Del I

Innledning

1

Hva er statistikk?

1.1 Bokas innhold

1.2 Noen viktige begreper

1.3 Modeller

1.4 Innsamling av data

Moderne statistikk er noe annet enn de kjedelige tallene og tabellene du kanskje tenker på. Sannsynlighetsteorien ble utviklet på 1600–1700-tallet for å analysere vinnersannsynligheten i forskjellige pengespill. Faget har siden utviklet seg sterkt, og den generelle teorien har fått en rekke anvendelser i naturvitenskap, teknikk, medisin, økonomi og samfunnsvitenskap.

1.1 Bokas innhold

Vi bruker statistikk som hjelpemiddel når vi skal fatte beslutninger i situasjoner der noe er uforutsigbart eller tilfeldig. I en komplisert verden er det ofte slik: Vi vet om *flere mulige konsekvenser*, men vi vet *ikke sikkert* hvilken konsekvens vår beslutning vil få. Beslutninger *må* fattes selv om usikkerheten er til stede. Gjennom statistikkfaget skal vi se hvordan vi kan fatte gode beslutninger i situasjoner hvor noe er variabelt eller tilfeldig.

1.1.1 Noen eksempler

Som en liten smakebit på hva statistikkfaget dreier seg om, skal vi kort beskrive noen eksempler der faget kan være til stor nytte.

- **Legemiddeltesting**. Før nye legemidler slippes ut på markedet, må produsenten bevise at middelet virker, og at det har begrensede bivirkninger. Det blir derfor gjennomført kontrollerte forsøk, først på dyr og deretter på mennesker som har sagt seg villige til å delta i forsøket. Noen av personene får en narrepille (placebopille) slik at man kan sammenligne behandlingseffekten hos dem som fikk reell behandling, og hos dem som fikk en narrepille. Hvis legemiddelet gir gode resultater, er det bare tilfeldig, eller skyldes det faktisk at medisinen er god? Statistiske metoder kan hjelpe legene til å planlegge et godt forsøk og til å analysere resultatene etterpå.
- **Økonomiske risikovurderinger**. Svært mange bedrifter kjennetegnes av at både utgifter og inntekter er usikre: Råvareprisene varierer, rentekostnadene varierer, produksjonskostnadene er usikre, markedsprisen for sluttproduktet varierer, og man vet ikke hvor mange produkter man klarer å selge. For å sette opp et budsjett må bedriften anslå disse viktige størrelsene. Statistikkfaget kan hjelpe til med å gi et bilde av hvilket overskudd bedriften kan forvente, og hvor stor risikoen er for å gå med underskudd.
- **Aksjeanalyse og porteføljeforvaltning**. En klassisk anvendelse av statistikk er knyttet til aksjehandel, der børskursene varierer fra minutt til minutt og genererer store mengder data om verdien av alle børsnoterte selskaper. Investorer og analytikere bruker betydelige ressurser på å analysere disse dataene på jakt etter mønstre som kan indikere om det er lurt å selge eller kjøpe. Mye av dette er fånyttes når det gjelder enkeltaksjer, fordi det er tilnærmet umulig å spå den framtidige børskursen på basis av historikken. Men når det gjelder en portefølje av flere aksjer, er statistikk og sannsynlighetsregning uunnværlige verktøy som kan bidra til optimal sammensetning av en portefølje i forhold til hvor stor risiko investoren er villig til å ta.

- **Teknisk risikoanalyse**. Ved dimensjonering av bygninger (oljeplattformer, broer, boligblokker, demninger) stilles man hele tiden overfor spørsmålet: Hvor sterke skal konstruksjonene være, hvor sterk påvirkning skal de tåle? Hvor sannsynlig er det at bygningene blir utsatt for de sterkeste påkjenningene de er ment å tåle? Bør norske hus tåle jordskjelv, slik som hus i Los Angeles bør? Hva er sannsynligheten for jordskjelv i Norge i forhold til LA? Ingeniører må alltid foreta en avveining mellom hvor sterk konstruksjonen må være, og hvor mye penger man skal bruke på bygningen. I denne avveiningen kan statistikkfaget være til stor hjelp.

- **Sykdomsrisiko**. Hva er sannsynligheten for å få kreft? Er denne sannsynligheten påvirket av bestemte risikofaktorer, slik som røyking, asbest, radonstråling, kaffedrikking og lettbrus? Det fremmes en mengde resultater fra medisinsk forskning som vi må lære å forholde oss til. Ofte får slike resultater fordreide oppslag i tabloidpressen. Hva betyr egentlig resultatene? Er det sikkert at de er gyldige for hele befolkningen? Er det bare tilfeldig at forsøksresultatet ble som det ble? Slike kritiske spørsmål står helt sentralt i statistikkfaget, innenfor fagområdet *statistisk hypotesetesting*.

- **Kvalitetstesting**. Hvor mange defekte deler er det i et vareparti? Mange typer produkter lar seg ikke kvalitetsteste uten at produktene blir ødelagt. Da tvinges vi til å si noe om hele varepartiet på grunnlag av en tilfeldig utvalgt stikkprøve. Andre ganger er det mulig, men svært kostbart å teste alle produktene, slik at man ønsker å kontrollere så få produkter som mulig. Hvor mange enheter må vi teste? Hvor mange feil kan vi godta i stikkprøven og likevel akseptere varepartiet? Denne typen spørsmål er svært sentrale i mange industrier, for eksempel for produsenter og kjøpere av bildeler, flydeler og elektroniske komponenter. Sannsynlighetsregningen kan hjelpe til å gi gode svar på disse viktige spørsmålene.

- **Vedlikeholdsplanlegging**. Vedlikehold utføres vanligvis for å minimere kostnader knyttet til drift og for å tilfredsstille visse minimumskrav til sikkerhet. I norsk elforsyning brukes hvert år flere milliarder kroner på å vedlikeholde kraftstasjoner og ledninger. Hvis et kraftselskap reduserer vedlikeholdskostnadene, blir det flere feil. For å bestemme riktig nivå på vedlikeholdet må kraftselskapet vite hvilke konsekvenser feilene har, og hvor ofte de skjer. Manglende kunnskap om konsekvensene har i mange tilfeller ført til for høyt vedlikeholdsnivå – for sikkerhets skyld. En lignende problemstilling har man knyttet til *forsikring*. Vi vet hva forsikringen koster, men vi vet ikke om vi får noen nytte av den.

- **Lagerstyring**. Det koster penger å ha varer på lager, ved at kapital og areal bindes opp. Men det koster også penger å ikke ha varene på lager når man trenger dem. Det gjelder enten lageret inneholder råvarer og reservedeler til egen produksjon, eller det inneholder salgsvarer som skal leveres til kundene. Ingen vet nøyaktig når reservedelene trengs, eller nøyaktig hvor mange varer kundene vil etterspørre i neste uke. Kunnskap om kostnader og leveringstider kan sammen med statistiske data lede fram til regler for optimalt lagerhold.

- **Prosjektstyring**. Et prosjekt kan deles opp i delprosjekter. Vanligvis er det lettere å anslå varighet og kostnad for delprosjektene enn for selve hovedprosjektet. Basert på slike anslag kan sannsynlighetsregningen hjelpe oss å anslå varighet og kostnad for hele prosjektet. Et prosjekt trenger ikke være gigantisk stort for at slike vurderinger skal være nyttige.

- **Kunstig intelligens**. Bedrifter som Facebook, Amazon og Netflix samler store datamengder om sine brukere og har utviklet algoritmer som gir god innsikt i brukernes preferanser. De vet hva kundene liker, og hvordan de bør markedsføre produktene for å få maksimalt salgsvolum. Avansert statistikk er deres viktigste verktøy.

- **Valg av servicenivå**. Mange tjenesteytende virksomheter håndterer en uforutsigbar kundemasse. Man vet ikke på forhånd hvor mange kunder som vil melde sin ankomst neste dag. Problemstillingen er aktuell for matbutikken, banken eller storbedriften som skal velge antall kassepersonale, bankfunksjonærer og sentralbordbetjenter. Er det for få ansatte, blir køene for lange og irritasjonen stor. Er det for mange ansatte, blir kostnadene unødig store. Med kunnskap om typiske mønstre i kundenes ankomsttider kan man finne fram til et riktig servicenivå. Problemstillingen er like aktuell for nettbutikker som må dimensjonere kapasiteten i sin varedistribusjon. (Det er opplagt mange bedrifter som ikke har gjort leksen sin her!)

- **Juridiske spørsmål**. Et hovedprinsipp i retten er at man er uskyldig inntil det motsatte er bevist. All *rimelig* tvil skal komme tiltalte til gode. Med sannsynlighetsregning kan man i noen saker beregne hvor stor tvilen er, basert på visse antakelser og observasjoner. I den såkalte Landås-saken ble en hjelpepleier tiltalt for drap på en rekke pleiepasienter. Det var spesielt mange som døde når denne hjelpepleieren var på jobb. Retten besluttet likevel at hjelpepleieren skulle frifinnes, fordi den mente at det kunne være tilfeldig at så mange var døde på akkurat disse vaktene. (Retten baserte sin beslutning på sannsynlighetsberegninger presentert av forsvarerne.)

- **Spørreundersøkelse**. De fleste av oss har opplevd å få tilsendt et spørreskjema eller bli oppringt fra et meningsmålingsinstitutt. Spørreundersøkelser kan være nyttige for å kartlegge forhold som ikke lar seg registrere eller måle med vanlige «måleinstrumenter». Slike målinger brukes ofte for å kartlegge hvor tilfredse kunder og medarbeidere er, oppslutningen til politiske partier, preferanser til ulike produkter og generelle holdninger til aktuelle problemstillinger. Svært ofte er det unødvendig og altfor kostbart å spørre hele populasjonen av medarbeidere, kunder og andre involverte. Spørmålene stilles bare til et *utvalg* fra populasjonen. Resultatene av undersøkelsen danner basis for beslutninger om tiltak som skal iverksettes (av bedriftsledelsen, partiledelsen, markedssjefen). Statistikkfaget gir svar på hvor mange som bør spørres, og hvor sikre resultatene fra undersøkelsen er.

- **Optimalisering av produksjonsprosess**. I prosessindustrien foregår produksjonen vanligvis ved at en mengde ulike råstoffer blandes sammen. Kvaliteten på sluttproduktet avgjøres av blandingsforholdene mellom disse stoffene, rekkefølgen på sammenblandingen, temperaturer, trykk, luftfuktighet osv. I kompliserte produksjonsprosesser er det ofte slik at man ikke har full kontroll på alle inputparameterne (de varierer). De produksjonsansvarlige har muligheten til å endre blandingsforhold, temperatur og trykk, men kjenner ofte ikke konsekvensene av slike endringer. For å lære mer om produksjonsprosessen er det vanlig å foreta prøveproduksjoner. Statistikkfaget kan hjelpe til med å planlegge slike prøver og analysere hvilke endringer som vil føre til bedre produktkvalitet.
- **Mønstergjenkjenning**. En tomflaskeautomat tar et bilde av flasken som settes inn i automaten. Ingen av bildene blir helt like, siden alle flaskene plasseres litt forskjellig i automaten. Automaten må deretter beslutte hva slags flaske dette er. I oljeselskapene analyserer man geologiske testdata fra borehull i Nordsjøen. Hver testmåling «avbildes» med sine karakteristikker. Ingen karakteristikker er helt like. Noen steder er det mye olje, andre steder er det lite eller ingenting. Geologene må gi en anbefaling om hvor de tror det kan finnes olje (helst mye). Oljeselskapet må deretter ta en beslutning om hvor det skal søke konsesjoner og foreta nye boringer. Google og Facebook gjør noe lignende når de sorterer bildene dine i forhold til hvem som er avbildet.
- **Veddemål**. Er du villig til å satse 700 kroner på at Norge kvalifiserer seg til fotball-VM i 2026 hvis du kan vinne 1000 kroner dersom du får rett? Statistikken taler ikke til Norges fordel. Hvor stor sjanse har Norge for å kvalifisere seg? Det kan være vanskelig å svare på dette spørsmålet direkte, men basert på kampoppsettet i Norges kvalifikasjonsgruppe kan du anslå hvilke kampresultater du har størst tro på. Sannsynlighetsregningen kan hjelpe deg til å kombinere dine gjetninger på en ryddig måte, slik at du får fram en totalgjetning på hvilke sjanser Norge har.
- **Spill**. Mange spill er slik at spillerne må ta beslutninger underveis. I poker må du vurdere om du skal bytte ut kort. *Du vet hva du har, men ikke hva du får.* Ved ruletten må du velge om du skal satse på rødt eller sort. Du kan observere rulettens resultater i årevis; likevel vet du ikke hvilken farge som blir resultatet i neste runde. I yatzy må du underveis bestemme deg for om du skal samle til «hus» eller om du skal prøve å få «yatzy». Det siste er vanskeligere, men gir flere poeng. Du kan ikke vite om du vil lykkes. I alle disse situasjonene kan sannsynlighetsregningen hjelpe deg til å ta gode beslutninger.

1.1.2 Historie

Statistikk har lenge dreid seg om å samle inn store tallmengder og stille disse opp i tabeller, særlig med tanke på å belyse ulike samfunnsmessige forhold. Slik statistikk har vært et viktig hjelpemiddel for politikere og andre makthavere i tusenvis av år. Tenk bare på Josef og Maria som reiste til Betlehem for å skrive seg inn i manntallet. Alle siviliserte samfunn har etter hvert bygd opp et stort apparat for å samle inn slike opplysninger.

Arven fra romerne videreføres i Norge av Statistisk sentralbyrå (SSB). *Statistisk årbok* inneholder om lag 500 tabeller som beskriver naturressurser, sosiodemografi, sosioøkonomi, næringsøkonomi, samfunnsøkonomi og samfunnsorganisatoriske emner. Det finnes for eksempel en tabell som heter «*Hoteller og liknande overnattingsbedrifter. Overnattingar, etter gjestene sitt bustadland*». Denne tabellen baserer seg på registreringsskjemaer som samtlige hoteller, pensjonater og campingplasser må fylle ut. I skjemaet skal det opplyses om antall overnattingsgjester hver dag i sesongen, fordelt på nasjonalitet. Kan du forestille deg hvor mye arbeid som totalt sett går med til å lage denne statistikken? Årlig er det om lag 33 millioner overnattinger på ca. 900 overnattingssteder. Hovedutfordringen i denne typen datainnsamling er å motivere de som skal fylle ut registreringsskjemaene til å gjøre jobben samvittighetsfullt og riktig. Denne typen «gammeldags» tallinnsamling vil *ikke* være tema i denne boka.

Moderne statistikk, slik vi definerer statistikkfaget i denne boka, ble først tatt i bruk i samfunnsvitenskap, landbruksforskning og medisinsk forskning. Carl F. Gauss, Francis Galton og andre oppdaget systematiske mønstre i det som mange trodde var et ubeskrivelig kaos. Deres ideer var svært nyttige når ulike fenomener skulle beskrives. Ronald A. Fisher var blant de første som innså hvor viktig det var å planlegge eksperimenter på en systematisk måte, slik at det ble mulig å trekke gode konklusjoner etterpå. Utviklingen fram til i dag har resultert i en mengde statistiske metoder som kan benyttes i svært mange sammenhenger. I denne boka skal vi lære om de viktigste metodene, men du skal vite at det finnes flere. Mange flere.

Sannsynlighetsregningen har også en lang forhistorie, nær knyttet til terning- og kortspill. Navn som Pierre de Fermat, Blaise Pascal, Jacques Bernoulli og Thomas Bayes står sentralt i den tidlige utviklingen av det som er dagens matematisk baserte sannsynlighetsregning. Selv om utgangspunktet var nært knyttet til spill, viste det seg raskt at sannsynlighetsregningen også kunne brukes på andre områder. I Norge utga Axel Sophus Guldberg (1838–1913) boka «*Om Sandsynlighetsregningen og dens Anvendelse paa Hazardspil og Forsikringsvæsen*» så tidlig som i 1873. I denne boka ser han sannsynlighetsregningen som «*Kilden til en ny Videnskab, som kunde erholde Anvendelser, hvorom man neppe skulde have nogen Anelse, og som griber ind i Forhold som synes at ligge udenfor al menneskelig Beregning*». Gjennom bokas mange eksempler skal vi se at Guldberg fikk rett i sin spådom: Sannsynlighetsregningen har usannsynlig mange anvendelsesområder.

På 2000-tallet har statistikkfaget opplevd en renessanse. Ny sensorteknologi og stadige framskritt innen mobile kommunikasjonsløsninger, har ført til en eksplosjon av tilgjengelige data. Både bedrifter og offentlige virksomheter samler store datamengder, som kan brukes til et mangfold av beslutninger. «Data er det nye gullet» eller den nye «oljen». New York-børsen beviser langt på vei dette. De siste års vinnere er nesten utelukkende selskaper som sitter på store, eksklusive datamengder som gir grunnlag for framtidig profitt. Velkjente eksempler er Google, Apple, Facebook og Amazon. Dette har gitt høy etterspørsel etter fageksperter som kan brukes til analyse av store datasett, og dyktige statistikere verdsettes i stadig flere bransjer og stillingskategorier.

1.1.3 Bokas oppbygning

Etter å ha lest denne boka bør du beherske de viktigste statistiske metodene slik at du kan bruke dem selv. Du skal også være i stand til å tolke statistiske opplysninger. Boka fokuserer mye på ideer og intuisjon framfor på matematikk og beviser. Hvert kapittel innledes med tekst og illustrasjoner som du kan forstå uten å måtte sette deg inn i matematiske formler.

Statistikkfaget inneholder både *statistikk* og *sannsynlighetsregning*. Dette er egentlig to ulike disipliner som griper sterkt inn i hverandre, på den måten at den ene delen ofte er en forutsetning for den andre. Vi trenger statistikk for å skaffe kunnskap til bruk i sannsynlighetsberegningene. Og vi trenger sannsynlighetsregningen for å lage gode modeller som vi bruker når vi gjør statistiske analyser. I denne boka vil vi veksle litt fram og tilbake mellom statistikk og sannsynlighetsregning.

Boka består av fem deler som bør leses i den rekkefølgen de er oppført. Mye av stoffet i del I (og noe av del II og del III) inngår i matematikkpensum for den videregående skolen. Selv om du har vært borti dette stoffet før, bør du lese igjennom stoffet for å sikre deg at definisjoner, begreper og symboler betyr det som du tror.

I. **Innledning**. Du har allerede begynt å lese innledningen, som er ment å være en myk start på statistikkfaget. Viktige begreper skal defineres, og vi skal se hvordan forsøksdata kan presenteres.

II. **Sannsynlighetsteori**. I denne delen skal vi for alvor gå løs på sannsynlighetsregningen, med fokus på tilfeldighet og hvilke mønstre som tross alt finnes i denne tilfeldigheten. Vi skal lære å regne ut sannsynligheten for at en hendelse vil inntreffe. Vi skal lære om tilfeldige variabler og bli kjent med de vanligste sannsynlighetsfordelingene.

III. **Grunnleggende statistikk**. Her skal vi lære om hvordan vi kan trekke konklusjoner om hele grupper basert på opplysninger om deler av gruppen. Vi skal se hvordan statistikk kan benyttes for å teste om hypoteser er sanne.

IV. **Statistisk analyse**. I bokas tre første deler lærer du en del grunnleggende teknikker og begreper som kan brukes i ulike sammenhenger. I del IV skal vi ta disse teknikkene i bruk på nye områder. Vi skal lære å *analysere sammenhenger* mellom ulike tallstørrelser (er det en sammenheng mellom utetemperaturen og solgt mengde iskrem?) og å *sammenligne ulike grupper* (virker medisin A bedre enn medisin B?). Her skal vi også lære om spørreundersøkelser og dataanalyse, og analyse av store datamengder.

V. **Andre anvendelser**. Vi skal lære om statistisk *kvalitetskontroll*. Til slutt skal vi se hvordan vi kan *simulere* tilfeldigheter ved hjelp av en datamaskin.

Den faglige dybden blir ikke like stor i de to siste delene av boka som i de foregående. Her er hovedhensikten å gi deg en smakebit på hva statistikken kan brukes til. Kanskje du da får lyst til å lære mer? Underveis i boka finner du enkelte avsnitt som er merket med en stjerne ($*$). Dette er stoff som normalt ikke vil inngå i pensum.

1.2 Noen viktige begreper

Hensikten med statistiske undersøkelser er ofte å skaffe kunnskap om en stor mengde *enheter* (individer eller objekter). Alle de enhetene som er av interesse for oss utgjør til sammen en *populasjon*. Istedenfor å undersøke alle enhetene velger vi ut noen få. Ut fra svaret vi får fra dette *utvalget*, forsøker vi å si noe om enhetene i hele populasjonen.

1.2.1 Populasjon og utvalg

Tenk over hva de følgende utsagnene betyr. Er det noen forskjell på innholdet?

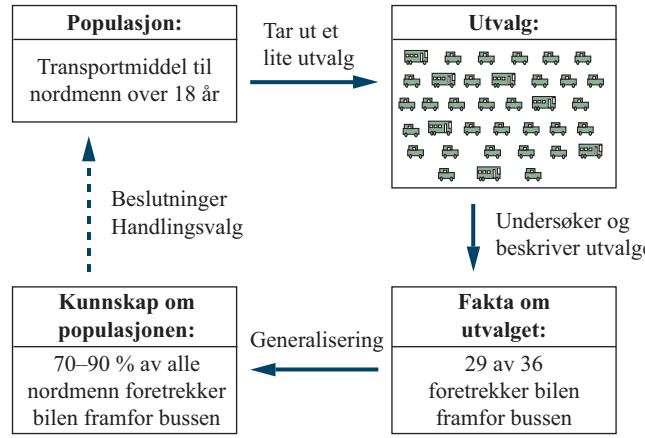
- Jonas, Kristoffer, Håvard og Amy foretrekker privatbilen. Hanna reiser kollektivt.
- 4 av 5 nordmenn foretrekker privatbilen.
- 29 av 36 nordmenn foretrekker privatbilen.
- 80 % av alle nordmenn foretrekker privatbilen.

Mange aviser har en 5-på-gaten-spalte der fem tilfeldige personer stilles et spørsmål. I onsdagsavisen uttaler Jonas, Kristoffer, Håvard og Amy at de foretrekker bilen, mens Hanna kjører kollektivt. Fire av de fem foretrekker bilen. Gir det oss grunnlag for å påstå at 80 % av alle nordmenn foretrekker bilen? Nei! Avisen må spørre flere enn fem personer for å trekke en slik konklusjon. Hvor mange må spørres for at avisen skal få et brukbart anslag på andelen som foretrekker bil? 10, 50, 100 eller 1000 personer? Vi skal senere i boka se hvordan vi kan gi et fornuftig svar på dette spørsmålet.

En populasjon er stor hvis den består av mange enheter. Det er ingenting i veien for at populasjonen kan være uendelig stor. Hver enhet kan beskrives ved verdien av en eller flere *variabler*, eksempelvis alder, vekt og bosted. I tabell 1.1 er det vist noen eksempler på populasjoner, enheter og variabler. Av eksemplene i tabellen er det bare barna i klasse 2b som er en liten populasjon. De andre populasjonene er store.

Tabell 1.1 Populasjon, enheter og variabler

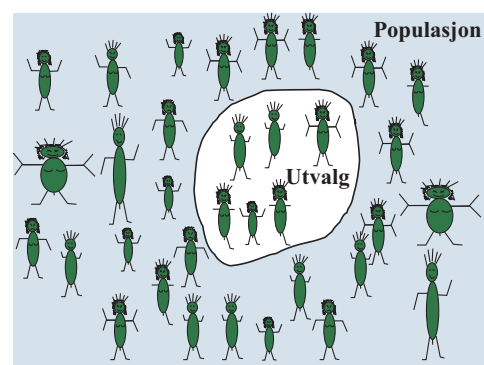
Populasjon	Enhet	Variabel (eksempler)
Alle biler i Norge	En bil	Motorytelse, Kjørt lengde, Farge, Antall seter, Antall eiere
Alle borehull i Nordsjøen	Ett borehull	Olje (ja/nei), Trykk, Dybde
Alle kvinner over 18 år	En kvinne	Høyde, Antall barn, Inntekt, Skostørrelse, Politisk parti
Alle ferskvann i Troms	Ett ferskvann	pH-verdi, Antall ørret per m^3
Alle vanndråper i Mjøsa	En vanndråpe	Antall tarmbakterier, Antall alger
Alle barn i klasse 2b	Ett barn	Vekt, IQ, Blodtype, Religion
Alle Rema-kunder	En kunde	Tilfredshet, Omsetning, Postadresse
Alle ungdommer i Norge	En ungdom	Antall sexpartnere per år, Antall sigaretter per dag



Figur 1.1 En viktig del av statistikkfaget: Å trekke konklusjoner om en hel populasjon basert på et lite, tilfeldig utvalg. Legg merke til at kunnskapen om populasjonen inneholder et element av usikkerhet.

Vi må vanligvis begrense oss til å undersøke noen av enhetene, et såkalt *utvalg* av enheter. Det skyldes at populasjonen er så stor at det er enten umulig, for kostbart eller for tidkrevende å undersøke alle enhetene. Men det kan også skyldes at enhetene må ødelegges for å bli undersøkt; smakstester av pølser, kollisjonstester av biler eller prøveutskytning av raketter er eksempler på slike destruktive tester. Faktaopplysninger om utvalgets enheter er startpunktet for alle statistiske undersøkelser. Målet med statistikkfaget er å si mest mulig (og riktig) om hele populasjonen basert på fakta fra utvalget. Å overføre fakta fra utvalget til kunnskap om populasjonen, kalles *generalisering*. Dette er illustrert i figur 1.1.

Under en arkeologisk utgravning fant man et hulemaleri som vist i figur 1.2. Det meste av maleriet var ødelagt, men en liten del av maleriet viser seks huleboere (utvalget). Arkeologene vet fra tidligere funn at dette antakelig var et maleri av alle huleboerne i stammen (populasjonen). Arkeologenes utfordring er å si mest mulig om populasjonen basert på bildet av de seks huleboerne. Arkeologene kan si ganske mye, men de mangler viktig informasjon om populasjonen. De kan jo ikke se resten av populasjonen, slik vi kan. Det er derfor stor risiko for at de trekker feil



Figur 1.2 Hva kan du si om populasjonen hvis du bare studerer utvalget?

konklusjoner av typen: «50 % av huleboerne hadde mistet en fot.» (Legg merke til at dette utvalget *ikke* er et tilfeldig utvalg, siden det bare inneholder «naboer» i huleboerpopulasjonen.)

Hvis enhetene i populasjonen er svært forskjelligartede, med stor innbyrdes variasjon, er det vanskelig å si noe om populasjonen basert på et lite utvalg. Sjansen vil være stor for at man trekker feil konklusjoner. For å få et godt bilde av slike populasjoner må vi ha opplysninger fra et stort utvalg som er representativt for populasjonen. Spørsmålet om hvordan man best gjør et utvalg, er svært sentralt i statistikkfaget. Hvis vi velger enheter fra populasjonen helt tilfeldig, får vi et *tilfeldig utvalg*. Det å foreta et utvalg og gjennomføre observasjoner for hver enhet, vil vi med en samlebetegnelse kalle et *forsøk*.

1.2.2 Variasjon

Hvis alle enheter var helt like, ville det være tilstrekkelig å undersøke *en* enhet. Så enkelt er det sjelden eller aldri. Vi skal kort nevne de tre viktigste typene variasjon som vi vil studere i statistikkfaget.

- Populasjonenes enheter er forskjellige. Hvis vi måler samme variabel for flere ulike enheter, vil vi vanligvis oppleve stor variasjon. Forskjellen fra menneske til menneske, fra bil til bil, fra snøkrystall til snøkrystall vil vi kalle *populasjonens variasjon*.

- Verdien til hver enkelt enhet kan endre seg. En terning viser forskjellig antall øyne fra kast til kast, og et fotballag oppnår forskjellig resultat fra kamp til kamp. Aksjekursene beveger seg fra dag til dag, og temperaturen i en smelteovn går opp eller ned fra time til time. Dette kalles *prosessvariasjon*. Noen ganger er det mulig å gi en prognose for utviklingen framover i tid, men i mange prosesser er det tilsynelatende tilfeldig hva som skjer fra ett tidspunkt til det neste.

- En målemetode kan gi forskjellig svar fra gang til gang når vi måler samme variabel for samme enhet flere ganger. Hvis vi måler lengden av et jernbanespor tre ganger, får vi tre forskjellige svar. Når optikeren måler øyets brytningsindeks fem ganger, får han fem forskjellige svar. Denne variasjonen kalles *måleusikkerhet*.

Når vi trekker ut en tilfeldig valgt enhet fra populasjonen på et tilfeldig valgt tidspunkt, kan vi ikke på forhånd vite verdien av en aktuell variabel. Denne verdien vil på et vis være tilfeldig, på grunn av de tre variasjonstypene som er beskrevet. Dette er grunnen til at vi senere i boka vil snakke om *tilfeldige variabler*, for å få tydelig fram at verdien til variabelen er påvirket av tilfeldigheter. Årsakene til variasjonen kan være mange, men statistikkfaget overlater til filosofi- og religionsfagene å diskutere kilden til denne variasjonen.

Det kan være vanskelig å skille de forskjellige variasjonstypene fra hverandre. Det gjelder særlig prosessvariasjonen og måleusikkerheten. Tenk på måling av blodtrykk hos en hjertepasient. Det er velkjent at blodtrykket kan variere betydelig

fra minutt til minutt. Vi vet også at de manuelle målemetodene har stor måleusikkerhet. Hvis vi foretar flere målinger etter hverandre, vil både selve blodtrykket *og* kvaliteten på målingene variere. En endring i målt blodtrykk betyr derfor ikke nødvendigvis en endring i faktisk blodtrykk. I den statistiske variansanalysen er det utviklet metoder for å beregne størrelsen på de forskjellige variasjonstypene. Dette vil vi komme tilbake til.

I fortsettelsen vil vi oftest ta utgangspunkt i populasjoner. Hver gang vi skriver «populasjon», kan du godt tenke på en prosess eller måleserie dersom det er mer dekkende for den situasjonen du studerer. Et tilfeldig utvalg fra en populasjon svarer da til «observasjoner av prosessen på uavhengige tidspunkt», eller til «gjentatte uavhengige målinger av samme fenomen».

1.2.3 Variabeltyper

Enhetene i utvalget og populasjonen kan som nevnt beskrives av visse *variabler*. Enheten «menneske» kan for eksempel beskrives av variablene alder, vekt, hvilepuls, kjønn, hårfarge, bosted, hobbyer, sosial status, antall barn, inntekt, yrke, politisk parti, skostørrelse osv. De mulige variabelverdiene kalles *kjennetegn*. Variabelen hårfarge kan for eksempel ha kjennetegnene hvit, lys blond, blond, mørk blond og svart. Legg merke til at det ikke finnes en objektivt riktig inndeling i kjennetegn. Vi kunne jo også ha inkludert hårfargene gul og brun. Inndeling i aktuelle kjennetegn er en viktig del av forberedelsene til en statistisk undersøkelse.

En måling eller registrering av kjennetegn for en bestemt enhet kalles en *observasjon*. Hans Georg Schwensen er en enhet fra populasjonen av mennesker. Vi kan observere at alderen er 35 år, vekten er 85 kg og at yrket er politibetjent. Vi har tre typer variabler:

1. **Kategoriske** variabler. Dette er kvalitative variabler som ikke naturlig kan beskrives ved et tall, men hvor enheten tilhører en bestemt kategori (gruppe med samme kjennetegn). Eksempler: kjønn, hårfarge, bosted, hobbyer, sosial status, yrke. Mye av den samfunnsvitenskapelige forskningen dreier seg om slike kategoriske data.

2. **Diskrete** variabler. Dette er kvantitative variabler der bare enkelte tall langs tallskalaen er aktuelle som kjennetegn. Tallene som ligger mellom disse på tallskalaen, gir ikke noen mening. Eksempler: skostørrelse og antall barn. Ofte vil vi bare bruke heltallene, men det kan like godt være andre tall som er aktuelle. Tenk bare på skostørrelsene oppgitt på amerikansk måte (8, $8\frac{1}{2}$, 9, $9\frac{1}{2}$ osv.). Antall barn er et eksempel på en *tellevariabel*, der bare heltallene er relevante.

3. **Kontinuerlige** variabler. Dette er kvantitative variabler der alle tallverdier innenfor et gitt intervall kan brukes for å angi et kjennetegn. Alle mulige desimaltall innenfor intervallet er altså tillatt. Eksempler: alder (år med desimalangivelse), høyde, vekt, hvilepuls og inntekt. Mye av den naturvitenskapelige forskningen dreier seg om slike kontinuerlige variabler.

I datamaskinalderen vil vi ofte bruke tallkoder også for kategoriske variabler, for eksempel slik at mann = 1 og kvinne = 2. Men her har ikke tallene i seg selv noen mening. Det er meningsløst å si at en kvinne er større enn en mann fordi 2 er større enn 1. Det gjelder generelt for kategoriske variabler at det er vanskelig å sortere verdiene. Ofte er det umulig, andre ganger lar det seg gjøre under tvil. Man kan sortere politiske partier langs en høyre-venstre-akse, men det er ikke alle som vil være enige i valget av en slik akse. Selv om vi skulle klare å sortere verdiene, er det ikke mulig å si noe om avstanden mellom de ulike variabelverdiene. Er avstanden mellom hobbyene frimerkesamling og myntsamling mindre enn avstanden mellom frihåndstegning og vektløfting? Svaret er ikke opplagt, siden det ikke finnes noe avstandsmål for kategoriske variabler. Av denne grunn er det umulig å beregne gjennomsnitt for kategoriske variabler. Det gir eksempelvis liten mening å snakke om gjennomsnittlig kjønn i en populasjon.

I mange situasjoner vil skillet mellom diskrete og kontinuerlige variabler være lite merkbart. Ta pulsmåling som eksempel: De fleste pulsmålere oppgir resultatet som et heltall. Unøyaktigheten i målemetoden er såpass stor at det ikke har noen hensikt å angi målingen med desimaler. Det fører til at den kontinuerlige variabelen framstår som en diskret variabel. Forskjellen på diskrete og kontinuerlige variabler har bare praktisk interesse hvis den diskrete variabelen har få mulige verdier. Ellers kan vi tenke på de fleste kvantitative variabler som kontinuerlige.

1.3 Modeller

En modell er en etterligning av virkeligheten. Modeller *kan* gi oss verdifull innsikt som det enten er umulig eller kostbart å skaffe direkte fra virkeligheten. Statistikkfaget gjør flittig bruk av modeller, fordi modellene kan være svært nyttige for oss når vi skal fatte beslutninger.

Fysiske modeller som et skip i miniatyr eller en prototyp av en bil, har lenge vært viktige for ingeniørene. Modellene kan utsettes for påkjenninger, og vi kan se hvordan modellen påvirkes av dette. Vi observerer, forandrer modellen og utsetter den for nye påkjenninger. Etter hvert er det vanligere med simuleringsmodeller, som lar oss eksperimentere med datamaskinen. Du har kanskje forsøkt deg som bilfører i et dataspill, eller som pilot i en flysimulator? Dette er også modeller av de systemene som simuleres. Tilsvarende finnes det spill som simulerer samfunnsøkonomien i et tenkt samfunn basert på matematiske modeller utviklet av sosialøkonomene. Avanserte 3D-modeller og visualiseringsteknikker, eksempelvis VR-briller, gjør modellene stadig mer realistiske.

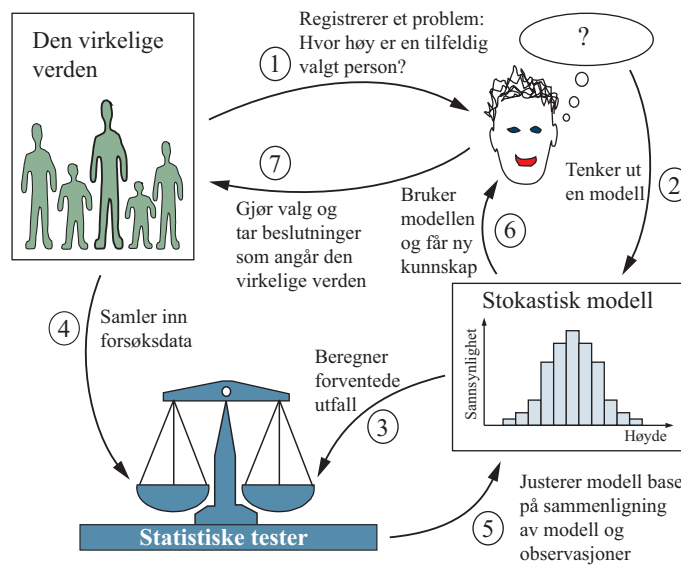
Du må aldri glemme at en modell er en *forenkling* av virkeligheten. Modellen baserer seg alltid på visse antakelser og forutsetninger. I visse situasjoner vil modellen stemme svært godt overens med virkeligheten, mens den i andre situasjoner oppfører seg helt annerledes enn virkeligheten ville gjort. Når vi skal fatte beslutninger basert på kunnskap fra modeller, må vi vite når modellen er god, og når den er ubrukelig.

Det finnes to hovedtyper av modeller:

- En *deterministisk modell* beskriver et fenomen som er forutsigbart. I matematikk- og fysikkfagene har du kanskje lært formler for kinetisk energi $\left(\frac{1}{2}mv^2\right)$ og for volumet av en kule $\left(\frac{4}{3}\pi r^3\right)$. Slike formler gir samme resultat hver gang. Det er ingen tilfeldigheter involvert! De fleste naturlovene du har lært i din tidligere skolegang er slike deterministiske teorier og modeller.

- En *stokastisk modell* beskriver et fenomen som involverer tilfeldigheter, som derfor er uforutsigbart. Modeller som beskriver antall oppringninger til sentralbordet i morgen, eller antall timer til neste strømstans, eller vekten av morgendagens fiskefangst, vil vanligvis være stokastiske. En stokastisk modell kan ikke si hva som kommer til å skje, men den kan angi hvilke utfall som er mulige, og hvilken sannsynlighet de ulike utfallene har.

Statistikkfaget omhandler usikkerhet. Rent deterministiske modeller er derfor ikke av interesse i statistikkfaget. Vi kan godt si at statistikkfaget er læren om utvikling og bruk av stokastiske modeller – som hjelp til å fatte beslutninger. Figur 1.3 viser en litt abstrakt skisse av statistikkfagets viktigste elementer. I figuren ser du at en stokastisk modell står sentralt plassert. Modellen blir hele tiden oppdatert på bakgrunn av data om virkeligheten. Og modellen danner basis for de beslutningene vi skal fatte.



Figur 1.3 En generell beskrivelse av statistikkfaget. Hovedpoenget er å være til hjelp når beslutninger skal fattes.

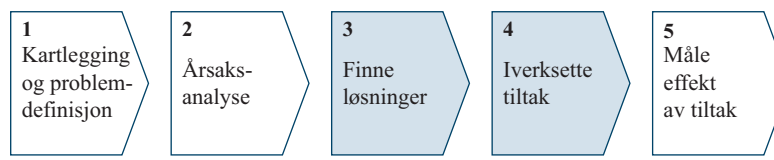
1.4 Innsamling av data

Utgangspunktet for hele statistikkfaget er at det foreligger et problem eller spørsmål. Se figur 1.3. Vi lager en modell for å ta gode beslutninger som knytter seg til problemet. For å lage brukbare modeller må vi ha god kunnskap om virkeligheten. Denne kunnskapen skaffer vi oss ved å samle inn data. Statistikkfaget kan benyttes for å analysere de innsamlede dataene, men faget kan *ikke* hjelpe deg til å stille de riktige spørsmålene og til å velge hvilke data som skal samles inn. Du må bruke din egen sunne fornuft!

1.4.1 Problemløsning krever data

En problemløsningsprosess kan deles inn i fem faser: kartlegging, årsaksanalyse, problemløsning, iverksetting og effektmåling, som vist i figur 1.4. Datainnsamling spiller en sentral rolle i fasene 1, 2 og 5. I praksis vil du ofte oppleve at man går direkte løs på fase 3, uten å gjøre en skikkelig jobb på de to første fasene, noe som lett kan resultere i dårlige beslutninger.

I fase 1 må vi definere noen variabler som er egnet til å beskrive problemets omfang. Slike variabler kalles gjerne *kvalitetsindikatorer* eller *nøkkeltall*. Vi må samle inn måleverdier for en eller flere slike variabler. Disse verdiene vil vise om vi har et reelt problem som fortjener videre oppmerksomhet. Hovedoppgaven i fase 1 er å finne fakta om problemet – uten å være forutinntatt. Det kan være lurt å stille spørsmål som starter med NÅR, HVOR, HVEM.



Figur 1.4 En vanlig prosess for problemløsning.

Som eksempel skal vi se på et pizzabudfirma i Oslo som har problemer med forsinkede leveranser. Firmaet velger leveringstiden som kvalitetsindikator. I første halvår har de hatt hele 27 % forsinkede leveranser i forhold til sin garanti om å levere innen en time. I fase 1 stiller vi spørsmålene nedenfor:

- *Når* oppstår problemet? Er det i rushtiden, i snøvær eller på varme sommerdager? Hvor store er forsinkelsene?
- *Hvem* er involvert? Er nyansatte mer forsinket enn andre? Er det spesielle kundegrupper som opplever problemet?
- *Hvor* oppstår problemet? Er det i bestillingen, på kjøkkenet eller i budbilen at tidsforbruket er for høyt? Er det spesielle deler av byen som har problemer?

I fase 2 spør vi HVORFOR. En idédugnad kan være nyttig for å få fram mange *hypoteser* om mulige årsaker. Som hjelp til å framsette hypoteser finnes 5M-regelen, som sier at årsakene kan knyttes til Menneske, Miljø, Materiell, Metode eller Maskin. Vi forsøker å grave oss stadig dypere ned i årsakens mysterium. Det kan vi gjøre ved hele tiden å spørre hvorfor, med utgangspunkt i hypotesene fra forrige hvorfor-spørsmål. Vi må også framsette hypoteser om hvilke årsaker som er viktigst, slik at vi i fase 3 kan forsøke å eliminere de viktigste delene av problemet. Datainnsamlingens funksjon i denne fasen er å skaffe fram fakta slik at vi kan teste om hypotesene våre er riktige.

Første fase viste at de nyansatte sjåførene ofte var involvert i forsinkelsene. I andre fase er det da naturlig å spørre:

- *Hvorfor* er de nyansatte sjåførene involvert? Hypoteser: finner ikke fram i byen, kjører de lengste avstandene, har de dårligste bilene.
- *Hvorfor* finner de ikke fram? Hypoteser: har ikke kjentmannskurs, blir sendt til de vanskeligste adressene, har ikke GPS, har liten erfaring.
- *Hvorfor* blir de nyansatte sendt til de vanskeligste adressene? Hypoteser: erfarne ansatte får velge turer først, erfarne ansatte flytter til de mest attraktive bydelene.

Fase 5 er viktig, men mangler i mange beskrivelser av en problemløsingsprosess. I denne fasen skal vi vende tilbake til den kvalitetsindikatoren som ble definert i starten av prosessen. Hensikten med fase 5 er å fortsette målingene for å se om de tiltakene som ble iverksatt virkelig har hjulpet på problemet. Pizzabudfirmaet må følge med for å se om de klarer å overholde garantitiden sin! Det er kanskje ikke nok å gi de vanskeligste adressene til de erfarne sjåførene.

Husk at en sammenheng mellom to variabler ikke nødvendigvis er det samme som en årsakssammenheng. Kaffedrikkere kan ha et dårligere sexliv enn andre, men det trenger ikke nødvendigvis å skyldes kaffen. Kanskje det skyldes at de står opp for tidlig om morgenen?

1.4.2 Generelt om datainnsamling

Kvaliteten av våre analyser kan aldri bli bedre enn kvaliteten av de dataene vi samler inn. Kvaliteten av dataene er nøye knyttet til svarene på følgende spørsmål:

- Er utvalget stort nok?
- Er utvalget representativt for hele populasjonen?
- Har målingene pågått lenge nok?
- Har observatøren påvirket målingene?
- Er variabler og kjennetegn entydig og fornuftig definert?
- Er målingene nøyaktige?

Det kan være stor forskjell på *planlagte forsøk* og mer *vilkårlige registreringer*. Planlagte forsøk vil normalt være å foretrekke, men det finnes mange problemstillinger der det er umulig eller uetisk å gjennomføre forsøk. Fordelen med planlagte forsøk er at vi kan kontrollere forsøksbetingelsene, slik at det blir lettere å trekke konklusjoner.

Tenk deg et forsøk der noen tusen personer ble bedt om å kjøre med sikkerhetsbelte bare annenhver dag, for å kunne vurdere effekten av beltebruk. Risikoen for personskader gjør at dette vil være uaktuelt. Vår eneste mulighet er å observere resultatet av bilulykker og se hvordan det går med førerne. Det kan tenkes at de som kjører uten bilbelte, i større grad enn andre har nye, kollisjonssikre biler med airbag. Hvis disse klarer seg bra uten bilbelte, betyr det ikke nødvendigvis at bilbelter er unødvendige.

Ett av statistikkfagets hovedmål er å gjøre deg til en kompetent mottaker av statistisk informasjon. Ikke ta alt du leser for «god fisk»! Still alltid spørsmålet: Kan resultatene ha en annen forklaring enn den som blir påstått?

Media påvirkes av mange aktører som har politiske eller økonomiske baktanker med sine budskap. Tenk alltid på hva de er ute etter å oppnå, og hvem som har finansiert, utført og publisert forsøksresultatene. Forestill deg et forsøk som utføres for å påvise at et nytt produkt er bedre enn andre produkter på markedet. Produsenten av det nye produktet vil neppe offentliggjøre forsøksresultater som viser at han ikke er best eller nest best. Han kan betale ulike seriøse forskningsinstitutter for å gjøre forsøkene, og så velge å offentliggjøre bare de rapportene han liker. Husk i tillegg å stille de seks spørsmålene ovenfor når du mottar spektakulære nyhetsmeldinger. Da vil du ofte oppdage at nyhetsmeldingene ikke er verdt papiret de er skrevet på.

1.4.3 Spørreundersøkelser

Senere i boka, i kapittel 9, skal vi se grundig på spørreundersøkelser. Her er en liten smakebit og noen advarsler.

Objektive målinger er normalt bedre enn subjektive utsagn. Det er bedre å måle sentralbordets responstid elektronisk enn å spørre innringeren om han/hun måtte vente litt, lenge eller veldig lenge. Spørreundersøkelser bør derfor ikke brukes ukritisk. Undersøk alltid om du kan få bedre informasjon på en annen måte.

Svarene fra undersøkelsen må være representative: Spørreskjemaene må sendes til et representativt utvalg, *og* et representativt utvalg av disse igjen må svare på skjemaet. Det samme gjelder selvsagt hvis undersøkelsen foretas via telefon, sms, e-post eller Internett. Et eksempel på særdeles lite representative målinger er slike «ring inn og gi din stemme under debatten»-målinger som etter hvert har blitt populære i TV. Hvem tror du det er som ringer inn?

På samme måte som vi må bruke riktig innstilt måleutstyr i eksperimenter, må også spørsmålsstillingen i et spørreskjema være presis, forståelig og minst mulig ledende. Det er gjennomført en mengde undersøkelser som viser hvor mye svarene kan endres bare ved en liten endring i spørsmålsteksten. Ved større justeringer i teksten kan man oppnå omtrent de svarene man vil ha.

La oss tenke oss at Greenpeace ønsker å utføre en spørreundersøkelse for å kartlegge andelen av nordmenn som ønsker å stanse hvalfangsten. Nedenfor er det foreslått to spørsmål som begge tar sikte på å anslå denne andelen. Hvilket spørsmål tror du miljøorganisasjonen ville valgt? Hva tror du hvalfangernes organisasjon ville valgt om de skulle gjøre en egen undersøkelse? Tror du de ville funnet samme svar?

1. Er du enig i at hvalfangst, drap på verdens største, utryddingstruede pattedyr, må forbys snarest mulig?
2. Er du enig i at hvalfangst må tillates i begrenset omfang, for å sikre bærekraftig ressursforvaltning og kystbefolkningens livsgrunnlag?

I spørreundersøkelser kan folk ha en tendens til å svare det de tror spørreren ønsker seg, og/eller ha en frykt for å avsløre egne synspunkter i betente saker. Derfor anbefales det vanligvis at *respondenten* (den som fyller ut skjemaet) får være anonym. Tidspunktet for når målingene utføres, kan også være av stor betydning. Tenk på en spørreundersøkelse om folks holdninger til prøveløslatelser fra fengslene. Anta at undersøkelsen utføres dagen etter at en voldtektsdømt har begått en ny, brutal voldtekt mens han var på permisjon. Svarene på denne undersøkelsen vil garantert bli annerledes enn om de samme spørsmålene ble stilt en uke tidligere. På samme måte kan også rekkefølgen av spørsmålene spille en stor rolle, siden alle spørsmålene styrer tankene til respondenten.

Hvorfor skjule opplysninger?

Publisert: 27.mai. 2009 19:00 Oppdatert: 27.mai. 2009 14:27

I en kronikk i Aftenposten 26. mai forteller Kari Vigeland at det er «flertall for dødshjelp» i det norske folk. Hun viser til en undersøkelse fra Synovate til støtte for dette budskapet, og hun gjengir noen av spørsmålene og svarprosentene som er kommet frem.

Som statistiker reagerer jeg umiddelbart på fremstillingsmåten. Det er svært enkelt å lage undersøkelser som gir denne typen resultater. Erfaring med utforming av spørreskjemaer viser at ordvalgene i spørsmålene, og ikke minst rekkefølgen av dem, kan manipulere folk i overraskende stor grad. Her benyttes det klassiske trikset med å spørre først om litt (bare få ørlite hjelp til selv å ta sitt eget liv), som jo kanskje ikke høres så ille ut, og deretter om litt mer, og til slutt mye. Denne gradvise overgangen fanges ikke opp og skaper ikke tilstrekkelig refleksjon hos den som blir stilt overfor spørsmålene. Dersom spørsmålene ble stilt i motsatt rekkefølge, ville svarene garantert blitt annerledes! Ordvalgene er minst like viktige. De som utformer spørsmålene vil, enten bevisst eller ubevisst, påvirke de som skal svare. Kari Vigelands kronikk pirret min nysgjerrighet. Både fordi den handler om et viktig tema som fortjener skikkelig behandling før det tas store beslutninger – og fordi den umiddelbart gir meg følelsen av at her misbrukes statistikkfaget. Det var derfor ikke på noen måte overraskende å finne ut via Google at undersøkelsen Synovate har utført, ble gjort på oppdrag for foreningen «Foreningen retten til en verdig død». Og at Kari Vigeland er oppført som styremedlem i foreningen. Hvorfor skjuler hun disse to opplysningene for Aftenpostens lesere?

GUNNAR G. LØVÅS

Figur 1.5 Vær varsom når du leser i avisen om spørreundersøkelser. (Kilde: Aftenposten.no)

I denne boka vil du lære hvordan svarene fra en spørreundersøkelse kan analyseres. Selve utformingen av spørsmål og skjemaer er en egen vitenskap som vi ikke vil se nærmere på. Som hovedregel anbefales det å gjennomføre en forundersøkelse blant et lite utvalg respondenter, en såkalt *pilotstudie*, der respondentene selv får formulere sitt svar. Basert på disse svarene kan man definere noen standard svaralternativer og luke bort misforståelser og tvetydigheter. Disse svaralternativene brukes i hovedundersøkelsen, der respondenten bes om å velge ett av svaralternativene. Vanligvis anbefales det at respondenten også tilbys å krysse av i en vet-ikke-boks. Uten et slikt svaralternativ vil mange gi uttrykk for flere meninger enn de egentlig har.

1.5 Oppsummering

Når du har lest dette kapitlet, bør du

- kjenne statistikkfagets basisdefinisjoner: populasjon, utvalg, variabel, variasjon, observasjon
- vite forskjell på kategoriske, diskrete og kontinuerlige variabler
- forstå at observasjoner fra et tilfeldig utvalg ikke sier alt om en populasjon
- være klar over at datainnsamling må planlegges godt
- vite at statistikkfaget kan hjelpe deg til å ta gode beslutninger