Arbeidsnotater fra Høgskolen i Buskerud nr. 73

ARBEIDSNOTAT

ARBEIDSNOTAT

Statistikk og SPSS for enkle undersøkelser

Knut W. Hansson

HØGSKOLEN i Buskerud

Arbeidsnotater fra Høgskolen i Buskerud

Nr. 73

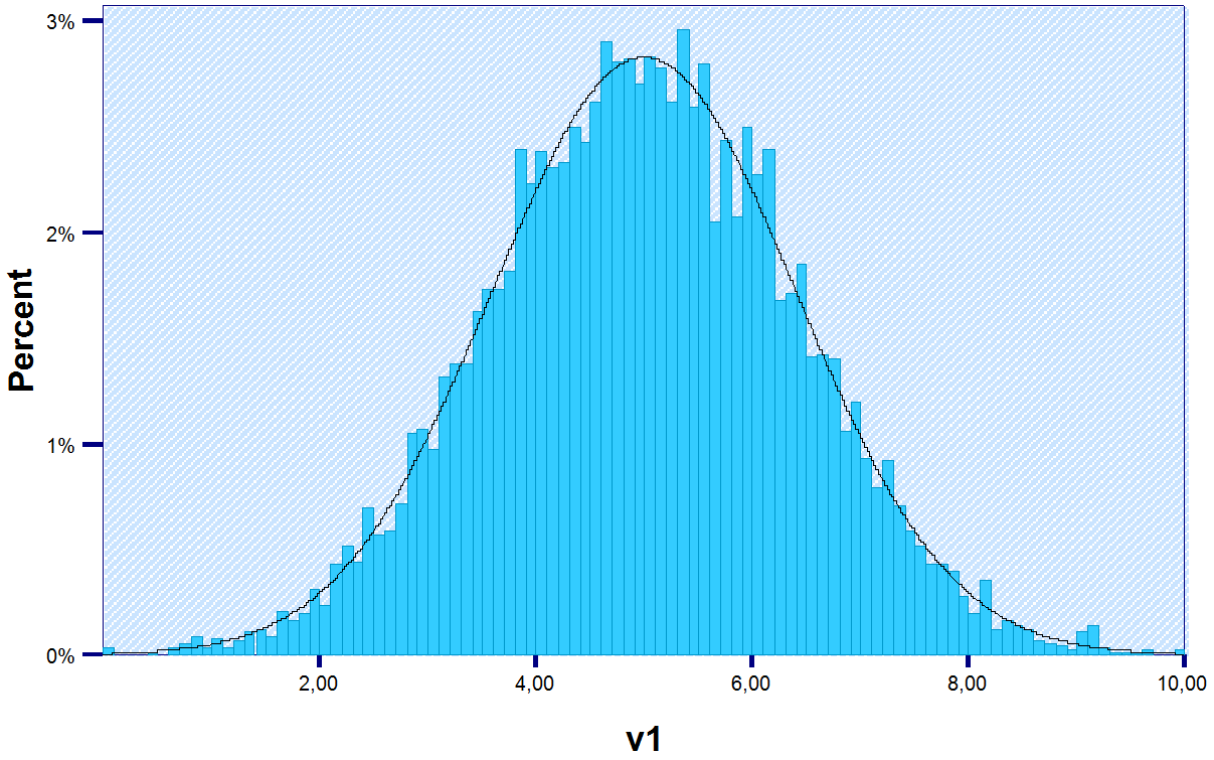
Statistikk og SPSS for enkle undersøkelser

Av

Knut W. Hansson

Hønefoss 2013

ISSN 1893-2398 (online)



Statistikk og SPSS

for enkle undersøkelser

Knut W. Hansson
Førstelektor IT
Høgskolen i Buskerud

Hønefoss, 2002/2013

Emneord:
statistikk
SPSS
undersøkelse

English keywords:
statistics
SPSS
survey

Sammendrag
Ved høyskolen i Buskerud, bachelorstudiene i IT, gjøres det noen ganger en undersøkelse som avsluttende Bacheloroppgave. Den gir da 7,5 studiepoeng.

Studentene har, enten parallelt eller på forhånd, hatt et relativt omfattende kurs i samfunnsvitenskapelig forskningsmetode (7,5 studiepoeng). Studentene blir noen ganger litt forvirret av dette kurset, da metode-kurset viser mange avanserte metoder som det ikke er grunnlag for i en enkel undersøkelse med lite utvalg.

Dette kompendiet er ment som en veiledning nettopp for slike enkle undersøkelser. Bare de mest aktuelle statistiske mål og teknikker er derfor tatt med.

En varm takk til min kollega og tidligere statistikklærer, høgskolelektor Jon Reinertsen, som har brukt tid på å lese igjennom dette kompendiet. Han har påpekt mange feil og uklarheter som jeg har forsøkt å rette etter beste evne.

Synopsis in English
At Buskerud College, as part of the bachelor IT education, sometimes students do a survey as their final Bachelor thesis. This gives them 7.5 ECTS credits.

The students have, either in parallel or earlier, had a fairly comprehensive course in social science research methods (7.5 ECTS credits). The students are sometimes a little confused by this course, since the research course shows them many advanced methods of which there is no basis in a simple survey with a small sample.

This compendium aims to be a guide for just such simple surveys. Therefore, only the most relevant statistical measurements and techniques are included.

Warm thanks to my colleague and former statistics teacher, lecturer Jon Reinertsen, who has spent time reading through this compendium. He has made me aware of many errors and ambiguities all of which I have tried to correct to the best of my ability.

Innhold

Introduksjon til kompendiet

Målgruppe

Dette kompendiet er skrevet for studenter som gjør enkle, kvantitative undersøkelser. Den er begrenset til det jeg tror studenter har bruk for i den anledning. Dette er *ikke* en lærebok for emnet "Samfunnsvitenskapelig forskningsmetode" som går mye lenger og dypere inn i statistikken, statistiske mål og teknikker og har med avansert bruk av SPSS eller tilsvarende.

Senere, i en jobbsituasjon, kan det være aktuelt med kvantitative undersøkelser enten for å få svar på et IT-spørsmål (valg av verktøy, valg av system) eller for å finne ut hva et litt større antall brukere har av ønsker og krav til et nytt system. Da skal man neppe gå dypt inn i en teoretisk forskning, men finne ut av enklere ting. Allikevel er det nyttig med strukturerte teknikker.

Det er for slike enkle, kvantitative undersøkelser at dette kompendiet er skrevet. Hvis det er få brukere eller de er godt representert, er det antakelig riktigere med kvalitative undersøkelser med intervju, fremvisning av prototyper eller annet.

Kompendiet egner seg ikke for "lesing" fra perm til per, men heller til oppslag og som "idébok". Jeg har lagt vekt på den praktiske anvendelsen av teknikkene og enkel bruk av SPSS.

Forkunnskaper

Kvantitative forskningsmetoder bygger på statistikk. Dette kompendiet forutsetter ikke slike kunnskaper, men bør kunne brukes greit uten. Det meste er forklart på et enkelt nivå uten å gå inn i teoriene bak.

Kapittel 1: Enheter, variable og verdier

Sentralt i samfunnsvitenskapelig forskning står empiri, altså data fra virkeligheten. Sentrale begreper er *enheter*, *variable* og *verdier*.

Enheter

Enhetene er de objektene du vil undersøke. Som oftest er det individer eller sosiale systemer som f.eks. organisasjoner, nasjoner eller grupper. (Enhetene kan sammenliknes med entiteter i datamodellering.)

Variable

Variablene er de egenskapene ved objektene som du vil studere. (Det kan sammenliknes med attributtene i datamodellering.) Variablene kan være

1) absolutte og referere til objektet selv, f.eks. individets vekt
2) relative og referere til andre objekter på samme nivå, f.eks. individets ektefelle
3) kontekstuelle og referere til metaenheter, f.eks. hvilken organisasjon individet er medlem av
4) kollektive og referere til subenheter, f.eks. hvilken del av nasjonen som stemmer ved Stortingsvalg

Variablene kan også inndeles i kvantitative som kan måles, og kvalitative som ikke er målbare. I samfunnsvitenskapelig forskning etterstrebes kvantitative variable.

Verdier

En *verdi* er den målte kvantiteten av en variabel og kan sammenliknes med domener i datamodellering. Man skiller mellom to hovedtyper av variable, etter hva slags verdier de kan ha:

1) Kontinuerlige variable som teoretisk kan ha en hvilken som helst verdi innen et intervall (i statistikk ofte omtalt som *variasjonsområde*), f.eks. vekt og alder. To enheter kan ikke ha nøyaktig samme verdi (hvis vi måler nøyaktig nok)
2) Diskrete variable som bare kan ha bestemte verdier fra en mengde (innen sitt variasjonsområde), f.eks. antall barn. Mange enheter kan ha samme verdi.

Ofte gjøres kontinuerlige variable diskrete i praksis, f.eks. når du måler alder i hele antall år, eller deler inne respondentenes alder i ”Barn, Ungdom, Voksen”. Andre ganger foretar du målingen på én måte, og endrer den senere: Du måler høyden så nøyaktig som du kan (tilnærmet kontinuerlig) men grupperer den senere til en grovere inndeling, f.eks.

Opp til 160 cm = ”Lav”
Over 160 opp til 180 cm = ”Middels”
Over 180 cm = ”Høy”

fordi du finner det mer ”fruktbart”.

De diskrete variablene kan igjen inndeles i følgende undertyper:

1) Nominalnivå der verdiene bare angir en klassifisering, f.eks. kjønn og nasjonalitet. Da kan du ikke snakke om ”størst” e.l.
2) Ordinalnivå der verdiene kan ordnes etter størrelse, f.eks. sosial status som ”Lav, Middels eller Høy”. Det er ingen fast avstand mellom verdien, f.eks. er ”Høy” større enn ”Middels”, men du vet ikke hvor meget høyere.
3) Intervallnivå der det er fast avstand mellom verdiene, f.eks. årstall. Det er like langt fra 1750 til 1850 som fra 1850 til 1950, men du kan ikke si at året 1800 er dobbelt så stort som året 900 (men kanskje ”dobbelt så bratt” jfr Øystein Sunde☺?). Bruker du muhammedansk tidsregning er det ikke lenger samme forhold mellom de aktuelle årstallene.
4) Forholdstallnivå der det er lik avstand mellom tallene, og de har et naturlig, absolutt nullpunkt. F.eks. er lengden av et veistykke målt i hele meter, fot eller favner ikke

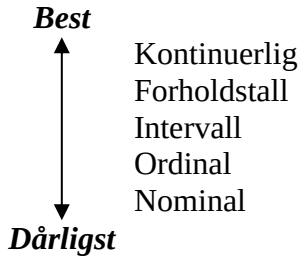
kontinuerlig (fordi den måles i hele meter) og den har et naturlig nullpunkt (et veistykke kan jo ikke ha en lengde mindre enn null). Da kan du si at et veistykke på 610 meter (= 2000 fot = 333 favner) dobbelt så langt som et på 305 meter (= 1000 fot = 167 favner). Det vil det være uansett hvilken skala du måler med og variabelen er da på forholdstallnivå. Tilsvarende er alder målt i hele år, måneder eller en annen tidsenhet, på forholdstallnivå. F.eks. er 20 år (240 måneder) dobbelt så gammel som 10 år (120 måneder).

Men temperatur målt i hele Celsiusgrader eller Fahrenheit-grader er ikke på forholdstallsnivå. F.eks. er 20^0C (=68^0F) er ikke dobbelt så varmt som 10^0C (=50^0F) fordi det avhenger av hvordan den måles. Amerikanere kan hevde at 64^0F (18^0C) er dobbelt så varmt som 32^0F (0^0C), men du som bruker Celsius ville nok stusse.

Det kan være avgjørende for nivå hvilken måleenhet du bruker til å måle variabelen med. Årstall kan f.eks. måles på mange måter, slik:

1) 1=Før Kristus, 2=Etter Kristus: Nominalnivå (ren klassifisering).
2) 1=Førhistorisk tid, 2=Middelalder, 3=Moderne tid: Ordinalnivå (kan ordnes etter størrelse).
3) Tallet angir et helt årstall etter Kristen tidsregning: Intervallnivå (fast avstand mellom tallene, men ikke noe rimelig forhold mellom verdiene)
4) Tallet angir et helt årstall siden ”The Big Bang”: Forholdstallnivå (som intervallnivå, men her finnes det et naturlig, absolutt nullpunkt).
5) Tallet angir et tidspunkt (et årstall med fritt antall desimaler): Kontinuerlig (en hvilken som helst verdi innen et intervall).

Av hensyn til de statistiske analysene bør du tilstrebe å måle variablene på så ”høyt” nivå som mulig. Rekkefølgen er da



Det holder imidlertid ikke å lage variable på "kunstig" høyt nivå. Hvis spørsmålet er et ja/nei-spørsmål kan du ikke be respondentene svare med et tall fra 1 til 7 (Lickertskala). Slike spørsmål er og blir nominelle. "Angi med et tall fra 1 til 7 om du noen ganger bruker PCen til spill" blir da helt feil.

Når variabelen ikke lar seg måle direkte, må du operasjonere. Det vil si at du måler noe annet, som du mener gir uttrykk for det du egentlig skal måle. Hvis du f.eks. vil måle sosial status, kan du kanskje bruke inntekt og/eller utdanningsnivå, skal du måle tilfredshet kan du spørre enhetene om det, hvis du vil måle etniske fordommer kan du spørre enhetene om hvor mange ”fremmedkulturelle” det bor i deres boligområde og sammenlikne resultatet med offentlig statistikk.

Graden av samsvar mellom den ønskede verdien og det du faktisk måler, kalles validitet (gyldighet = du måler det du faktisk ønsker å måle). Graden av nøyaktighet i målingene, kalles reliabilitet (= kan du stole på målingene).

Kapittel 2: Tips om SPSS før analysen

Som "IT-eksperter" bør dere selv kunne finne ut av bruken av SPSS. Blant annet er det en "Tutorial" innebygget som viser prosedyrene trinn for trinn. Likevel vil jeg gi noen tips.

1) Før du begynner med SPSS, så se nøye igjennom svarskjemaene du har fått. (Hvis du samler dataene automatisk i en database, bør du ta en utskrift.) Dette kalles å "kode" svarene. Noen svar må forkastes, fordi de åpenbart er bare tull, eller totalt misforstått. Videre kan det være aktuelt å gruppere svar, f.eks. åpne svar, i kategorier. Du bør tenke igjennom hvilke koder du skal bruke for flervalgspørsmål, hva variablene skal kalles osv. Det er også ofte aktuelt å gi koder for "Ikke besvart", "Vet ikke" og "Uaktuelt" – det siste for spørsmål som respondenten ikke skal svare på.
2) Hvis dataene er samlet inn i en database, kan de overføres til SPSS. Det blir bedre enn å taste dem inn pånytt. Det enkleste er å eksportere dataene til et regneark, og så importere dem til SPSS derfra (*File|Open|Data*). Hvis dataene er samlet på papir (spørreskjema, intervjureferat o.l.), bør du vente med inntastingen i SPSS til du har definert variablene.
3) Definer variablene omhyggelig. Du kan sette
   a) **Variabelnavn ('name')** som er det navnet det henvises til i SPSS, f.eks. i formler.
   b) **Type ('type')** er datatypen. SPSS har datatypene *tall* i forskjellige formater, *dato* som i realiteten er et tall og *string*. Den siste regner SPSS som nominal eller ordinalnivå. Hvis du vil regne gjennomsnitt e.l. må du altså kode tekstene som tall.
   c) **Bredde ('width')** som er det antall tegn/siffer som maksimalt skal vises. Tekster blir trunkert. Tall vises likevel i sin helhet utover dette, hvis det er plass i kolonnen.
   d) **Desimaler ('decimals')** som er antall plasser bak komma i tall.
   e) **Etikett ('label')** som er den teksten som står i tabeller og andre analyser.
   f) **Verdier ('values')** som er de tekstene som hører til en gitt kode, f.eks. 3="Ung". Jeg har ikke møtt noen grense for antall verdier, men hvis det er mer enn 24 (kan endres i *Edit|Options|Interactive*), vil SPSS gjøre om målenivået til 'Scale' (se nedenfor).
   g) **Manglende verdier ('missing')** som er de verdiene som skal regnes som ubesvart. Du kan kategorisere inntil tre verdier som "Missing". De vil bli tatt med i noen sammenhenger, f.eks. når dataene telles opp, men blir ikke med i statistiske analyser, f.eks. i gjennomsnitt, prosentfordeling osv. Du kan endre dette under analysen.
   h) **Kolonner ('columns')** som er bredden på kolonnen i antall tegn/siffer.
   i) **Tekstjustering ('align')**, altså hvordan teksten skal justeres i kolonnene: Venstre, høyre eller sentrert.
   j) **Målenivå ('measure')** der du velger mellom
      i) *'Scale'*, dvs verdier som enten er
         - ✓ kontinuerlige, der ethvert tall innen et intervall er lovlig, f.eks. tidspunkt
         - ✓ diskrete på intervallnivå, der tallene har fast avstand, f.eks. årstall eller
         - ✓ diskrete på forholdstallsnivå, der tallene har fast avstand og et absolutt nullpunkt, f.eks. høyde 1=0-50 cm, 2=51-100 cm, 3=101-150 cm osv, alder i hele år og inntekt i hele tusen
      ii) *'Ordinal'*, dvs verdier som har en naturlig sortering, f.eks. "lav, middels, høy" eller "barn, ungdom, voksen".
      iii) *'Nominal'* dvs verdier som ikke har noen naturlig sortering, f.eks. nasjonalitet og kjønn.

Tekstverdier kan være på ordinal- eller nominalnivå, men rekkefølgen av ordinalverdiene vil være den alfabetiske, f.eks. "Høy, Lav, Middels". Det vil sjelden passe, og det er derfor meget tryggere å bruke tallkoder. Evt. kan du kalle verdiene "01-Lav", "02-Middels" og "03-Høy". (Det må være like mange sifre i tallene, ellers blir "21" mindre enn "3".)

4) Hvis dataene er innsamlet på skjemaer kan du nå taste dem inn i SPSS. Etter inntastingen bør du ta utskrift og kontrollere mot skjemaene – det er fort gjort å taste feil!
5) Første del av analysen bør være å analysere variablene hver for seg. Sammenlikn dine resultater med andre kilder der det er relevant (f.eks. antall kvinner i utvalget mot antall kvinner i populasjonen).
6) Deretter sammenlikner du to og to variable[1]. Hvis du bare har ett utvalg, sammenlikner du to og to variable for å finne interessante ting (f.eks. kjønn mot IT-bruk). Hvis du har to, tilfeldige utvalg fra hver sin populasjon, bør du også lete etter interessante forskjeller/likheter mellom utvalgene (f.eks. IT-bruken til hvert av de to utvalgene eller samvariasjonen mellom kjønn og IT-bruk for hvert utvalg).

[1] Statistisk og samfunnsvitenskapelig sett er en slik "fisking" ikke bra. Du skal først ha en begrunnet formodning om samvariasjon mellom variablene, deretter lager du en undersøkelse/eksperiment for å kontrollere det, og så til slutt sjekker du om samvariasjonen virker sannsynlig ut fra utvalget ditt. I praksis er likevel slik "fisking" ganske vanlig.

Dette bør du passe på når du skriver rapporten – forklar først hvorfor det er grunn til å tro noe, deretter hva du gjorde (målte) for å finne det ut og tilslutt hva du fant. **OBS! Jeg vil ikke siteres på dette rådet i rapporten☺!**

Kapittel 3: Beskrivelse av utvalget generelt

Før du går løs på å beskrive variablene, bør du diskutere selve utvalget. Du vet jo ingenting om populasjonen som utvalget er trukket fra, men om utvalget vet du en god del. Hvis diskusjonen av hva du fant i utvalget skal være interessant, må du først overbevise om at utvalget er *representativt*. Da kreves det for det første at utvalget er tilfeldig.

Videre er det sjelden slik at alle som ble forespurt faktisk svarte. Spørsmålet er da hvorfor de ikke gjorde det – en såkalt *frafallsanalyse*. Før et amerikansk presidentvalg forutså meningsmålinger at den republikanske kandidaten ville vinne klart, men ved valget vant den demokratiske kandidaten. Det viste seg at undersøkelsen var gjort (med tilfeldig utvalg) med telefon. Da republikanske velgere hadde høyrere gjennomsnittsinntekt og derfor også flere av dem hadde telefon, ble demokratiske velgere klart underrepresentert i utvalget. Altså: Hvis det kan tenkes at ikke alle i populasjonen er kommet med i en "riktig" andel, blir resultatet feil. Du kan ofte undersøke litt om frafallet ved å se om utvalget har samme andel kjente karakteristikker som f.eks. kjønn, alder som i populasjonen. Ellers blir dette mest en drøfting utfra hvordan utvalget er fremkommet.

Kapittel 4: Beskrivelse av utvalget – én og én variabel i utvalget

Du ønsker nok gjerne å si noe om *populasjonen*. Det kan du bare gjøre med usikkerhet. Hvor stor usikkerheten er, varierer med hvor stort utvalget er evt. hvor stor andel av populasjonen du har med i utvalget. Du kan imidlertid uttale deg med sikkerhet om utvalget, for det har du faktisk sjekket. Husk riktignok på at det kan være feil også her, f.eks. kan respondentene ha misforstått spørsmål, svart feil ved et uhell eller med vilje, du kan ha tastet feil i SPSS og annet.

Når dataene er lagt inn i SPSS, er tiden kommet for analysen. Du skal da beskrive dine funn fra empirien på en ryddig og oversiktlig måte, og diskutere det. Hvis du bare har undersøkt tre enheter, blir jo en slik beskrivelse triviell, men normalt vil du ha et betydelig antall enheter og da gir det ingen oversikt å liste opp alle verdiene for hver variabel. En liste med 200 aldre får f.eks. ingen noe ut av. Da må du oppgi forskjellige statistiske mål, der detaljene skjules men mønstre kommer klarere frem. Det kan f.eks. være betydelig mer interessant å få vite at "gjennomsnittsalderen var 32,7 år, ingen var yngre enn 15 år og ingen eldre enn 48, og at 78 % var mellom 28 og 37 år".

SPSS har en utmerket *Statistics Coach* under hjelpmenyen som du absolutt bør prøve.

SPSS skiller mellom frekvensanalyser (*Analyze*|*Descriptive statistics*|*Frequencies*) og beskrivende statistikk (*Analyze*|*Descriptive statistics*|*Descriptives*).

Beskrivelse med en tabell

Diskrete variable kan beskrives i en tabell, med absolutte tall og prosenter[2], f.eks.:

ALDER

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Ung	150	25,0	27,3	27,3
	Voksen	350	58,3	63,6	90,9
	Gammel	50	8,3	9,1	100,0
	Total	550	91,7	100,0	
Missing	Ubesvart	50	8,3		
Total		600	100,0		

Legg merke til at "Ubesvart" er satt opp her som "missing" og at det beregnes prosent både med og uten de "Ubesvarte". Den kumulative prosenten gjør det enklere å se hvor mange som totalt er "mindre enn", f.eks. at 90,9 % av dem som svarte er "Voksen" eller yngre.

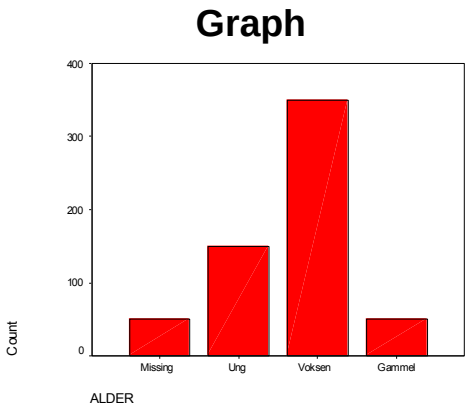
Kontinuerlige variable kan ikke beskrives direkte i en tabell.

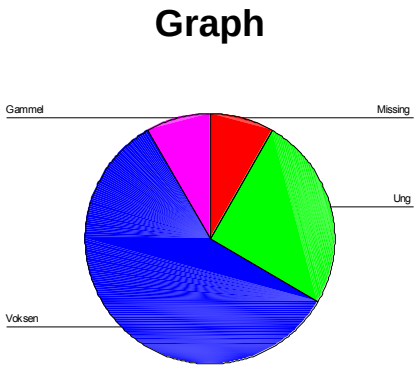
Beskrivelse med en graf

Diskrete variable kan grafes, f.eks. med stolper ('bar') eller med "bløtkake-diagram" ('pie')[3] som viser antall eller prosentverdi i hver kategori.

[2] Denne tabellen er laget med SPSS, eksportert i HTML-format og hentet til Word med *Sett inn*|*Fil*.

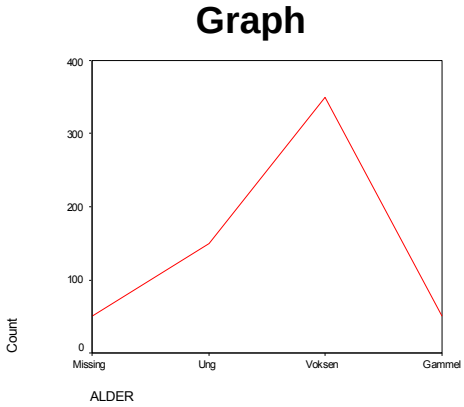
[3] Her skar det seg med eksport av grafikken til HTML, så isteden har jeg kopiert "objektene" i SPSS og limt inn i Word.





Diagrammene blir like, enten du bruker antall eller prosenter, men i stolpediagrammer kan prosenter være uheldig fordi det skjuler usikkerhet. Hvis du f.eks. har spurt bare fire personer er det mer opplysende å få vite at tre av de fire var ”fornøyd” enn at 75 % var det. Tilsvarende er det f.eks. bedre å få vite at det ble solgt fire biler i år mot tre i fjor, enn å bli fortalt at salget er gått opp med 33 %. Med store tall, kan det fort være omvendt, fordi det er vanskelig å forholde seg til dem: ”2.306 er fornøyd og 1.254 er misfornøyd” gir leseren mindre enn at ”65 % er fornøyd og 35 % misfornøyd”. Uansett vil du ofte i teksten gi begge tallene: ”65 % (2.306) er fornøyd og 35 % (1.254) er misfornøyd”. Selve diagrammet blir imidlertid likt, det er bare skalaen på venstre side som endres fra antall til prosent.

SPSS kan også skrive ut grafen slik:



men det er ikke bra i denne forbindelse, fordi det ser ut som om variabelen er kontinuerlig (riktignok med brudd), hvilket jo er usant her.

Kontinuerlige variable kan ikke grafes alene, men to kontinuerlige variabel kan plottes mot hverandre i et koordinatsystem slik det er vist nedenfor i avsnittet ”Sammenlikning av to kontinuerlige variable”.

Beskrivelse med statistiske mål

Antall i utvalget ('N')

Gir bare antallet. Det er viktig for å kunne si noe om i hvilken grad utvalget kan anses som representativt. Det kan også være interessant å se det i forhold til populasjonen, gjerne angitt som en prosentdel hvis populasjonen er begrenset[4].

Svarprosent

Svarprosenten angir prosenten av dem du fikk undersøkt (N) i prosent av dem du ville undersøke (det ønskede utvalget). Dette er interessant som en indikasjon på reliabilitet. Hvis svarprosenten er lav, må du diskutere hva grunnen kan ha vært til at så mange ikke svarte, og om det kan bety at utvalget ditt ikke er representativt på en eller annen måte. Lav svarprosent skaper en mistanke om at utvalget ditt er "skjevt". F.eks. er det vel ikke urimelig at mange som er syke, lar være å svare på om de "får eller har fått behandling for en psykiatrisk sykdom"? De syke vil da være underrepresentert i ditt utvalg. Du bør vel heller ikke spørre i et spørreskjema om respondenten "kan lese og skrive norsk".

Svarprosent gjelder både for undersøkelsen som helhet og for hver variabel for seg.

Sum av verdiene ('sum')

For kontinuerlige variable kan dette være interessant, f.eks. når "den utkjørte distansen for alle bilene i undersøkelsen var 23.507 km til sammen".

For diskrete variable som representerer et intervall, må du bruke summen med stor forsiktighet. Det har f.eks. liten interesse å vite at summen av alle alderskategoriene (0, 1, 2 og 3 i eksemplet ovenfor) er 1000, men hvis du vet at dataene er kodet slik:

Kode	Tekst	Alder[5]	Midt
0	Ubesvart	-	-
1	Ung	[0..20>	10
2	Voksen	[20..60>	40
3	Gammel	[60..80>	70

kan du benytte midten av klassen og summere. (Summen blir her 19.000 og det sier deg ikke så meget – bare at respondentene var 19.000 år gamle tilsammen). For å få til dette, ber du SPSS kode om (*Transform*|*Recode*) – her ber du om at $1 \Rightarrow 10$, $2 \Rightarrow 40$ og $3 \Rightarrow 70$. Det er ikke alltid like lett, hvis f.eks. kode 3 = "gammel" angir "over 60 år", er det vanskelig å vite hvor midten er. Videre er forutsetter dette at verdiene innen hver klasse fordeler seg likt omkring midten, altså at gjennomsnittet for klasse 1="ung" er 10. Det vet du jo ikke noe om, hvis du da ikke målte variabelen kontinuerlig og grupperte selv. Her vil jeg si at det er svært tvilsomt om de i gruppe 3="gammel" faktisk er gjennomsnittlig 70 – sannsynligvis er de nærmere 60 i gjennomsnitt.

For nominelle variable blir summen vanligvis meningsløs. Det er vel ikke særlig interessant at summen av alle kjønn er 28.706 (1="Mann" og 2="Kvinne")?

Gjennomsnitt ('mean')

Gjennomsnittet defineres som summen av alle verdiene, delt med antallet. Det er ett av flere sentralmål[6] og angir det som er "typisk" eller "sentralt". Hvis summen kan beregnes og har

[4] Hvis populasjonen er svært stor, blir prosentandelen alltid svært liten og gir lite informasjon. Du kan ikke regne prosenter av en uendelig stor populasjon.

[5] Matematisk skrives ofte intervaller med komma mellom grenseverdien, slik: [0, 20>. Jeg velger å bruke prikker mellom grenseverdiene for å unngå problemer med desimaltall, f.eks. [5,5, 3,2>

[6] De sentralmålene som omtales her, er gjennomsnitt, median og typetall.

mening, har sannsynligvis gjennomsnittet det også, og kanskje har bare gjennomsnittet mening. Hvis summen på den annen side ikke kan beregnes, kan heller ikke gjennomsnittet det.

For kontinuerlige variable kan gjennomsnittet beregnes, og har som oftest mening, f.eks. når " hver bil i utvalget ble kjørt gjennomsnittlig 2.350,7 km".

For variable på minst ordinalnivå, kan også gjennomsnittet være interessant. F.eks. er gjennomsnittsalderen i eksemplet ovenfor 34,5 år. Forbeholdet om at "midt i klassen" er fornuftig gjelder fortsatt fullt ut.

Nominelle variable bør ikke beskrives med gjennomsnitt. Hva betyr det f.eks. at "gjennomsnittskjønn er 1,62"?

Som nevnt under avsnittet om konfidensintervall og hypotesetesting ovenfor, kan gjennomsnittet i utvalget gi utgangspunkt for å anta noe om (estimere) populasjons-gjennomsnittet. SPSS kan gi deg et mål for hvor mye du må regne med at gjennomsnittet vil variere hvis du hadde tatt mange utvalg av den størrelsen du har brukt. Variasjonen måles med standardfeil for gjennomsnittet ('S.E.Mean' = 'Standard Error of Mean'). Du kan omtrent regne med at populasjonsgjennomsnittet ligger innenfor utvalgets gjennomsnitt ±2*'S.E.Mean'[7]. I eksemplet med alder, er 'S.E.Mean' ca 0,74 og antallet stort (N = 550). Du kan da tillate deg å anta med rimelig sikkerhet at populasjonens forventning ligger innenfor konfidensintervallet

$$34{,}5 \pm 2 * 0{,}74 \approx 34{,}5 \pm 1{,}5 = [33 \,..\, 36]$$

Dette er altså 95 % konfidensintervall for populasjonsgjennomsnittet. Dette gjelder bare hvis utvalget er tilfeldig!

Typetall ('mode')

Typetallet er den verdien det er flest av ('mode' er fransk for mote, altså den verdien som "er på moten"). I alderseksemplet ovenfor er typetallet "voksen" = 40, fordi det er flest i denne gruppen.

Median ('median')

Medianen er den verdien som kommer i midten, hvis verdiene sorteres etter størrelse. For alderseksemplet ovenfor er medianen 40. Halvparten av aldrene er altså mindre/lik 40, den andre halvparten større/lik. Hvis antallet i utvalget er et partall, finnes det to verdier i midten, da brukes ofte gjennomsnittet av de to.

Kvartiler ('quartiles'), percentiler ('percentiles')[8]

Verdiene sorteres stigende etter størrelse. Den verdien som er f.eks. 35 % fra start, kalles "35 %-percentilen". I et utvalg på 1.000 enheter er det altså verdi nr 350. 25 %- og 75 %-percentilen er så vanlige at de har fått eget navn, nemlig kvartilene, og 50 %-percentilen heter altså median. Medianen er et sentralmål, men andre percentiler gir et inntrykk av hvordan verdiene fordeler seg.

Hvis verdiene dine er "midtverdier" i klasser, kan du be SPSS beregne en antatt median eller andre percentiler (klikk for *'Values are group midpoints'*). Antakelsen bygger på at verdiene er jevnt fordelt innen klassen, noe du bør drøfte særskilt.

[7] Egentlig varierer denne faktoren. Hvis utvalget er rimelig stort, større enn 30, er tallet svært nær 1,96, men ved mindre utvalg er tallet større. Faktoren 2 er altså bare en "tommelfingerregel".

[8] Benevningen varierer litt. Hvis det brukes andeler (⅓, ½, ¾ osv) brukes ofte uttrykket fraktiler (av eng. 'fraction' = brøk). 25 %-percentil kalles også "25 %-fraktil".

Her er noen percentiler, kvartilene og medianen i alderseksemplet, beregnet på den måten:

Statistics

Alder utfra midten av hver klasse

N	Valid	550
	Missing	50
Median		34,0000[a]
Mode		40,00
Percentiles	25	17,5000[b]
	35	24,1000
	40	27,4000
	50	34,0000
	75	53,1250

a. Calculated from grouped data.

b. Percentiles are calculated from grouped data.

Legg merke til advarslene som står under tabellen!

Minimum/maksimum ('minimum'/maximum')

Den minste og den største verdien i utvalget, gir et inntrykk av hvor spredte verdier du fant. F.eks. vil mange oppfatte disse to undersøkelsene av rekrutter som ganske forskjellige:

Mål	Undersøkelse 1	Undersøkelse 2	Undersøkelse 3
Gjennomsnittshøyde	180 cm	180 cm	180 cm
Minste høyde	165 cm	175 cm	179 cm
Største høyde	201 cm	185 cm	189 cm

De viser samme gjennomsnitt, men undersøkelse 1 spenner over størst intervall. Minimum og maksimum viser bare absolutte ytterpunkter for verdiene og kan være nokså avhengig av tilfeldigheter.

Variasjonsbredden ('range')[9]

Variasjonsbredden er forskjellen på minimum og maksimum. Det skjuler evt. skjevheter. I rekrutteksemplet ovenfor er variasjonsbredden henholdsvis 36, 10 og 10. Det siste tallet skjuler en åpenbar skjevhet; mange flere er nok under 180 enn over siden gjennomsnittet ligger så nær minimumsverdien.

Variasjonsbredden forteller noe om hvor spredt verdiene er, og er et av flere spredningsmål[10]. Variasjonsbredden er et "primitivt" spredningsmål, fordi det bare tar hensyn til to av verdiene.

Kurtose[11] ('kurtosis')

Kurtose angir hvor "spiss" fordelingen er i forhold til normalfordelingen. Etter den definisjonen som SPSS benytter[12], har normalfordelingen kurtose 0. Hvis kurtosen er positiv, er verdiene fordelt "spissere" (altså mindre spredt) enn normalfordelingen. Hvis kurtosen er negativ, er verdiene fordelt "flatere" (mer spredt) omkring gjennomsnittet enn normalfordelingen. Også her vil SPSS oppgi kurtosens standardavvik, og du kan regne med at

[9] Variasjonsbredden kalles også amplituden (= utslaget til hver side i en svingning)
[10] De spredningsmål som omtales her er percentiler, variasjonsbredde, kurtose og varians/standardavvik.
[11] Er i noe litteratur kalt "Eksess".
[12] Normalfordelingens kurtose er pr definisjon 3, men mange bruker en variant der de trekker 3 fra den beregnede kurtose, som da blir 0 for normalfordelingen. SPSS beregner den siste varianten.

også populasjonen har positiv/negativ kurtose hvis absoluttverdien av utvalgets kurtose er mer enn to ganger sitt eget standardavvik. I alderseksemplet ovenfor er kurtosen litt negativ, så verdiene er noe mindre spredt enn om de hadde vært normalfordelt[13].

Asymmetri ('skewness')

Asymmetri eller skjevhet viser om verdien er skjevt fordelt rundt gjennomsnittet. Hvis verdiene er helt symmetrisk fordelt er symmetrien 0. Hvis den er positiv, er den skjevfordelt mot større verdier. Hvis den er negativ er den skjevfordelt med flere mindre verdier. Som en tommelfingerregel, kan du anta at også populasjonen er skjevt fordelt hvis absoluttverdien av utvalgets asymmetri er mer enn dobbelt så stor som sitt eget standardavvik (det oppgis av SPSS). Alderseksemplet ovenfor har en liten, positiv asymmetri (ganske mange verdier er jo 40 og 70, altså større enn gjennomsnittet 34,5)[14].

Varians/standardavvik ('variance'/'standard deviation')

Varians og standardavvik er de mest brukte, og på mange måter beste, spredningsmål. De angir spredningen rundt gjennomsnittet. Det er tett samvariasjon mellom de to, da standardavviket er kvadratroten av variansen[15]. SPSS beregner begge på forespørsel.
Du kan regne med at minst 95 % av verdiene i utvalget ligger rundt utvalgets gjennomsnitt og innenfor to ganger standardavviket på hver side, altså at 95 % av verdiene ligger i intervallet

[Gjennomsnitt - 2 * Standardavvik .. Gjennomsnitt + 2 * Standardavvik]

Hvis du f.eks. i et utvalg har funnet at gjennomsnittlig høyde på rekrutter er 180 cm og standardavviket er 10 cm, så er minst 95 % av rekruttene i utvalget mellom 160 cm (180 - 2 * 10) og 200 cm (180 + 2 * 10), altså i intervallet [160..200].

Hvis variansen/standardavviket er beregnet for et tilfeldig utvalg, kan de også være utgangspunkt for å tro noe om hele populasjonens varians/standardavvik på samme måte som for gjennomsnittet. Jo større utvalget er, desto nærmere populasjonens varians/standardavvik kan du regne med å komme, men SPSS beregner ikke varians/standardavvik for populasjonens varians/standardavvik (en slags 'Standard Error of Variance') selv om den finnes for normalfordelte verdier.

---

[13] Du kan imidlertid ikke forkaste en antakelse om at populasjonen likevel er normalfordelt.

[14] Du kan imidlertid ikke forkaste en antakelse om at populasjonen likevel er symmetrisk fordelt.

[15] Variansen beregnes ved å se på hver verdi og hvor meget den avviker fra gjennomsnittet. Disse avvikene kvadreres til et "kvadratavvik" som så summeres sammen. Det gjennomsnittlige kvadratavviket, kalles variansen. Slik ser formelen ut:

$$VAR = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Her er n antallet i utvalget og $\overline{X}$ er gjennomsnittet av verdiene. For å beregne gjennomsnittet, skulle du jo egentlig delt med n og ikke (n-1), men begrunnelsen for det skal jeg ikke gå nærmere inn på her.

Standardavviket er kvadratroten av variansen, altså

$$STD = \sqrt{VAR} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Kapittel 5: Å si noe om populasjon basert på utvalget

Hvis utvalget er stort nok, *kan* det være aktuelt å prøve å si noe om populasjonen utfra det du fant i utvalget. For enklere undersøkelser er det – i motsetning til forskning – er det lite aktuelt. Jeg tar allikevel med litt om det her, men du bør søke veiledning før du benytter deg av disse teknikkene.

Konfidensintervall og hypotesetesting

Når du har undersøkt enhetene, kan du beskrive det du har funnet på mange måter. Du kan oppgi hvor mange det er, hvordan verdiene fordeler seg for hver variabel du har målt og du kan angi mange statistiske mål som f.eks. gjennomsnitt. Det du da beskriver er sikkert og eksakt – i den grad verdiene er reliable.

Hvis du har undersøkt samtlige enheter som finnes, hele populasjonen, gjelder beskrivelsene dine også for den. Ofte må du imidlertid nøye deg med å undersøke bare en del av populasjonen, et utvalg. Kanskje er det for mange enheter i populasjonen til at du kan undersøke dem alle, kanskje er noen av dem ikke til å få tak i, eller kanskje har noen nektet å la seg undersøke. Selv om du bare har undersøkt noen enheter fra populasjonen, vil du likevel gjerne si noe om den[16]. Oftest er det selve hensikten med undersøkelsen.

Det er ikke helt enkelt. Når du bare undersøker et utvalg, må du jo regne med at de statistiske målene kan falle litt tilfeldig ut, avhengig av tilfeldigheter i utvalget. Selv om det er 52,3 % kvinner i Norge, er det jo ikke sikkert at du får 52,3 % kvinner i ditt utvalg, og omvendt.

De fleste statistiske målene som beskrives her, er såkalt forventningsrette. Det vil si at hvis du tar mange utvalg fra populasjonen og beregner det samme statistiske målet for hvert utvalg, vil resultatene variere på begge sider omkring populasjonens mål. F.eks. er utvalgets gjennomsnitt forventningsrett for populasjonsgjennomsnittet. Du kan da "gjette" på at populasjonsgjennomsnittet er likt det gjennomsnittet du fant i ditt utvalg. Fordi du ikke undersøkte hele populasjonen, må du finne deg i at du ikke kan angi populasjons-gjennomsnittet nøyaktig, bare som en antakelse (estimat) basert på utvalget ditt.

Det er imidlertid, i mange tilfelle, mulig å si noe om hvor nøyaktig du kan angi populasjonens statistiske mål. Det er kjent at jo større utvalget er i absolutte tall eller i forhold til populasjonen, desto mer vil f.eks. utvalgets gjennomsnitt nærme seg populasjonsgjennomsnittet[17]. Helt tilsvarende vil andel kvinner i utvalget, nærme seg kvinneandelen i populasjonen etterhvert som utvalgets størrelse øker. Legg merke til at det står "nærme seg"; når som helst kan det dukke opp et utvalg som skiller seg sterkt fra populasjonen som helhet. Men sjansen for det blir mindre når utvalgsstørrelsen øker.

Du må altså nøye deg med å si noe slikt som at "det statistiske målet for populasjonen er sannsynligvis omtrent X" der X er det samme målet i henhold til utvalget. Ved hjelp av statistisk analyse, kan det også være mulig å si noe mer presist om hvilke grenser populasjonens mål antakelig ligger innefor. Populasjonsmålet er jo fast, men fordi du bare har målt et utvalg, er du usikker på hva populasjonsmålet egentlig er. Du angir derfor ofte populasjonsmålet med et konfidensintervall med et visst konfidensnivå, som "X ± Y på Z % konfidensnivå", der X er hentet fra utvalget ditt og Y har en fast, angitt størrelse avhengig av hvor sikker du vil være (Z %) på at intervallet omfatter det virkelige populasjonsmålet. I et

[16] Det å mene noe om populasjonen basert på undersøkelse av et utvalg, kalles **statistisk induksjon**.
[17] Dette kalles "**Bernoullis lov om de store tall**" eller bare "**de store talls lov**".

bestemt tilfelle kan du f.eks. kanskje si at "på 95 % konfidensnivå er kvinneandelen i Norges befolkning er 51,6 % ± 1,4 % basert på mitt utvalg". Det betyr at ut fra din undersøkelse, er du 95 % sikker på at kvinneandelen i befolkningen er mellom 50,2 % og 53 %, og det er 5 % sjanse for at kvinneandelen i virkeligheten ligger utenfor dette intervallet[18]. Jo sikrere du vil være på at den virkelige kvinneandelen i befolkningen ligger innenfor intervallet, desto bredere må intervallet være. Det er vanlig å bruke 95 % eller 99 % konfidensnivå. Etter de store talls lov, vil konfidensintervallet bli smalere og smalere for et gitt nivå med økende antall i utvalget.

Forutsetningen for å bruke statistisk metode på denne måten, er at utvalget er valgt ut helt tilfeldig. Ordet "tilfeldig" brukes litt løst i norsk, f.eks. "vi spurte noen tilfeldige kunder på handlesenteret". Her er det imidlertid snakk om statistisk tilfeldighet, kalt stokastisk, som innebærer at det er foretatt en form for loddtrekning som er slik at alle enheter i populasjonen i utgangspunktet har samme sannsynlighet for å komme med i utvalget. Det er neppe slik det er gjort på handlesenteret og utvalget er derfor *ikke* tilfeldig i statistisk forstand. Utvalget kan da statistisk sett ikke si noe som helst om populasjonen "alle kunder på handlesenteret". Selv om du klarer å trekke lodd mellom alle kunder på handlesenteret på en viss tid, en viss dag eller lignende er utvalget ikke stokastisk fordi bare noen av handlesenterets kunder var der på denne tiden. Hvis 80 % svarer på en spørreundersøkelse, er det heller ikke lenger stokastisk – det er ikke tilfeldig hvem som svarte. Her syndes det dessverre mye i samfunnsvitenskapelig forskning.

Hvis ikke utvalget er gjort helt tilfeldig, altså *stokastisk*, faller alle muligheter for å si noe om populasjonen bort[19]. Ofte er det også andre forutsetninger for å anvende en statistisk teknikk f.eks. at populasjonen er normalfordelt, at populasjonen er uendelig eller i alle fall meget stor, eller at utvalget er stort (absolutt eller i forhold til populasjonen). Nedenfor vil slike andre betingelser bli uttrykkelig angitt.

I andre tilfeller kan det være interessant å spørre om det er sannsynlig at et gitt statistisk mål gjelder for populasjonen, når du vet det du fant i ditt utvalg. Er det f.eks. rimelig å hevde at det er like mange menn som kvinner i Norge, når du fant 51,6 % kvinner i ditt tilfeldige utvalg? Dette kontrolleres ved såkalt hypotesetesting (eller hypoteseprøving). Du starter da med en forutsetning om populasjonen, en hypotese, kalt nullhypotesen H_0 og en alternativ hypotese H_1. I dette eksemplet kan det formuleres slik:

> H_0: "Det er like mange kvinner som menn i den norske befolkning (og det avviket fra dette som du observerte i ditt utvalg er dermed helt tilfeldig)."
> H_1: "Det er ikke like mange kvinner som menn i den norske befolkning"

Hvis du ikke har spesiell grunn til å tro på en alternativ hypotese, må H_1 være det stikk motsatte av H_0 slik det er ovenfor. Hvis du mener å ha en slik grunn, må den være basert på noe annet enn at du fant det i ditt utvalg. Hvis den eneste grunnen er at ditt utvalg viser det, vil du alltid ende med å forkaste H_0, fordi H_1 alltid vil virke mer sannsynlig.

[18] Det vil ikke være urimelig da å hevde at det er flere kvinner enn menn i befolkningen, fordi konfidensintervallet ikke omfatter 50 %. Men sikker er du altså ikke – du kan jo ha tilfeldigvis ha fått et litt spesielt utvalg!

[19] Det er vel ikke helt sant at du ikke kan si noe – du kan jo f.eks. si at "det finnes noen i populasjonen som…". Det er en vits om det: Tre professorer sitter på toget i Skottland. Den ene ser en sort sau og sier: "Nei, se – sauene i Skottland er sorte!". Den andre svarer: "Nei, du vet jo bare at de har minst én sort sau i Skottland!" Den tredje: "Nei, egentlig vet vi bare at de har minst én sau i Skottland med minst én sort side!". Det er altså ikke mye som skal til for at vi kan si *noe* om populasjonen – men særlig *nyttig* er det kanskje ikke.

Ved hjelp av statistiske metoder kan du finne ut hvor rimelig H_0 er i forhold til det du fant i ditt utvalg. Du beregner på en eller annen måte et avvik mellom det du kunne forvente etter H_0 og det du faktisk fant i utvalget. Hvis avviket er stort, vil sjansen for å få et slikt utvalg (gitt at H_0 er sann) være liten, og da forkaster du H_0. Da mener du å vite at H_1 er rimeligere. Du kan da f.eks. ikke tro på at det er like mange kvinner som menn i befolkningen, det blir for usannsynlig utfra utvalget. Grensen for når H_0 bør forkastes settes oftest til 5 %, men noen ganger brukes 1 %. Det kalles "5 % nivå" eller "1 % nivå"[20]. SPSS regner ut slike tall, og kaller dem 'Sig' = 'Significance'. Du vil se dette mange ganger nedenfor. H_0 bør forkastes hvis signifikansen er lavere enn 5 % eller 1 % avhengig av hvor sikker du vil være. En viss risiko for å konkludere feil er det alltid - se nedenfor om "feil av type I og "type II".

Hvis H_1 er som her at det ikke er like mange menn som kvinner, men du har ingen grunn til å vite hvem det er flest av, må du regne "tosidig" ('2-tailed') så du får med både sjansen for at det skal være flere menn og sjansen for at det er færre. Hvis H_1 – basert på annen informasjon – er at "Det er flere kvinner enn menn i den norske befolkning", kan du regne "ensidig" ('1-tailed').

Hvis du sammenlikner to tilfeldige utvalg fra hver sin populasjon, kan du f.eks. finne at gjennomsnittet er forskjellige i de to utvalgene. Du vil vite om forskjellen er så stor at det er grunn til å tro at også populasjonene har forskjellige gjennomsnitt. Da er nullhypotesen H_0 at det ikke er forskjell på populasjonsgjennomsnittene (det er det du helst vil forkaste). Den alternative hypotesen er da at det faktisk er forskjell på populasjonsgjennomsnittene, og evt. hvilken som er størst hvis du har grunn til å tro på det.

Andre måter å teste hypoteser på, er omtalt nedenfor, men prinsippet er det samme: Anta noe om populasjonen (H_0 og H_1) og beregn signifikansen ut fra H_0 (ensidig eller tosidig avhengig av H_1). Hvis denne signifikansen er under 5 % eller 1 % nivå, forkaster du H_0 og "foretrekker" H_1. Da har du antakelig funnet noe interessant å si om populasjonen.

Feil av type I og type II i hypotese-testing

Når du undersøker en hypotese (H_0), kan du gjøre to, prinsipielt forskjellige feil. De kalles "type I" feil og "type II" feil.

Type I feil gjør du hvis du forkaster H_0, enda den faktisk er riktig. Tilfeldigvis ble ditt utvalg så forskjellig fra det forventede at sannsynligheten ble svært lav. Hvis du forkaster H_0 på 5 % nivå, har du 5 % sjanse for å gjøre feil av type I, 1 % nivå gir 1 % sjanse for type I feil osv.

Type II feil gjør du når du aksepterer (altså nekter å forkaste) H_0, enda den faktisk er gal. Tilfeldigvis fant du ikke store nok forskjeller fra H_0 til at du turte å forkaste H_0. Sjansen for å gjøre feil av type II, avhenger helt av hva som faktisk er sant, og det vet du ikke – du har jo bare en alternativ hypotese.

Populasjonens fordeling av én variabel – kjikvadrat

Hvis du har en antakelse om hvordan verdiene fordeler seg i populasjonen og vil sjekke antakelsen, benytter du den såkalte kjikvadrat-testen. Anta f.eks. at du har kastet en terning 1.000 ganger og fant følgende fordeling:

[20] Konklusjoner etter 95 % konfidensintervall tilsvarer 5 % signifikansnivå – 99 % konfidensintervall tilsvarer 1 % signifikansnivå.

TERNING

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	183	18,3	18,3	18,3
	2	161	16,1	16,1	34,4
	3	142	14,2	14,2	48,6
	4	174	17,4	17,4	66,0
	5	181	18,1	18,1	84,1
	6	159	15,9	15,9	100,0
	Total	1000	100,0	100,0	

Som du ser, er ikke fordelingen helt som forventet – du "burde" vel fått ca 166,7 (=16,7 %) av hvert antall øyne. Men er avviket fra det forventede så stort at du bør forkaste en antakelse om at terningen egentlig er "rettferdig" og at avvikene bare skyldes tilfeldigheter? Formelt lager du følgende hypoteser:

H_0: "Terningen er "rettferdig" – det er like stor sjanse for å få hver av de seks antall øyne (og avvikene du har observert skyldes bare tilfeldigheter)."
H_1: "Terningen er ikke "rettferdig" – det er ikke like stor sjanse for å få hver av de seks antall øyne".

Når du ber SPSS sjekke dette, velger du
Analyze|Non-parametric Tests|Legacy Dialogs|Chi-Square og setter *Expected Range* til *Get from data* (alle verdiene i utvalget blir med – det lages en klasse/gruppe for hver verdi) og *Expected values* til *All categories equal* (alle klassene/gruppene er like sannsynlige – det følger av H_0), og får følgende tabeller:

TERNING

	Observed N	Expected N	Residual
1	183	166,7	16,3
2	161	166,7	-5,7
3	142	166,7	-24,7
4	174	166,7	7,3
5	181	166,7	14,3
6	159	166,7	-7,7
Total	1000		

Test Statistics

	TERNING
Chi-Square[a]	7,352
df	5
Asymp. Sig.	,196

a. 0 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is 166,7.

Den første sammenlikner ditt utvalg ('Observed N') med det du skulle forvente etter H_0 og beregner forskjellen ('Residual') for hver verdi. Resultatet oppsummeres i den andre tabellen, der du kan se at det er beregnet et kjikvadrat-tall og et antall "frihetsgrader" ('df' = 'Degrees of Freedom'). Disse tallene betyr ikke så meget for deg[21]. For deg er det viktig at ('asymp.

[21] Kjikvadrat-tallet angir hvor stort avvik det er totalt mellom det som er observert og det forventede etter H0. Antall frihetsgrader (her antall klasser – 1, altså 5) definerer en teoretisk fordeling for dette kjikvadrat-tallet.

sig' = 'asymmetric significance') er 0,196 = 19,6 %[22]. Konklusjonen i dette tilfellet blir altså at du ikke kan forkaste H_0 hverken på 1 % nivå eller 5 % nivå – det du observerte er ikke uforenlig med at alle antall øyne er like sannsynlig.

Merk at SPSS også forteller at ingen celler har forventet antall mindre enn 5. Det er et krav for å bruke kjikvadrat-test at det er slik. Hvis det forventede antallet av en verdi er mindre enn fem, kan ikke kjikvadrat-testen brukes. Du må da slå sammen noen klasser/grupper slik at det forventede antallet for den sammenslåtte klassen/gruppen er minst fem.

Du kan benytte kjikvadrat-testen også om den forventede fordeling er skjev. Du kan f.eks. tro – på grunnlag av andre undersøkelser og fakta – at terningen ikke er "rettferdig, men derimot slik at

H_0: "Det er dobbelt så stor sjanse for å få ett eller to øyne, som hver av de andre antall øyne[23]."
H_1: "Det er ikke slik som beskrevet i H_0".

Du legger da inn følgende *Expected values*: 2, 2, 1, 1, 1 og 1. SPSS bruker forholdet mellom disse tallene til å beregne forventede antall og du får følgende tabeller:

TERNING

	Observed N	Expected N	Residual
1	183	250,0	-67,0
2	161	250,0	-89,0
3	142	125,0	17,0
4	174	125,0	49,0
5	181	125,0	56,0
6	159	125,0	34,0
Total	1000		

Test Statistics

	TERNING
Chi-Square[a]	105,496
df	5
Asymp. Sig.	,000

a. 0 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is 125,0.

Du ser at det nå forventes dobbelt så mange enere og toere som de andre. Avvikene blir større og dermed også Kjikvadrat-tallet, og signifikansen beregnes til mindre enn 0,000, dvs mindre enn 0,5‰. Konklusjonen blir at H_0 må forkastes, både på 1 % og på 5 % nivå – dette kan du dog ikke tro på.

Pass på at H_0 ikke er basert på utvalget, men på noe annet, ellers vil det aldri bli mulig å forkaste H_0 – utvalget passer jo da perfekt med hypotesen. Det gir feil konklusjon. Hvis vi i terningseksempelet f.eks. formulerer følgende H_0 basert på utvalgets resultat: "H_0: Det er 18,3 % sjanse for å få én, 16,1 % sjanse for å få to, 14,2 % sjanse for å få tre... osv." så vil signifikansen nødvendigvis bli 1,0 (100 %).

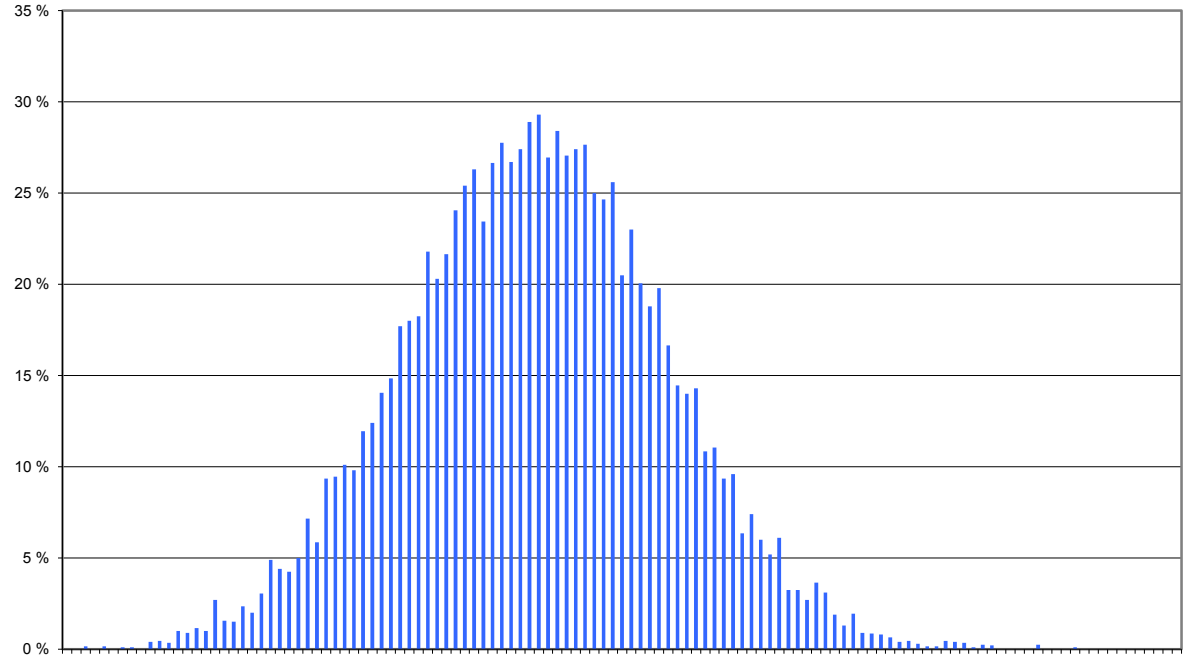
[22] Hvis du gjorde dette eksperimentet om og om igjen, måtte du regne med at nesten 20 % av gangene fikk du minst et så stort kjikvadrat-tall som her.
[23] Det er alltid underforstått i H_0 at de avvikene fra H_0 som du har observert, bare skyldes tilfeldigheter.

Populasjonens fordeling av én variabel – Normalfordeling

Normalfordelingen er en teoretisk fordeling som egentlig forutsetter en uendelig populasjon, at variabelen er kontinuerlig, og at verdiene kan være fra $-\infty$ til $+\infty$. Normalfordelingen er symmetrisk. Selv om disse forutsetningene svikter på ett eller flere punkter, sies ofte fordelingen å være "tilnærmet normalfordelt" og brukes likevel.

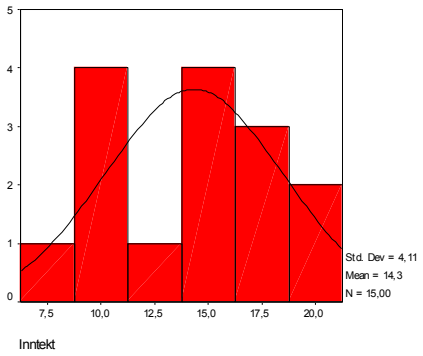
Tilnærmede normalfordelinger påtreffes forbausende ofte i naturen, f.eks. vekt av individer, høyde og lengden av fisk. Ofte kan også variable i samfunnsvitenskapen tilnærmet normalfordelt, f.eks. inntekt, karakterer og kjørehastighet på et gitt sted. Alle disse er bare tilnærmet normalfordelt, bl.a. fordi de ikke kan ha verdier mindre enn null og har en faktisk grense oppad, noen av dem kan bare ha heltallsverdier, populasjonen er ikke egentlig uendelig stor osv.

Figuren nedenfor viser resultatet av en simulering der tallene er tilnærmet normalfordelt med forventning 5 og varians 2 (standardavvik 1,4142). Ingen verdier i simuleringen kan bli mindre enn null, og tallene er gruppert i grupper med bredde 0,1 (f.eks. fra 2,0 til 2,1). Begge deler gjør at denne populasjonen bare er tilnærmet normalfordelt, selv om det i prinsippet ikke er noen grense for hvor mange enheter simuleringen kunne produsert.



En enkel måte å sjekke mot normalfordelingen, er å tegne histogram (*Graphs|Legacy Dialogs|Histogram*) og krysse av for *Display normal curve*. Her er inntekter i et utvalg[24] grafet sammen med normalfordelingen:

[24] Tallene er fra eksemplet som brukes lenger ned.



De fleste vil vel anta at grafen ikke tyder på at inntektene kan være normalfordelt i populasjonen. Det er imidlertid lett å bli lurt av små utvalg. Siden bare 15 personer er undersøkt her, kan du ikke regne med at fordelingen av utvalget vil likne særlig på normalfordelingen, selv om utvalget er fra en normalfordelt populasjon. Avvikene blir lett større enn de fleste vil tro.

Vil du teste mer nøyaktig, kan du velge *Analyze|Descriptive Statistics|Explore* der du krysser av for *Normality Plot with tests* under *Plots*.

H_0: "Inntektene i populasjonen er tilnærmet normalfordelt"
H_1: "Inntektene i populasjonen er ikke tilnærmet normalfordelt"

I inntekstseksemplet gir SPSS følgende:

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Inntekt	,125	15	,200*	,972	15	,859

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Tabellen viser at inntektene godt kan være normalfordelt i populasjonen. Signifikansen er minst 0,200 = 20 % (20 % er angitt som 'lower bound') etter den ene testen og 0,859 = 85,9 % etter den andre testen, som bare beregnes for utvalg på mindre enn 50. Du kan altså ikke forkaste H_0, så inntektene i populasjonen kan godt være normalfordelt. Det er den motsatte konklusjonen av den som er naturlig når du bare ser på histogrammet ovenfor.

For terning-eksemplet ovenfor, gir SPSS følgende resultat:

H_0: "Terningen er tilnærmet normalfordelt, dvs at sjansen for å få 1, 2, 3 osv er tilnærmet normalfordelt."
H_1: "Terningen er ikke tilnærmet normalfordelt."

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
TERNING	,149	1000	,000

a. Lilliefors Significance Correction

Signifikansen er her 0,000 (det vil si mindre enn 0,5‰) så du bør forkaste H_0 uansett nivå. Det ville jo også være svært uvanlig med en tilnærmet normalfordelt terning der 1 og 6 er svært sjeldne mens 3 og 4 forekommer ofte, og det er ingenting i utvalget som tyder på det.

Populasjonens samvariasjon mellom to variable

Det kan være aktuelt å sammenholde to variable fra samme utvalg for å se om det kan være samvariasjon i populasjonen.

Samvariasjon innebærer at de to variablene "varierer i takt". Da er det slik at hvis du kjenner verdien av den ene variabelen, kan du si noe om verdien av den andre for samme enhet. Det er sjelden snakk om en funksjonell sammenheng. Heller er det slik at når du vet verdien av én variabel, så endrer sannsynlighetene seg for den andre (såkalt betinget sannsynlighet). Det er f.eks. kjent at de med blå øyne har større sannsynlighet for å være blonde – det er flere blonde blant blåøyde enn blant andre i befolkningen og blant befolkningen generelt. Det er da en samvariasjon mellom øyen- og hårfarge. Hvis du vet at en person har blå øyne, kan du ikke vite med sikkerhet at personen er blond, men sjansen for det er større.

At to variable samvarierer er ikke et bevis på årsakssammenheng. Det er en ganske vanlig feilslutning blant ikke-statistikere (og dessverre også blant samfunnsforskere). Man finner f.eks. i mange undersøkelser en sterk samvariasjon mellom etnisitet og fengselsstraffer i USA og det kan være nær å slutte at visse etniske grupper er mer forbryterske enn andre (visse etniske grupper er overrepresentert i fengslene). Kriminologer, som sjekker andre variable i tillegg, hevder imidlertid bestemt at sosial status og inntekt også samvarierer med fengselsstraff, og de "forklarer" fengselsstraffene meget bedre en etnisitet. Visse etniske grupper har generelt lavere sosial status og inntekt enn resten av befolkningen, og det er disse faktorene som tenderer både til mer forbrytelser og til lengre straffer, helt uavhengig av etnisitet. Den direkte årsaken er altså å finne andre steder enn i etnisiteten. En tilsvarende debatt har vi hatt i Norge vedrørende første generasjons innvandrere, som er overrepresentert i våre fengsler.

Et annet, populært eksempel er å vise til en dansk undersøkelse, der man fant at på steder med mange storker, ble det født flere barn pr innbygger. Mange storker gir altså flere barn!

Du må også huske på at i et tilfeldig utvalg, kan to variable samvariere pga rene tilfeldigheter. Slike samvariasjoner kalles ofte spuriøse ('spurious').

Det er vanlig anerkjent at den eneste riktige måten å vise årsakssammenheng på, er med eksperimenter[25]. I eksperimenter har man først en hypotese om årsakssammenheng. Så setter man opp et eksperiment og forutsier resultatet i henhold til hypotesen. Deretter gjennomfører man eksperimentet en rekke ganger idet man bevisst endrer én variabel av gangen. Resultatet av eksperimentet måles og sammenliknes med forutsigelsen. Hvis resultatene samsvarer med hypotesen, anser man årsakssammenhengen som sannsynliggjort. Sikker kan man aldri bli.

[25] Noen ganger kan man anvende et såkalt "naturlig eksperiment". F.eks. har man konkludert med at forholdene i Australia – sannsynligvis solen – gir flere som får sykdommen føflekkreft. Konklusjonen er basert på at den hvite befolkning i stor grad er genetisk lik befolkningen i Storbritannia, men forekomsten av føflekkreft er vesentlig større. (Det er i tillegg funnet flere andre risikofaktorer på tilsvarende måte.) Tvillingforskning – der man sammenlikner to tvillinger som har vokst opp fra hverandre i hvert sitt miljø – brukes ofte som naturlige eksperimenter.

En annen type "eksperiment" som er interessant for IT-folk og som brukes mer og mer ettersom datamaskinene er blitt kraftigere, er simulering. Enkelte ting lar seg jo ikke eksperimentere med f.eks. fordi de tar for lang tid (hvordan ble solsystemet til, hvordan utvikler arter seg, hvordan varsler man været). For å teste en teori, bygger man da en kjørbar modell basert på teorien. Hypotesen er da at teorien kan forklare dagens observerte situasjon utfra en kjent situasjon i fortiden. Modellen kjøres og resultatet sammenliknes med empirien. Hvis resultatet stemmer dårlig med virkeligheten, så forkastes hypotesen dvs. teorien som den bygger på, ellers anser man teorien styrket.

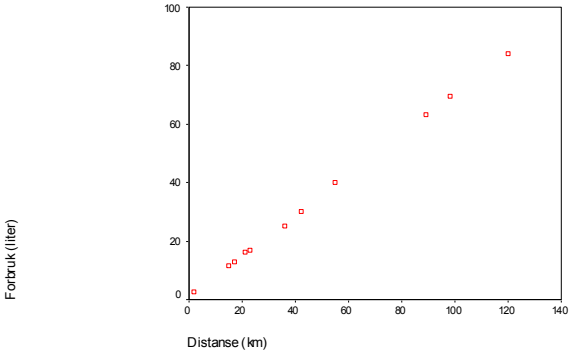
Som Einstein skal ha sagt: "Ingen eksperimenter kan noen gang bekrefte at jeg har rett – når som helst kan ett eksperiment vise at jeg tok feil". Det kan også være problemer med oppsettet av eksperimentet og med målefeil. Vanligvis godtar man derfor ikke et resultat før det er replikert av flere forskere/laboratorier.

Jeg anbefaler deg å omhyggelig unngå uttrykk i rapporten som tyder på at du mener at det er årsakssammenheng[26]. Hold deg til begrepet "samvariasjon". Evt. kan du tillate deg å spekulere på mulige årsakssammenhenger.

Samvariasjon mellom to kontinuerlige variable

Samvariasjonen mellom to kontinuerlige variable, kan analyseres ganske nøyaktig. Det er derfor du bør tilstrebe at variablene er kontinuerlige. Det er bedre å spørre respondentene om alder og høyde, enn å be dem krysse av for et intervall (0-10 år, 11-20 år osv).

Du har f.eks. fått en bil prøvekjørt, og målt kjøredistanse og bensinforbruk for hver tur. Pga tilfeldigheter – opp-/nedoverbakker, trafikk, stans, hastighet osv – kan du ikke regne med noen nøyaktig samvariasjon mellom de to, men etter inntasting i SPSS ser det slik ut:



Målingene er "prikket inn" i et koordinatsystem. Grafen er laget av SPSS med *Graphs|Legacy Dialogs|Scatter/Dots*, velg Distanse langs X-aksen og Forbruk langs Y-aksen. Punktene ligger omtrent langs en rett linje. Det kan se ut som forbruket varierer med distansen – tilsynelatende er det en lineær, funksjonell sammenheng mellom distanse og forbruk. Det er jo også andre grunner til å tro noe slikt.

H_0: "De to variablene (distanse og forbruk) er uavhengige av hverandre – det er ingen samvariasjon mellom dem."
H_1: "Det er samvariasjon og den beste lineære funksjonen[27] som kan brukes for å estimere Forbruk når Distanse er kjent, er

*Forbruk = a * Distanse + b*"

Du ønsker nå å finne den "beste" rett linjen som bør stå i H_1. Da må du finne konstantene *a* og *b*. Den "beste" linjen er den som gir minst totalt avvik fra dine data[28]. Denne linjen kalles regresjonslinjen og siden linjen er rett, kalles dette lineær regresjon.

Du prøver altså å finne en rimelig, funksjonell avhengighet som kan tenkes å gjelde for populasjonen. Avvikene fra denne, som du har observert, antar du skyldes tilfeldigheter,

[26] Du kan lese mye mer om dette på https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation
[27] Det vi ser på er samvariasjonen. For å vise funksjonell sammenheng, dvs. årsakssammenheng, må det foretas eksperimenter.
[28] Rent teknisk beregnes først avvikene fra den rette linjen. Disse avvikene kvadreres og summeres. Den "beste" linjen er da den linjen som gir minst slik sum. Hvordan du da finner a og b i uttrykket, kan du trygt overlate til SPSS.

feilmålinger e.l. Hvis feilene er systematiske, f.eks. fordi du har en klar tendens til å måle for kort kjørelengde, holder ikke dette.

SPSS beregner konstantene *a* og *b* i funksjonen og gir flere tabeller (*Analyze*|*Regression*|*Linear* og velg *Include constant in equation* under *Options*). Denne tabellen er den viktigste for deg:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1,135	,226		5,014	,001
	Distanse (km)	,696	,004	1,000	184,168	,000

a. Dependent Variable: Forbruk (liter)

Her er det på første linje angitt en konstant som tilsvarer din *b*, som er foreslått til 1,135. SPSS har beregnet et standardavvik på 0,226, hvilket vil innebære et konfidensintervall på 95 % nivå for konstanten *b* omtrent lik 1,135 ± 2 * 0,226 = [0,683 .. 1,587]. Videre er konstanten *a* beregnet i andre linje til 0,696 med standardavvik på bare 0,004. Signifikansen er oppgitt i siste kolonne. Som du ser er den svært liten. H_0 bør forkastes både på 5 % og 1 % nivå. Den endelige funksjonen, som altså kalles regresjonslinjen, ble slik, og skal stå i H_1:

*R1: Forbruk = 0,696 * Distanse + 1,135*, eller tilnærmet

*R1: Forbruk ≈ 0,7 * Distanse + 1,1*

Det er altså, etter disse beregningene og forutsetningene, tilsynelatende en direkte samvariasjon mellom distanse og bensinforbruk, gitt ved denne funksjonen. Det betyr at du kan temmelig sikkert estimere Forbruk ut fra Distanse, men noen årsakssammenheng er ikke vist her. I dine utvalg må du regne med tilfeldige avvik fra dette. Du må også regne med at den egentlige samvariasjonen mellom forbruk og distanse, i populasjonen, er noe annerledes enn den du har kommet frem til her. Du har jo beregnet konstantene *a* og *b* bare på grunnlag av et utvalg. Funksjonen i H_1 er bare et estimat for den funksjonen som eventuelt gjelder for populasjonen.

SPSS kan også beregne et forholdstall som angir hvor "god" den lineære regresjonslinjen er i forhold til ditt utvalg. Forholdstallet kalles korrelasjonskoeffisient, og er i intervallet [-1 .. +1]. Hvis koeffisienten er 0, er ikke linjen "best" på noen måte – faktisk vil uendelig mange andre rette linjer gi like bra resultat. Hvis absoluttverdien er 1, ligger alle verdiene i ditt utvalg nøyaktig på regresjonslinjen, og angir en "perfekt" tilpasning. Negativ koeffisient angir at den ene verdien går ned når den andre går opp (og omvendt), mens positiv verdi angir at enten går begge opp eller begge ned samtidig. For eksemplet gir SPSS følgende (*Analyze*|*Correlate*|*Bivariate* – det er to variable, og be om *Pearson*):

H_0: "Det ikke er ingen lineær samvariasjon[29] mellom variablene, korrelasjonskoeffisient i populasjonen er 0."
H_1: "Det er en lineær samvariasjon mellom distanse og forbruk, korrelasjonskoeffisient er forskjellig fra 0".

Correlations

		Distanse (km)	Forbruk (liter)
Distanse (km)	Pearson Correlation	1,000	1,000**
	Sig. (2-tailed)	,	,000
	N	11	11
Forbruk (liter)	Pearson Correlation	1,000**	1,000
	Sig. (2-tailed)	,000	,
	N	11	11

**. Correlation is significant at the 0.01 level (2-tailed).

Tabellen viser, som du kunne vente, at det er perfekt samvariasjon mellom distanse og distanse, og mellom forbruk og forbruk (det er nesten pussig at tallet oppgis). Det som er mer interessant, er at det også er korrelasjon +1 mellom distanse og forbruk, med andre ord en meget tett samvariasjon. Signifikansen er mindre enn 0,5‰, og H_0 må forkastes.

Du vet altså nå at det sannsynligvis er det samvariasjon mellom distanse og bensinforbruk, men er det sikkert at forbruket ikke er null når distansen er null? (Hvis den funksjonen du har funnet er riktig, bruker jo bilen ca 1,1 liter bensin selv om distansen er null.) Kanskje er dette riktig – det kan koste vel en liter bensin bare å starte bilen – eller kanskje skyldes det tilfeldigheter, eller kanskje er ikke samvariasjonen lineær likevel. Du kan først se på lineær regresjon uten konstanten *b*. Når konstanten *b* er null, blir det jo ikke noe bensinforbruk når distansen er null. SPSS gir flere tabeller (*Analyze*|*Regression*|*Linear* og velg ikke *Include constant in equation* under *Options*). Denne tabellen er den viktigste for deg:

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	Distanse (km)	,711	,004	1,000	164,622	,000

a. Dependent Variable: Forbruk (liter)

b. Linear Regression through the Origin

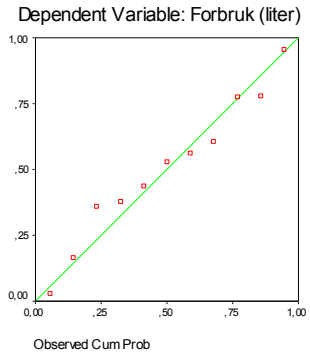
Konstanten *a* oppgis nå til 0,711, og funksjonen i H_1 blir nå

*R2: Forbruk = 0,711 * Distanse*, eller tilnærmet

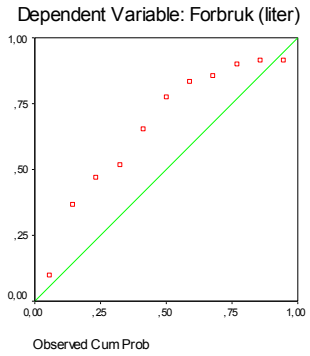
*R2: Forbruk ≈= 0,7 * Distanse*

[29] Korrelasjonskoeffisienten kan bare beregnes for lineære, funksjonelle sammenhenger (rette linjer). Derfor er dette tatt med i H_0.

Hvilken av de to funksjonene R1 og R2 som er best, er ikke lett å si ettersom begge har en oppgitt signifikans på 0,000, men en plot som viser kumulative prosenter kan være til hjelp. Forventede prosenter etter funksjonen er gitt langs y-aksen og observerte prosenter i utvalget langs x-aksen. Du ser ganske tydelig at R1 (til venstre) er best:

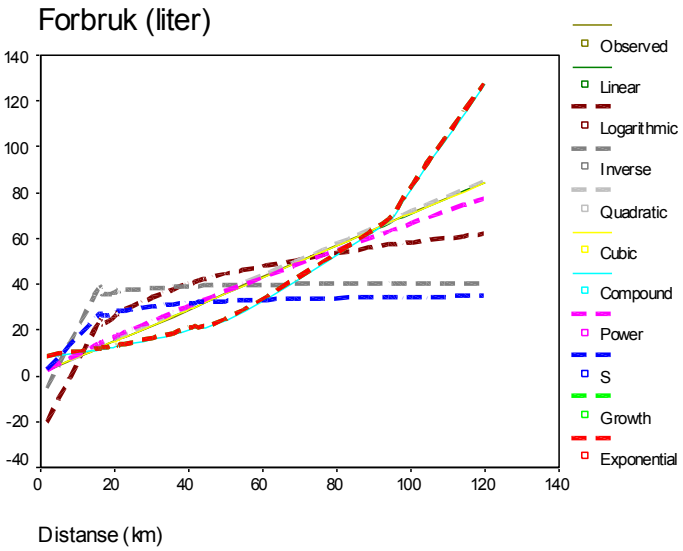


Regresjonslinjen R1

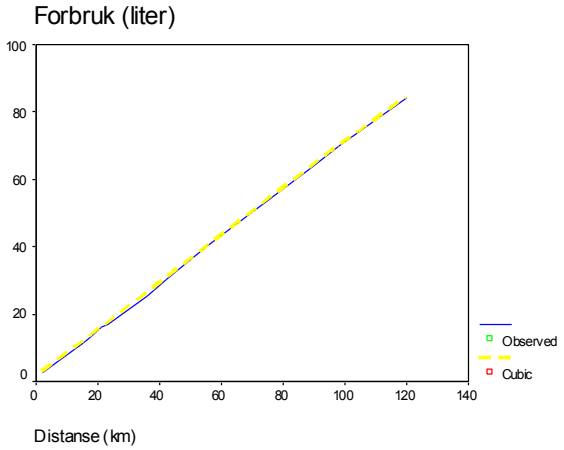


Regresjonslinjen R2

Det kan jo også hende at samvariasjonen ikke er lineær. Hvis du har en slik mistanke, kan du gå "på fisketur"[30] med SPSS og be om *Analyze|Regression|Curve Estimation* der du velger et antall funksjoner som du vil prøve (hvorfor ikke alle?):



Det er neppe lett å se her, men de observerte data følger en nesten rett linje oppover mot høyre. Når du fjerner alle som passer dårlig, sitter du igjen med denne:



[30] **Ikke mine ord, igjen☺!**

Det er den ”kubiske”, dvs tredjegradsfunksjonen som ser best ut, og SPSS beregner konstantene slik:

Dependent	Mth	b0	b1	b2	b3
FORBRUK	CUB	1,3811	,6634	,0008	-5,E-06

der *b0* er konstanten foran Distanse^0, *b1* konstanten foran Distanse^1 osv, så funksjonen blir nå:

*R3: Forbruk = 1,3811 + 0,6634*Distanse + 0,0008*Distanse^2 – 0,000005*Distanse^3*

Ettersom konstantene foran annen- og tredjegradsleddet er så små, er dette nesten lik en rett linje – svært nær den som ble beregnet i sted:

*R1: Forbruk = 1,135 + 0,696 * Distanse*

Hvis du vil satse på denne tredjegradsfunksjonen, bør du nok ha andre, gode grunner i tillegg!

Husk at hvis du bruker en annen funksjon enn den rette linje, kan ikke korrelasjons-koeffisienten beregnes.

Samvariasjon mellom én kontinuerlig og én diskret variabel

En kontinuerlig variabel kan ikke direkte sammenliknes med en diskret (så vidt jeg vet). En av dem må gjøres om.

Det er umulig å gjøre om en diskret variabel til en kontinuerlig, uten å vite noe – eller ta forutsetning om noe – ekstra. Du vet jo ikke hvilke verdier som egentlig ligger bak. I alderseksemplet som er brukt ovenfor, er alder fordelt slik:

ALDER

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Ung	150	25,0	27,3	27,3
	Voksen	350	58,3	63,6	90,9
	Gammel	50	8,3	9,1	100,0
	Total	550	91,7	100,0	
Missing	Ubesvart	50	8,3		
Total		600	100,0		

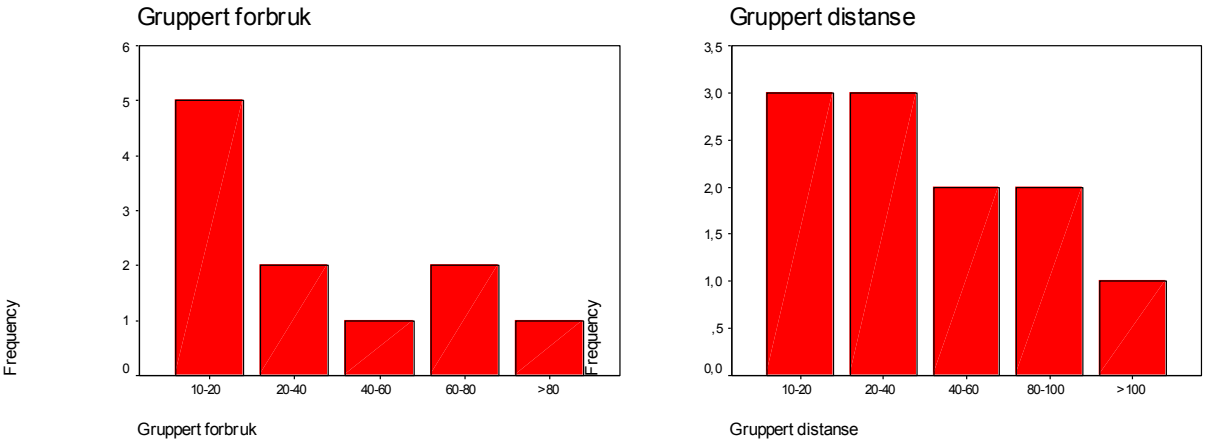
Hvordan aldersfordelingen er for hver av de 150 respondentene i gruppen ”Ung” kan du ikke vite.

Derimot kan en kontinuerlig variabel lett gjøres diskret, f.eks. ved å avrunde verdien til nærmeste ”tier” eller gruppere den i intervaller. Begge deler gjør SPSS lett (*Transform|Recode|Into different variables* – det er også mulig å omkode til samme variabel, men da mister du de opprinnelige data).

Løsningen blir altså at du gjør den kontinuerlige variabelen diskret, og får to, diskrete variabler å sammenlikne, slik det er gjort i neste avsnitt.

Samvariasjon mellom to, diskrete variable

Samvariasjonen mellom to, diskrete variable, analyseres i en tabell. Her er bileksemplet gjort om ved gruppering, og da ser fordelingene slik ut:



Begge variablene er nå diskrete på forholdstallsnivå. Du antar nå at de to variablene er uavhengige av hverandre:

H_0: "Forbruk og distanse er uavhengige av hverandre."
H_1: "Forbruk og distanse er ikke uavhengige av hverandre."

SPSS setter opp en tabell, som f.eks. kan se slik ut (*Analyze*|*Descriptives*|*Crosstabs* og velg Gruppert Forbruk mot Gruppert distanse):

Gruppert forbruk * Gruppert distanse Crosstabulation

			Gruppert distanse					Total
			10-20	20-40	40-60	80-100	>100	
Gruppert forbruk	10-20	Count	3	2	0	0	0	5
		Expected Count	1,4	1,4	,9	,9	,5	5,0
		% of Total	27,3%	18,2%	,0%	,0%	,0%	45,5%
	20-40	Count	0	1	1	0	0	2
		Expected Count	,5	,5	,4	,4	,2	2,0
		% of Total	,0%	9,1%	9,1%	,0%	,0%	18,2%
	40-60	Count	0	0	1	0	0	1
		Expected Count	,3	,3	,2	,2	,1	1,0
		% of Total	,0%	,0%	9,1%	,0%	,0%	9,1%
	60-80	Count	0	0	0	2	0	2
		Expected Count	,5	,5	,4	,4	,2	2,0
		% of Total	,0%	,0%	,0%	18,2%	,0%	18,2%
	>80	Count	0	0	0	0	1	1
		Expected Count	,3	,3	,2	,2	,1	1,0
		% of Total	,0%	,0%	,0%	,0%	9,1%	9,1%
Total		Count	3	3	2	2	1	11
		Expected Count	3,0	3,0	2,0	2,0	1,0	11,0
		% of Total	27,3%	27,3%	18,2%	18,2%	9,1%	100,0%

Tabellen er laget slik at først er verdiene fordelt til riktig kolonne og rad, og tellet opp. Det er f.eks. 2 enheter i utvalget som har forbruk 10-20 liter og kjørte 20-40 km. For hver kolonne er andelen av utvalget angitt – det er f.eks. 27,3 % som kjørte 10-20 km. Tilsvarende ser du for hver rad, at f.eks. 18,2 % av utvalget hadde et forbruk på 60-80 liter. Hvis de to variablene er uavhengige av hverandre, skal alle cellene på raden 60-80 liter ha 18,2 % av verdiene i kolonnen sin. Tilsvarende gjelder for alle cellene i kolonnen for 10-20 km, som skal ha 27,3 % av alle verdiene på raden sin. Ut fra dette beregnes det ett forventet antall for hver

celle – f.eks. forventes det altså i cellen for 10-20 km og 60-80 liter at 27,3 % * 18,2 % av alle 11 i utvalget "havner" der, dvs omtrent 4,97 % av de 11 = 0,55 enheter ≈ 0,5 enheter.

Er det rimelig at du får de observerte avvikene, hvis H_0 stemmer? Du kan la SPSS beregne Kjikvadrat-tallet for å sjekke.

H_0: "Forbruk og distanse er uavhengige av hverandre." H_1: "Forbruk og distanse er ikke uavhengige av hverandre."

I dette eksemplet er det egentlig for lite data til å bruke kjikvadrat-testen, fordi det (fortsatt) kreves at det forventede tallet er minst 5 for alle celler. Her er det ingen celler som oppfyller dette kravet, og hele testen blir svært tvilsom. SPSS foretar likevel beregningene og finner:

Chi-Square Tests

	Value	df	Asy mp. Sig. (2-sided)
Pearson Chi-Square	30,617[a]	16	,015
Likelihood Ratio	24,522	16	,079
Linear-by-Linear Association	9,299	1	,002
N of Valid Cases	11		

a. 25 cells (100,0%) have expected count less than 5. The minimum expected count is ,09.

Kjikvadrat-tallet er 30,617 og det gir en signifikans på 0,015 = 1,5 %. Hvis du valgte å stole på disse tallene, burde du forkaste H_0 på 5 % nivå, men ikke på 1 % nivå. Legg imidlertid merke til at SPSS advarer mot at 25 celler (dvs samtlige) har mindre enn fem som forventet antall etter H_0. Da er kjikvadrat-testen ubrukelig; Dataene må grupperes ytterligere og kjikvadrat-testen kjøres igjen.

Selv om du grupperer disse dataene maksimalt, vil du få en tabell omtrent slik:

Ekstra gruppert forbruk * Ekstra grupper distanse Crosstabulation

			Ekstra grupper distanse		
			0-40	>40	Total
Ekstra gruppert forbruk	0-20	Count	5	0	5
		Expected Count	2,7	2,3	5,0
		% of Total	45,5%	,0%	45,5%
	>20	Count	1	5	6
		Expected Count	3,3	2,7	6,0
		% of Total	9,1%	45,5%	54,5%
Total		Count	6	5	11
		Expected Count	6,0	5,0	11,0
		% of Total	54,5%	45,5%	100,0%

Fortsatt har alle fire cellene mindre enn fem forventet, så det hjalp ikke. Du må konkludere med at denne undersøkelsen rett og slett er for liten til å bruke kjikvadrat-test. Du vil uansett ikke kunne si noe særlig om samvariasjonen mellom kjørt distanse og bensinforbruk. Da du analyserte begge variablene som kontinuerlige, kunne du imidlertid si ganske meget om samvariasjonen. Det illustrerer hvor meget bedre det er med kontinuerlige variable. På den annen side må du passe på at ikke reliabiliteten ødelegges – det er f.eks. ikke sikkert at respondentene husker sin nøyaktige skattbare inntekt for to år siden, men kanskje kan de angi

det med ganske stor sikkerhet i et intervall. Ofte må du altså bruke diskrete verdier, selv om kontinuerlige ville gitt bedre analyse.

Samvariasjon mellom tre eller flere variable i ett utvalg

Du har kanskje en mistanke om at det er samvariasjon mellom to, tre eller flere variable til sammen på den ene side, og en variabel på den annen, f.eks. at fedme, alder og arv samvarierer med hyppigheten av hjerteinfarkt:

Hyppighet av hjerteinfarkt = f (Fedme, Alder, Arv)

Det finnes måter å teste det på, der du finner den "beste" funksjonen *f*, og får en antakelse om hvor meget hver faktor (fri variable) "bidrar" med. Det er slike analyser kriminologene har brukt, når de har analysert etnisitet, sosial status og inntekt mot fengselsstraffer. Dette området er imidlertid vanskelig og jeg tar det ikke opp her. Hold deg til sammenlikning av to og to variable av gangen.

Avslutning: Noen råd til slutt

Det er meget arbeid å analysere kvantifiserte data, selv med et verktøy som SPSS! Imidlertid er det noen farer forbundet med å få så mange, tilsynelatende nøyaktige tall med tre desimaler eller mer.

Her er noen helt generelle råd til undersøkelsen og rapporten:

1) **Analysen blir aldri bedre enn dataene**. De aller fleste undersøkelser har problemer med reliabiliteten og ofte også validiteten. ”Når utgangspunktet er galest, blir resultatet ofte originalest”.

2) **Glem ikke å ta forbehold**. Si ikke at ”det er sammenheng mellom variable x og y” men bare at ”dataene i utvalget tyder på at det kan være en samvariasjon mellom x og y”.

3) **Snakk ikke om årsakssammenheng annet enn som spekulasjoner**. Det er både lovlig og interessant å drøfte slike sammenhenger, men det er og blir *spekulasjoner* så lenge du ikke har gjennomført eksperimenter. Vær kreativ og tenk også på bakenforliggende årsaker som kan gi samvariasjonen.

4) **Pass på de forutsetningene som ligger til grunn for den statistiske analysen**. Hvis teknikken f.eks. forutsetter at dataene er fra et tilfeldig utvalg og at populasjonen er normalfordelt, må du også uttrykkelig drøfte (gjerne sjekke) om det virkelig forholder seg slik. Bruk f.eks. heller ikke en teknikk beregnet på to, *uavhengige* utvalg til å analysere *ett utvalg* som du kjekt deler inn *etter* innsamlingen. SPSS merker ikke slik ”juksing”, men sensorene ser det kanskje – og feil blir det jo, uansett.

5) **Kontroller at ditt utvalg faktisk er representativt for populasjonen**. Hvis du f.eks. har spurt 30 kvinner og tre menn om noe, og det er flest menn i populasjonen, virker det dumt å analysere som om ditt utvalg er representativt. Da hjelper det ikke at respondentene er valgt helt tilfeldig – utvalget er og blir ”skjevt”.

6) **Ikke ta munnen for full**. Jo færre enheter du har i utvalget, desto mindre grunnlag har du for å uttale deg generelt. Det er bare barn som generaliserer ut fra ett tilfelle: ”Joda, det er farlig å gå på rødt, f.eks. en gang da bestemoren min...” Studenter (ikke du, naturligvis☺!) gjør noen ganger det samme etter å ha spurt 10 kunder i en butikk.

7) **Sammenlikn dine resultater med andre kilder**. Kanskje er en liknende undersøkelse gjort tidligere eller i utlandet?

8) **Ikke tro at nøyaktigheten i svarene er fornuftig!** SPSS viser svarene beregnet med mange desimaler, men med små antall blir det fort meningsløst. F.eks. er det ingen grunn til å oppgi en andel som 42,85714 %. Derimot er "noe under halvparten", "omtrent fire av 10" eller helst "3 av de 7" bedre uttrykk i en slik forbindelse.

9) **Ikke ta alt som SPSS produserer for god fisk!** SPSS har små muligheter til å kontrollere at forutsetningen for teknikken er til stede. SPSS beregner det du ber om, selv om det kan være helt tåpelig. Jeg har f.eks. sett en student som henviste til at "gjennomsnittskjønn er 0,6" etter at SPSS hadde regnet snittet av kode 0 = mann og kode 1 = kvinne.

10) **Bruk ikke prosenter på små antall**. Når det er fire respondentene som er kvinner over 45 år som studerer IT på deltid, er det "tre av de fire" som er fornøyd – ikke "flesteparten" og slett ikke "hele 75 %". Tilsvarende er det ikke en "dobling" eller "100 %" mer hvis det var én i fjor og to i år. Det er lov å runde av tallene, men det må gjøres ærlig: 50,001 % er ikke "over halvparten" men faktisk "omtrent halvparten". Det er inntrykket du skaper hos leseren som teller. Tenk deg hvilket nøyaktige tall leseren ville gjettet på: Hva er f.eks. "omtrent hver tredje respondent" som nøyaktig prosentandel?

11) **Kanskje kan du få mer ut av en intensiv, kvalitativ undersøkelse**. Med grundige intervjuer, observasjoner og beskrivelse av noen få kan du ofte få vite mer om et emne enn med en *ekstensiv, kvantitativ* undersøkelse med mange enheter og mange tall.

12) **Grafiske fremstillinger gir leseren meget, mens tabeller gir mindre**. En gammel regel sier at fremstillingen aldri skal være avhengig av figurer – det som er interessant ved figuren skal nevnes uttrykkelig i teksten. Tilsvarende kan jeg legge til at du ikke bør omtale en figur som ikke vises i rapporten, selv om den kan finnes i vedlegg, på nett e.l.

Lykke til med forskningsarbeidet!

Ekstra: Sammenlikning av to populasjoner basert på to utvalg

Merk: *I enkle undersøkelser er dette sjelden aktuelt. Det er mange forutsetninger for når teknikkene i dette kapitlet kan brukes. Det er tvilsomt om de er til stede i din (enkle) undersøkelse.*

Hvis du har målt samme variabel for to tilfeldige utvalg fra hver sin populasjon, og har fått forskjellige resultater, kan du lure på om det også er forskjell på gjennomsnitt, varians/standardavvik eller annet også i de to populasjonene. (De to populasjonene er jo helt sikkert forskjellige, men likevel kan de statistiske målene være like.)

Måten du gjør dette på, likner på det som er beskrevet ovenfor angående kontroll av om et utvalg kan være fra en populasjon med en bestemt fordeling. Du lager en H_0 som går ut på at selv om utvalgene ble forskjellige, så er ikke populasjonene det (i alle fall når det gjelder denne ene variabelen) – samt en motsatt H_1. Så lar du SPSS beregne signifikansen. Hvis den er liten – mindre enn 5 % eller helst 1 % - forkaster du H_0.

Merk at utvalgene må være tilfeldige (stokastiske) og at populasjonene må være store (rent teoretisk skal de være uendelig store og variabelen skal være normalfordelt). Merk også at utvalgene skal være fra hver sin populasjon. Da er det ikke lov å dele inn utvalget *etterpå* og påstå at de er fra hver sin populasjon. Hvis du f.eks. har valgt ut noen studenter fra tredje klasse (populasjonen) og så etterpå deler dem etter kjønn, er de *ikke* fra forskjellige populasjoner. For at de skal være det, må du lage et tilfeldig utvalg av kvinner i tredje klasse (populasjon A) og et annet tilfeldig utvalg av menn i tredjeklasse (populasjon B).

Sammenlikning av gjennomsnittene (t-test)

Anta at du har undersøkt inntekt for kvinner og menn, ved å spørre et tilfeldig utvalg menn og et tilfeldig utvalg kvinner. Utvalgene er valgt hver for seg, og du fant følgende:

Group Statistics

	Kjønn	N	Mean	Std. Deviation	Std. Error Mean
Inntekt	Mann	10	16,000	3,543	1,121
	Kvinne	5	11,000	3,162	1,414

Du ser at det – helt sikkert – er forskjell på gjennomsnittsinntektene i dine to utvalg. Men er forskjellen så stor at det antakelig er forskjell på gjennomsnittsinntektene i de to populasjonene? Utgangspunktet er følgende:

H_0: "Menn og kvinner i populasjonene har samme gjennomsnittsinntekt."
H_1: "Menn og kvinner i populasjonene har ikke samme gjennomsnittsinntekt."

Siden du bare har to populasjoner, kan du bruke den såkalte Student's t-test. SPSS beregner (*Analyze|Compare Means|Independant Samples T Test* med inntekt som *Test variable* og kjønn som *Grouping variable* med *Cut-point* 1.5 som skille mellom kjønnene) følgende:

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Inntekt	Equal variances assumed	,151	,704	2,661	13	,020	5,000	1,879	,941	9,059
	Equal variances not assumed			2,771	9,019	,022	5,000	1,804	,920	9,080

Første linje i tabellen viser beregningene basert på en forutsetning om at populasjons-variansene er like (selv om dine utvalg – tilfeldigvis – viser noe annet). SPSS har beregnet et tall "t" som har gitt navnet til testen: 'Student's t-test'[31]. Du behøver ikke bry deg så mye om dette tallet. Det er viktigere for deg å se at den tosidige signifikansen er beregnet til 0,02 = 2 %. Altså bør du forkaste H_0 på 5 % nivå, men ikke på 1 % nivå. Du ser også at SPSS har beregnet 95 % konfidensintervall for forskjellen mellom de to gjennomsnittene, og funnet [0,941..9,059]. Siden konfidensintervallet ikke omfatter forskjellen 0, må du også her forkaste H_0 på dette nivået. 99 % konfidensintervall (som tilsvarer signifikansnivå 1 %) er ikke beregnet.

Du ser av første linje også at signifikansen (beregnet med F – se neste avsnitt) for at populasjonsvariansene er like, er hele 0,704 = 70,46 %. En hypotese om like populasjonsvarianser kan altså ikke forkastes. Det gjør det naturlig å bruke første linje her.

På den andre linjen i tabellen vises beregningene basert på en forutsetning om at populasjons-variansene ikke er like. Signifikansen (2,2 %) – og dermed også konklusjonen – blir omtrent den samme. Det gjelder også hvis du ser på konfidensintervallet.

Sammenlikning av variansene og gjennomsnittene med ANOVA (F-test)
Enveis[32] ANOVA ('ANalysis Of VAriance') er en annen måte å undersøke om gjennomsnittene i populasjonene er like, men ANOVA kan brukes til å sammenlikne mer enn to populasjoner, f.eks. gjennomsnittsinntekten i fem byer. ANOVA forutsetter egentlig at alle populasjonene er normalfordelt, men er godt brukbar også i andre tilfeller bare ikke fordelingen er særlig skjev i forhold til gjennomsnittet.

I ANOVA er det et krav at alle populasjonene har **samme varians**. Det kan vi sjekke med følgende hypoteser:

H_0: "Populasjonene har samme varians."
H_1: "Populasjonene har ikke samme varians."

Det kan du teste med Levene's test for lik varians ('Levene's Homogeneity-of-variance test'), som gir følgende resultat for inntekstesksemplet ovenfor (*Analyze|Compare Means|One-Way*

[31] Også denne, teoretiske fordelingen er gitt ved et antall "frihetsgrader" som her er antall i utvalgene til sammen – 2 = 13.

[32] Det finnes også en toveis ANOVA, der du trekker inn to faktorer for å forklare forskjellene. Hvis du undersøker bensinforbruk på 15 biler av 5 typer med 3 sjåfører, kan det være interessant å se på både forskjellige biltyper og sjåfør som forklaringsfaktor for forskjeller i forbruk: Noe av forskjellen forklares best ved at det faktisk er forskjell på biltypene, noe ved at det er forskjell på sjåførene, og noe ved tilfeldige avvik innen hver kombinasjon av sjåfør og biltype. Toveis ANOVA kaller SPSS *Analyze|General Linear Model|Univariate* og du oppgir hvilken variabel som er avhengig (*Dependant variable)* og hvilke variable som angir populasjonen (*Covariates*). Du kan også utføre mer enn toveis analyse ved å trekke inn mer enn to forklaringsvariable, men det er alltid bare én avhengig variabel.

ANOVA, sett inntekt som *dependant* og kjønn som *factor* og be om *Homogeneity-of-Variance* under *Options*):

Test of Homogeneity of Variances

Inntekt

Levene Statistic	df1	df2	Sig.
,151	1	13	,704

Igjen viser SPSS en del beregnede tall som du ikke behøver å ta hensyn til. Det viktige for deg er tallet 'Sig' som er hele 0,704 = 70,4 %, så H_0 kan ikke forkastes. Du kan følgelig fortsette analysen basert på den forutsetning at populasjonsvariansene er like. (Du fikk samme resultater med Student's t-test – se forrige avsnitt.)

For å kontrollere om **gjennomsnittene** synes like, setter vi opp følgende hypoteser (gjentatt):

Gjentatt:
H_0: "Menn og kvinner i populasjonene har samme gjennomsnittsinntekt."
H_1: "Menn og kvinner i populasjonene har ikke samme gjennomsnittsinntekt."

SPSS gir (*samme menyer som for varians*) følgende tabell for inntekstеksemplet ovenfor, der du bare har to kjønn:

ANOVA

Inntekt

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	83,333	1	83,333	7,081	,020
Within Groups	153,000	13	11,769		
Total	236,333	14			

Igjen har SPSS beregnet og viser en rekke tall som nok er interessante for en statistiker, men neppe for deg[33], bl.a. tallet F som er avledet av testens navn Fisher's F-test. Det du bør merke deg er tallet 'Sig' = 0,020 = 2 % i den siste kolonnen. På 5 % nivå bør du altså forkaste nullhypotesen om at populasjonene er like, men ikke på 1 % nivå. Det innebærer at du ikke kan være særlig sikker på at det er forskjell. (Student's t-test ga akkurat samme resultat, og det er ikke tilfeldig – det er jo samme H_0 og de samme data du analyserer.)

ANOVA er mest "effektiv" (det vil si at det er lettest å forkaste H_0) når utvalgene er like store. Hvis det er mulig, bør du altså tilstrebe det, men den kan også brukes når utvalgene ikke er like store, som her.

Sammenlikning av datapar ('paired data')

Ovenfor gjaldt sammenlikningen gjennomsnittene for to forskjellige, tilfeldige utvalg fra to populasjoner, for å se om det er grunn til å tro at det er forskjell på populasjonene. Noen ganger har du målt en variabel to ganger fra samme utvalg, på forskjellige tidspunkt. Da blir

[33] ANOVA-tabellen har to linjer. Den første analyserer gjennomsnittene i hvert utvalg i forhold til et felles gjennomsnitt for samtlige, den andre linjen viser enkeltdataenes avvik fra gjennomsnittet for sitt utvalg. Det første kalles forklart ('explained between groups') avvik fordi det kan forklares med at det er forskjell på gjennomsnittet i de to populasjonene. Den andre kalles uforklart ('unexplained within groups') fordi den må forklares med tilfeldige avvik fra gjennomsnittet i hver populasjon. Videre er det beregnet tallet F som er et mål for det totale avviket mellom det observerte og det forventede etter H_0. Antall frihetsgrader (antall forskjellige populasjoner – 1) definerer en fordeling av tallet F.

situasjonen helt annerledes. Du må forvente mindre variasjon, fordi tilfeldighetene som oppstår når du velger et tilfeldig utvalg blir borte. Likevel er jo utvalgene på en måte fra to populasjoner, siden populasjonen tydeligvis har forandret seg. Dette fant du f.eks. i en undersøkelse av fem studenter:

Navn	**Høstscore**	**Vårscore**	**Diff**
A	64	54	10
B	66	54	12
C	89	70	19
D	77	62	15
E	73	-	-
Snitt	**74**	**60**	**14**
St.avvik	**11,52**	**7,66**	**-**

H_0: "Det er ingen forskjell på score høst og vår."
H_1: "Det er forskjell på score høst og vår." (Tosidig test)
alternativt
H_2: "Høstscore er større enn Vårscore." (Ensidig test)

SPSS (*Analyze*|*Compare Means*|*Paired Samples T Test*) ser bort fra student E, siden det mangler Vårscore for vedkommende:

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	HØST	74,00	4	11,52	5,76
	VÅR	60,00	4	7,66	3,83

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	HØST & VÅR	4	,997	,003

Den første tabellen oppsummerer noen sentral- og spredningsmål. Den andre angir for det første en korrelasjon på 0,997. Dette tallet vil bli nærmere forklart senere. For det andre angir SPSS en signifikans på 0,003 = 0,3 % så H_0 bør forkastes både på 5 % og på 1 % nivå. Denne signifikansen er ensidig, basert på H_2 ovenfor. Nærmere detaljer gis i enda en tabell:

Paired Samples Test

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1 HØST - VÅR	14,00	3,92	1,96	7,77	20,23	7,151	3	,006

Her kan du legge merke til at den tosidige signifikansen (basert på H_1) er 0,006, altså dobbelt så stor som den ensidige angitt i forrige tabell. Legg også merke til at konfidensintervallet for differansen mellom populasjonsgjennomsnittene er [7,77 .. 20,23][34]. Hvis de to populasjonene skulle ha samme gjennomsnitt, måtte denne differansen ha vært 0, men det ligger jo langt utenfor konfidensintervallet. Da er det svært lite sannsynlig, og en slik antakelse (H_0) må forkastes. Bruken av signifikans og konfidensintervall gir igjen samme konklusjon.

[34] Jeg har tidligere gitt "tommelfingerregelen" at konfidensintervallet er utvalgets gjennomsnitt ± 2*'S.E.Mean' som her blir 14 ± 2 * 1,96 ≈ [10 .. 18]. Det stemmer dårlig med det SPSS har beregnet, fordi faktoren 2 blir svært gal for så små utvalg (N = 4) som her. Den riktige faktoren her er faktisk hele 3,182.

Denne diskusjonen startet med følgende hypoteser:

Gjentatt:
H_0: ”Det er ingen forskjell på score høst og vår.”
H_1: ”Det er forskjell på score høst og vår.” (Tosidig test)
alternativt
H_2: ”Høstscore er større enn Vårscore.” (Ensidig test)

Du ser ovenfor at H_0 bør forkastes uansett, men bør du da velge H_1 eller H_2? Det avhenger, som nevnt tidligere, av hva du ellers mener å vite om populasjonen. Du skal ikke velge etterpå, men grunngi den alternative hypotesen på forhånd[35].

[35] I alle fall skal det se slik ut i rapporten. **OBS! Sitatforbud☺!**

Høgskolen i Buskerud
Postboks 235
3603 Kongsberg
Telefon: 32 86 95 00

www.hibu.no

ISSN 1893-2398 (online)

HØGSKOLEN
i Buskerud