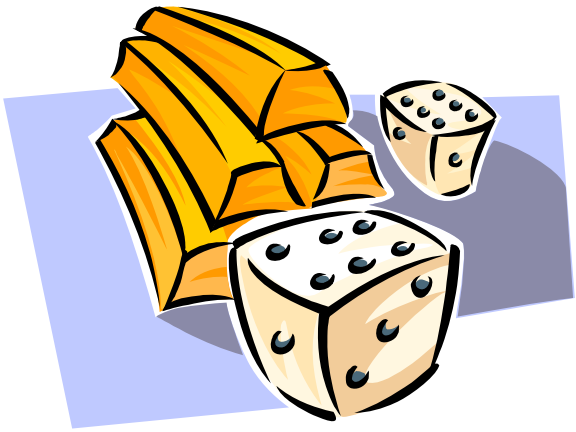
Universitetet i Stavanger



Bruk statistikk riktig!

Jan Terje Kvaløy

Universitetet i Stavanger

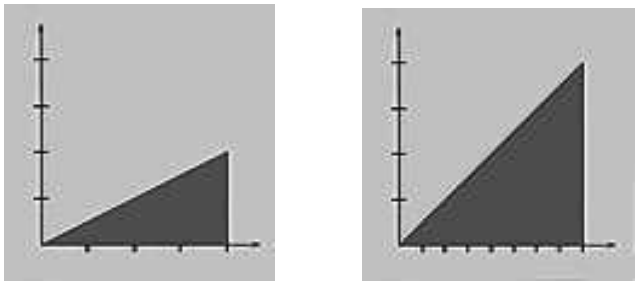

GODE RESULTATER, KARER! IFØLGE SPØRRE-UNDERSØKELSEN ER DET BARE ÉN AV TI ANSATTE I TESTGRUPPA SOM OPPLEVER OSS SOM EN MANNSDOMINERT BEDRIFT!

Innledning

Dette notatet omhandler en del viktig ting som ofte ikke nevnes eksplisitt i lærebøker i statistikk, men som det er viktig å være oppmerksom på når man bruker statistikk i praksis. Notatet er både ment som en hjelp til å gjøre ting riktig når du selv skal gjøre en statistisk undersøkelse, og som et utgangspunkt for å kunne gjøre kritiske vurderinger av statistiske undersøkelser gjort av andre og av fremstillingen av slike undersøkelser i media. Takk til Hilde Grude Borgos og Bjørn Auestad for gode innspill i arbeidet med notatet.

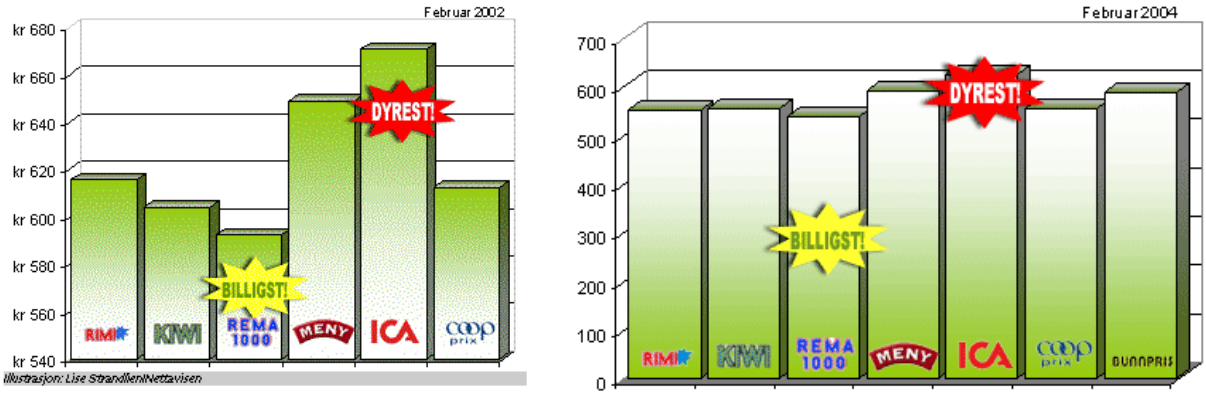
Grafisk fremstilling

Ved grafiske fremstillinger er det lett å manipulere med aksene slik at ting visuelt sett ser mer eller mindre dramatiske ut enn de i virkeligheten er. Et enkelt eksempel er gitt i Figur 1, der den samme kurven er fremstilt med ulik akseinndeling på *x*-aksen.



Figur 1 Den samme matematisk kurven fremstilt med to ulike akseinndelinger

Et annet eksempel er gitt i Figur 2. Forholdene mellom matvareprisene i de to undersøkelsene som legges frem er noenlunde like, men forskjellene ser mye mer dramatiske ut i det venstre plottet hvor *y*-aksen er kuttet ved 540 enn i det høyre plottet hvor hele søylehøyden er vist.



Figur 2 Fremstilling av matvarepriser i Nettavisen i hhv februar 2002 og februar 2004.

Noen råd for å unngå å villede med grafisk fremstilling:

- Bruk samme skala på ulike plott så langt det er rimelig.
- Ta med origo i plottet så sant det er rimelig.
- Vis akseinndelingen

Absolutte versus relative tall?

Den kanskje enkleste form for statistisk undersøkelse man kan gjøre er enkel opptelling av antall hendelser. Men selv i slike undersøkelser må man tenke gjennom hvordan man bruker tallene. Særlig må man vurdere om man skal bruke absolutte eller relative tall.

Eksempel: En avis i Stavanger så på data over hvor mange som skadet seg på trampoline og på skateboard. Det viste seg at antall registrerte skader på trampoline var større enn antall registrerte skader på skateboard – og basert på dette konkluderte avisa med at trampoline var farligere enn skateboard… I en annen reportasje i samme avis ble de konkludert med at det er farligere å være sivil enn soldat i krig basert på observasjonen at flere sivile enn soldater ble drept i krig…

Problemet med disse undersøkelsene er at man baserer konklusjonene på absolutte tall framfor relative tall. Med samme type resonnement kan man for eksempel lett konkludere med at det er mye farligere å være hjemme enn å drive basehopping fordi det skjer mange flere ulykker i hjemmet enn under basehopping. For å få en mer realistisk vurdering av risikoen ved ulike typer aktivitet må man se på relative tall, for eksempel antall skader i forhold til antall utøvere av aktiviteten.

Hvordan man skal regne ut relative tall er imidlertid ikke alltid helt opplagt. Dersom man ønsker å vurdere om trampoline eller skateboard er mest risikofylt, skal man se på antall skader i forhold til antall brukere, i forhold til total tid brukt på aktiviteten eller kanskje i forhold til antall trampoliner/skateboard? Uansett hva man velger er det viktig at man tenker gjennom og er bevisst på hva man gjør – og at det formidles presist hva som er gjort.

Skal man bare bruke relative tall? Nei, avhengig av fokus kan både absolutte og relative tall være av interesse. Dersom vi igjen bruker skader ved ulike aktiviteter som eksempel vil det for den enkelte utøver være mest av interesse å vite noe om relative tall, mens det fra et samfunnsmessig synspunkt også vil være av interesse å vite noe om totalt antall ulykker. Kanskje kan den samfunnsmessige gevinsten være større ved å forsøke å få ned antall skader ved en aktivitet som mange driver med og som dermed har et stort total skadeantall, fremfor ved en aktivitet med høyere skaderisiko men som så få driver med at det totale antall ulykker likevel er lavt.

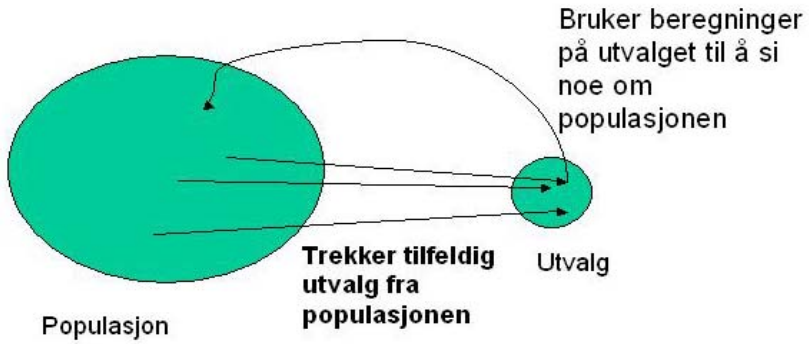
Oppsummering:

- Tenk gjennom om relative eller absolutte tall gir den mest hensiktsmessige beskrivelsen.

Representative data

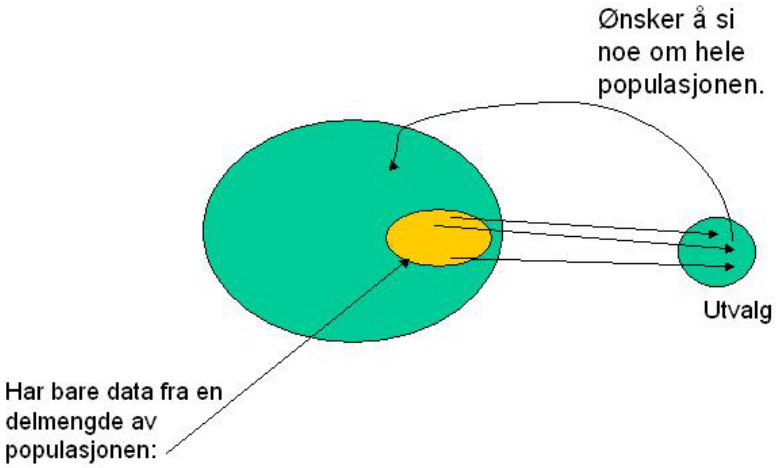
En grunnleggende forutsetning for at resultatet av en statistisk undersøkelse skal være korrekt er at de innsamlede dataene er et *representativt utvalg* fra den populasjonen/det fenomenet man ønsker å si noe om. For å få et slikt representativt utvalg må vi trekke et tilfeldig utvalg fra populasjonen/fenomenet.

En skisse av hva vi ønsker å gjøre er vist under.



Figur 3

Dersom vi har et utvalg som er tilfeldig trukket fra populasjonen kan vi basert på dataene i dette utvalget trekke konklusjoner angående hele populasjonen. Dessverre er situasjonen noen ganger heller som i figuren under.



Figur 4

Dersom vi bare har et utvalg fra en bestemt delmengde av populasjonen er det vanskelig å trekke konklusjoner for hele populasjonen basert på dataene i utvalget.

Eksempel: Anta at vi ønsker å undersøke hva Norges befolkning mener om EU. Dersom vi spør 1000 høyrevelgere bosatt i Bærum hva de mener om EU er det ganske opplagt at vi ikke kan bruke resultatet til å si noe om hva Norges befolkning generelt mener om EU. Problemet er at vi har tatt utvalget vårt kun fra en bestemt delmengde av populasjonen ”Norges befolkning”.

Dersom vi har valgt de 1000 tilfeldig fra delmengden ”høyrevelgere i Bærum” kan vi bruke resultatet til å si noe om hva denne delmengden av populasjonen mener om EU, men altså ikke noe om hva hele populasjonen mener. □

Eksempel: Følgende sak er hentet fra Stavanger Aftenblad.

Byen får stryk på internett

Publisert 12. oktober 2001 06:25

Folk i Stavanger er ikke nådige i kritikken av sin egen by i en ny undersøkelse på internett. De ansattes serviceinnstilling og kompetanse får strykkarakter. Det samme gjør åpningstider, parkeringstilbud og kollektivtilbud.

Bernt Eirik Rød og Jone Østebø

Folk som bruker Stavanger by har falt: De fleste er svært misfornøyd med hva sentrum tilbyr dem av service, parkering og åpningstider. Siddiser er ikke stolte av byen sin, og veldig få oppsøker sentrum for opplevelsens skyld.

En undersøkelse utført av Markedsføringshuset har gitt folk muligheten til å si sin mening om Stavanger by. Undersøkelsen er gjort via internett på oppdrag fra Stavanger kommune, Byen, Destinasjon Stavanger, Stavanger Næringsforening og Gårdeierforeningen. I tre uker poppet et spørreskjema opp på skjermen til tilfeldig utvalgte brukere av de fire nettsidene.

916 personer svarte. Nesten 800 av disse supplerte med utfyllende kommentarer om hvordan de opplever sentrum, og om hva de mener må til for å gjøre byen til en bedre plass å ferdes i.

Merk at undersøkelsen som ligger til grunn er utført ved at et spørreskjema poppet opp på skjermen til tilfeldige brukere av fire bestemte nettsider slik at disse kunne besvare spørsmålene i skjemaet dersom de ønsket. Vil man få et representativt utvalg på denne måten? Neppe! For det første er det neppe et tilfeldig utvalg av Stavangers befolkning som bruker internett og jevnlig er innom akkurat de fire aktuelle nettsidene. For det andre er det blant de som er inne på disse sidene og får spørreskjemaet opp på skjermen frivillig om man vil svare eller ikke. Er det da tilfeldig hvem som svarer og hvem som ikke svarer? Eller kan det for eksempel tenkes at de som er misfornøyde her ser en anledning til å si ifra om det de er misfornøyde med, mens de som er greit fornøyde ikke gidder å svare på undersøkelsen? □

Generelt er spørreundersøkelser på internett, ved innringing til debattprogram og andre former for undersøkelser der deltakerne må gjøre noe aktivt for å delta ikke til å stole på. Ved slike undersøkelser får man ikke et representativt/tilfeldig utvalg av befolkningen noe eksemplet under, hentet fra Dimakos et al. (2004), er en klar illustrasjon av.

Eksempel: I TV2s debattprogrammer ”Holmgang” stilles et spørsmål til seerne med to svaralternativer. Spørsmålet er knyttet til temaet som debatteres, og seerne kan gi sin tilbakemelding ved å ringe et bestemt telefonnummer. Som en oppfølger til spørre-undersøkelser gjort blant seerne i to slike program utførte Opinion A/S en representativ undersøkelse (publisert i VG 28. april 1999) av de samme spørsmålene. Resultatene ble:

”Skal vi stenge grensene for flyktninger fra fjerne land?”	***Ja***	***Nei***
Holmgang	89%	11%
Opinion A/S (representative tall)	17%	83%

”Reagerer du negativt på svart arbeid?”	***Ja***	***Nei***
Holmgang	30%	70%
Opinion A/S (representative tall)	62%	38%

Tabell 1

Disse resultatene kan forklares ved at det ikke er et tilfeldig utvalg av befolkningen som ser på Holmgang, og det er ikke et tilfeldig utvalg av de som ser på programmet som ringer inn. □

Ofte er ikke situasjonen at man gjør en undersøkelse fra starten av, men heller at man får et ferdig innsamlet datasett å jobbe med. Da er det viktig at man startet med å bringe på det rene hva disse dataene er representative for – hva er populasjonen/fenomenet disse dataene kan sies å være et tilfeldig utvalg fra?

Eksempel: I læreboka Walpole et al. (2002) finner man følgende datasettet som viser diameteren på sylindriske metalldeler produsert av en maskin: 1.01, 0.97, 1.03, 1.04, 0.99, 0.98, 0.99, 1.01 og 1.03. Basert på disse dataene kan man si noe om forventet diameter for metalldelene og variansen i diameter fra del til del (man kan for eksempel lage konfidensintervall for forventningsverdien og for standardavviket). Men hvor langt kan man generalisere resultatet som fremkommer?

- Kan man for eksempel si at resultatet vil gjelde for alle metalldeler av denne typen produsert ved denne fabrikken? Dersom vi har målinger fra et tilfeldig utvalg av metalldeler produsert ved maskinen og fabrikken kun har en slik maskin sier resultatet oss noe om all totalproduksjonen av slike deler ved fabrikken. Men dersom fabrikken har flere maskiner bør vi heller velge ut deler tilfeldig fra totalproduksjonen og ikke bare fra en maskin. Det kan godt være systematiske forskjeller i produktkvalitet ved ulike maskiner.
- Kan man i det minste si at resultatet vil gjelde for all metalldeler produsert på denne maskinen? Dersom maskinen fungerer med jevn presisjon over tid, uavhengig av operatør, temperatur, råvarekvalitet osv (eller at alle slike faktorer er stabile) burde det være trygt å generalisere resultatet til all produksjon ved denne maskinen. Men dersom ulike ytre faktorer påvirker kvaliteten på produktet må man vite noe om hvordan dataene er samlet inn for å vite hvor langt de kan generaliseres. Var for eksempel delene man målte alle produserte i ferien, med en uerfaren operatør, med lang tid siden maskinen hadde hatt service og med uvanlig høy temperatur i produksjonslokalet? Eller var kanskje alle delene produserte under ”idealbetingelser” som man egentlig sjelden i praksis har? Eller har kanskje den som gikk og hentet ut deler til testen gått og sett etter deler som såg mest mulig rette ut i dimensjon? Forhåpentligvis ingen av delene, dersom noe av dette var tilfelle kan man ikke bruke resultatene til å si noe generelt om produksjonen. Det kan man derimot dersom det vi har er et tilfeldig utvalg av deler produsert ved maskinen. □

Det er med andre ord av avgjørende betydning å vite mest mulig om bakgrunnen for det datamaterialet man har. Særlig er det viktig å vite hvordan dataene er samlet inn og hvilke eventuelle begrensninger som ligger i måten innsamlingen er gjort på. Hva er dataene representative for? Men det er også viktig å ha kunnskap om det fenomenet som dataene er innhentet fra for å kunne si noe om generaliseringer av resultatene.

Når man hører at en undersøkelse ett sted i verden har vist noe som er det tvert motsatte av hva en annen undersøkelse utført et annet sted har vist er nok i mange tilfeller forklaringen at man, dersom man ser nærmere etter, stengt tatt ikke har målt det samme i de to undersøkelsene.

Oppsummering:

- I planleggingen av en undersøkelse tenk grundig gjennom hvordan du skal få tak i representative data (dvs tilfeldig utvalg) fra den populasjonen/det fenomenet du ønsker å si noe om.
- Dersom du får i hendene et ferdig innsamlet datasett, undersøk grundig hva slags populasjon/fenomen dataene er representative for.

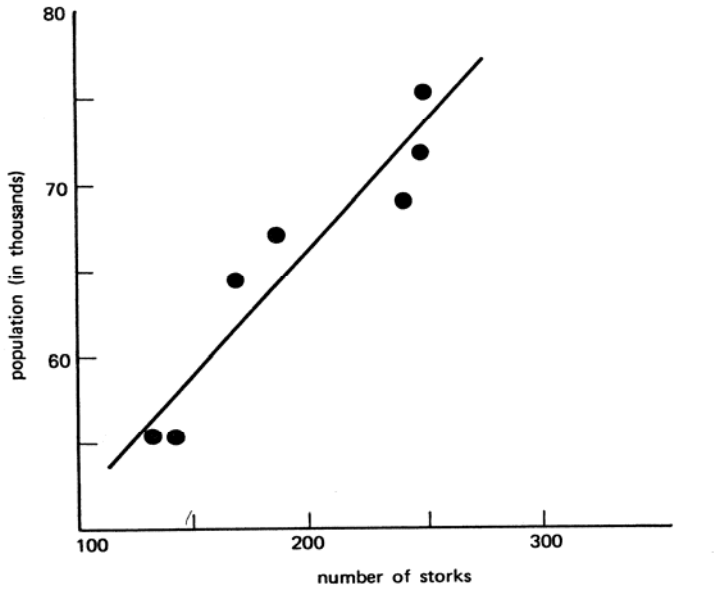
Statistisk sammenheng versus årsak/virkning.

En påvist statistisk sammenheng mellom to eller flere variable betyr generelt *ikke* at det er påvist et årsak-/virkningsforhold mellom variablene.

Eksempel: Det er en smal sak å vise ved bruk av statistiske undersøkelser at det er en sammenheng mellom en persons høyde og hvilket skonummer han bruker – høye personer tenderer til å ha stort skonummer og lave personer tenderer til å ha lite skonummer. Dette betyr imidlertid selvfølgelig ikke at det er et årsak-/virkningsforhold mellom høyde og skonummer. Det er verken rett å si at en persons høyde påvirker vedkommendes skonummer, eller at en persons skonummer påvirker personens høyde. Grunnen til den statistiske sammenhengen man likevel ser er en annen felles underliggende faktor mellom variablene – i dette tilfellet personers størrelse. Store personer er høye og har typisk høyt skonummer, mens små personer er lave og har typisk lite skonummer. Dermed får man en sammenheng mellom høyde og skonummer uten at det er et årsak-/virkningsforhold. □

Eksempel: Et eksempel på en situasjon der man opplagt har et årsak-/virkningsforhold mellom variable er dersom man ser på gjødsel og avling. Inntil en viss grense vil det være slik at jo mer gjødsel man tilfører jo mer avling får man. Mengden gjødsel påvirker mengden avling. □

Eksempel: Plottet i figur 5 er kopiert fra Box, Hunter og Hunter (1978) og viser et plott av befolkningstall mot antall storker i den tyske byen Oldenburg i årene 1930-1936. Årsak/virkning…? □



Figur 5 Plott av folketall mot storkebestand for hvert år i perioden 1930-1936 i den tyske byen Oldenburg.

Generelt vil det være slik at dersom man har et årsak-/virkningsforhold vil dette også gi en statistisk sammenheng, men man kan ha en statistisk sammenheng uten at det er et årsak-/virkningsforhold.

Har man noen måte å avgjøre når en påvist statistisk sammenheng er et årsak-/virkningsforhold og når sammenhengen bare skyldes andre underliggende årsaker? Generelt

er dette vanskelig, men dersom man utfører en undersøkelse på rette måten kan det være mulig å si noe om årsak-/virkning. For eksempel kan man si noe om årsak-/virkning ved enkelte typer randomiserte forsøk, eller ved kontrollerte forsøk. Andre ganger kan man ut fra kunnskap om fenomenet man betrakter si noe om årsak-/virkning.

Eksempel: Dersom man ønsker å avgjøre om en bestemt type medisin hjelper mot en bestemt lidelse er en måte å avgjøre dette på ved å *tilfeldig* tildele pasienter med lidelsen enten den aktuelle medisinen eller en virkningsløs placebomedisin (uten at pasienten vet hva han får). Etter en tid kan man så sammenligne hvor mange som har oppnådd bedring i henholdsvis gruppen som har fått medisin og gruppen som har fått placebo. Dersom det er forskjell mellom de to gruppene har man vist en årsak/virkning mellom medisin og lindring. Merk at det her er svært viktig at tildelingen av placebo versus medisin skjer tilfeldig, ellers ville forsøket fort ødelegges av andre faktorer som kunne være systematisk forskjellige i de to gruppene. □

Oppsummering:

- Vær oppmerksom på at en påvist statistisk sammenheng ikke nødvendigvis betyr at vi har et årsak-/virkningsforhold.

Naturlig tilfeldig variasjon

Et fenomen man jevnlig ser blant annet i media er det jeg vil kalle dårlig forståelse for naturlig tilfeldig variasjon. Et typisk eksempel var Dagbladet som i en årrekke ukentlig publiserte ”formtabeller” for Lotto-tallene. Dette var selvfølgelig helt meningsløst siden hver ny Lotto-trekning er uavhengig av de tidligere trekningene.

Et annet eksempel er all tolkningen som legges i små endringer i partioppslutning fra en måling av partioppslutning til en annen. Slike meningsmålinger utføres ved at man spør et begrenset utvalg velgere hva de ville ha stemt dersom det var valg i dag. Selv om partioppslutningen er uendret i befolkningen vil det på være naturlige variasjoner i måleresultatene på grunn av naturlig tilfeldige variasjoner i hvem som blir valgt ut. Meningsmålingsbyråene gjør noen grep for å redusere den tilfeldige variasjonen i partimålingene fra mellom ±1 prosentpoeng (for partier med liten oppslutning) til opp mot ±3 prosentpoeng (for partier med stor oppslutning) som man i utgangspunktet vil ha dersom man spør et tilfeldig utvalg på 1000, men en del variasjon vil der alltid være.

Et gullkorn omtrent som følgende sto å lese på NRKs tekst-tv for en del år tilbake: ”Resultatene av målingene vekker stor bekymring. Nær halvparten av alle målinger viste en verdi over gjennomsnittet.” Nå kan forklaringen her være at det var en ren skrivefeil, for eksempel at det i stedet for ”gjennomsnittet” skulle ha stått ”tiltaksverdien” – men det illustrerer noe man tidvis ser, at en gjennomsnittsverdi oppfattes som det normale (kalles også ofte normaler) og at avvik fra gjennomsnittet oppfattes som ”unormalt”. Imidlertid er det normale i de fleste situasjoner hvor man sammenligner med et gjennomsnitt at resultatene varierer mer eller mindre tilfeldig fra måling til måling. Har vi et fenomen med stor naturlig variasjon, for eksempel månedsnedbøren en bestemt måned på Sola, er det helt normalt at vi får målinger som avviker mye fra gjennomsnittet. I mange sammenhenger ville det trolig vært bedre og mer illustrativt om man i stedet for å sammenligne med gjennomsnitt/”normaler” sammenlignet med ”normalintervaller”, for eksempel intervaller hvor 90% av tidligere måledata har falt innen.

Eksempel: Figur 6 under viser et forsideoppslag i VG som sto på trykk i februar 2000. Her slås det stort opp at ei nyfødt jente fra Sekken utenfor Molde i følge VG har 29 år lenger forventet levetid enn en nyfødt gutt i Berlevåg: ”I Norge i år 2000 godtar norske myndigheter at den lille gutten til høyre statistisk sett skal ha 29 år kortere forventet levetid enn jenta. VG setter i dag fokus på de oppsiktsvekkende resultatene.” VG brukte fire sider av avisa på denne saken. Bakgrunnen var tall fra Statistisk sentralbyrå over gjennomsnittlig levealder for henholdsvis kvinner og menn i alle kommuner i Norge som døde i perioden 1993-1997. I forsideoppslaget har VG så dratt frem den kommunene med høyeste gjennomsnittlig levealder for kvinner og sammenlignet med en av kommunene med lavest gjennomsnittlig levealder for menn. Det viser seg at forskjellen blir 29 år.

For det første er neppe gjennomsnittlig levealder for de som har levd fra tidlig på 1900-tallet en god estimator for forventet levealder for de som skal vokse opp i dagens samfunn – jamfør delkapitlet om representative data! Dessuten viser VG svært dårlig forståelse for naturlig tilfeldig variasjon når de slår den observerte forskjellen på 29 år stort opp. Det finnes mange små kommuner i Norge og i løpet av en femårsperiode vil det være mange kommuner der det kun er noen få kvinner eller menn som dør. Tilfeldig variasjon vil da kunne slå sterkt ut på tallene. Siden man har brukt gjennomsnittlig levealder (og ikke for eksempel median) vil ting som for eksempel en ulykke hvor unge omkommer eller et tilfelle eller to av krybbedød dra gjennomsnittet sterkt ned i en liten kommune hvor noe slikt skjer. I andre små kommuner har man kanskje få unge innbyggere slik at det som dør stort sett bare vil være gamle. Når man så ser på hele landet og trekker ut de meste ekstreme observasjonene er det helt naturlig at man ser store forskjeller.

Figur 6 Forsiden på VG 26. februar 2000.

I reportasjen inne i avisa nevner VG også tall for hele landet og for noen fylker. Dette er tall basert på så mange observasjoner at man kan bruke dem til å si noe om reelle forskjeller. Tallene for hele landet viste en forskjell på seks år i forskjell på menns og kvinner levealder. Fylkesvise forskjeller var små, Finmark skilte seg mest ut – her levde kvinner i gjennomsnitt to år kortere og menn i gjennomsnitt fire år kortere en i landet sett under ett. Med andre ord forskjeller av en helt annen størrelsesorden enn de som slås stort opp på forsiden. □

På tilsvarende måte som at et gjennomsnitt blir assosiert med ”det normale”, ser man av og til også at modalverdien (den verdien som opptrer hyppigst) blir ansett som en slags ”normalverdi”. Tilsvarende som med gjennomsnitt er imidlertid det viktige å vite noe om hvordan variasjonen omkring modalverdien er.

Oppsummering:

- Vær oppmerksom på den naturlig tilfeldige variasjonen vi ofte har og hvordan den slår ut.
- Gjennomsnitt eller modalverdier er ikke det samme som ”det normale” – ”det normale” er at vi har tilfeldig variasjon omkring gjennomsnitt/modalverdiene.

Sjeldne hendelser

Fra tid til annen ser man sensasjonsoppslag i media om ”usannsynlige” hendelser som har inntruffet. For eksempel at samme person har vunnet topp-premien i lotto to ganger, at samme person har blitt truffet av lynet flere ganger, en familie som har ti barn hvor alle er gutter, osv. Og man opplever selv fra tid til annen ting som virker ”usannsynlige”, for eksempel at man tenker på en bestemt person man ikke har møtt på lenge for så å tilfeldig treffe på vedkommende på gata dagen etterpå. Man bør imidlertid være forsiktig med å legge for mye tolkning i at det har inntruffet slike ”usannsynlige” hendelser. Spesielt bør man tenke gjennom hvor usannsynlig er dette egentlig - hvor mange ganger har noe slikt som dette hatt muligheten for å inntreffe?

Eksempel: Et enkelt eksempel som illustrerer hvordan man kan tenke i forbindelse med ”usannsynlige” hendelser får man ved å betrakte Lotto-trekningen i Norge. Dersom du tipper ei rekke i Lotto er sannsynligheten for at du vinner førstepremien $1/\ 5379616 = 1.86 \cdot 10^{-7}$. Til sammenligning er sannsynligheten for å få yatzy (fem like terninger) på ett kast 1/1296. Selv om du tipper flere rekker så er fremdeles sannsynligheten for at du skal vinne førstepremien i Lotto forsvinnende liten. Likevel er det nesten hver uke noen som vinner førstepremien. Forklaringen er selvfølgelig at det leveres inn så veldig mange rekker, vanligvis over 20 millioner rekker. Så selv om hver enkelt rekke har svært liten sannsynlighet for å gi 7 rette, vil vi vanligvis, på grunn av det store antallet rekker som leveres, observere at det finnes innleverte rekker med 7 rette. □

Sagt litt mer generelt illustrerer dette eksemplet at når det gjøres mange ”forsøk”, så vil man jevnlig se at en hendelse inntreffer selv om sannsynligheten for at den skulle inntreffe i hvert av enkeltforsøkene er svært liten. Så det viktige spørsmålet å stille når man hører om noe ”usannsynlig” som har inntruffet er: ”Hvor mange ganger har noe som dette kunne ha skjedd?”

Eksempel: På side 162-163 i Aalen et. al (2006) finner vi følgende eksempel som gjelder forekomst av en sjelden misdannelse i sentralnervesystemet hos nyfødte. I 1980-81 ble det i Bømlo kommune observert hele 3 tilfeller av en bestemt type misdannelser hos nyfødte barn i løpet av et halvt år. Forekomsten av den aktuelle typen misdannelse er så sjelden at man normalt i det lange løp ville vente rundt 1 tilfelle hvert 4. år i Bømlo. Når man da observerte så mange som 3 tilfeller på et halvår ble det satt i verk undersøkelser for å se om det kunne være spesielle årsaker til den overhyppigheten man registrerte. Man fant imidlertid ikke noen slike årsaker, og konkluderte med at det hele var et resultat av tilfeldigheter.

Kunne folk i Bømlo slå seg til ro med denne konklusjonen om at det hele skyldtes tilfeldigheter? Aalen et. al (2006) gjør følgende beregninger: Basert på informasjon om hvor hyppig misdannelsen forekommer (ca 1.6 per 1000 fødsler) og hvor mange barn som blir født på Bømlo (ca 80 per halvår) kan det regnes ut at sannsynligheten for tre eller flere misdannelser i Bømlo per halvår er ca 0.00033. Man kan da umiddelbart være fristet til å tro at det hele vanskelig kan skyldes tilfeldigheter. Sannsynligheten for at noe slik skulle inntreffe ved en tilfeldighet er jo tilsynelatende forsvinnende liten. Men dersom man ser på situasjonen i et litt større perspektiv så vet man at det finnes mange kommuner på størrelse med Bømlo og man har opplevd mange halvårsperioder. Det er med andre ord gjort mange ”forsøk”, det har vært mange muligheter for at noe slikt som dette kunne skje. Aalen et. al (2006) regner på eksemplet at man ser på 50 kommuner på størrelse med Bømlo og følger disse over en femårs periode – man gjør da totalt 500 ”forsøk” av typen observere antall misdannelser blant nyfødte i kommuner på størrelse med Bømlo i en halvårsperiode. Sannsynligheten for at man i minst en av kommunene i en av halvårsperiodene skal komme til å observere minst 3 misdannelser ved en tilfeldighet er $1-(1-0.00033)^{500}=0.15$. Dersom man ser på flere kommuner og/eller lengre tidsrom blir sannsynligheten for å observere noe slikt som dette et eller annet sted på ett eller annet tidspunkt enda større.

Resonnementet over viser at det ikke er noe merkelig i å av og til observere en slik opphopning som den som ble sett på Bømlo. Men merk at selv om det er vist at opphopningen *kan* skyldes en tilfeldighet så er selvfølgelig ikke det et bevis på at den virkelig skyldes en tilfeldighet – det var nok fornuftig å sjekke om det var spesielle årsaker som låg til grunn for opphopningen på Bømlo, men når det ikke ble funnet noen slike årsaker viser resonnementet over at man kan slå seg til ro med at slik ting av og til vil inntreffe ved en tilfeldighet. □

Når man står ovenfor tilsynelatende usannsynlige hendelser bør man altså gjøre en vurdering av hvor mange ganger noe slik som det man har observert kunne ha inntruffet før man legger for mye tolkning i observasjonen. Hvor mange ganger har men tenkt på en eller annen person uten å tilfeldig treffe på vedkommende neste dag? Hvor mange sjanser har det vært for at en eller annen person i Norge kunne ha vunnet førstepremien i Lotto to ganger? Osv.

I juridiske sammenhenger er også slike resonnement viktige. Dersom eneste indisium mot en person er noe som kunne ha skjedd ved en tilfeldighet må det tenkes gjennom hvor mange sjanser det har vært for at noe slikt kunne skje ved en tilfeldighet med en eller annen person på ett eller annet tidspunkt.

Eksempel: Følgende historie sto i Nettavisen i november 2003.

Norges heldigste mann

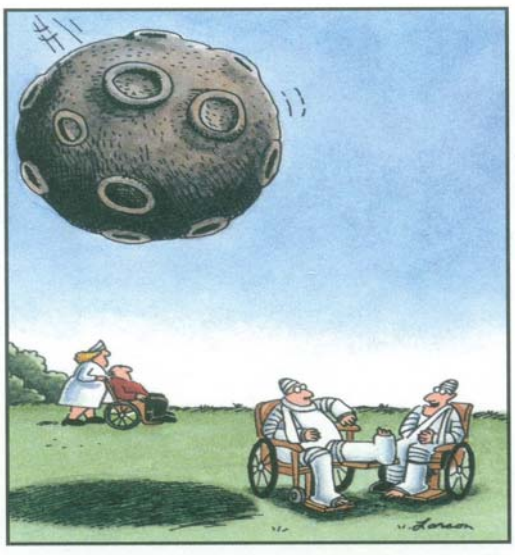
Onsdag kunne en mann fra Nord-Trøndelag innkassere millionpremie for tredje gang. Sjansen for trippelgevinsten er 1 til 1.000.000.000.

Denne gang har den usedvanlig heldige mannen vunnet 4,4 millioner kroner i Viking Lotto. Tidlig på 1990-tallet vant mannen 1,3 millioner kroner i Tipping, så vant han 1,7 millioner i Viking Lotto i 1999.
- Det var ikke så verst dette, sa mannen da Norsk Tipping ringte ham etter trekningen.
Representanten fra Norsk Tipping spurte om mannen hadde vunnet noe særlig før, og da måtte han innrømme at han faktisk hadde det.
- Det er faste tall jeg har spilt i mange år, så jeg kjente dem igjen, forklarte mannen, som spiller et system på 10 tall. Det koster 630 kroner å levere, men nå betalte innsatsen seg veldig mange ganger.
I tillegg til seks rette, fikk mannen et utall 3., 4. og 5. premier, slik at premien økte fra førstepremien på drøyt 4,2 millioner kroner til totalpremien 4,4 millioner kroner.
Det er én til 12,3 millioner sjanse til å få seks rette i Viking Lotto, og det har den heldige mannen klart to ganger. I løpet av 17 år har 3000 nordmenn blitt lotto-millionærer.
Mannen hadde levert kupongen sin hos Coop på Høylandet i Nord-Trøndelag.

Fotballtipping er vanskelig å regne på siden det til en viss grad er et kunnskapsspill, men du kan jo prøve å regne ut hvor stor sannsynlighet en som tipper et system på 10 tall (=210 rekker) i Viking-Lotto hver uke over for eksempel en periode på 10 år har for å vinne to førstepremier – og så vurdere hvor mange du tror tipper omtrent så mye eller mer og deretter regne ut sannsynligheten for at minst en av disse vil oppleve en dobbel toppgevinst…

Oppsummering:

- Når du hører om en tilsynelatende ”usannsynlig” hendelse, tenk gjennom hvor mange ganger noe slikt har hatt mulighet for å skje.



"You're kidding! ... I was struck twice by lightning too!"

Problemer knyttet til *p*-verdier

Hypotesetesting brukes i mange sammenhenger, og svært ofte rapporteres resultatet av en hypotesetest i form av en *p*-verdi. Rapportering av p-verdier for tester av ulike ting er en vanlig framgangsmåte i mange undersøkelser. Det er imidlertid noen begrensninger ved hypotesetesting man må være oppmerksom på for å unngå å gjøre ting som er feil eller lite meningsfulle. Vi skal spesielt se på det såkalte "fisketurproblemet" og distinksjonen mellom signifikant og interessant effekt, men først litt om tolkning av hva en *p*-verdi egentlig er.

Tolkning

En måte å si hva en *p*-verdi er for noe er "sannsynligheten for å oppservere noe som motsier nullhypotesen minst like mye som det vi har observert, gitt at nullhypotesen er korrekt". Eller litt mer løselig, "sannsynligheten for å observere noe minst like ekstremt som det vi har observert, gitt at nullhypotesen er korrekt".
Et viktig poeng her er at vi alltid regner ut *p*-verdien under antagelsen om at nullhypotesen er korrekt ("gitt at nullhypotesen er korrekt"). I hypotesetesting antar vi i utgangspunkt at nullhypotesen er korrekt, og så regner vi ut hvor sannsynlige de observerte dataene er under denne antagelsen – dersom de er svært lite sannsynlige (liten *p*-verdi) konkluderer vi med at antagelsen (nullhypotesen) er gal.

Liten, versus stor *p*-verdi:

- En liten *p*-verdi (typisk mindre enn 0.05) betyr at vi forkaster nullhypotesen og påstår at alternativ hypotese er korrekt.
- En stor *p*-verdi betyr bare at vi ikke forkaster nullhypotesen – **både** nullhypotese **og** alternativ hypotese er mulige.

Vi kan altså aldri "bevise" at en nullhypotese er rett – bare at den er gal!

En vanlig misforståelse er at *p*-verdien sier oss noe om "sannsynligheten for at nullhypotesen er rett". I vanlig klassisk statistikk er dette et meningsløst utsagn. Nullhypotesen er enten rett eller gal, vi skal bare avgjøre hva som er tilfelle. I Bayesiansk statistikk derimot kan vi kombinere bakgrunnsinformasjon om hvor sannsynlige man mener de ulike alternativene er med informasjon i observerte data til å regne ut slike sannsynligheter. Dette kalles imidlertid ikke *p*-verdier (men aposteriorisannsynligheter).

Fisketurproblemet

Et vanlig og alvorlig problem i analyse av data er at man drar på "fisketur" etter lave *p*-verdier. Dette problemet oppstår når man gjør mange hypotesetester.

Anta at vi bruker 0.05 som forkastningsgrense. Selv om nullhypotesen er korrekt er det 5% sannsynlighet for at *p*-verdien blir mindre enn 0.05 ved en ren tilfeldighet. Dette betyr at når vi gjør mange tester er det stor sannsynlighet for at minst en av dem gir en *p*-verdi på mindre enn 0.05 bare pga en tilfeldighet (uten at det er noen reell effekt).

Dersom man "leter rundt i dataene" etter *p*-verdier som er mindre enn 0.05 (drar på fisketur etter lave *p*-verdier) er det stor fare for at det man finner ikke er reelle effekter! Eventuelle interessante funn man gjør med en slik fremgangsmåte bør undersøkes nærmere ved en oppfølgingsundersøkelse.

Ideelt sett bør man gå frem på følgende måte i situasjoner der man ønsker å teste mange forskjellige ting:

1. Spesifiser på forhånd hvilke hypoteser man ønsker å teste.
2. Når man utførere testene, bruk α/k som grense for hvor lav en *p*-verdi skal være før man forkaster en nullhypotese. Her er k antall tester som skal utføres, mens vi typisk bruker $\alpha=0.05$. (Denne typen korreksjon kalles en Bonferroni-korreksjon)

Dvs dersom man f.eks. skal utføre 20 tester og normalt ville bruke 0.05 som grense for *p*-verdien ved en enkelt test bør vi nå bruke 0.05/20=0.0025 som grense for når vi forkaster nullhypotesen. Vi er da på "trygg grunn" i den forstand at den totale sannsynligheten for å gjøre en såkalt type-I feil (feilaktig forkaste) fremdeles er maksimalt 0.05. Uten denne korreksjonen ville sannsynligheten for å gjøre en type-I feil i minst en av de 20 testene øke til inntil 64%!!!

Eksempel: Jeg genererte 200 tilfeldige tall som jeg kalte *y*-verdier, deretter 200 nye tilfeldige tall som jeg kalte x_1-verdier, 200 tilfeldige x_2-verdier, osv…, opp til og med 200 tilfeldige x_{20}-verdier. Jeg prøvde så å tilpasse en regresjonsmodell for *y* som inkluderte alle disse 20 *x*-variablene (som altså består av helt tilfeldige tall som ikke har noe med *y* å gjøre!) som mulige forklaringsvariable, dvs modellen

$$Y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+_+\beta_{20}x_{20}+\varepsilon$$

Estimat for hver parameter og *p*-verdi for test av nullhypotesen at parameteren har verdien 0 mot alternativet ulik 0 er vist i tabell 2. Dersom nullhypotesene om verdi 0 forkastes for en eller flere av parametrene tolker vi det vanligvis som at den/de tilhørende *x*-variabelen/variablene har en reell sammenheng med *Y*-variabelen. I dette eksemplet vet vi at det ikke er noen slik sammenheng mellom noen av variablene, men likevel ser vi i tabell 1 at vi får en svært lav *p*-verdi for variabel 1 og variabel 19! Uten å foreta noen korreksjon for det faktumet at vi her gjør mange tester kunne vi fort komme til å dra den her feilaktige konklusjonen at x_1 og x_{19} har en sammenheng med *Y*! □

i	b_i	*p*-verdi		*i*	b_i	*p*-verdi		*i*	b_i	*p*-verdi		*i*	b_i	*p*-verdi
1	-0.196	0.008		6	0.064	0.375		11	0.022	0.762		16	-0.013	0.855
2	0.054	0.484		7	-0.016	0.827		12	0.044	0.578		17	-0.004	0.960
3	0.093	0.196		8	-0.018	0.812		13	0.016	0.837		18	0.046	0.532
4	0.040	0.565		9	-0.004	0.952		14	-0.001	0.988		19	0.201	0.004
5	-0.044	0.534		10	-0.061	0.401		15	0.012	0.874		20	0.015	0.838

Tabell 2

Ulempen med Bonferroni-korreksjon er at korreksjonen kan bli litt vel konservativ, en rekke alternative fremgangsmåter for å skille ut hvilke tester som bør gi forkastning når man gjør mange tester finnes. Se for eksempel Schweder og Spjøtvoll (1982), Simes (1986) eller Benjamini og Hochberg (1995).

Signifikant versus interessant

Ved hypotesetesting er det også viktig å være bevisst på at en signifikant effekt (dvs forkastning av en nullhypotese) ikke behøver være det samme som en interessant effekt. Dersom man har store datamengder eller svært lav varians kan selv marginale avvik fra en nullhypotese detekteres. I praksis kan imidlertid slike små avvik kanskje være nokså uinteressante.

Eksempel: Anta at man i en medisinsk studie er interesserte i å undersøke kolesterolnivået i to ulike befolkningsgrupper. La oss videre anta at man har målt kolesterolverdier hos 1000 tilfeldig valgte personer i hver gruppe, og at disse målingene gav et gjennomsnitt og empirisk standardavvik på henholdsvis 6.82 og 1.21 i gruppe 1 og på 6.96 og 1.14 i gruppe 2. Dersom man tester en nullhypotese om lik forventningsverdi mot alternativet ulik forventningsverdi for kolesterolmålinger i de to gruppene gir testen klar forkastning, *p*-verdien blir 0.008. Vi har med andre ord påvist en signifikant forskjell mellom kolesterolnivåene i de to gruppene. Men er dette en interessant forskjell? Neppe! Den estimerte forskjellen i forventningsverdi er 0.14, noe som er av marginal interesse når vi ser at standardavviket til målingene er over 1. På grunn av det store antallet målinger er vi imidlertid i stand til å påvise en så (uinteressant) liten forskjell. □

Eksemplet over illustrerer at vi ikke bare må se blindt på om vi får forkastning eller ikke, men at vi også må vurdere størrelsen på "effekten" (avviket fra nullhypotesen). Dette kan vi for eksempel gjøre ved å rapportere størrelsen på estimert effekt i tillegg til *p*-verdi, eller ofte enda bedre, ved å rapportere et konfidensintervall.

Oppsummering:

- Vær oppmerksom på "fisketurproblematikken" når du gjør mange hypotesetester.
- Når du får forkastning av en nullhypotese, se også på hvor stort det estimerte avviket fra nullhypotesen er og vurder om dette avviket er av praktisk interesse.

Gjennomføring av en statistisk undersøkelse

Til slutt noe tips om ting å tenke på dersom du skal gjennomføre en statistisk undersøkelse.

Planlegg undersøkelsen

Legg ned et grundig arbeid i planleggingen av undersøkelsen. Dersom planleggingen er for dårlig kan man ende opp med å gjennomføre en undersøkelse som viser seg å være verdiløs. Noen punkter å tenke på er listet opp under.

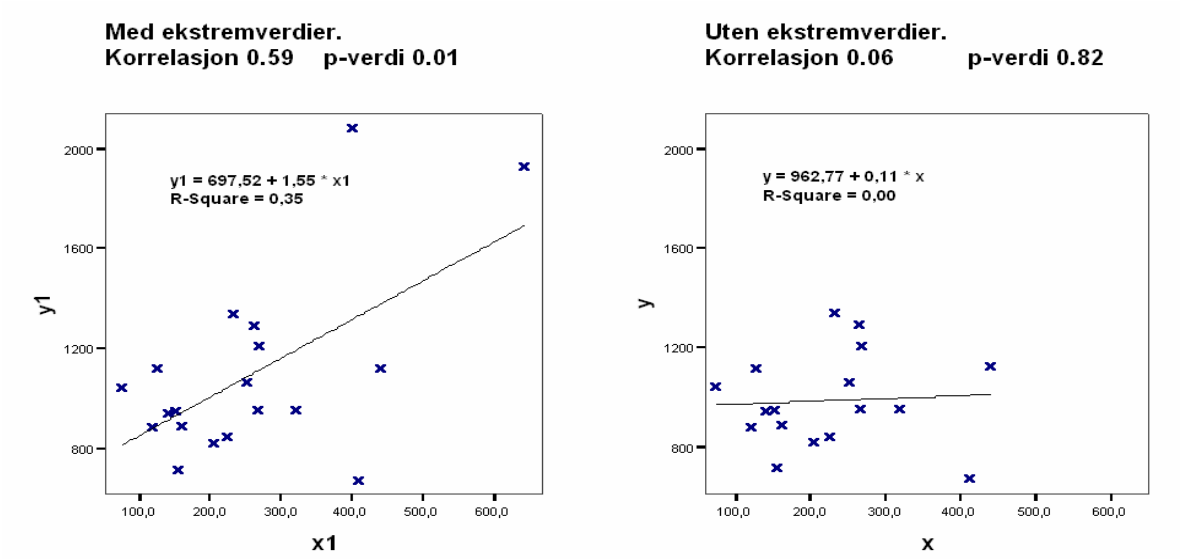
- Tenk grundig gjennom hva du ønsker å undersøke. Skriv ned de hypotesene du eventuelt ønsker å teste.
- Tenk gjennom hvordan du kan få samlet inn representative data (tilfeldig utvalg) for det du ønsker å undersøke.
- Tenk gjennom hvordan dataene skal analyseres etter at de har blitt samlet inn. Hva slags statistiske modeller og metoder kan brukes? Hvilke antagelser bygger disse modellene på og hva slags krav stiller dette eventuelt til datainnsamlingen? Kan du for eksempel regne med å få uavhengige data dersom metoden krever det?

- Beregn nødvendig utvalgsstørrelse – dvs hvor mye data du trenger å samle inn for med rimelig stor sannsynlighet å kunne få svar på det du ønsker å finne ut av med undersøkelsen. Typisk må du da ta stilling til hva som er minste "effekt" du synes er viktig å avdekke og hvor stor sannsynlighet du ønsker å for å avdekke en slik effekt.

Plott dataene

Når du har samlet inn dataene, lag alle slags typer relevante plott av dataene for å bli kjent med dataene. Mange viktige aspekter kan avdekkes ved plotting. For eksempel hovedeffektene i dataene og forskjellige typer problemer som for eksempel avhengigheter eller uteliggere/ekstremverdier.

Eksempel: Figur 7 under viser to plott av et datasett med en regresjonslinje tilpasset dataene. I plottet til venstre er hele datasettet vist. I dette plottet ser man at to av observasjonene avviker sterkt fra de øvrige observasjonene. Dersom man tar bort disse to observasjonene og beregner regresjonslinja på nytt får vi resultatet vist i plottet til høyre. Legg merke til den enorme innflytelsen de to ekstremverdiene har på regresjonlinja. Når disse to verdiene er tatt ut ser vi ingen sammenheng mellom *x* og *y*. Ekstremverdier kan ha stor innflytelse på regresjonslinjer, men lar seg lett avdekke ved plotting av dataene. Dersom vi har bare hadde sett på analyseresultatene og ikke plottet dataene kunne vi lett ha kommet til å dra feilaktige konklusjoner om sammenhengen mellom variablene. □



Figur 7 Regresjonslinje tilpasset datasett med og uten to ekstremverdier.

Sjekk modellantagelsene

I forbindelse med utføringen av selve dataanalysen bør du alltid sjekke så godt det lar seg gjøre at modellantagelsene som analysen bygger på holder. Ved plotting av residualer, normalplott eller andre fordelingsplott, ulike plott av selve dataene etc. kan ofte mange av de viktigste antagelsene sjekkes. Skriv ned alle antagelsene dataanalysen bygger på, særlig de antagelsene som er vanskelige å sjekke.

Andre problemer

Tenk gjennom om noen av de andre tingene diskutert i dette notatet er relevante å ta hensyn til for analysen av dataene, for tolkningen av resultatene eller for presentasjonen av resultatene. Tenk også gjennom om det er andre aspekter som kan ha betydning for analyse, tolkning og presentasjon.

Referanser

Aalen, O.O. (red) et. al. *Statistiske metoder i medisin og helsefag*. Gyldendal Akademisk, Oslo, 2006.

Benjamini, Y. og Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57: 289-300, 1995.

Box, G. E. P., Hunter, W. G. og Hunter, J. S. *Statistics for experimenters.* Wiley, New York, 1978.

Dimakos, X., Haff, I.H. og Løland, A. *A short course in statistics.* Notat fra Norsk Regnesentral. NR note no.: SAMBA/07/04, 2004.

Schweder, T. og Spjøtvoll, E. Plots of P-values to evaluate many tests simultaneously. *Biometrika*, 69: 493-502, 1982.

Simes, R. J. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73: 751-754, 1986.

Walpole. R.E., Myers, R.H., Myers, S.L. og Ye, K. *Probability and statistics for engineers and scientists*. Prentice Hall, New Jersey, 2002.