Statistikk og dataanalyse

Njål Foldnes,
Steffen Grønneberg,
Gudmund Horn Hermansen og
Einar Christopher Wellén

Statistikk og dataanalyse

En moderne innføring

2. utgave

CAPPELEN DAMM AKADEMISK

ISBN 978-82-02-83512-5

1. utgave, 1. versjon, 2024
E-utgaven baserer seg på 2. trykte utgave, 1. opplag, 2024,
Cappelen Damm Akademisk

Elektronisk tilrettelegging: HAVE A BOOK
Omslagsdesign: Kristin Berg Johnsen
Illustrasjoner: David Keeping og HAVE A BOOK. Illustrasjon s. 21 og s. 24: Einar Wellén s. 36 (Figur 2.1): Njål Foldnes, s. 51 (Figur 4.1) iStock/Getty Images

www.cdu.no
akademisk@cappelendamm.no

FSC
www.fsc.org
MIX
Papir fra bærekraftige kilder
FSC® C047286

Miljøfyrtårn®
Cappelen Damm AS
er Miljøfyrtårn-sertifisert

Forord

Statistikk som et verktøy til å forstå verden

Velkommen til statistikkfaget! Her skal du lære om hvordan du kan analysere data og bruke det til å ta bedre avgjørelser.

Vi får stadig tilgang til mer og mer data, og etterspørselen etter folk som kan tolke kvantitativ informasjon, er økende.

Denne boken er en introduksjon til hvordan man kan bruke tall og data til å finne ut hvordan verden er skrudd sammen.

Verden blir mer og mer kvantitativ – det samles inn mer og mer digitalisert informasjon. Ta for eksempel Wal-Mart, en amerikansk kjøpesenterkjede med en årlig omsetning på størrelse med Norges brutto nasjonalprodukt. Wal-Mart lagrer informasjon om mer enn én million handler hver eneste time! Alle slags organisasjoner lagrer data om sin virksomhet, og de beste bruker dette for alt det er verdt for å maksimere verdiskapingen sin. Her hjemme samler Statistisk sentralbyrå inn store mengder kvantitative data fra alle samfunnsområder. Denne informasjonen blir brukt til å analysere sammenhenger og trender. Dette er viktig informasjon til politikere og andre beslutningstakere.

Kvantitativ Informasjon: Informasjon som omhandler tall og målinger. I motsetning til kvalitativ informasjon. Alt som kan måles og telles er kvantitativ informasjon. Kvantitative metoder er altså matematiske og statistiske teknikker som hjelper oss å forstå tallmateriale.

Med moderne datateknologi kan du selv gjøre kvantitativt arbeid og oppdage nye sammenhenger. Noen få (men smarte) tastevalg i regneark som Excel eller statistisk programvare som JMP, R eller SPSS kan gi ny informasjon som kan hjelpe din bedrift. Etterspørselen etter folk som kan tolke og «knuse» tall kommer garantert til å vokse også i årene framover.

Vi som underviser i statistikk vet godt at mange studenter har et anstrengt forhold til tall og matematikk. Mange studenter i samfunnsfagene føler seg distansert fra tall og regning, og mange har dessverre negative erfaringer fra skolen. Om dette skulle gjelde deg, også – prøv likevel å være åpen! I denne boken kreves det ingen ekstraordinære matematikkferdigheter. Likevel viker vi på ingen måte unna dette viktige faget, for målet vil alltid være å forstå dataene, og da trenger vi noe matematikk, rett og slett.

I denne boken legger vi vekt på tenkning og forståelse. Du vil lære å tolke resultatene av statistisk analyse i den sammenhengen du vil bruke dem i.

Formler og formelle statistiske prosedyrer vil være en del av kurset, men til syvende og sist er det din forståelse av analysen som skal være grunnlaget for avgjørelser du tar.

Enkelte overskrifter i boken er merket med (*). Dette betyr at teksten som følger er fordypningsstoff, og ikke behøver å leses grundig ved første gjennomgang av boken.

Slik er boken organisert

Boken har fire hovedbolker. Det kan være greit å få en oversikt over disse bolkene før man begynner å lese. De første tre bolkene fokuserer på statistisk analyse av én variabel mens den siste og fjerde bolken handler om analyse av to eller flere variabler.

Vi starter i **bolk 1** med å introdusere hovedideene i en statistisk analyse. Hovedmålet med boken er å få deg til å utføre slike analyser på egenhånd. Vi presenterer en skjematisk oversikt over det vi kaller de fire elementene i en statistisk analyse. Dette gir et godt utgangspunkt for å lese boken. Du vil også bli gjort kjent med de mest grunnleggende begrepene i statistikk og herunder bli kjent med de ulike variabeltypene.

Vi tar deg deretter gjennom de tre første stegene i en statistisk analyse. Det første steget er å definere problemstillingen du ønsker å løse. Deretter lærer vi deg å trekke et representativt utvalg (andre steget i en statistisk analyse). Til slutt tar vi deg gjennom det tredje steget i en statistisk analyse, som handler om å beskrive data fra utvalget med nøkkeltall (f.eks. gjennomsnitt) og grafer (f.eks. histogram).

Det siste steget er inferens, som betyr hvordan vi kan gå fra informasjon om utvalget til å si noe om hele populasjonen utvalget er trukket fra. Dette steget kommer først i bolk 3, og er det mest sofistikerte steget. For å kunne gjøre og forstå inferens må vi først studere sannsynlighet. Dette gjøres i bolk 2 av boken.

Bolk 2 tar for seg sannsynlighetsregning og består av regneregler og bruk av matematiske teknikker. Her er det altså mange formler. Vi lærer regler om sannsynligheten av hendelser, og vi introduserer begrepet tilfeldig variabel, og dens sannsynlighetsfordeling. Kunnskap om sannsynlighetsregning er nødvendig for å kunne gå fra steg 3 (beskrivende statistikk) til steg 4 (inferens) i en statistisk analyse, men vi lærer også sannsynlighetsregning fordi vi ofte trenger å bruke sannsynlighetsregning direkte på problemstillinger innen økonomi, finans, markedsføring og andre fag.

Bolk 3 handler om inferens: hvordan kan vi gå fra informasjon om utvalget til å si noe om hele populasjonen som utvalget er trukket fra. Inferens er steg 4 og det siste steget i den statistiske analysen. Når vi utfører inferens svarer vi på problemstillingen fra steg 1 i den statistiske analysen. For å kunne utføre inferens kreves kjennskap til sannsynlighetsregning (broen mellom steg 3 og steg 4 i en statistisk analyse) og bruk av nøkkeltall fra den beskrivende analysen (steg 3 i en statistisk analyse). Inferens gjøres ved å beregne konfidensintervaller og utføre hypotesetester.

I **bolk 4** lærer vi å analysere sammenhengen mellom to variable. Vi studerer samvariasjon mellom to variable via simultane sannsynlighetsfordelinger, regresjon og khikvadrattester. Kan du regresjon, har du et godt grunnlag for å studere mer avanserte statistiske modeller i videregående kurs.

Kapitteloversikt

Innhold

KAPITTEL 22

Enkel regresjon 536

KAPITTEL 23

Samvariasjon for to kategoriske variable 570

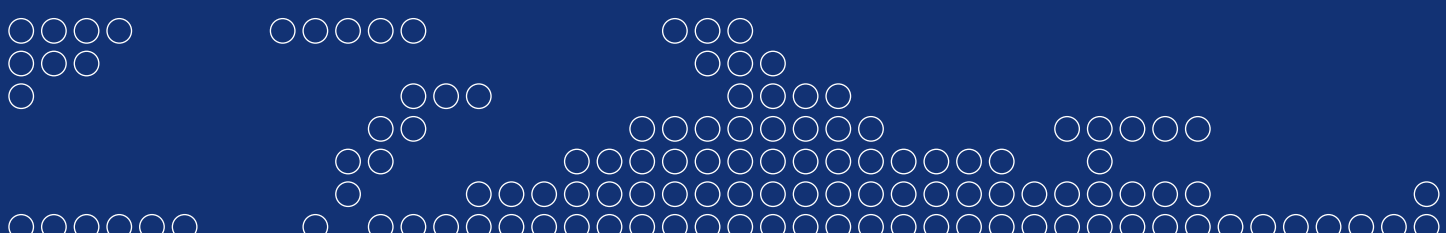
VEDLEGG 583

STIKKORDREGISTER 592

Bolk

1

Introduksjon til statistikk



1 Introduksjon til statistikk

I dette kapitlet vil du få en kort introduksjon til hvordan en statistisk analyse kan utføres samtidig som du lærer noen grunnleggende begreper i statistikk. Hovedmålet med denne boken er nettopp å gjøre deg i stand til å gjennomføre statistiske analyser på egenhånd.

Statistiske analyser består av to ting: Vi må ha noe å analysere, og dette kalles data, og vi må ha noen analyseteknikker, og dette kalles statistiske metoder. Du vil gjennom denne boken bli kjent med ulike statistiske metoder (analyseteknikker), samt hvordan data bør samles inn.

Statistiske utregninger, som inngår i en statistisk metode, overlater vi i stor grad til programvare, men et statistisk program gir ingen forståelse. Det er i vår tolkning av datamaskinens resultater at vi kan oppnå ny kunnskap som er gyldig for problemstillingen vår. Vi vil derfor, i denne boken, gi deg nok innsikt til å kunne forstå og tolke statistiske utregninger på en god måte. Dette vil vi blant annet gjøre ved å gjøre deg kjent med og anvende mange av formlene, som datamaskinen bruker i de statistiske utregningene, på små datamengder.

Vi må kjenne til mulighetene og begrensningene som ligger i statistikkfaget. Se på dette kapitlet som en første grunnstein til å bygge opp en god forståelse av fagets karakter og særpreg.

1.1 Hva er statistikk og hvorfor er det viktig

Statistikk er vitenskapen om hvordan vi samler inn og analyserer data for å få kunnskap til å ta best mulige avgjørelser. I mange bedrifter og organisasjoner er det viktig å få svar på spørsmål som:

- Hva er gjennomsnittlig køtid i vår fast-food-restaurant?
- Hvilken aldersgruppe er mest lojal mot merkevaren vår?
- Hvor mye har utdanning å si for lønnen vår?

Statistikk er et «verktøyfag» som kan hjelpe deg å svare på slike spørsmål, men du definerer spørsmålene selv, og hva du vil gjøre med svarene. Mange bedrifter sitter på store mengder data som bare venter på å bli utforsket. Statistiske metoder kan hjelpe til å oppdage og tallfeste sammenhenger som kan gi bedre grunnlag for å ta beslutninger. Hvis man ikke har data tilgjengelig, kan statistiske metoder fortelle hvordan vi bør hente inn dette og hvordan man tolker og trekker konklusjoner basert på dataene, samt hvordan man bør presentere dette.

Statistisk analyse er et sentralt verktøy i realfag og samfunnsvitenskap, som økonomi, finans, markedsføring og psykologi. I offentlig forvaltning og i politikken tas beslutninger ofte med henvisning til statistiske analyser. Og ikke minst kan du bedre forstå saker i media dersom du har satt deg inn i hvordan statistisk analyse fungerer. Politiske debatter føres ofte med henvisning til tall og analyser, og det er viktig at du da kan gjøre deg opp en egen kvalifisert mening om hvor relevante disse henvisningene egentlig er for den politiske argumentasjonen.

En skulle kanskje tro at det å analysere og tolke data er en ukomplisert affære. Men dataene vil sjelden fortelle hele historien om den problemstillingen eller den situasjonen vi ønsker å forstå. Det er nesten alltid en viss grad av tilfeldighet i dataene vi har samlet inn, eller har fått tilgang til. Dersom vi samler inn data to ganger, vil resultatene bli litt forskjellige, siden *tilfeldighetene* alltid spiller inn. Et særpreg med statistikkfaget er at det hele tiden prøver å filtrere ut det tilfeldige for å finne ut hvordan ting virkelig er.

I denne boken gis kun en innføring til de aller mest grunnleggende og sentrale statistiske metodene som brukes. Disse metodene er viktige i praksis, men mange metoder og datasituasjoner er ikke dekket i denne boken, for eksempel metoder for å analysere økonomiske datasett med tidsvariasjon slik som aksjekurser og maskinlæringsmetodikk. I videre kurs kan du lære om mer avanserte metoder, og disse bygger på teorien beskrevet i denne boken. I Seksjon 1.5 sier vi litt mer om hvordan dette henger sammen.

1.2 Variabler: Envariabelstatistikk og flervariabelstatistikk

En variabel i statistikk er noe man kan måle for de enhetene man analyserer. Hvis man jobber med kundedata er enhetene personer. Her er for eksempel kjønn og salgssum variabler, det vil si informasjon man kan ha for hver kunde. Hvis man jobber med økonomiske datasett er det ofte bedrifter som er enhetene, og variabler kan være fortjenestemargin og driftskapital per bedrift for et gitt år. En grundigere introduksjon til variabler gis i kapittel 2.

Boken er delt opp slik at de tre første bolkene handler om å utføre statistikk på en enkelt variabel. I den fjerde og siste bolken vil vi se på samspillet mellom to variable.

Gjennomsnittlig køtid i en fast-food-restaurant avhenger kun av én variabel: køtiden per kunde. Det er kun én ting man trenger å registrere per kunde. Men hvor mye utdanning har å si for lønn er et spørsmål om hvordan to variabler er koblet sammen: utdanningsnivå og lønn. For å undersøke dette må vi ha begge disse to tallene for hver person i undersøkelsen for å si noe om hvordan de henger sammen.

Tovariabelstatistikk er startpunktet for mer avanserte statistiske analyser som dekkes i senere metodekurs, slik som økonometri: En analyse av utdanningsnivå og lønn kan virke som at det kun omhandler disse to variablene, men for å forstå sammenhengen mellom utdanningsnivå og lønn ordentlig må man ta hensyn til ytterligere variabler som bosted, kjønn, og liknende. Dette er praktisk viktig, men en omfattende analyse av denne problemstillingen er utenfor bokens omfang.

1.3 Elementene i en statistisk analyse – boken oppsummert i et bilde

I denne boken skal vi fokusere på en viktig klasse statistiske problemer: situasjoner der datasettet vi jobber med er et utvalg fra en populasjon. Dette er den enkleste situasjonen der statistikk benyttes. Statistikk brukes også i mange andre situasjoner, og vi skal skissere noen på slutten av dette kapitlet, men for enkelhets skyld skal vi stort sett snakke om statistiske analyser som om de alltid omhandler utvalg fra en populasjon.

1.3.1 Populasjon og tilfeldige utvalg

Vi definerer nå begrepene populasjon og tilfeldige utvalg, som vi skal diskutere mer inngående i kapittel 4.

En **populasjon** består av samtlige studieobjekter (for eksempel alle personene eller firmaene du ønsker å studere).

Et **tilfeldig utvalg** (ofte kalt stikkprøve) er den delen av populasjonen som du har samlet inn data om. Utvalget utgjør altså vanligvis bare en liten del av populasjonen, og den skal **trekkes tilfeldig**. En måte å trekke tilfeldig fra en populasjon er å gi alle i populasjonen et tall, skrive ned alles tall på små papirlapper i en stor krukke, blande godt, og så trekke opp like mange lapper som du vil ha i utvalget ditt. Datamaskiner kan trekke tilfeldig for oss, og det er dette som gjøres i praksis.

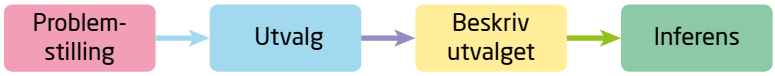
Grunnen til at vi ikke innhenter dataene fra alle (spør hele populasjonen), er at det koster for mye og tar for lang tid, eller at det rett og slett er umulig.

Utvalget tas altså fra en populasjon og som oftest er målet å få mest mulig sikker kunnskap om hele populasjonen, basert på informasjonen i utvalget. Vi studerer hvordan ting henger sammen i utvalget, og prøver så å generalisere dette til hele populasjonen. Statistisk analyse dreier seg altså om å lære mest mulig om hele populasjonen, basert på informasjonen fra et mindre utvalg.

Det er to viktige grunner til at vi tar et **tilfeldig utvalg**, og ikke bare et utvalg som vi tror blir bra: For det første vil vi ved et tilfeldig utvalg få en jevn blanding av folk fra hele populasjonen – for eksempel vil man ikke bare velge de man kjenner selv, som kanskje ikke er typiske for resten av populasjonen. Dette skal vi diskutere grundig i kapittel 4. Den andre grunnen til at man vil ta et tilfeldig utvalg er at vi da kan bruke sannsynlighetsregning for å analysere hvordan aspekter ved utvalget, slik som gjennomsnittet av en variabel, gjenspeiler aspekter ved populasjonen.

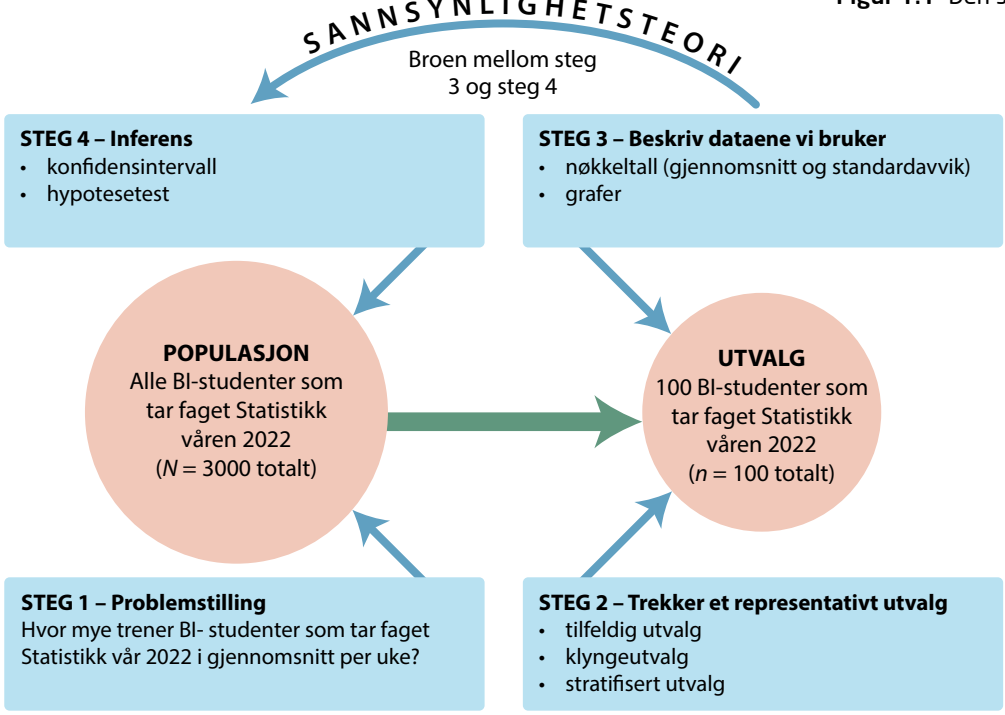
Et utvalg som er tilstrekkelig stort og trukket helt tilfeldig, eller med høy grad av tilfeldighet, sier vi er representativt, fordi det avspeiler populasjonen på en god måte.

1.3.2 De fire stegene (elementene) i en statistisk analyse



Hovedmålet med denne boken er å gjøre deg i stand til å gjennomføre en statistisk analyse på egenhånd. Målet med en statistisk analyse er å få kunnskap om populasjonen selv om vi ikke har muligheten til å spørre alle i populasjonen. Den typen statistisk undersøkelse vi skal jobbe med i boken foregår i fire hovedsteg (studer figur 1.1, Den statistiske analysen, nøye når du leser om de fire stegene nedenfor):

Figur 1.1 Den statistiske analysen



1) Vi definerer en problemstilling – se figuren over

I en statistisk analyse må vi først definere problemstillingen vi ønsker svar på. For eksempel ønsker vi å finne ut hvor mye BI-studenter, som tar faget Statistikk våren 2022, trener i gjennomsnitt per uke. MERK: Problemstillingen gjelder faget alltid populasjoen, som her er alle BI-studenter som tar faget Statistikk våren 2022.

I kapittel 3 vil vi se nøye på dette første elementet i en statistisk analyse.

2) Vi trekker et representativt utvalg – se figur 1.1

Etter å ha definert problemstillingen må vi trekke et representativt utvalg. For eksempel trekker vi 100 BI-studenter som tar faget Statistikk våren 2022. Disse 100 studentene utgjør utvalget vårt. Det er mange tenkelige utvalg, men i en gitt situasjon sitter du med data bare fra ett utvalg. Du må klare deg med det ene utvalget og håpe at det inneholder nok informasjon om hele populasjonen. Problemet er som sagt at det er en del tilfeldigheter som avgjør akkurat hvilket utvalg du har. Hvert utvalg avspeiler populasjonen på sin egen måte. Men hvis utvalget er tilfeldig og stort nok, inneholder det svært presis informasjon om populasjonen.

I kapittel 4 vil vi se nøye på hvordan vi kan trekke utvalget og ulike utvalgsmetoder.

3) Vi beskriver utvalget ved hjelp av nøkkeltall og grafer (beskrivende statistikk) – se figur 1.1

Vi gjør nå beregninger og lager grafer basert på utvalget. Vi beregner for eksempel gjennomsnittet til antall treningstimer for de hundre studentene i utvalget. Gjennomsnittet vi får her sier kun hva gjennomsnittet til de hundre studentene i utvalget er. Det er viktig å merke seg at gjennomsnittet i utvalget ikke sier hva gjennomsnittet i populasjonen er, som jo er det problemstillingen vår ønsker svar på. I punkt tre i en statistisk analyse er det altså utvalget vi fokuserer på. Dette fokuset utdypes i kapittel 5 der vi beregner ulike nøkkeltall for dataene våre og viser hvordan vi kan lage grafer som framstiller informasjon fra utvalget visuelt.

4) Inferens (konfidensintervaller og hypotesetester) gir informasjon om populasjonen – se figur 1.1

I det siste og fjerde steget i en statistisk analyse svarer vi på problemstillingen som ble formulert i steg 1. Vi kan for eksempel gjøre dette ved å lage et konfidensintervall som angir med stor grad av sikkerhet at BI-studenter som tar faget Statistikk våren 2022, trener i gjennomsnitt et sted mellom 2.3 og 2.8 timer i uka.

Dette siste trinnet kalles *inferens* og består i å **generalisere fra utvalget til hele populasjonen**. For å utføre inferens trenger vi kun noen nøkkeltall, som beregnet i steg 3 over, og noe sannsynlighetsteori (broen mellom steg 3 og steg 4 i en statistisk analyse).

Hvordan vi gjør inferens (lager konfidensintervaller og utfører hypotesetester) omhandles i kapitlene 13 og utover.

Broen mellom steg 3 og steg 4 i en statistisk analyse (sannsynlighetsteori) – se figur 1.1

For å utføre inferens trenger vi kun noen nøkkeltall som beregnet i steg 3 over, og noe sannsynlighetsteori (broen mellom steg 3 og steg 4 i en statistisk analyse). Forklaringen på hvorfor og hvordan sannsynlighetsteori er viktig for å kunne utføre inferens venter vi med å fortelle til vi kommer til bolk 2.

1.3.3 Hvor sikker er informasjonen vi får om populasjonen fra den statistiske analysen?

Det er viktig å forstå at kunnskapen, som vi får om populasjonen gjennom en statistisk analyse ikke er hundre prosent sikker. Siden utvalget bare utgjør en liten del av populasjonen, vil det alltid være en viss grad av usikkerhet når vi generaliserer til hele populasjonen. Usikkerheten skyldes at utvalget ikke dekker hele populasjonen. Utvalget er bare ett av mange tenkelige utvalg. Hadde vi samlet inn data på nytt, ville vi fått et annet utvalg og litt andre tall og grafer for utvalget. Med andre ord inneholder utvalget noe støy, i tillegg til å avspeile reelle forhold i populasjonen. Men magien er at vi kan forstå og minimere denne usikkerheten slik at vi faktisk får nyttig kunnskap om populasjonen. Informasjon fra utvalget, for eksempel nøkkeltall som gjennomsnitt og prosentandeler eller grafer som histogrammer og stolpediagram, er sikker informasjon som bare gjelder utvalget. Informasjon fra utvalget om populasjonen vil derimot alltid være usikker, og angir kun anslag for gjennomsnitt og prosentandeler. I tillegg angir vi hvor sikre vi er på disse anslagene, gjerne i form av feilmarginer.

1.3.4 Parameter og estimat

En **parameter** er et tall som beskriver en kvantitativ størrelse eller egenskap til populasjonen. Siden parameteren beregnes ved å spørre hele populasjonen, noe som vanligvis ikke lar seg gjøre , kan vi ikke regne den ut helt nøyaktig. Men vi kan anslå den via et estimat. Et **estimat** er et tall regnet ut i et utvalg, som anslår verdien til en ukjent parameter i populasjonen.

La oss studere forskjellen mellom en parameter og et estimat. For eksempel ønsker vi å vite gjennomsnittlig antall treningstimer for alle BI-studenter som tar faget Statistikk våren 2022. Vi kunne spurt hele populasjonen (samtlige BI-studenter som tok faget Statistikk våren 2022) og så beregnet gjennomsnittet. Dette gjennomsnittet kalles da en **parameter** (fasitsvaret for populasjonen) og denne parameteren er et fast tall, som **er uforanderlig.** Det er, som vi vet, vanligvis umulig å få spurt hele populasjonen og vi må nøye oss med å **spørre et mindre utvalg**. Vi spør for eksempel 100 tilfeldig valgte BI-studenter som tok faget Statistikk våren 2022 om hvor mange timer de trener per uke. Beregner vi gjennomsnittet av disse 100 observasjonene

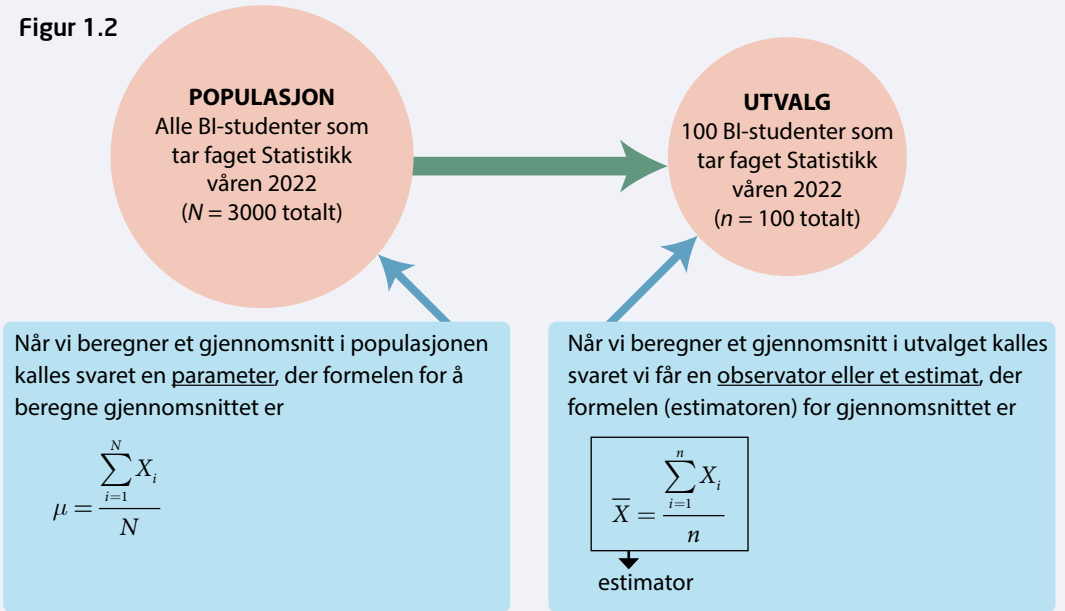
kaller vi dette tallet et estimat. Det er viktig å merke seg at estimatet vi akkurat fant kunne blitt et annet tall dersom utvalget besto av en eller flere andre personer enn de 100 som tilfeldigvis ble trukket ut i utvalget vårt. Et estimat vil med andre ord typisk **variere** fra utvalg til utvalg, og denne variasjonen er **tilfeldig** siden vi trekker tilfeldig.

I en statistisk analyse regner vi alltid ut estimatet i steg 3. Estimatet regnes ut via en formel som generelt kalles en **estimator**. Estimatet anslår verdien til den tilsvarende parameter i populasjonen, se eksemplene nedenfor.

I praksis kjenner vi ikke parameterverdien. Da er kunsten å bruke utvalgsestimatet sammen med sannsynlighetsteori til å si mest mulig om parameterverdien.

EKSEMPEL 1.1

Figur 1.2



EKSEMPEL 1.2

Vi ser på et tilfeldig utvalg av 3 nyfødte barn i 2020. Barna hadde følgende fødselsvekter: 3100 g, 3600 g og 4400 g. Gjennomsnittlig fødselsvekt i dette utvalget er:

$$\overline{X} = \frac{3100 + 3600 + 4400}{3} = 3700 \text{ gram}$$

(dette er altså observatoren eller estimatet på populasjonsparameteren.)

I dette tilfellet går det an å finne gjennomsnittet i populasjonen (alle nyfødte barn i 2020) ved å gå inn på SSB sine hjemmesider. Der finner du at gjennomsnittet i populasjonen er:
μ = 3500 gram (dette er altså populasjonsparameteren)

1.3.5 Forutsetninger for å kunne utføre en statistisk analyse

Det hender at noen gjennomfører en fancy dataanalyse uten å forstå grunnleggende statistisk tenkning, og da blir resultatet ofte ubrukelig. Det er for eksempel viktig at utvalget er trukket tilfeldig. Tenk deg at du ønsker å finne ut hvor mange timer BI-studenter som tar faget Statistikk våren 2022 trener i gjennomsnitt per uke. Forestill deg videre at du trener på et elitelag i Hockey og at du og 9 andre på laget ditt går på BI på faget Statistikk våren 2022. Dersom du nå av bekvemmelighetsgrunner trekker utvalget ditt slik at det består av deg og disse 9 andre hockeyspillerne vil gjennomsnittlig antall treningstimer til disse 10 være svært lite representativt for gjennomsnittet i populasjonen (BI-studenter som tar faget Statistikk våren 2022), og gjennomsnittet blir altfor høyt i forhold til slik det ville vært med et representativt utvalg. Du har brutt en av de viktigste forutsetningene for å kunne utføre en god statistisk analyse og dine resultater fra en statistisk analyse vil nå være ubrukelige.

Pass derfor alltid på at forutsetningene for metodene og prosedyrene som du lærer om i denne boken er oppfylt før du faktisk benytter dem.

1.3.6 Et reelt eksempel: Spørreundersøkelse om bankbransjen

Våren 2016 foretok Norsk Kundebarometer og Barcode en spørreundersøkelse blant norske bankkunder. Undersøkelsen ble gjort på oppdrag av norske banker. Her skal vi se på data for fem store banker: Handelsbanken, Nordea, SpareBank1, Danske Bank og DNB. Formålet med undersøkelsen er å finne ut hvor tilfredse kundene er, og hvor lojale de er mot banken sin. Det er også viktig for bankene å finne ut hvilke faktorer som resulterer i kundetilfredshet. Er prising av banktjenester viktig for kundene? Hva med kvaliteten på nettbanktjenestene? Er kvaliteten på telefonsupport en viktig faktor for tilfredshet? Dette er steg 1 i en statistisk analyse, se figur 1.1, der vi formulerer problemstillinger.

Trekk av et representativt utvalg (steg 2 i en statistisk analyse) skjedde ved telefonintervju og ble gjennomført av markedsanalysebedriften Norstat. Et intervju foregår ved at en intervjuer trekker et tilfeldig telefonnummer fra en digital telefonkatalog. Dersom innehaveren av nummeret er minst 18 år gammel og høyst 85 år, gjennomføres intervjuet. Intervjuene foregikk helt til undersøkelsen hadde 200 kunder for hver bank. Så totalt ble 1000 bankkunder intervjuet. Det er interessant å merke seg at kundene ble valgt tilfeldig.

Tilfeldig utvelging (eller tilfeldig utvalg) spiller en viktig rolle i sannsynlighetsteori, som omhandles i bokens bolk 2. Her nøyer vi oss med å hevde at tilfeldig trekning av respondenter – forutsatt at alle de spurte velger å delta i undersøkelsen, og at vi har et rimelig stort antall respondenter – gir et utvalg som med stor sannsynlighet vil være representativt for bankens kunder.

Når utvalget ikke er større enn 200 kunder per bank, skyldes det at det er dyrt å foreta lange telefonintervju. Så vi har å gjøre med et kompromiss mellom kostnader og presisjonsgrad. Undersøkelsen har som formål å gi en pekepinn til bankene om hvor tilfredse kundene er, og hvilke faktorer som er viktige for å øke denne tilfredsheten. For dette formålet har Barcode funnet at et utvalg på 200 er tilstrekkelig stort til å få nyttige svar. Selvsagt ville det gitt mer presis informasjon om en undersøkte holdningene til 400 kunder, men dette ville jo kostet oppdragsgiverne (bankene) dobbelt så mye.

I markedsanalyser stilles en rekke spørsmål for å kartlegge hvor fornøyd kunden er med et produkt. I vår undersøkelse ble det stilt rundt 50 spørsmål (50 variabler). Noen av disse spørsmålene beskriver respondentens bakgrunn, som kjønn, alder og husstandens brutto årsinntekt. De resterende spørsmålene handler om bruk, tilfredshet og lojalitet. Disse ble målt på en skala 1–10.

Etter datainnsamling kan vi begynne på steg 3 i den statistiske analysen ved å oppsummere data i grafer og nøkkeltall.

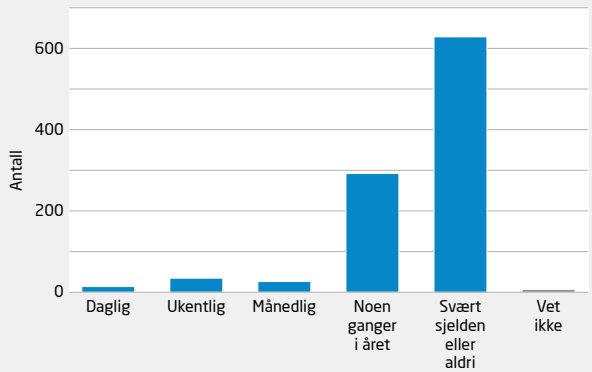
Stolpediagrammet består av en stolpe for hver svarkategori. Høyden på stolpen viser hvor mange som brukte svarkategorien.

Ett av spørsmålene handler om hvor ofte kunden besøker bankkontoret. Stolpediagrammet på figur 1.3 viser fordelingen av de 1000 respondentene på de seks svaralternativene. Vi kan bruke diagrammet til å grovtelle. For eksempel ser vi at om lag 600 av respondentene besøker bankkontoret svært sjelden eller aldri. Vi ser også at i 2016 er det svært få av kundene som oppsøker bankkontoret månedlig eller oftere. Dette er potensielt viktig kunnskap for bankbransjen i Norge. Men vi bør være litt varsomme med å tolke dette lave tallet. Skyldes det at kundene foretrekker å bruke internett til sine bankgjøremål, eller skyldes det at bankene har lagt ned så mange bankkontor at kundene ikke har noe annet alternativ enn internett eller telefon? Sannsynligvis er svaret en blanding av disse to utviklingstrendene. Et annet spørsmål var hvor tilfreds du er med banken din, og her har vi for eksempel nøkkeltallet 7.67 som er gjennomsnittscore for Nordeakundene.

Til slutt, i steg 4 i en statistisk analyse, kan vi prøve å svare på problemstillinger som ble formulert i steg 1. For eksempel kan vi prøve å si noe om gjennomsnittlig tilfredshet blant alle Nordeakunder. Det kan gjøres med å beregne et konfidensintervall som med stor grad av sikkerhet inneholder gjennomsnittet i hele populasjonen av Nordeakunder. I eksempel 14.4 viser vi hvordan dette gjøres, og konklusjonen blir at vi er rimelig sikre på at

gjennomsnittlig tilfredshet blant Nordea kunder ligger mellom 7.52 og 8.04, der skalaen går fra 1 til 10. Dette er et eksempel på steg 4, der vi gjør inferens fra utvalget til hele populasjonen. Legg merke til at vi ikke kan være helt sikre på konklusjonen i en inferens, men at vi er rimelig sikre. Hva som menes med «rimelig» skal vi komme tilbake til i bolk 3.

Figur 1.3 Besøk av bankkontor. 1000 respondenter



1.4 Datamaskinens rolle i statistisk analyse

Datamaskinen er uunnværlig i statistikk. Grafer lages og utregninger blir gjort i et av de mange statistiske programmene som finnes. Dataene lagres som datafiler, og rapportene skrives selvfølgelig på en datamaskin. Tilfeldig utvalg fra populasjonen og tilfeldig fordeling i grupper i randomiserte eksperimenter kan foretas av datamaskinen. Datamaskinen kan også generere (simulere) kunstige datasett som etterlikner utvelging av data i virkeligheten, slik at vi kan undersøke hvordan metodene våre fungerer.

Datamaskinen er kort og godt et fantastisk verktøy. Alle bedrifter og organisasjoner har store mengder data liggende, og med statistikk og dataanalyse kan vi bruke denne informasjonen til å finne viktige trender og sammenhenger. Det fører til at statistikk og dataanalyse er en svært ettertraktet kompetanse i arbeidsmarkedet, og kraftige verktøy krever høy kompetanse hos brukeren.

Av gode grunner gjennomføres eksamen ved mange læresteder uten datamaskinen som hjelpemiddel. Men ofte er det tillatt å bruke kalkulator. Selv om kalkulatoren har et begrenset bruksområde sammenliknet med en datamaskin med statistisk programvare, kan det være nyttig å lære å utføre utregninger på kalkulator. Vær likevel klar over at en kalkulator ikke