnittet i populasjonen, forsøker vi altså å finne et tall på hvor stort gjennomsnittet er. Problemet med å estimere gjennomsnittet i populasjonen er at estimatet ville blitt annerledes med et annet utvalg.

Målet med kapittelet er å skjønne begrepet **standardfeil**. Deretter er det forholdsvis enkelt å skjønne hva et **konfidensintervall** er. Med disse begrepene vil du kunne forstå flere resultater i forskningsartikler, samt skjønne magien[17] statistikere støtter seg til når de uttaler seg sikkert om populasjonen uten å ha samlet noe særlig data fra den. Det vil også gjøre det lettere for deg å forstå statistisk hypotesetesting, noe de fleste innføringsbøker i kvantitativ metode bruker mye plass på, men som jeg ikke ser meg tjent med i denne sammenhengen.

[17] Vel, matematikken, men siden matematikk har et image-problem velger jeg å benytte ordet *magi*.

Men jeg ønsker å lære om statistisk hypotesetesting!

Statistisk hypotesetesting er en metode som bruker data og beregninger til å vurdere om en antakelse om en populasjon kan forkastes. Et eksempel på en slik antakelse er at «gutter og jenter har lik lesehastiget i populasjonen». Noen lærerutdannere (og kanskje noen studenter) ønsker gjerne å dekke hypotesetesting i innføringskurs i kvantitative metoder. Argumentene *for* å lære om hypotesetesting er at det opptrer svært ofte i forskningslitteraturen og at noen sensorer kanskje vil forvente hypotesetesting i en kvantitativ masteroppgave. Argumentene *mot* er svært mye bedre.

- statistisk hypotesetesting passer til *svært* få forskningsspørsmål
- statistisk hypotesetesting blir svært ofte misforstått av studenter, forskere (Lytsy et al., 2022) og lærebokforfattere i kvantitativ metode (Cassidy et al., 2019)
- statistisk hypotesetesting innbyr til tanketom bruk (Gigerenzer, 2018)
- å lære om andre ting (utvalgsfordeling og standardfeil) gjør deg bedre rustet til å lære annen statistikk

Statistisk hypotesetesting er så ofte misbrukt at foreningen for amerikanske statistikere advarer mot å bruke det (Wasserstein & Lazar, 2016).

7.1 Fire typer usikkerhet

Når en forsker rapporterer et resultat, er det alltid beheftet med usikkerhet. Det er sikkert en million forskjellige kilder til usikkerhet, og for oss vil det være nyttig å skille mellom minst fire:

- **Måleusikkerhet.** Måleinstrumentene våre er aldri perfekte. En lesehastighetsprøve fanger ikke opp lesehastighet helt presist; dagsform, konsentrasjon og tilfeldigheter ved akkurat den teksten eleven leser, spiller inn. Vi drøftet dette i Kapittel 2..
- **Modellusikkerhet.** Når vi analyserer data, gjør vi en rekke antakelser: at utvalget er tilfeldig, at en effekt er den samme for alle, og så videre. Hvis antakelsene ikke stemmer overens med virkeligheten, er også konklusjonene usikre.
- **Utvalgsusikkerhet.** Vi har bare et utvalg, ikke hele populasjonen, og et annet utvalg ville gitt et annet estimat.
- **Tilordningsusikkerhet.** I eksperimenter fordeles deltakerne tilfeldig på en intervensjons- og en kontrollgruppe. Da vil den tilfeldige tilordningen være en kilde til usikkerhet: En annen tilfeldig inndeling av de *samme* deltakerne ville gitt et litt annet resultat. Dette ligner utvalgsusikkerheten, men oppstår under prosessen med gruppetilordning, ikke utvalg.

Vi skal lære inferensiell statistikk som tar hensyn til de to siste usikkerhetene, utvalgs- og tilordningsusikkerhet. Men begrepene du lærer, standardfeil og konfidensintervall, gjelder ikke bare dem; de vil være grunnlag for å lære mye annen statistikk også. Måle- og modellusikkerheten ser vi bort fra, slik det dessverre ofte gjøres i utdanningsforskning. Selvsagt er det usikkerhet i målingene og selvsagt er det usikkerhet i de statistiske modellene, men det

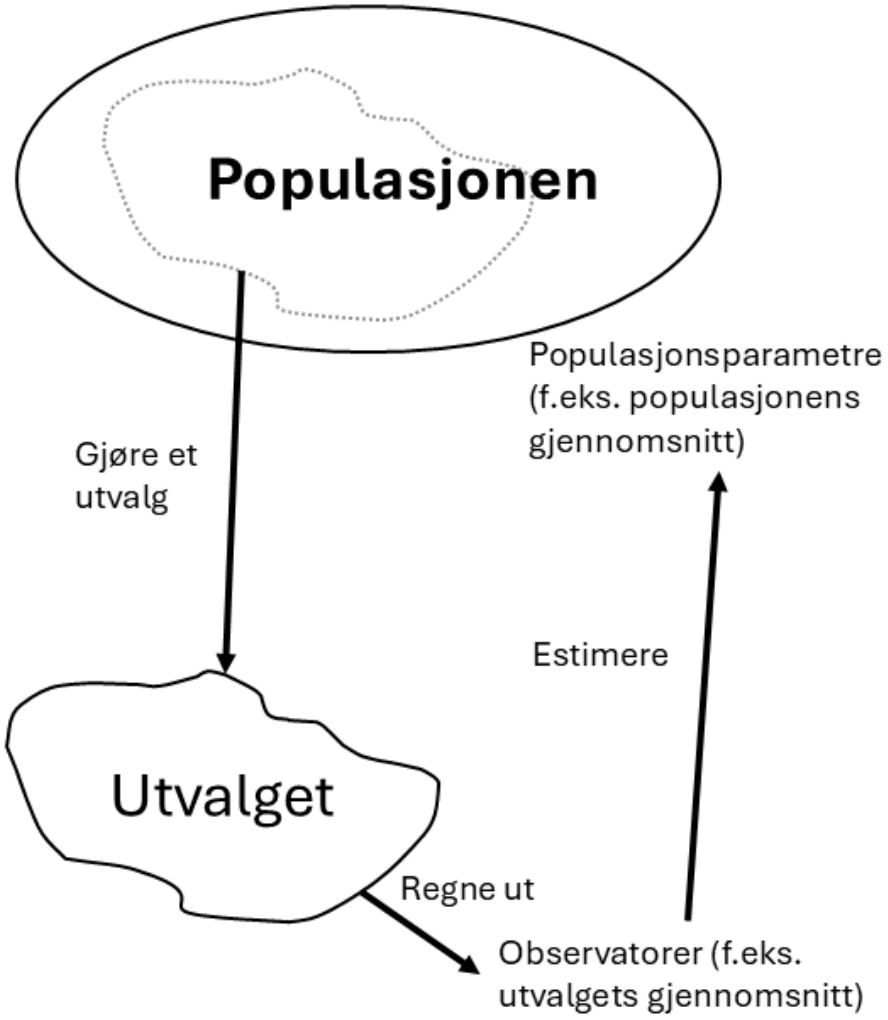
tas for sjeldent hensyn til disse usikkerhetene. Grunnen er neppe at forskerne ikke vet om dem, men at det ikke er del av vanlig statistisk praksis, og at det krever mer faglig skjønn enn en formel kan gi.

7.2 Populasjonsparametre og observatorer

Gjennom hele kapittelet skal vi bruke estimering av lesehastighet som eksempel. Konstruktet er altså lesehastighet, og vi skal måle den i antall ord per minutt. Lesehastighet måles i de svært mye brukte Carlstenprøvene, som finnes i varianter fra 1. til 10. trinn. Carlstenprøven benyttes i nesten 80% av norske barneskoler (Arnesen et al., 2019). Jeg husker selv at jeg tok denne prøven på barneskolen. Hele klassen ble bedt om å lese i et lesehefte, læreren sa fra når vi skulle stoppe å lese, vi markerte hvor langt vi kom med en strek, og så svarte vi på spørsmål om teksten så langt vi var kommet. I likhet med nesten alle kartleggingsprøvene som benyttes i norsk skole, har Carlstenprøven ingen informasjon om validitet og reliabilitet, og ingen studier har undersøkt dette (Arnesen et al., 2019).

Vi tenker oss en forskergruppe som vil estimere den gjennomsnittlige lesehastigheten for alle norske 4.-klassinger.[18] Prosessen vi nå skal forstå, er vist i Figur 7.1, som er en presisering av Figur 3.3.

[18] Merk at dette er en *forskningsbruk* av Carlstenprøven. I klasserommet bruker læreren prøven som en kartlegging av sine egne elever, ikke for å trekke slutninger om en populasjon.



Figur 7.1: Logikken bak estimering av populasjonsparametre

Populasjonen er hele gruppa forskerne vil si noe om, altså alle norske 4.-klassinger. For å beskrive den bruker statistikere en **fordeling**: en kurve som viser hvilke lesehastigheter som er vanlige blant disse elevene, og hvilke som er sjeldne. Egenskapene ved denne fordelingen kalles **populasjonsparametre**. De to viktigste er populasjonsgjennomsnittet og populasjonsstandardavviket. Disse kjenner forskerne ikke, og det er nettopp dem de vil finne ut av.

Siden forskerne ikke kan teste alle de rundt 60 000 fjerdeklassingene, trekker de et utvalg og gir dem Carlstenprøven. Da får de mange størrelser de *kan* regne ut, slik som gjennomsnittet og standardavviket i utvalget sitt. Slike størrelser, som beregnes fra dataene, kalles **observatorer**. Observatorene ligner gjerne på populasjonsparametrene, men de er ikke de samme. Vi benytter altså observatorene til å anslå populasjonsparametrene.

Populasjonsparametre beskriver *populasjonen*. Observatorer beskriver *utvalget*.

Det vi skal lære nå, er den siste pila i Figur 7.1, hvordan man estimerer populasjonsparametre. Vi skal også lære hvordan man estimerer utvalgsusikkerheten. (Og for ordens skyld: Legg merke til at figuren ikke nevner måleusikkerhet eller andre former for usikkerhet. Livet som utdanningsforsker blir mer komfortabelt hvis man bare lukker øyene for slikt.)

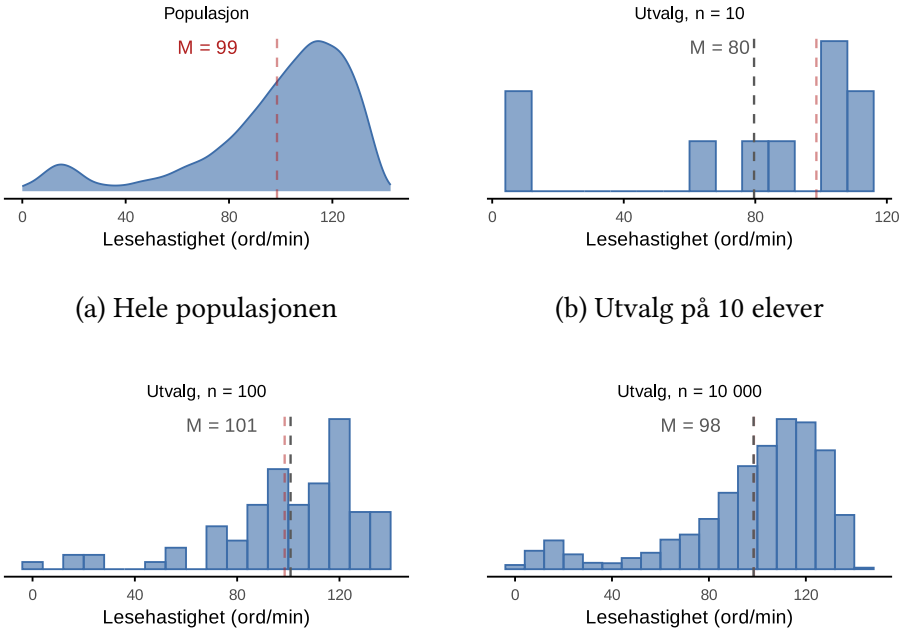
7.3 Populasjonsfordelingen og utvalg trukket fra den

Til vanlig har man ikke kunnskap om hele populasjonen, men nå, for å forklare estimering av populasjonsparametre, skal vi ha det. Det beste hadde vært om vi hadde et datasett der Carlstenprøven var gitt til hele populasjonen av norske fjerdeklassinger. Da hadde vi kunnet sett hvor godt populasjonsparameteren ble estimert fra et utvalg. Det nest beste hadde vært å simulere et datasett som likner, så det er det jeg har gjort. Dessverre er jeg ikke lesedidaktiker, men jeg har gjort mitt beste for å lage et datasett som virker ekte.

Populasjonsfordelingen til det simulerte datasettet vises i panel (a) i Figur 7.2. Legg merke til at fordelingen har to humper. Helt til venstre, nær 0–20 ord/min, er det en liten hump av elever som ennå ikke har knekt lesekoden. Hovedtyngden av elevene leser raskere, med en topp litt over 100 ord/min, men denne humpen er skjev: den har en lengre hale av elever med lavere lesehastighet.[19]

For å få gode estimater må utvalget være representativt for populasjonen, og den enkleste måten å få til representative utvalg på er ved tilfeldige utvalg. Se Seksjon 3.6.2 hvis du trenger å friske opp enkle tilfeldige utvalg. Figur 7.2 (b), (c) og (d) viser tre enkle tilfeldige utvalg med forskjellig størrelse. Alle tre har omtrent samme form som populasjonen, men de store utvalgene likner mer på populasjonen enn de små. Gjennomsnittene for utvalgene (observatorene) er vist som en grå vertikal strek og med `M =`. Populasjonsgjennomsnittet er vist i rødt og er 99. Legg merke til at utvalgenes gjennomsnitt blir bedre og bedre estimater på populasjonsgjennomsnittet med større utvalg. For det minste utvalget ($n = 10$) er gjennomsnittet langt unna populasjonsgjennomsnittet, det mellomste ($n = 100$) er nære, og det største utvalget ($n = 10000$) er nesten eksakt.

[19] Et poeng som er verdt å huske fra Kapittel 5.: Siden fordelingen er skjev, er ikke populasjonsgjennomsnittet og medianen like. Gjennomsnittet er på 99 ord/min, og er lavere enn populasjonsmedianen, som er på 106 ord/min, som igjen er lavere enn den vanligste lesehastigheten, som er rundt 110 ord/min. Gjennomsnittet er altså noe vi *velger* å regne ut og estimere, ikke nødvendigvis et godt sammendrag av en typisk elev, jamfør den økologiske feilslutningen, Seksjon 6.4.3.



(a) Hele populasjonen

(b) Utvalg på 10 elever

(c) Utvalg på 100 elever

(d) Utvalg på 10 000 elever

Figur 7.2: Lesehastighet på 4. trinn i hele populasjonen og i tre tilfeldige utvalg trukket fra den. Den røde stiplede streken er populasjonsgjennomsnittet; i utvalgspanelene viser den grå streken utvalgets eget gjennomsnitt.

Hvis vi hadde vært lite ambisiøse kunne kapittelet sluttet her. I så fall hadde du lært følgende:

Hvis vi gjør et enkelt tilfeldig utvalg fra en populasjon vil

- utvalgsfordelingen tilnærme seg populasjonsfordelingen med større utvalg
- observatorene tilnærme seg populasjonsparameteren med større utvalg

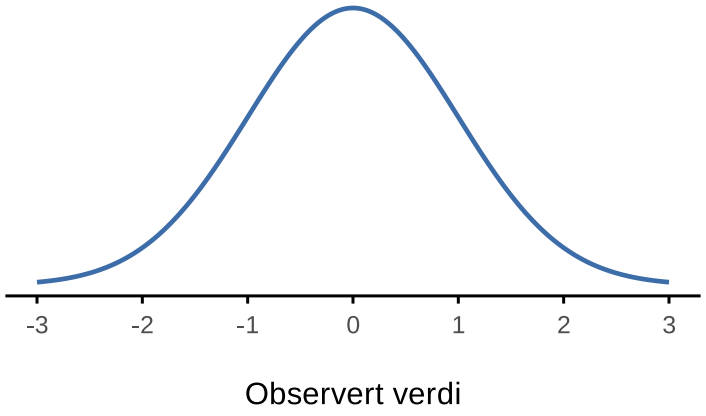
Men vi skal ikke slutte her, for i virkeligheten er det ikke nok å vite at vi kan estimere populasjonsparameteren bedre hvis vi gjør et større utvalg; vi må vite hvor godt vi har estimert populasjonsparameteren med det utvalget vi har. Hvis vi bare har råd til å gjøre Carlstenprøven på 100 elever, hvor godt får vi da estimert gjennomsnittet?

Målet nå er å sette et tall på hvor usikre estimatene er, men vi trenger litt teori først. Et naturlig sted å begynne er normalfordelingen.

7.4 Normalfordelingen

Normalfordelingen er berømt, så du har kanskje hørt om den allerede. Du skal ikke lære så mye om den, annet enn at den har bjelleform og er fullt ut beskrevet av sitt gjennomsnitt og standardavvik.

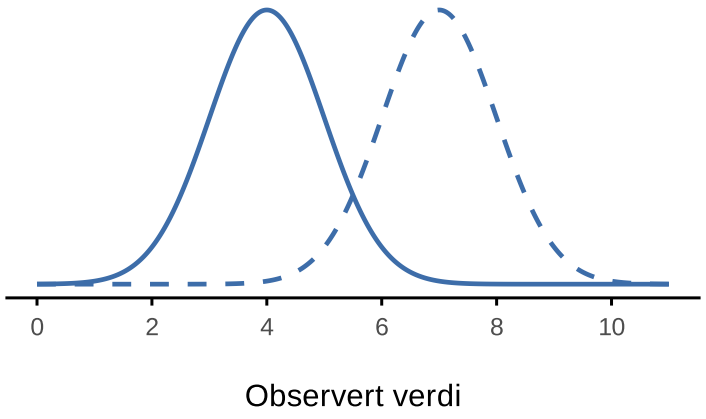
Figur 7.3 viser en normalfordeling med gjennomsnitt 0 og standardavvik 1. Denne normalfordelingen kalles en standard normalfordeling. På den vannrette aksen leser vi av verdien til en eller annen variabel, og høyden på kurven forteller hvilke verdier av variabelen som er sannsynlige.



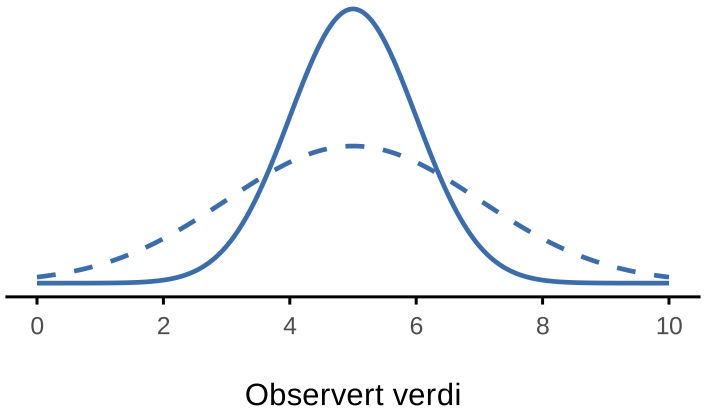
Figur 7.3: Normalfordelingen med gjennomsnitt 0 og standardavvik 1. På x-aksen finner man verdien til en variabel, og høyden på kurven forteller hvor sannsynlig den verdien er

Normalfordelinger kan ha andre gjennomsnitt og standardavvik enn det som ble vist over, så nå skal vi forandre normalfordelingens gjennomsnitt og standardavvik for å se hvordan kurven forandrer seg. Jeg tenker på gjennomsnittet og standardavviket som to knotter vi kan skru på for å endre fordelingen. Endrer vi gjennomsnittet, *flytter* hele kurven til venstre eller høyre uten å endre formen. Endrer vi standardavviket, blir kurven bredere eller smalere, men den blir værende på samme sted. En bred kurve betyr stor spredning; en smal kurve betyr at verdiene ligger tett rundt gjennomsnittet.

Figur 7.4 viser hva som skjer når vi endrer gjennomsnittet, og Figur 7.5 hva som skjer når vi endrer standardavviket.



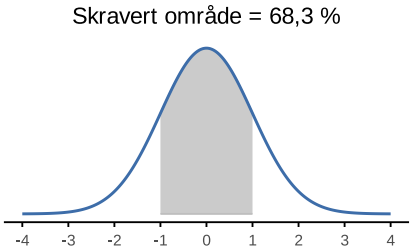
Figur 7.4: Hva som skjer når vi endrer gjennomsnittet. Den heltrukne linjen har gjennomsnitt 4, den stiplede 7. Begge har standardavvik 1. Formen er den samme; kurven er bare forskjøvet



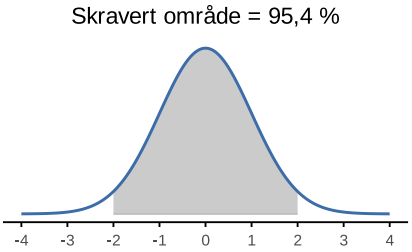
Figur 7.5: Hva som skjer når vi endrer standardavviket. Begge kurvene har gjennomsnitt 5. Den heltrukne har standardavvik 1, den stiplede 2. De er sentrert på samme sted, men den stiplede er bredere og lavere

Nå vet du hvordan alle mulige normalfordelinger ser ut; de kan være lenger til høyre og venstre eller være smalere og bredere.

For alle normalfordelingene gjelder en enkel tommelfingerregel, **68–95–99,7-regelen**. Uansett hvilket gjennomsnitt og standardavvik fordelingen har, vil rundt 68 % av observasjonene ligge innenfor ett standardavvik fra gjennomsnittet, rundt 95 % innenfor to standardavvik, og hele 99,7 % innenfor tre. Figur 7.6 illustrerer de to første tilfellene.



(a) 68,3 % av observasjonene faller innenfor ett standardavvik fra gjennomsnittet



(b) 95,4 % faller innenfor to standardavvik

Figur 7.6: Andelen av en normalfordeling som ligger innenfor ett og to standardavvik fra gjennomsnittet.

Kanskje lurer du på hvorfor du må lære dette om normalfordelingen. Lesehastigheten i populasjonen var jo *ikke* normalfordelt; den har blant annet to humper, slik vi nettopp så i Figur 7.2. Vi forventer ikke at fordelinger i den virkelige verden er normalfordelte, så hvorfor er normalfordelingen da så viktig for oss? Bare les videre.

7.5 Utvalgsfordelingen til gjennomsnittet

Vi skal nå forstå noe litt vanskelig, så ta det rolig, les sakte, og forsøk å henge med. Hvis du ikke helt forstår, kan det være fint å lese videre uansett, men deretter returnere til disse delkapitlene senere. Du skal lære at *utvalgets gjennomsnitt* har sin egen fordeling. Dette er rart. Vi har jo bare ett utvalg vi kan estimere populasjonsgjennomsnittet med, så vi får bare én verdi for utvalgets gjennomsnitt. Hvordan kan denne ene verdien ha en fordeling?

For å forstå dette må vi gjøre et tankeeksperiment: Vi må se for oss alle tenkelige utvalg vi kunne fått. Tenk deg at vi gjør studien vår med 5 elever, altså $n = 5$. I Tabell 7.1 vises 10 forskjellige tilfeldige utvalg vi kunne fått. Vi ser de fem elevene i utvalget har forskjellige lesehastigheter, og at dette gir forskjellige gjennomsnitt (kolonnen helt til høyre). Det er dette man mener med at utvalgets gjennomsnitt har en fordeling. Dette er bare 10 utvalg; det finnes enormt mange forskjellige slike utvalg man kunne trukket.

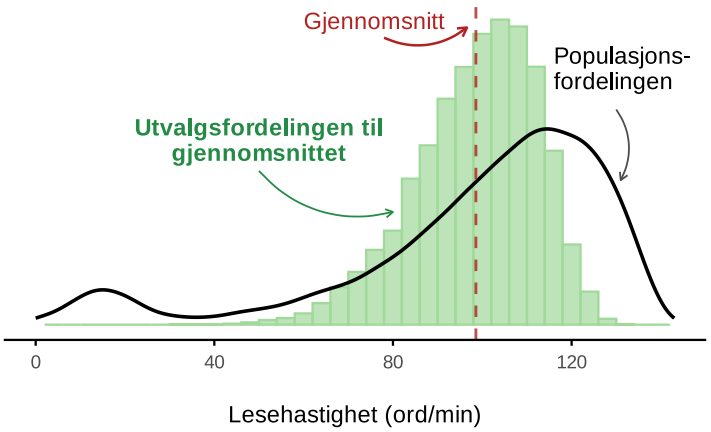
Tabell 7.1: Ti forskjellige utvalg av en liten lesehastighetsstudie, hver med et utvalg på $n = 5$ elever. Utvalgsgjennomsnittet varierer en god del fra utvalg til utvalg

	Elev 1	Elev 2	Elev 3	Elev 4	Elev 5	Utvalgets gjennomsnitt
Utvalg 1	102	99	112	87	45	89
Utvalg 2	110	79	97	103	96	97
Utvalg 3	127	110	134	17	100	98
Utvalg 4	93	94	125	103	122	107
Utvalg 5	115	92	120	50	96	95
Utvalg 6	119	124	87	106	66	100
Utvalg 7	49	109	86	112	134	98
Utvalg 8	110	135	11	102	100	92
Utvalg 9	100	60	30	92	91	75
Utvalg 10	94	130	115	111	113	113

Legg merke til hvordan gjennomsnittet varierer fra utvalg til utvalg. Lesehastighetene til hver enkelt elev varierer fra ca. 10 til 130, men gjennomsnittene varierer fra ca. 75 til 115. Gjennomsnittene varierer altså mindre enn lesehastighetene til enkeltelever. Hvis jeg hadde vist mange flere slike utvalg, med tilhørende gjennomsnitt, hadde vi fått mange forskjellige gjennomsnittsverdier. Da kunne jeg vist **utvalgsfordelingen til gjennomsnittet**, altså fordelingen til alle utvalgsgjennomsnitt man kan få. Denne fordelingen vil bestå av tallene i kolonnen lengst til høyre i Tabell 7.1, i hvert fall hvis kolonnen hadde vist alle mulige utvalg man kunne fått.

Men – hey! – vi har jo simulert et datasett med hele populasjonen, det som ble vist i Figur 7.2, så jeg *kan* faktisk vise dette. Nyt og beundre Figur 7.7, der

jeg har tegnet opp gjennomsnittene man kan få fra populasjonen når man trekker utvalg på $n = 5$ elever.



Figur 7.7: Utvalgsfordelingen til gjennomsnittet for studier med $n = 5$ elever, vist som et grønt histogram. Til sammenligning viser den svarte linjen populasjonsfordelingen av lesehastighet. Den røde stiplede streken er gjennomsnittet, som er det samme i begge fordelingene

Jeg har tegnet inn populasjonsfordelingen som en svart linje, slik at vi kan sammenlikne den med utvalgsfordelingen til gjennomsnittet. Legg merke til det følgende i figuren:

- Utvalgsfordelingen til gjennomsnittet er jevnere og ser mer normalformet ut enn populasjonsfordelingen. De to humpene i populasjonsfordelingen har forsvunnet i utvalgsfordelingen, og de grønne søylene ser mer ut som en normalfordeling.
- Utvalgsfordelingen til gjennomsnittet og populasjonsfordelingen har samme gjennomsnitt; den røde stiplede streken er gjennomsnittet til begge fordelingene. Derfor kan vi bruke utvalgets gjennomsnitt til å estimere populasjonens gjennomsnitt.
- Utvalgsfordelingen til gjennomsnittet er mindre spredt enn populasjonsfordelingen. Tenk på forskjellen mellom to måter å trekke på. Trekker du én elev om gangen, kan du få nesten hva som helst: noen elever leser svært sakte, andre svært raskt. Trekker du derimot fem elever og regner ut gjennomsnittet deres, betyr det mindre om én av dem er en uvanlig rask eller uvanlig langsom leser, for de andre fire drar gjennomsnittet mot midten. Gjennomsnittet av fem tilfeldige elever er derfor mer stabilt enn lesehastigheten til én tilfeldig elev.

Dette delkapittelet er verdt å skjønne ordentlig, for utvalgsfordelingen til gjennomsnittet er nøkkelen for resten. Her er oppsummeringen:

- Vi prøver å estimere populasjonens gjennomsnitt ut fra utvalgets gjennomsnitt.
- Vi har bare ett utvalg og derfor bare ett gjennomsnitt, og det lager ingen fordeling. Men vi kan tenke oss gjennomsnittene til alle mulige utvalg vi kunne trukket; de utgjør utvalgsfordelingen til gjennomsnittet.

- Denne utvalgsfordelingen likner mer på normalfordelingen og har samme gjennomsnitt, men mindre spredning, enn populasjonsfordelingen.

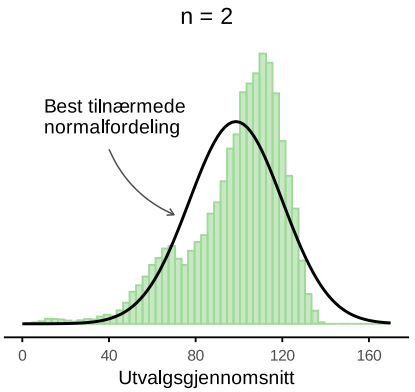
Når vi har vårt ene utvalg og regner ut vårt ene gjennomsnitt, ser vi på dette som et trekk fra denne utvalgsfordelingen.

Men dersom ulike tilfeldige utvalg kan gi ulike gjennomsnitt, hvor mye kan vi egentlig stole på det gjennomsnittet vi har observert? Og hvordan kan ett enkelt utvalg brukes til å si noe om den ukjente populasjonen? For å svare på disse spørsmålene må vi se nærmere på utvalgsfordelingens form og spredning, og det er beskrevet av sentralgrenseteoremet og standardfeilen.

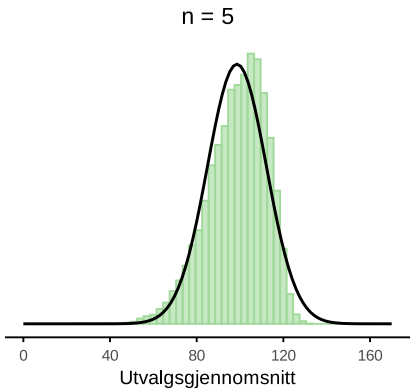
7.6 Sentralgrenseteoremet og standardfeilen

I forrige delkapittel så vi at utvalgsfordelingen så litt normalfordelt ut, men det var kun for utvalg på fem elever, $n = 5$. Hvordan vil utvalgsfordelingen se ut for større utvalg?

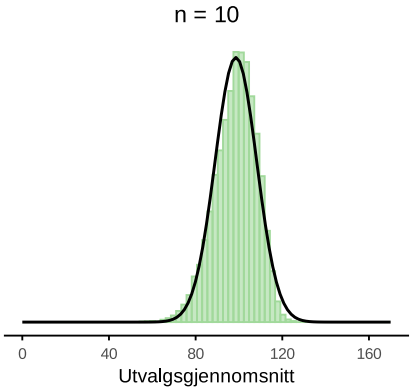
Vi finner ut av det ved å tegne samme figur, bare for flere forskjellige n. I Figur 7.8 viser jeg utvalgsfordelingen for utvalg med 2 elever, 5 elever, 10 elever og 100 elever. Hvert grønt histogram er altså utvalgsfordelingen til gjennomsnittet for én bestemt utvalgsstørrelse. Denne gangen er den svarte kurven den best tilpassede normalfordelingen, slik at man ser tydeligere hvorvidt utvalgsfordelingen er lik normalfordelingen.



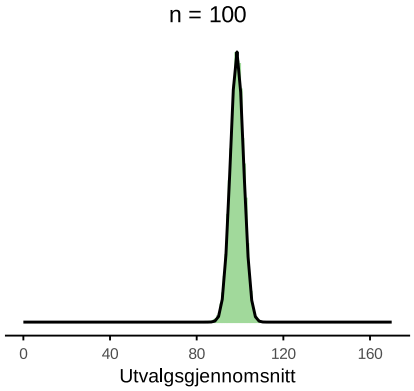
(a) For $n = 2$ ser man fremdeles de to humpene, og normalfordelingen passer dårlig



(b) Allerede ved $n = 5$ ligner fordelingen mer på en normalfordeling, og den er smalere



(c) Ved $n = 10$ er normalfordelingen en meget god tilnærming



(d) Ved $n = 100$ er fordelingen svært smal og praktisk talt normal

Figur 7.8: Sentralgrenseteoremet demonstrert med lesehastighets-eksempelet. Hvert panel viser utvalgsfordelingen til gjennomsnittet (grønne søyler) for én bestemt utvalgsstørrelse n, med den best tilpassede normalfordelingen lagt over (svart linje). Selv om populasjonens fordeling er humpete, blir gjennomsnittets fordeling mer og mer normal når n øker. Den blir også mindre spredt.

Selv om populasjonsfordelingen var aldri så humpete, ser vi at utvalgsfordelingen til gjennomsnittet fort blir normalfordelt. Allerede ved $n = 10$ er utvalgsfordelingen nesten en perfekt bjellekurve, og ved $n = 100$ er den smal og praktisk talt helt normal. Dette er innholdet i sentralgrenseteoremet:

Sentralgrenseteoremet:
Uansett hvilken fordeling populasjonen har, gjelder følgende om gjennomsnittets utvalgsfordeling:

- den blir tilnærmet normal når utvalget er tilstrekkelig stort
- den har samme gjennomsnitt som populasjonsfordelingen
- den blir smalere og smalere (får mindre og mindre standardavvik) når n øker.

Standardavviket til utvalgsfordelingen kalles **standardfeilen**, og sentralgrenseteoremet gir oss en enkel formel for den:

$$\text{standardfeil} = \frac{\text{populasjonens standardavvik}}{\sqrt{n}} \tag{7.1}$$

Men her stusser du kanskje: Formelen krever populasjonens standardavvik, og det er jo nettopp et av tallene vi ikke kjenner. Løsningen er den samme som ellers i kapittelet: Siden vi sjelden kan undersøke hele populasjonen, bruker vi variasjonen i utvalget som beste anslag på variasjonen i populasjonen. Vi setter altså inn utvalgets standardavvik i formelen.

Vanligvis skrives dette med forkortelser, SE for standardfeilen (av engelsk *standard error*) og SD for standardavviket (*standard deviation*), altså $\text{SE} = \frac{\text{SD}}{\sqrt{n}}$. Det er slik du vil se den i forskningsartikler, og slik vi kommer til å regne i Seksjon 7.7.

> ⚠ En unødvendig presisering: Bessels korreksjon
>
> Å bruke utvalgets standardavvik i stedet for populasjonens er en tilnærming, og her innfrir jeg et løfte fra Kapittel 5.. Regner man ut standardavviket slik vi gjorde der, ved å dele på antallet observasjoner n, får man en litt for lav verdi når målet er å *anslå* populasjonens standardavvik. Verdien må derfor ganges med **Bessels korreksjon**, som for standardavviket er $\sqrt{\frac{n}{n-1}}$. Det er det samme som å dele på $n-1$ i stedet for n inne under rottegnet, og det er nettopp dette jamovi gjør, så du får korreksjonen gratis. Den gjør standardfeilen bittelitt høyere enn den ellers ville blitt. For alt annet enn ørsmå utvalg er forskjellen helt ubetydelig, så du trenger ikke tenke mer på den.

Dette gjør normalfordelingen nyttig for oss. Selv om populasjonen vår slett ikke er normalfordelt, vet vi at utvalgsfordelingen til gjennomsnittet er normalfordelt (så lenge utvalget er stort nok). Og ikke nok med det, vi vet presisjonen vi anslår gjennomsnittet med, siden vi vet standardfeilen.

Legg merke til kvadratroten i nevneren. Den har en viktig konsekvens: for å *halvere* standardfeilen må du *firedoble* utvalget, og for å gjøre den en tidel så stor må utvalget bli hundre ganger større. Presisjon er dyrt, og det er grunnen til at studier sjelden har så store utvalg som vi skulle ønske.

7.7 Et eksempel

La oss se hvordan dette faktisk benyttes. Vi har søkt penger til studien vår for å anslå lesehastigheten til norske fjerdeklassinger. Vi ble svært skuffet da vi bare fikk nok penger til å samle data fra $n = 64$ tilfeldig trukkede elever. Huff, det er jommen dyrt å reise land og strand rundt for å gi Carlstenprøven

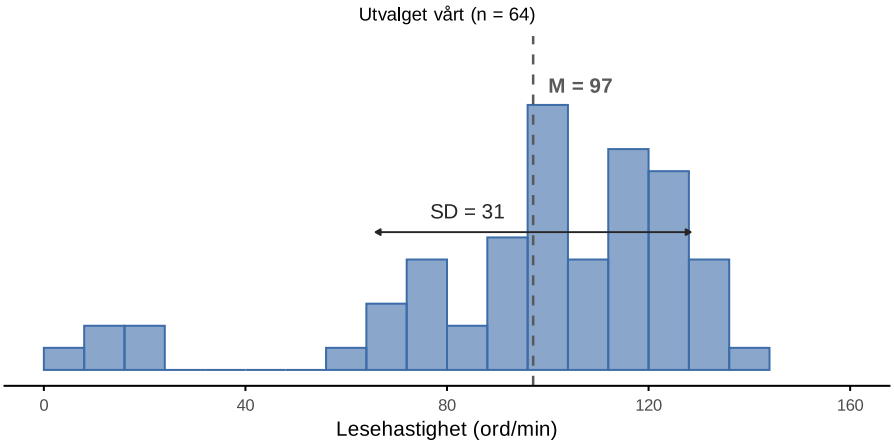
til tilfeldige elever. Hvordan skal det gå, når vi anslår lesehastigheten til 60 000 elever ut fra bare 64?

Etter å ha samlet dataene, gjør vi litt deskriptiv statistikk. Husk at i virkeligheten ville vi ikke visst populasjonsfordelingen, det er bare i dette kapittelet vi faktisk vet «fasiten», så nå må du late som om du ikke aner noe om den. Figur 7.9a viser lesehastigheten til de 64 elevene vi testet: gjennomsnittet ble M = 97 ord/min og standardavviket SD = 31 ord/min. Disse tallene gjelder utvalget. Kanskje blir vi bekymret av det rare histogrammet. Har vi hatt uflaks som har så mange elever med lav lesehastighet? Er utvalget vårt blitt lite representativt, slik at gjennomsnittet (M = 97) bommer grovt på populasjonen? Ingen grunn til bekymring, sentralgrenseteoremet kommer til unnsetning!

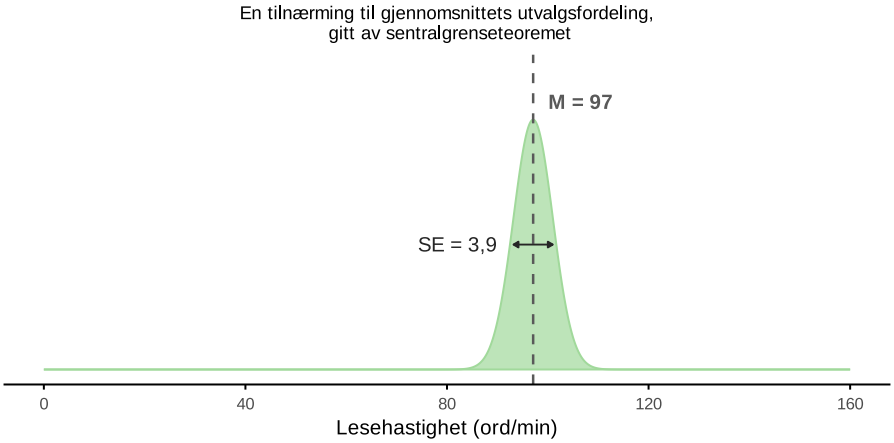
Ut fra de tre tallene, (M, SD og n), gir sentralgrenseteoremet oss en normalfordeling som er en svært god tilnærming til utvalgsfordelingen til gjennomsnittet. Denne er tegnet i Figur 7.9b. Som vi lærte i forrige delkapittel, er den sentrert på samme gjennomsnitt som populasjonen, men er mye smalere. Standardfeilen er $\text{SE} = \frac{\text{SD}}{\sqrt{n}} = \frac{31}{\sqrt{64}} = 3{,}9$. Nå er vi i mål!

I mål med hva da? Med å estimere populasjonsgjennomsnittet ut fra utvalget vårt. La oss tenke over hva vi har oppnådd. Utvalget vårt, Figur 7.9a, hadde en rar fordeling og høyt standardavvik. Sentralgrenseteoremet forteller oss at gjennomsnittet til utvalget er trukket fra en annen fordeling, vist i Figur 7.9b, en normalfordeling med mindre standardavvik som vi husker kalles standardfeilen. Vi har derfor svært god grunn til å tro at gjennomsnittet i populasjonen er ca. 97. Og på grunn av 68-95-99,7-regelen (Seksjon 7.4) er vi 68 % sikre på at gjennomsnittet er innenfor én standardfeil fra estimatet vårt.

Haha! For en berusende følelse! Vi visste ingenting om lesehastigheten til de 60 000 fjerdeklassingene, og så målte vi kun 64 av elevene, men likevel kan vi være 68% sikre på at snitthastigheten er $97 \pm 3{,}9$!!! Og siden to standardavvik er $3{,}9 \cdot 2 = 7{,}8$, så kan vi være 95% sikre på at gjennomsnittet er $97 \pm 7{,}8$. Wow!



(a) Lesehastigheten til hver enkelt elev, med gjennomsnittet (M) og standardavviket (SD)



(b) Utvalgsfordelingen til gjennomsnittet slik sentralgrenseteoremet spår den: en normalfordeling sentrert på det samme gjennomsnittet, men med en spredning, standardfeilen (SE), som er mindre enn standardavviket

Figur 7.9: En sammenlikning av fordelingen til utvalget på 64 elever, og den tilhørende utvalgsfordelingen til gjennomsnittet. Fordelingene har samme gjennomsnitt, men svært ulik spredning.

Jeg pepret forrige avsnitt med utropstegn, fordi dette virkelig er helt fantastisk. Vi har en hemmelig ingrediens som, krydret med litt matematikk, gir oss muligheten til å si hvor godt vi har estimert gjennomsnittet til 60 000 elever ut fra bare 64. Og hva er den hemmelige ingrediensen?

Tilfeldig utvalg.

Uten tilfeldige utvalg fungerer ingenting av dette; vi har ikke peiling på om utvalgets gjennomsnitt er en tilnærming av populasjonsgjennomsnittet og vi har i hvert fall ingen peiling på usikkerheten.

Men, for å ikke bli alt for beruset av lykke over sentralgrenseteoremet, så kan det være lurt å komme med noen innvendinger:

- Tenk tilbake på Seksjon 3.6.2 om hvor enkelt det var å hente inn et tilfeldig utvalg. Det var så vanskelig at jeg skrev, litt satt på spissen, at enkle tilfeldige utvalg ikke finnes! Oops.
- Husk at statistiske resultater er heftet med mange typer usikkerhet. Enkelt tilfeldig utvalg lar oss kun anslå utvalgsusikkerheten. Hva med måleusikkerheten, for eksempel? Kanskje har vi ikke anslått elevenes lesehastighet så godt som vi tror.

7.8 Konfidensintervall

Statistics means never having to say you're certain

Å rapportere et estimat sammen med en øvre og nedre grense er så vanlig at det har fått sitt eget navn, **konfidensintervall**. Det vanligste er å oppgi 95% konfidensintervall.

Oppskriften er enkel når vi først har standardfeilen. Vi vet fra sentralgrenseteoremet at gjennomsnittets utvalgsfordeling er normal, og vi vet fra 68–95–99,7-regelen at 95 % av en normalfordeling ligger innenfor (nesten nøyaktig) to standardavvik fra sentrum.[20] Standardavviket til utvalgsfordelingen er jo standardfeilen, så derfor blir et **95 % konfidensintervall** ganske enkelt:

$$\text{estimat} \pm 2 \cdot \text{standardfeil} \tag{7.2}$$

La oss regne det ut for studien vår. Vi trakk 64 elever, fikk gjennomsnittet M = 97 ord/min og standardfeilen SE = 3,9 (regnet ut i forrige delkapittel). Et 95 % konfidensintervall har da

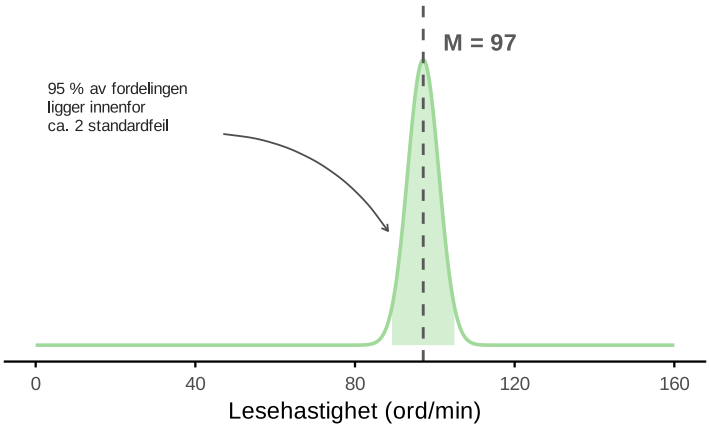
- nedre grense på $97 - 2 \cdot 3{,}9 = 89{,}3$
- øvre grense på $97 + 2 \cdot 3{,}9 = 104{,}9$.

Konfidensintervallet strekker seg dermed fra 89,3 til 104,9 ord/min.[21]

Figur 7.10 viser konfidensintervallet tegnet på gjennomsnittets utvalgsfordeling. Det skraverte området utgjør 95 % av fordelingen og er alt som ligger innenfor 2 standardfeil av estimatet. Det er nettopp dette midtpartiet konfidensintervallet fanger. Legg merke til hvor smalt det er sammenlignet med spennet av lesehastigheter som ble vist i Figur 7.9a.

[20] Det presise tallet er 1,96 standardavvik, ikke 2,0. Jeg regner med 2 fordi det er lett å gjøre overslag med i hodet, og fordi forskjellen er ubetydelig.

[21] Se bort fra små avrundingsfeil; de oppstår fordi jeg ikke viser alle desimaler.



Figur 7.10: 95 % konfidensintervall for studien vår, tegnet på gjennomsnittets utvalgsfordeling og på samme vannrette akse som forrige figur, slik at man ser hvor smal fordelingen er. Det skraverte området er de 95 % av fordelingen som ligger innenfor 2 standardfeil til hver side av estimatet (M); det er konfidensintervallet

I en forskningsartikkel ser du gjerne estimatet og konfidensintervallet skrevet svært kompakt, ofte med engelsk *CI* for *confidence interval*:

> Gjennomsnittlig lesehastighet var 97 ord/min, 95% CI [89.3, 104.9].

I jamovi: konfidensintervall for et gjennomsnitt

Analyses → Exploration → Descriptives. Legg variabelen du vil undersøke i Variables-boksen. Åpne så seksjonen Statistics og kryss av for «Confidence interval for the mean» (du finner den i bolken om gjennomsnitt). Standardinnstillingen er 95 %, men du kan endre prosenten i feltet ved siden av. Samme sted kan du krysse av for «Std. error of Mean» hvis du vil se standardfeilen direkte.

7.9 Tolking av konfidensintervallet

Det vanskeligste med konfidensintervaller er å tolke dem riktig, og her snubler mange, også erfarne forskere. Den riktige tolkningen er:

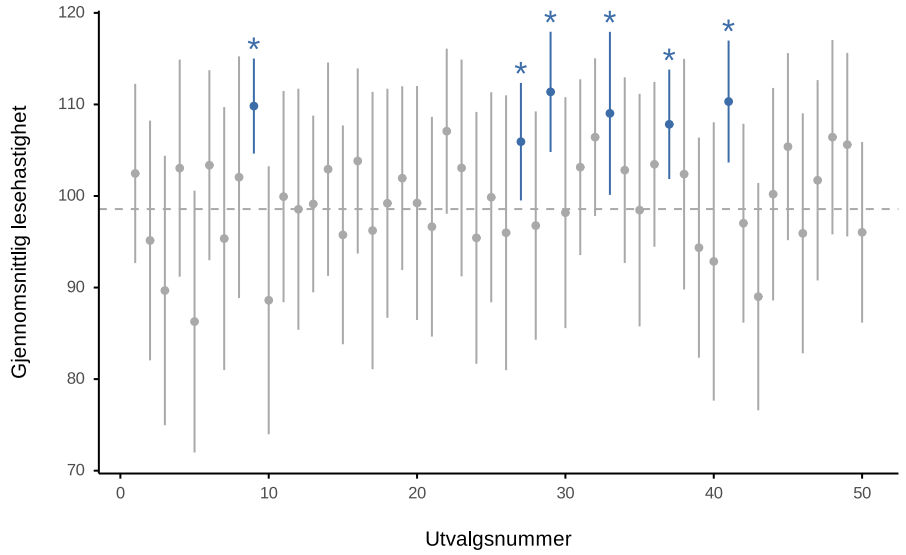
> Ved tilfeldig utvalg fra en populasjon, vil et 95% konfidensintervall inneholde populasjonsparameteren i 95% av tilfellene.

Husk at populasjonsparameteren er en tallverdi som beskriver hele populasjonen. Hvis populasjonsparameteren er gjennomsnittet, sier tolkningen altså at det 95 % konfidensintervallet inneholder populasjonsgjennomsnittet i 95 % av tilfellene.

Men hva betyr «95 % av tilfellene» her? Vi har jo bare utført studien én gang. Disse tilfellene viser til et tankeeksperiment. Se for deg at du gjentar

studien mange, mange ganger med nye tilfeldige utvalg og regner ut konfidensintervallet hver gang. I 95 % av disse intervallene vil man finne det sanne gjennomsnittet. Men i virkeligheten har man bare ett intervall og man vet ikke om det er blant de 95 % som traff, eller blant de 5 % som bommet. Akkurat ditt konfidensintervall kan jo ha bommet grovt.

Figur 7.11 illustrerer tankeeksperimentet. Det sanne populasjonsgjennomsnittet ligger fast (den stiplede linjen i midten). De loddrette strekene viser 50 gjentakelser av studien, altså samme studie utført 50 ganger med forskjellig tilfeldig utvalg, som gir forskjellige estimat på gjennomsnittlig lesehastighet. De fleste konfidensintervallene fanger det sanne gjennomsnittet, men noen få, markert i blått og med stjerne over, bommer. Når du gjør én enkelt studie, vet du ikke om du har fått et av de grå konfidensintervallene eller et av de blå, altså om konfidensintervallet inneholder populasjonsgjennomsnittet eller ikke. Det du vet, er at ved å utføre prosedyren, altså ved å gjennomføre en studie med tilfeldig utvalg, så vil konfidensintervallet inneholde populasjonsparameteren 95% av gangene.



Figur 7.11: 95 % konfidensintervaller fra 50 forskjellige tilfeldige utvalg i en lesehastighetsstudie med $n = 25$. Prikken er utvalgets gjennomsnitt og linjen er konfidensintervallet. Den stiplede vannrette linjen er det sanne populasjonsgjennomsnittet. De fleste intervallene inneholder det sanne gjennomsnittet, men noen få, i blått og markert med stjerne, gjør det ikke

Så når du møter et 95 % konfidensintervall i en forskningsartikkel, må du tenke at du ikke vet om dette intervallet er blant de 95% som traff eller blant de 5% som bommet.

7.10 Tilordningsusikkerhet i randomiserte kontrollstudier

Så langt har usikkerheten vår handlet om *tilfeldig utvalg*: vi trakk elever tilfeldig fra en populasjon, og spørsmålet var hvor godt utvalgsgjennomsnittet

estimerte populasjonsgjennomsnittet. Men konfidensintervaller dukker også opp ved *tilfeldig tilordning* i randomiserte kontrollstudier.

Vi tenker oss en **randomisert kontrollstudie** (se Kapittel 3.), for eksempel der forskerne rekrutterer 100 elever på 4. trinn fra én skole og deler dem tilfeldig i to grupper. Den ene gruppa, intervensjonsgruppa, får prøve ut en ny lesestrategi; den andre, kontrollgruppa, fortsetter som før. Alle elevene tar Carlstenprøven noen uker etterpå. Forskjellen i lesehastighet mellom de to gruppene er effekten av intervensjonen.

Anta at intervensjonsgruppa i snitt leste 4,2 ord/min mer enn kontrollgruppa etter intervensjonen. Denne *forskjellen i gjennomsnitt* er også et estimat som har sin egen utvalgsfordeling og sin egen standardfeil, akkurat slik som gjennomsnittet i forrige eksempel. Forskere kan dermed finne et konfidensintervall, for eksempel et 95 % konfidensintervall på 0,3 til 8,1.

Forskjellen på dette konfidensintervallet og det forrige er hva usikkerheten beskriver. Her er det ingen tilfeldig trekning fra en populasjon av norske 4.-klassinger; det er bare disse 100 elevene, og de er ikke trukket tilfeldig fra noe sted. Det tilfeldige er *tilordningen* til gruppene. Konfidensintervallet svarer derfor ikke på spørsmålet «hva er effekten av leseintervensjonen for norske 4.-klassinger generelt?», men på spørsmålet «hva er effekten av leseintervensjonen for akkurat disse 100 elevene?». Det er en usikkerhet knyttet til den tilfeldige tilordningen, siden estimatet av effekten hadde blitt annerledes med en annen tilordning. Den randomiserte kontrollstudien sier i utgangspunktet bare noe om elevene i utvalget, ikke om at funnet skal gjelde andre elever. Den ytre validiteten er altså ikke noe konfidensintervallet gir svar på for randomiserte kontrollstudier.

7.11 Hva konfidensintervallet ikke fanger

Kapittelet begynte med å skille mellom fire typer usikkerhet, men har hittil kun vist hvordan konfidensintervallet kan dekke utvalgs- og tilordningsusikkerhet. Det er fordi konfidensintervallet i dette kapittelet, som er det som vanligvis brukes i forskning, ikke tar hensyn til de andre typene usikkerhet. Det er lett å glemme, fordi konfidensintervallet er det eneste tallet vi faktisk regner ut, men de andre kildene til usikkerhet er der likevel, og er ofte større.

- **Måleusikkerhet.** Carlstenprøven måler ikke lesehastighet helt presist. Dagsform, teksten som benyttes i prøven, og tilfeldigheter i selve gjennomføringen gir feil som konfidensintervallet ikke fanger.
- **Modellusikkerhet.** Hele regnestykket bygger på antakelser: at utvalget er et enkelt tilfeldig utvalg, målingen er god for alle elevene, og så videre. Hvis antakelsene er gale, for eksempel ved selektiv dropout (se Seksjon 3.6.2), så kan konfidensintervallet være misvisende.
- **Ikke-tilfeldige utvalg.** Dette er kanskje den viktigste begrensningen i lærerstudenters egne masteroppgaver. Nesten ingen masteroppgaver bygger på et enkelt tilfeldig utvalg fra en veldefinert populasjon; de bruker de elevene og lærerne forskeren får tilgang til. Da hviler konfidensintervallet på en forutsetning som ikke er oppfylt, og det er strengt tatt meningsløst. Det betyr ikke at studien er verdiløs, men at usikkerheten må vurderes med faglig skjønn og teoretiske argumenter, ikke med et konfidensintervall regnet ut uten at forutsetningene er oppfylt.

7.12 Oppsummering

I dette kapittelet har jeg forklart hvordan man estimerer populasjonsparametre. Jeg har bare vist hvordan man estimerer populasjonsgjennomsnittet, men liknende tenkning fungerer for alle mulige populasjonsparametre. For eksempel, hvis man skulle estimert populasjons*medianen*, så måtte man funnet utvalgsfordelingen til medianen, som kanskje ikke er normalfordelt. Da hadde man trengt noe som likner på sentralgrenseteoremet, men som gjelder for medianer. Deretter kunne man regnet ut et konfidensintervall for medianen. Alle begrepene du har lært, utgjør altså et språk for estimering og måling av usikkerhet som kan brukes mer generelt. Du har lært:

- Fire typer usikkerhet: måle-, modell-, utvalgs- og tilordningsusikkerhet. Kapittelet handlet mest om de to siste.
- Populasjon og utvalg: forskjellen på populasjonsparametre og utvalgsobservatorer.
- Normalfordelingen: en bjelleformet fordeling definert av to tall, gjennomsnitt og standardavvik. Den følger 68–95–99,7-regelen.
- Utvalgsfordelingen til gjennomsnittet: Også gjennomsnittet har en fordeling.
- Sentralgrenseteoremet og standardfeilen: Utvalgsfordelingen til gjennomsnittet blir tilnærmet normalfordelt for store nok utvalg uansett hvordan populasjonsfordelingen ser ut. Standardfeilen måler usikkerheten i gjennomsnittet.
- Konfidensintervallet: den vanligste måten å angi usikkerhet på. Vanligst med 95% konfidensintervall, regnet ut som $\pm 2 \cdot SE$. Vi lærte hvordan det skal tolkes.
- At samme matematikk ligger bak konfidensintervaller i randomiserte kontrollstudier, men at usikkerheten der kommer fra tilordning, ikke utvalg.
- At konfidensintervallet ikke fanger all usikkerhet: Det fanger ikke måleusikkerhet, og det trenger tilfeldig utvalg eller tilordning for å fungere.

Etterord

«Begin at the beginning», the King said, very gravely, «and go on till you come to the end: then stop»
– Lewis Carroll, *Alice in Wonderland*

Som nevnt i forordet, er denne læreboken en betydelig omarbeidet og forkortet versjon av en bok skrevet for et kurs som krevde et firedels semesters arbeid, altså 7,5 studiepoeng. Fra over 500 sider har jeg kuttet og komprimert materialet ned til omtrent 130 sider, samtidig som jeg har gjort det relevant for utdanningsforskning. Komprimeringen betyr naturligvis at mange detaljer, og faktisk også mye som ikke er detaljer, har blitt utelatt for å holde omfanget håndterbart. Skulle du ønske å lese mer om en metode, kanskje særlig hvis du planlegger å benytte den i en fremtidig masteroppgave, vil du finne et vell av ytterligere informasjon og detaljer i den originale utgaven av boken (Navarro & Foxcroft, 2025).

Noe av hovedpoenget med å skrive denne boka er å gjøre alt materialet relevant for lærere i skolen. Dette arbeidet er langt fra i mål. Til fremtidige utgaver håper jeg å ha flere og bedre eksempler, og at eksemplene samlet sett viser noe av bredden i kvantitativ forskning på utdanningsfeltet. Jeg ønsker å inkludere egne kapitler om store internasjonale undersøkelser, nasjonale prøver, kartleggingsprøver og summative vurderinger i skolen.

Målet er at alle lærere skal tenke «jeg er jommen glad jeg har lært dette» når de kommer til slutten av boka.

Referanseliste

Bibliografi

Anscombe, F. J. (1973). Graphs in Statistical Analysis. *American Statistician*, *27*, 17–21. https://doi.org/10.1080/00031305.1973.10478966

Arnesen, A., Braeken, J., Ogden, T., & Melby-Lervåg, M. (2019). Assessing Children's Social Functioning and Reading Proficiency: A Systematic Review of the Quality of Educational Assessment Instruments Used in Norwegian Elementary Schools. *Scandinavian Journal of Educational Research*, *63*(3), 465–490. https://doi.org/10.1080/00313831.2017.1420685

Bergesen, H. O. (2006). *Kampen om kunnskapskolen*. Universitetsforlaget.

Bettinger, E., Ludvigsen, S., Rege, M., Solli, I. F., & Yeager, D. (2018). Increasing Perseverance in Math: Evidence from a Field Experiment in Norway. *Journal of Economic Behavior & Organization*, *146*, 1–15. https://doi.org/10.1016/j.jebo.2017.11.032

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex Bias in Graduate Admissions: Data from Berkeley. *Science*, *187*, 398–404. https://doi.org/10.1126/science.187.4175.398

Booher-Jennings, J. (2005). Below the Bubble: ``Educational Triage'' and the Texas Accountability System. *American Educational Research Journal*, *42*(2), 231–268. https://doi.org/10.3102/00028312042002231

Campbell, D. T. (1979). Assessing the Impact of Planned Social Change. *Evaluation and Program Planning*, *2*(1), 67–90. https://doi.org/10.1016/0149-7189(79)90048-X

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and Quasi-Experimental Designs for Research*. Houghton Mifflin.

Cassidy, S. A., Dimova, R., Giguère, B., Spence, J. R., & Stanley, D. J. (2019). Failing Grade: 89% of Introduction-to-Psychology Textbooks That Define or Explain Statistical Significance Do So Incorrectly. *Advances in Methods and Practices in Psychological Science*, 2515245919858072. https://doi.org/10.1177/2515245919858072

Eilertsen, T. (2020). *Trine Eilertsen Beklager Misbruk Av Statistikk i A-magasinet*.

Fraillon, J., Friedman, T., & Fraillon, J. (Red.). (2024). *ICCS 2022 Technical Report* [Technical report].

Gigerenzer, G. (2018). Statistical Rituals: The Replication Delusion and How We Got There. *Advances in Methods and Practices in Psychological Science*, *1*(2), 198–218. https://doi.org/10.1177/2515245918771329

Greenland, S. (2022). *What Does Statistics Have to Offer the Epidemiologist?*.

Jensen, F., Kjærnsli, M., Knain, E., Løvgren, M., & Pettersen, A. (2024). Sammenhengen Mellom Utforskende Undervisning i Naturfag Og Elevers Prestasjoner Og Interesse for Og Trivsel Med Naturvitenskap. Norske Resultater Fra PISA 2015. *Nordic Studies in Science Education*, *20*(1), 103–118. https://doi.org/10.5617/nordina.9335

Koretz, D. (2017). *The Testing Charade: Pretending to Make Schools Better*. University of Chicago Press.

Leatham, K. R. (2006). Viewing Mathematics Teachers' Beliefs as Sensible Systems*. *Journal of Mathematics Teacher Education*, *9*(1), 91–102. https://doi.org/10.1007/s10857-006-9006-8

Lytsy, P., Hartman, M., & Pingel, R. (2022). Misinterpretations of P-values and Statistical Tests Persists among Researchers and Professionals Working with Statistics and Epidemiology. *Upsala Journal of Medical Sciences*, *127*, 10.48101/ujms.v127.8760. https://doi.org/10.48101/ujms.v127.8760

Mosvold, R., & Ohnstad, F. O. (2016). Profesjonsetiske Perspektiver På~Læreres Omtaler Av Elever. *Norsk pedagogisk tidsskrift*, *100*(1), 26–36. https://doi.org/10.18261/issn.1504-2987-2016-01-04

Navarro, D. (2011). *Learning Statistics with R* (0.6).

Navarro, D., & Foxcroft, D. (2025). *Learning Statistics with Jamovi: A Tutorial for Beginners in Statistical Analysis* (1. utg.). Open Book Publishers. https://doi.org/10.11647/OBP.0333

Roser, M. (2022). Millions of Children Learn Only Very Little. How Can the World Provide a Better Education to the next Generation?. *Our World in Data*.

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, *103*, 677–680. https://doi.org/10.1126/science.103.2684.677

Tangen, N. (2022). Bonus Episode: Grit with Angela Duckworth [Audio Podcast Episode]. *In Good Company*.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA Statement on P-Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108