Fra spørreskjemakonstruksjon til multivariat analyse av data:

En innføring i survey-metoden

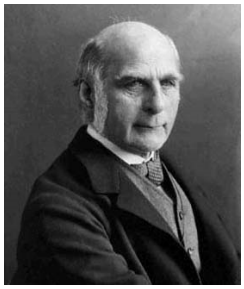
(2. utgave)

av

Leif Edvard Aarø

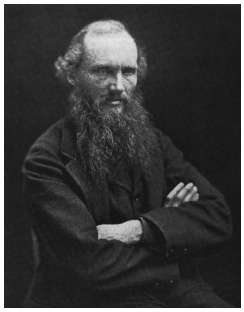
HEMIL-senteret
og
Grieg-akademiet

Universitetet i Bergen



... until the phenomena of any branch of knowledge have been submitted to measurement and number, it cannot assume the dignity of science.

Sir Francis Galton (1822-1911)



... one cannot understand a phenomenon until it is subjected to measurement.

Lord William Thomson Kelvin (1824-1907)

Whatever exists, exists in some amount, and can therefore eventually be subjected to measurement and counting

Edward Lee Thorndike (1874-1949)

Forord

Denne innføringen i survey-metoden ble opprinnelig skrevet med støtte fra Norges Forskningsråd, eller mer presist, det som tidligere het Norges allmennvitenskapelige forskningsråd (NAVF), Rådet for samfunnvitenskapelig forskning (RSF). Den er senere brukt i undervisningen på doktorgradsprogrammet ved Det psykologiske fakultet, i metode-undervisningen ved Institutt for utdanning og helse og ved masterprogrammet i musikkterapi ved Grieg-akademiet.

Teksten er forsøkt holdt i et enkelt, ikke-matematisk språk. En del enkle formler er likevel tatt med. Innledningskapittelet gir en introduksjon til survey-metoden og beskriver blant annet en del sentrale begreper, prinsipper for konstruksjon av spørreskjema og en del stoff om trekking av utvalg. Det andre kapittelet handler om den elementære, univariate statistikken. Etterfølgende kapitler tar for seg bivariat statistikk, variansanalyse, faktoranalyse (samt prinsipal komponentanalyse), analyse av reliabilitet og regresjonsanalyse.

Tanken med denne teksten er først og fremst at en skal få en forståelse av hvordan surveys kan planlegges og gjennomføres. Leseren skal også få en forholdsvis grundig innføring i den elementære statistikken og en del smakebiter på multivariat statistikk. Noen kan kanskje klare seg med dette. De som virkelig vil anvende statistikk på egne data, anbefales å gå videre og skaffe seg mer spesialiserte lærebøker. Det er rikelig med henvisninger til slike i teksten.

Jeg vil dessuten anbefale bruk av statistiske ressurser som finnes på internett. Jeg skal ikke oppgi noe bestemt nettsted. Men dersom en går inn på en nettleser og søker på statistiske ord og uttrykk, kommer det som regel opp en mengde adresser, noen av disse til gode nettsteder. Her kan en lære interaktivt og blant annet se hva som skjer med ulike statistiske størrelser når en endrer formen på fordelinger, utvalgsstørrelse, varians og liknende.

Jeg har underveis i arbeidet med denne teksten hatt stor nytte av å diskutere metode med kolleger og studenter. Jeg vil gjerne takke kolleger ved Psykologisk institutt, Universitetet i Oslo og kolleger ved Institutt for samfunnspsykologi, HEMIL-senteret og Grieg-akademiet, Universitetet i Bergen. Dessuten en takk til deltakerne ved survey-kursene på doktorgradsprogrammet ved Det psykologiske fakultet, Universitetet i Bergen for nyttige tilbakemeldinger og diskusjoner og for trivelig samvær. En særlig takk til min statistikklærer fra embetsstudiet i psykologi ved Universitetet i Oslo, Torleif Lund.

For ordens skyld vil jeg gjøre oppmerksom på at sitatene på forrige side ikke er et uttrykk for at undertegnede er en forstokket og sneversynt positivist. De er ment som et forsøk på en lett provokatorisk start på dette heftet.

Bergen august 2007

Leif Edvard Aarø

Innhold

Kap 1: Generelt om survey-metoden

1.1 Innledning

Svært mye av den forsknings- og utredningsvirksomhet som foregår i samfunnsvitenskapelige og psykologiske fagmiljøer i Norge i dag baserer seg på en eller annen variant av survey-metoden. Det har gått inflasjon i bruk av intervjuer og spørreskjemaer. Dette går ofte ut over kvaliteten. Det gjøres elementære feil når spørreskjemaer konstrueres, dataene analyseres ofte altfor overfladisk, og en ser ikke sjelden at det trekkes uholdbare konklusjoner. Samtidig er survey-metoden en kostnadseffektiv og informativ metode dersom den anvendes på forsvarlig måte.

Survey-metoden er også svært utbredt internasjonalt og har vært det lenge. Martin Bulmer skrev i 1984 at:

> The social survey dominates empirical social research in Western industrial societies. A very large proportion of social research is carried out using these methods, and the majority of textbooks on research methods devote most attention to aspects of research design, sampling, data collection and analysis for social surveys (s.53).

Survey-metoden er i en viss forstand allemannseie. Mens de fleste nok nøler med å sette igang undersøkelser som baserer seg på laboratorie-eksperimenter eller deltakende observasjon, uten å ha gjennomgått en opplæring i slik metode, ser det ut til at motforestillingene mot å lage spørreskjema og administrere surveys er svært små. Dette er selvfølgelig et problem fordi kompetansen ikke alltid står i forhold til lysten til å gjøre slik forskning. Samtidig er det positivt fordi flere involveres i faglig utviklingsarbeid der de benytter en metode som iallfall kan brukes vitenskapelig, og antakeligvis lærer de noe av denne erfaringen.

Dette manuset er ikke ment å dekke hele feltet survey-metode. Det eksisterer bra innføringstekster i survey-metode fra før. En av disse er Steinar Ilstads (1989) bok som er skrevet på en slik måte at den er tilgjengelig for de aller fleste. En annen er Helleviks (2005) lærebok om samfunnsvitenskapelig forskningsmetode, som dekker et bredere felt enn bare survey-metoden. Den foreliggende innføringen går noe lenger enn Ilstads og Helleviks bøker i retning av å presentere multivariate statistiske teknikker.

1.2 Hva er en survey?

Surveys er en av mange ulike forskningsmetoder som anvendes i samfunnsforskning og atferdsforskning. I sin innføringsbok lister Ilstad (1989) opp til sammen 12 slike metoder. Disse omfatter:

1. Laboratorie-eksperiment
2. Felteksperiment
3. Naturlige eksperiment
4. Surveys
5. Case-studier
6. Prospektive (longitudinelle) studier
7. Retrospektive studier
8. Panelstudier
9. Deltakende observasjon
10. Sammenlikning mellom geografiske områder
11. Tidsserie-studier
12. Prosess-studier.

Med survey undersøkelser mener Ilstad følgende:

> Det karakteristiske er et relativt stort, representativt utvalg fra en geografisk spredt populasjon, datainnsamling ved hjelp av spørreskjema (ved intervju, selvutfylling etc.), og en relativt rutinisert analyse av data, ordnet i avhengige og uavhengige variabler. Survey-undersøkelser er mye brukt i anvendt sosio-psykologisk forskning.

Ilstad presiserer at flere av de andre forskningsmetodene ligner på survey-metoden. Den inndelingen Ilstad foretar er nyttig nok for praktiske formål, men bygger på et noe uoversiktlig sett av kriterier.

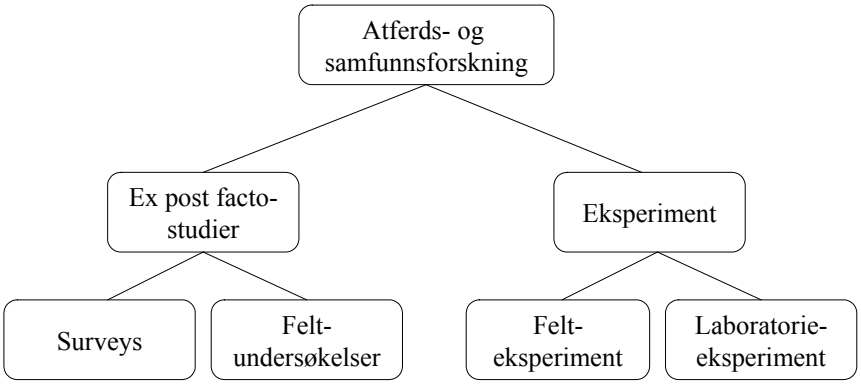
En av de bredeste definisjonene av hva surveys er, finner vi i en lærebok i samfunnsmedisin (Abramson, 1984). Der hevdes det at en lettest kan definere surveys ved å si hva en survey ikke er. En survey er ifølge denne forfatteren en ikke-eksperimentell undersøkelse.

Fred N. Kerlinger & Howard B. Lee (2000) nærmer seg survey-metoden på en mindre pragmatisk og mer prinsipiell måte og skiller innledningsvis mellom to ulike typer atferdsforskning:

- Eksperimentelle undersøkelser og
- Ex-post-facto-undersøkelser

Eksperimentelle undersøkelser innebærer at det gjennomføres systematiske tiltak eller intervensjoner med tanke på å skape bestemte virkninger. En ønsker kanskje å bedre livskvaliteten og den psykiske helsen hos enslige som har mistet sin ektefelle. For å finne ut om systematisk bruk av selvhjelpsgrupper har noen positiv effekt, kan en identfisere et antall som er villige til å delta, dele tilfeldig inn i en intervensjonsgruppe og en sammenlikningsgruppe og teste ut tiltaket blant de som havner i intervensjonsgruppen. Ved å måle endringer i begge gruppene over tid, kan en se om intervensjonen har hjulpet (Dalgard, 1996).

Fig. 1.1: Ulike typer undersøkelser innen samfunns- og atferdsforskningen (Kerlinger & Lee, 2000)



Ex-post-facto-undersøkelser innebærer at en studerer verden slik den er, uten å gripe inn på noen bestemt og planlagt måte for å skape systematiske endringer. I praksis er det vel umulig å gjennomføre forskning uten at en på en eller annen måte griper inn i virkeligheten til de en forsker blant. Datainnsamlinger vil alltid innebære en eller annen form for påvirkning og endring. Poenget er bare at dette er endringer som oppstår uten at de er resultatet av en bevisst plan med tanke på å skape bestemte effekter. I eksperimentell forskning handler det alltid om å planlegge og forsøke å få til bestemte endringer. I ex-post-facto studier forsøker en å unngå å påvirke systematisk, og heller studere verden slik den

Kerlinger & Lee går videre og sier at de to hovedkategoriene begge kan splittes i to undergrupper, slik som vist i Fig. 1.1. Ex-post-facto-undersøkelser kan deles inn i surveys og feltundersøkelser. Det som først og fremst skiller disse fra hverandre er hvordan informantene velges ut. I en survey vil informantene vanligvis være spredt tynt utover et større geografisk område. Et godt eksempel på en survey er Statistisk sentralbyrås røykevaneundersøkelser som

ble startet opp i 1973 etter oppdrag fra det som i dag heter Avdeling tobakk i Sosial- og helsedirektoratet. Røykevaneundersøkelsene var lenge del av et mer omfattende system av omnibus-undersøkelser[1] som Byrået administrerte. Røykevaneundersøkelsene har vært til stor nytte i arbeidet for å redusere tobakksskadene her i landet. De har vist hvordan røykevanene har endret seg både totalt sett og i bestemte grupper i befolkningen. Blant annet ble det registrert en oppgang i røyking blant yngre menn og kvinner (aldergruppen 16-19 år) på 1990-tallet (Kraft & Svendsen, 1997). Denne surveyen har også vært brukt til å se på befolkningens holdninger til ulike spørsmål, blant annet til tobakkslovgivningen. Røykevaneundersøkelsene er et eksempel på surveys som er blitt gjentatt årlig over en lang tidsperiode. En survey kan like gjerne være en enkeltstående undersøkelse som senere ikke blir gjentatt.

En feltundersøkelse innebærer at en gjennomfører en mer intensiv granskning innenfor et avgrenset sosialt system, som for eksempel et lokalsamfunn, en arbeidsplass, en skoleklasse eller tilsvarende. I feltundersøkelser er det for øvrig vanlig med helt andre framgangsmåter for innsamling av data, f.eks. deltakende observasjon, kvalitative intervjuer eller bruk av video. Et godt eksempel på en feltundersøkelse er Sverre Lysgaard (1976) og medarbeideres undersøkelse av "arbeiderkollektivet", der de gikk inn i en bestemt bedrift og gjennom deltakende observasjon og samtaler samlet informasjon om hvordan det blant de menige arbeiderne på golvet eksisterte et uformelt og "usynlig" sosialt system som var sterkt bestemmende for arbeidernes atferd og produktivitet. Studien er en klassiker i norsk samfunnsforskning.

Den eksperimentelle forskningen blir av Kerlinger & Lee delt opp i to hovedkategorier: Laboratorie-eksperiment og felteksperiment. Disse skiller seg fra hverandre først og fremst ved den grad av kontroll en har over betingelsene. Laboratorie-eksperimentet kjennetegnes av nær perfekt kontroll over situasjonen, mens felteksperimentelle undersøkelser foregår ute i det virkelige liv, noe som gjør det svært vanskelig å kontrollere alle de faktorene som kan påvirke utfallet av eksperimentet. Dette gjør ikke felteksperimentet til en mindreverdig metode forskningsmessig, men til en metode som kan benyttes for å gjennomføre forsøk under realistiske betingelser. Det felteksperimentet taper på manglende kontroll over betingelsene (redusert indre validitet), tar den igjen ved å gi gode generaliseringsmuligheter (økt ytre validitet)[2].

Et bra eksempel på et laboratorie-eksperiment er Muzafer Sherifs studier av normdannelse fra 1930-årene (Sherif & Sherif, 1969). Studiene er kjennetegnet av at en henter forsøkspersonene inn i en nokså kunstig situasjon, der en har svært god kontroll over de

---

[1] Omnibus er betegnelsen på en type større, gjerne landsrepresentative surveys som administreres av profesjonelle byrå, der kunder kan kjøpe seg inn med grupper av spørsmål og få adgang til både disse spørsmålene og til relevante demografiske opplysninger.

[2] Ordet validitet brukes her for å beskrive egenskaper ved et eksperimentelt design. Indre validitet dreier seg om i hvilken grad en med sikkerhet kan hevde at effektene en fant i et eksperiment kan tilskrives en bestemt uavhengig variabel. Ytre validitet handler om mulighetene til å generalisere fra en bestemt eksperimentell studie til et bredere spekter av situasjoner og omstendigheter. Disse måtene å definere validitet på må ikke forveksles med instrumenters validitet, som vil bli omtalt senere i denne teksten.

påvirkninger de utsettes for. Forsøkspersonene sitter i et rom som er fullstendig mørklagt. Så tennes et lyspunkt i horisontal høyde foran forsøkspersonen. Lyset vises bare i et halvt sekund. Lyspunktet står egentlig helt stille, men siden forsøkspersonene ikke har noen perseptuelle (sansemessige) holdepunkter, men sitter i et helt mørklagt rom, oppfatter de bevegelse. Forsøkspersonene skal så vurdere hvor langt punktet beveger seg. Hvor mye de oppfatter at punktet beveger seg, er ganske tilfeldig. Det viser seg at når en setter flere personer sammen, tilpasser de seg hverandres vurderinger og blir på en måte enige om hvor langt punktet beveger seg hver gang. Eksperimentet illustrerer hvordan vi tenker oss at sosiale normer dannes i det virkelige liv.

Et godt eksempel på et felteksperiment er HEMIL-senterets evaluering av Den Norske Kreftforenings intervensjon mot røyking blant skoleelever (VÆR røykFRI). I denne studien sammenliknes endringer i røykevaner i tre forskjellige grupper av skole-elever (som ble utsatt for tre ulike intervensjoner) med endringene som finner sted i en kontrollgruppe. I hver gruppe inngår skoler spredt over hele landet. Først ble det gjennomført en baseline-undersøkelse, deretter tre oppfølgende undersøkelser og til slutt flere langtids etterundersøkelser. Det viste seg at rekrutteringen av røykere ble sterkt redusert under en av intervensjonsbetingelsene. Det ble konkludert med at etter tre år var det om lag 30% færre røykere i denne beste gruppen, og en hadde redusert eksperimenteringen med hasj og marihuana med omtrent 50% (Jøsendal et al., 2005).

Kerlinger & Lee definerer survey-metoden på følgende måte:

> Survey research studies large and small populations (or universes) by selecting and studying samples chosen from the populations to discover the relative incidence, distribution, and interrelations of sociological and psychological variables.
>
> Kerlinger & Lee (2000), s. 377.

Kerlinger & Lee sier ikke at det nødvendigvis skal trekkes store utvalg for å kalle en undersøkelse for en survey, slik Ilstad gjør. De mener heller ikke at det skal være "rutinisert" data-analyse, og de sier heller ikke at det skal foreligge en spesiell måte å arrangere i uavhengige og avhengige variabler på. Det de legger vekt på er at surveys handler om å undersøke utvalg fra større populasjoner og at dataene som samles inn analyseres kvantitativt.

Kerlinger & Lee legger altså vekt på at det skal trekkes utvalg fra populasjoner. En kan da stille spørsmål om en survey går over til ikke å være en survey i det øyeblikk en tar med en hel populasjon. Er en spørreskjemaundersøkelse blant rådmenn i et 20 prosents utvalg av norske kommuner en survey, mens et studium av hele populasjonen av rådmenn en annen type undersøkelse? Vi kan vel si det slik at det Kerlinger & Lee gir er en definisjon som forteller oss hva som er en typisk survey. Det vil alltid være mulig å finne eksempler som ikke passer helt med definisjonene av ulike typer forskning. En slik klassifisering i typer forskning som Kerlinger & Lee foretar er likevel nyttig og informativ.

Brian Everitt (1996) deler inn de vanligste formene for psykologiske undersøkelser i fire grupper. Han skiller mellom (i) survey-undersøkelser, (ii) observasjonsstudier, (iii) kvasi-

eksperiment og (iv) eksperiment. Kvasi-eksperiment kjennetegnes ved at de foregår under forhold som gjør det vanskelig eller umulig å sikre sammenliknbare grupper gjennom randomisering, mens man i et "skikkelig" eksperiment kan plassere forsøkspersonene tilfeldig i ulike grupper som utsettes for forskjellig behandling eller påvirkning. Et felteksperiment kan både være et skikkelig eksperiment og et kvasi-eksperiment, avhengig av om det ble gjort en randomisering. At en undersøkelse karakteriseres som kvasieksperimentell, betyr ikke uten videre at den er dårlig. Kvasieksperimentelle studier kan noen ganger være det beste designet som er mulig å få til, og kan gi svært interessante og informative resultater.

Et interessant eksempel på et kvasi-eksperiment var studiene rundt effektene av innføringen av en avgift på sigaretter i California i 1988. Avgiften var så lav som 25 cent per sigarettpakke. En så liten prisøkning hadde sannsynligvis lite å bety for salget av sigaretter. Men avgiften gav en inntekt på 100 million dollar per år, og en betydelig del av disse pengene ble satt inn i arbeidet mot tobakk. Pengene ble brukt til kampanjer i media og til lokale tiltak. Det viste seg at i perioden 1989-1994 gikk tobakksforbruket i California ned med 28%, en nedgang som var dobbelt så høy som i resten av USA. Dette er tatt til inntekt for at de tiltakene som ble satt i gang hadde en betydelig effekt på tobakksforbruket i California (Pierce et al., 2006).

Problemet med dette designet er at en ikke sikkert kan vite at det var mediakampanjene og de lokale tiltakene som førte til den sterke nedgangen i røyking. Dersom en gjennomfører en undersøkelse der en har mange enheter, randomiserer disse (fordeler dem tilfeldig på intervensjonsgruppe og kontrollgruppe), intervenerer bare i den ene gruppen, og så ser at endringene i den ene gruppen jevnt over går i en annen retning enn i den andre gruppen, kan en med større sikkerhet si at intervensjonen var årsaken til forskjellene i endring. Det er imidlertid ikke så enkelt å gjennomføre noe slikt i praksis når det er snakk om å undersøke virkninger av tiltak i hele stater eller land. Stater og land lar seg ikke så lett randomisere og plassere i intervensjonsgrupper og kontrollgrupper. Og de lar ikke utenforstående diktere deres avgiftspolitikk og deres bruk av penger til forebyggende tiltak. Men la oss tenke oss at ikke bare California, men også andre stater etter hvert innførte tilsvarende avgift og lot pengene gå til tilsvarende typer tiltak. Dersom en i hver enkelt av disse statene kunne registrere en økt nedgang i sigarettforbruket etter at dette skjedde, og at nedgangen var større enn i andre stater, ville en etter hvert ha god grunn til å anta at det eksisterte en kausalsammenheng, med andre ord at tiltakene var årsak til nedgangen i røyking.

Mens en i surveys innhenter data ved å intervjue eller ved å la de som deltar fylle ut skjema (eller kombinasjoner av disse), vil en i observasjons-studier innhente data på andre måter. De statistiske teknikkene som anvendes i forbindelse med surveys kan imidlertid ofte komme til nytte ved analyse av data fra observasjonsstudier. Det samme gjelder både felteksperimentelle undersøkelser, kvasi-eksperimentelle undersøkelser og for den del også laboratoriestudier, der en anvender spørreskjema eller på andre måter innhenter informasjon som kan kodes og analyseres kvantitativt.

Grensene mellom de ulike typene forskning er ikke skarpe, og vi kan tenke oss et stort antall ulike kombinasjoner. Dersom vi f.eks. gjennomfører en intervju-undersøkelse ved hjelp av et strukturert og pre-kodet intervjuskjema blant alle arbeidstakerne i en middels stor bedrift,

med tanke på å studere ulike sider ved bedriftskulturen, er det en mellomting mellom en survey og en feltundersøkelse. Spørsmålet om hva som er en survey og hva som ikke er en survey er ikke noe avgjørende spørsmål. Vi kan løse problemet ved å beskrive hva som er en typisk survey-undersøkelse, men samtidig være klare over at grensene til andre typer undersøkelser er glidende. En typisk survey er en undersøkelse der en:

1. Definerer en undersøkelses-populasjon som består av personer. Noen ganger er det aktuelt å gjennomføre undersøkelsen blant alle disse. Oftest er det mest aktuelt å trekke et utvalg.
2. Ved trekking av utvalg følger en vanligvis bestemte prosedyrer for å sikre statistisk representativitet.
3. En innhenter det meste av informasjonen gjennom spørreskjema eller strukturerte intervjuer.
4. En analyserer informasjonen ved bruk av statistiske teknikker beregnet på å beskrive kvantiteter.

1.3 Hva kan en survey brukes til?

Det er vanlig å skille mellom to typer bruk av data fra survey-undersøkelser:

- deskriptiv og
- analytisk.

Deskriptiv bruk av survey-undersøkelser vil si å fortelle noe om hva som karakteriserer en hel befolkning eller deler av en befolkning. For eksempel kan en beskrive utbredelsen av psykiske og somatiske plager i en befolkning, og en beskriver gjerne også hvordan situasjonen ser ut i ulike subgrupper (blant menn og kvinner, i ulike aldergrupper etc.).

Analytisk bruk vil si å undersøke sammenhenger mellom variabler. Ofte sier en at formålet er å predikere. Dersom en har målt opplevde belastninger i arbeids-situasjonen blant arbeidstakerne i et utvalg, kan en korrelere dette målet med psykiske og somatiske plager og dermed forsøke å forklare en del av variasjonen i plager. Dersom sammenhengene er sterke, sier en gjerne at en har lyktes i å forklare mye av variasjonen i plager. Når en bruker data til analytiske formål, anvender en som regel teori eller begrepsmodeller. Ved deskriptiv bruk av data anvendes som regel ingen bestemt teori.

Grensene mellom deskriptiv og analytisk bruk av data fra survey-undersøkelser er ikke så klare som det kan synes. Når en for eksempel undersøker hvordan helseplager varierer på tvers av befolkningsgrupper, kan formålet være beskrivende. Men dersom gruppene defineres ved variabler som utdanning og inntekt, kan formålet være å undersøke samvariasjonen mellom indikatorer på sosioøkonomisk status og helseplager. I så fall kan det tenkes at en også gjør bruk av teori eller begrepsmodeller.

I en diskusjon av hva slags informasjon som trenges innen området forebyggende sosialpolitikk, skiller Hernes (1979) mellom fire typer:

- Probleminformasjon - den sier noe om utbredelsen av et problem i den befolkningen som undersøkes, f.eks. at det er en uakseptabelt høy forekomst av depressivitet i en spesiell subgruppe i befolkningen.

- Årsaksinformasjon - den sier noe om hva som er årsakene til utbredelsen av et problem, f.eks. at denne gruppen rapporterer om spesielt store belastninger i jobben.

- Tiltaksinformasjon - den sier noe om hva slags tiltak som har god effekt. En type tiltak kan f.eks. være økt grad av medbestemmelse over egen arbeids-situasjon.

- Kostnads-nytte-informasjon - den sier noe om effektene av tiltak vurdert mot kostnadene ved de samme tiltakene.

Innen epidemiologisk forskning skiller en mellom deskriptive studier (som tilsvarer det å framskaffe probleminformasjon), analytiske studier (som gir årsaksinformasjon) og eksperimentelle studier (som gir tiltaksinformasjon og noen ganger kostnads-nytte-informasjon) (Bakketeig og Magnus, 2003).

Surveys assosieres gjerne med den første kategorien. De sier noe om gjennomsnitt, prosenter og fordelinger. De brukes med andre ord deskriptivt (beskrivende). En kan imidlertid også bruke surveys til å si noe om sammenhenger mellom variabler. Noen ganger kan en med utgangpunkt i data fra surveys studere temmelig komplekse modeller. Når en gjennomfører slike analyser, er det som regel fordi en er på jakt etter å beskrive kausale prosesser. I slike tilfeller bruker en data fra surveys på en analytisk måte. Noen ganger brukes surveys i forbindelse med felteksperimentelle undersøkelser. Innen forskningen om forebyggende helsearbeid brukes surveys jevnlig for å evaluere kampanjer og aksjoner og til å skaffe informasjon som kan si noe om kostnadseffektivitet. Survey-metoden har med andre ord flere forskjellige anvendelsesområder. Den kan brukes deskriptivt, analytisk og eksperimentelt.

Ofte framheves survey-forskningens begrensninger. Samtidig glemmes lett survey-forskningens fortrinn. Innen deler av den psykologiske forskningen betraktes gjerne laboratorieeksperimentet som den aller beste forskningsmetoden. Dette fordi en her har god kontroll over faktorer som virker inn på de som deltar i undersøkelsen og fordi en med stor sikkerhet kan si noe om hva som er årsak og hva som er virkning. En åpenbar styrke ved survey-undersøkelser er at de ikke fjerner forsøkspersonene fra deres vanlige miljø. En unngår å skape en kunstig situasjon som gjør at folk kanskje oppfører seg og tenker på andre måter enn den vanligvis gjør. Dette siste er laboratorieforskningens store svakhet.

1.4 Kausalitet

Kausalitet er et viktig begrep i vitenskapen. Matematikeren og fysikeren Max Born[3] (1949) summerte opp tidligere forskning og teoretisering omkring kausalitet, og mente på bakgrunn

[3] Max Born (1882-1970), født av jødiske foreldre in Breslau, Tyskland, ble tildelt nobelprisen i fysikk i 1954, og han var bestefar (morfar) til Olivia Newton John, kjent australsk sanger og skuespiller.

av dette at det eksisterer tre kriterier som må være oppfylt for at en skal kunne si at det foreligger et årsaks-virknings-forhold:

1) Det må eksistere noe (B) som avhenger av at noe annet (A) opptrer. Vi kaller A for årsak og B for virkning. Dette "noe" kan være et fysisk objekt, en hendelse, en situasjon eller et fenomén.
2) A må inntreffe tidligere enn - eller i det minste samtidig med - B.
3) A og B må være i fysisk kontakt med hverandre, direkte eller indirekte.

Innen psykologisk og samfunnsvitenskapelig forskning er det sjelden at en bare er interessert i hvordan ett forhold påvirker ett annet. Som regel er det snakk om at det er mange forhold som virker inn, og det kan være snakk om ganske komplekse prosesser. Og ofte er det slik at påvirkningene går begge veier. Likevel er en ofte på jakt etter å finne ut om ett bestemt forhold, når alt annet holdes konstant, har konsekvenser for et annet forhold. Er det for eksempel slik at det å bli arbeidsledig gir økt risiko for psykiske problemer? Eller er det slik at musikkterapi virker bra på pasienter med schizofreni?

Som vi allerede har vært inne på, har survey-forskningen potensiale til også å kaste lys over årsaks-virkningsforhold. Her kan det være fordelaktig å bruke andre typer design enn den typiske tverrsnittsundersøkelsen (undersøkelser som gjennomføres på ett bestemt tidspunkt og ikke gjentas). Visser og medarbeidere (2000) skiller mellom surveys brukt i fire forskjellige sammenhenger:

- Enkle tverrsnittsundersøkelser
- Repeterte tverrsnittsundersøkelser
- Prospektive panelstudier
- Surveys innen eksperimentelle design

De hevder videre at til og med helt enkle tverrsnittsundersøkelser kan brukes til å teste kausalitet. Ved bruk av en teknikk som kalles to-trinns minste kvadraters metode (two-stage least squares), kan en beregne hvor sterkt variabel A virker inn på variabel B og samtidig B's innvirkning på A (Blalock, 1972). Slike analyser baserer seg imidlertid på antagelser om de kausale relasjonene mellom de to variablene. Disse antagelsene kan på sin side testes og endres underveis (James & Singh, 1978).

Videre kan en bruke noe som kalles sti-analyse for å se om sammenhengen mellom to variabler medieres (forklares) av en tredje variabel (Baron & Kenny, 1986; Kenny, 1979). Og endelig kan en i tverrsnitts-surveys også identifisere undergrupper der en sammenheng er til stede til forskjell fra andre undergrupper der sammenhengen ikke er til stede. Den variabelen en bruker for å identifisere undergruppene kalles en moderatorvariabel. Slike moderatorer kan identifiseres ved bruk av en rekke forskjellige statistiske teknikker. I følge Visser et al (2000) vil alle disse formene for statistiske analyser av data fra tverrsnittsundersøkelser (totrinns minste kvadraters metode og identifikasjon av mediatorer og moderatorer) bidra til å belyse spørsmålet om kausalitet.

Hvis endringer over tid i en variabel ledsages av endringer over tid i en annen variabel, er det et tegn på at variablene er kausalt relatert til hverandre. Slike endringer kan studeres ved bruk av repeterte tverrsnittsundersøkelser. Når en skal gjennomføre repeterte tverrsnittsundersøkelser, gjennomføres undersøkelsene med jevne mellomrom blant stadig nye utvalg av personer trukket fra den samme befolkningen. Endringene i de to variablene kan noen ganger være praktisk talt helt parallelle. I så fall er det snakk om en samtidig kausalitet, altså at endringer i den ene umiddelbart fører til endringer i den andre (dersom det da ikke er snakk om en felles, bakenforliggende årsak). Noen ganger kan det imidlertid være snakk om en viss tidsforskyvning, slik at endringer i den ene variabelen kan leses av i form av endringer i den andre som finner sted noe senere. I så fall bør endringene i den variabelen en mener er årsak komme først, mens endringene i den variabelen som er virkning komme noe senere.

Et kjent eksempel er relasjonen mellom røyking i befolkningen og forekomsten av lungekreft. Siden det ofte tar lang tid å utvikle lungekreft, vil en økning i røykingen i en befolkning først etter en tidsutsettelse på 25 – 30 år eller mer følges av en økning i insidensen (antall nye tilfeller) av lungekreft (Shibuya, Inoue & Lopez, 2005). Mens informasjon om røykevaner kan innhentes gjennom surveys, vil en kunne følge endringene i forekomst av lungekreft gjennom bruk av registerdata fra kreftregistre og dødsårsaksregistre.

La oss tenke oss at vi over tid måler befolkningens holdninger til røyking og samtidig måler deres røykevaner. La oss videre tenke oss at vi først registrerer en endring i holdningene. Flere er blitt negative til det å røyke. Etter en tid observerer vi dessuten at andelen som slutter har økt. I et slik tilfelle er det rimelig å tenke seg at endringene i holdninger er forklaringen på at flere har sluttet. Dette er imidlertid ikke noe sterkt funn. Det er for eksempel ikke vanskelig å tenke seg andre forklaringer. Kanskje er det de sosiale normene til røyking som har endret seg, noe som kan gi seg utslag både i endrede holdninger og endret atferd. For å sannsynliggjøre at den første forklaringen er riktig, er det en fordel også å ha målt andre forhold som kan tenke seg å spille inn, for å undersøke om vi kan se bort fra disse.

Både rene tverrsnittsundersøkelser og serier med tverrsnittsundersøkelser har imidlertid sine klare begrensninger når en skal belyse kausalitet. Det å registrere at to eller flere faktorer endrer seg parallelt over tid, er en svært svak indikasjon på at det eksisterer noe slags årsaks-virkning-forhold. Et stykke lenger kommer en dersom en har data fra en prospektiv panelundersøkelse. En prospektiv panelundersøkelse er en undersøkelse der en følger samme personer over tid med gjentatte målinger (minst to ganger). Som regel vil det være interessant å bruke i det minste noen av de samme spørsmålene og skalaene hver gang. Når en har slike data, kan en belyse kausalitet på minst to forskjellige måter. For det første kan en undersøke om endringer i to variabler over tid henger sammen på individnivå. Dersom det er slik at jo mer økning en finner i depressivitet, desto sterkere er økningen i bruk av alkohol, kan det tyde på at disse kausalt henger sammen, uten at en dermed vet i hvilken retning sammenhengen går (den kan i prinsippet også gå begge veier samtidig). For det andre kan en undersøke om en variabel målt på ett tidspunkt henger sammen med endringer i en annen variabel fra det samme tidspunkt til senere målinger. Ved å sammenlikne disse resultatene med det en får ved å bytte om på variablene, får en holdepunkter for hvilken som er årsak og hvilken som er virkning, eller om det kanskje er slik at begge er både årsak og virkning (at det altså er snakk

om et vekselspill). Slike analyser er blant annet gjort for å se i hvilken grad depressivitet kan være en årsak til røyking blant ungdom (Strønstad et al., 2001).

Men de aller klareste holdepunktene for kausalitet får en dersom en kombinerer surveys med eksperimentelle forskningsdesign. Som vi allerede har vært inne på ovenfor, klassifiserer Kerlinger & Lee surveys og eksperimentelle undersøkelser som ulike typer forskning. Men slik vi også har slått fast ovenfor, er ikke skillet mellom ulike typer forskning alltid like klart, og det er ikke noe i veien for å gjennomføre surveys innenfor rammene av eksperimentelle undersøkelser. I slike tilfeller vil det være snakk om felteksperimentelle undersøkelser. Det klassiske designet i slike undersøkelser består i at en først randomiserer (deler tilfeldig inn i intervensjonsgruppe og kontrollgruppe), deretter gjennomføres en eller flere såkalte baseline-undersøkelser (altså datainnsamlinger som blir gjort før det administreres noen intervensjon) i begge gruppene, deretter gjennomføres en intervensjon i den ene gruppen, og deretter en eller flere oppfølgende undersøkelser i begge gruppene. Dersom en finner statistisk sikre forskjeller i endringer mellom de to gruppene på den variabelen eller den faktoren en prøver å påvirke, er dette et sterkt tegn på kausalitet. Et eksempel på en slik felteksperimentell undersøkelse, der en gjorde utstrakt bruk av surveys, er HEMIL-senterets evaluering av Den Norske Kreftforenings program mot røyking blant ungdom som ble kalt "VÆR røykFRI" (Jøsendal et al., 2005). En behøver ikke begrense slike undersøkelser til bare en intervensjonsgruppe og en kontrollgruppe. Noen ganger kan det være interessant å ha flere ulike kontrollgrupper eller flere ulike intervensjonsgrupper.

I dette avsnittet har vi brukt en rekke begreper som vi så langt i teksten ikke har definert eller bare har forklart ganske summarisk, for eksempel variabel, mediator, moderator, og sti-analyse. Disse begrepene vil vi komme grundigere tilbake til senere i teksten.

1.5 Hva er det å måle noe?

I survey-undersøkelser skiller en mellom enheter og variabler. En enhet er vanligvis en person. En undersøkelse omfatter vanligvis et stort antall enkeltpersoner. For hver person registreres en rekke karakteristika og egenskaper.

Når slike karakteristiska og egenskaper er kodet som symboler eller tall og lagret på en datafil i datamaskinen, kalles de variabler. Innen eksperimentell forskning skiller en mellom uavhengige og avhengige variabler. De uavhengige er variabler som en mener påvirker andre variabler (for eksempel en eksperimentell manipulasjon). De avhengige er slike som blir påvirket (for eksempel holdninger, hvis eksperimentet handler om å påvirke holdninger). Innen survey-forskningen bruker en i stedet begrepene prediktorer (ekvivalent til uavhengige variabler) og kriterievariabler (ekvivalent til avhengige variabler). Noen lærebokforfattere og forskere er imidlertid ikke særlig konsekvente i sin begrepsbruk og går på tvers av disse tradisjonene. Aron & Aron (1999) bruker for eksempel begrepsparet "prediktorer" og "avhengige variabler". De går altså på tvers av de to tradisjonene, og hevder at mange andre gjør det samme.

Det å registrere karakteristika og egenskaper på en slik måte at de kan analyseres statistisk etterpå, kalles en måling. Måling defineres i samsvar med Stevens' klassiske redegjørelse fra 1951 vanligvis som det å knytte symboler eller tallverdier til objekter eller hendelser i samsvar med regler. Duncan (1984) har karakterisert denne definisjonen som ufullstendig og sammenliknet med det å definere pianospill som det å slå på tangentene i samsvar med et bestemt mønster. Han tilføyer at en måling handler om å knytte symbolene eller tallverdiene til objektene på en slik måte at det tilsvarer bestemte egenskaper eller ulike grader av en bestemt kvalitet.

I surveyforskningen er det mest aktuelt å knytte tallverdier til egenskaper eller karakteristika hos personer. Reglene det snakkes om i definisjonen ovenfor er framgangsmåter som forteller oss hva vi skal gjøre. En slik regel kan f.eks. være: "Dersom personen er en mann, knytt tallverdien 1 til denne personen, dersom personen er en kvinne, skal tallverdien være 2". Eller, dersom en respondent har svart "Helt enig" på et holdningsspørsmål som har fem svarkategorier, gir vi denne responsen tallverdien 1. Dersom vedkommende har svart "Helt uenig", gir vil tallverdien 5. Og så gir vi tallverdiene 2, 3 eller 4 avhengig av hvilken av de mellomliggende kategoriene det er satt kryss ved.

Ikke alle egenskaper er like enkle å registrere. Mange viktige egenskaper ved mennesket er vanskelige å finne gode mål på. Innen målingsteori brukes begrepet isomorfi om graden av overensstemmelse mellom virkeligheten og resultatet av en måling. Isomorfi betyr bokstavelig oversatt "identitet" eller "likhet i form". Et godt eksempel på isomorfi er overensstemmelsen mellom et geografisk område og et kart over det samme området. Det kan selvsagt innvendes at mennesker er altfor kompliserte til at de kan la seg kartlegge ved bruk av variabler og tall. Til det er å svare at målinger gjennom surveys ikke gir seg ut for å måle mennesket i all sin kompleksitet. Det en måler er bestemte aspekter eller egenskaper ved menneskene som inngår i undersøkelsen. Vi skal også huske på at selv om de enkelte spørsmålene i et spørreskjema eller et intervjuskjema kan (og bør) være ganske enkle, så kan en ved å sette sammen mange nok biter, danne seg et temmelig sammensatt bilde av de gruppene en undersøker.

Likevel vil en survey aldri kunne gi en så god beskrivelse og innsikt i enkeltindivider som en kan oppnå gjennom dybdeintervjuer. Det er da heller ikke hensikten med en survey. En survey er egnet til å beskrive egenskaper ved et større antall personer samlet, likheter og forskjeller mellom grupper av personer, samt mønstre av sammenhenger og interaksjonseffekter mellom de egenskapene en kartlegger. Hvordan en skal fortolke kravet om isomorfi må sees i lys av formålet med den aktuelle undersøkelsen.

Mange egenskaper som en prøver å fange opp gjennom survey-undersøkelser er det umulig å finne sikre og ”objektive” informasjoner om. Holdninger og personlighetstrekk kan vi bare indirekte slutte oss til. I slike tilfeller kan vi skaffe oss et inntrykk av hvor godt instrumentet fungerer ved å finne mange indikatorer på det samme fenomenet, og ved å se på overensstemmelsen mellom disse. Dette skal vi komme grundigere tilbake til i kapittel 2.

I situasjoner der vi faktisk ikke har direkte tilgang til sikker informasjon om den egenskapen vi gjerne vil måle, snakker vi om indikatorer. Dersom vi observerer at en person prater mye

med andre, kan det tas som en indikator på det underliggende personlighetstrekket sosiabilitet. Jo flere ulike indikatorer vi har på et fenomen, desto bedre har vi som regel klart å måle dette fenomenet.

Resultatet av målingene som gjennomføres i en survey er svært ofte en firkantet datamatrise som er organisert slik som på Fig.1.2. I en bestemt rad finner vi tallverdier som symboliserer alle de opplysninger som omhandler en bestemt person eller et bestemt subjekt (S). I en bestemt kolonne finner vi en bestemt opplysning om alle personene som inngår i materialet, vanligvis kalt en variabel (V).

Fig. 1.2: Datamatrisen

	V_1	V_2	V_3	.	V_m
S_1	x_{11}	x_{12}	x_{13}	.	x_{1m}
S_2	x_{21}	x_{22}	x_{23}	.	x_{2m}
S_3	x_{31}	x_{32}	x_{33}	.	x_{3m}
	.	.	.	.	
S_n	x_{n1}	x_{n2}	x_{n3}	.	x_{nm}

S - subjekt V - variabel x - verdi

$_{1,2,3 \ldots m}$ - variabelnummerering

$_{1,2,3 \ldots n}$ - subjektnummerering

1.6 Målenivå

Når vi måler noe, kan dette skje på ulike målenivå. Det er vanlig å skille mellom fire målenivåer:

- Nominal
- Ordinal
- Intervall
- Ratio

En måling på nominalnivå vil si at vi er i stand til å klassifisere i ulike grupper, men uten at det gir mening å plassere gruppene langs noen bestemt dimensjon eller i en bestemt rekkefølge. Gruppene bør være gjensidig utelukkende og alle bør kunne plasseres i en av kategoriene. Eksempler på nominalvariabler er tilknytning til religiøs organisasjon (statskirke, muslimsk trossamfunn, pinsemenighet, ikke tilknyttet noe kirkesamfunn etc.) eller hvilket

politisk parti en stemmer på. Yrke vil som regel også måtte betraktes som en nominalvariabel, selv om det ofte blir gjort forsøk på å sortere yrker langs en dimensjon fra lavstatus til høystatus.

Ordinalvariabler er variabler der en kan rangere observasjonene, men ikke kan si noe bestemt om avstanden mellom dem. La oss tenke oss at skoleelever plasserer seg selv på en skala der de vurderer sine egne skoleprestasjoner. Skalaen kan for eksempel se slik ut: ”Svært flink”, ”Flink”, ”Bedre enn gjennomsnittlig”, ”Omtrent gjennomsnittlig”, ”Dårligere enn gjennomsnittlig”. Det er ganske klart at det her foreligger en bestemt rekkefølge. Dersom vi bytter om på kategoriene slik at de endrer rekkefølge, blir skalaen kaotisk og vanskelig å bruke. Samtidig er det klart at vi ikke kan si at det er like stor avstand fra en kategori til den neste som fra en annen kategori til den neste. Vi kan ikke med sikkerhet si at avstanden mellom ”Svært flink” og ”Flink” er like stor som avstanden mellom ”Omtrent gjennomsnittlig” og ”Dårligere enn gjennomsnittlig”. Dermed er det her snakk om en ordinalvariabel, og ikke det vi nedenfor kaller en intervallvariabel.

Når en skala er såpass grov at vi får flere observasjoner i samme kategori, snakker vi gjerne om ”ties”. Hvis skalaen er svært detaljert, som for eksempel når en rangerer idrettsutøvere etter en konkurranse, kan det noen ganger være null ties.

Intervallvariabler har vi når alle intervaller på skalaen er like lange, men uten at skalaen har et absolutt nullpunkt. Skalaer til måling av intelligens (IQ) betraktes gjerne som intervallskalaer. Dette fordi det er nedlagt et betydelig arbeid i å lage en skala som tilfredsstiller bestemte krav om verdier og fordeling. Det gir imidlertid ikke noe særlig mening å si at skalaen har et absolutt nullpunkt, og det gir heller ingen mening å si at en med IQ på 110 har 10% høyere IQ enn en med IQ på 100. Et annet ofte brukt eksempel på intervallskalaer er temperatur målt på en Celsius-skala (eller like gjerne en Fahrenheit-skala). Null grader på en celcius-skala betyr egentlig ikke at temperaturen ikke kan bli lavere. Dersom vi i stedet bruker Kelvin-skalaen, kan en derimot snakke om et absolutt nullpunkt. Dermed er det ikke lenger bare en intervallskala, men en ratioskala.

Ratioskalaer har med andre ord i tillegg til kravet om like store avstander mellom etterfølgende punkter på skalaen også krav om at det skal eksistere et absolutt nullpunkt. Høyde målt i centimeter eller vekt målt i kilo er eksempler på intervallvariabler. Det gir mening å si at en person på 100 kg er dobbelt så tung som en person på 50 kilo. Ratiovariabler ser en ikke ofte i psykologisk forskning, men det finnes likevel eksempler. Reaksjonstid på en stimulus er en slik ratio-variabel.

En betydelig del av de statistiske teknikkene som er utviklet er basert på at en har variabler som er målt på intervallnivå eller rationivå. Med en fellesbetegnelse kaller en slike variabler for metriske (Weisberg, 1993).

Dikotomier (for eksempel kjønn) er egentlig kategorielle variabler. Dikotomiene står imidlertid i en særstilling. Siden det bare finnes en enkelt distanse på en dikotom variabel (avstanden mellom de to kategoriene), kan en godt si at den er en intervallvariabel. Alle distansene på skalaen (nemlig bare den ene) er like. Alle variabler, uansett målenivå, kan

forenkles til en dikotomi eller til flere dikotomier. Dersom en dikotomiserer en metrisk variabel, vil en imidlertid miste en del informasjon.

Når en variabel er målt på et bestemt målenivå, inneholder den alltid informasjon om de lavere målenivåene. Dersom en går på et idrettsstevne og noterer alle tidene (for eksempel på en 1500 meter på skøyter), kan en lett rangere listen slik at den beste får tallet 1, den nest beste tallet 2 etc., akkurat slik det skjer når en setter opp en resultatliste. Tiden målt i minutter, sekunder og hundredeler er en metrisk variabel, nærmere bestemt en ratiovariabel. Men siden en med utgangspunkt i resultater målt på denne skalaen kunne sette opp en rangering, inneholder den med andre ord også ordinalinformasjon. Dersom en har målinger gjort på en intervallskala eller en ordinalskala, innebærer dette at en kan klassifisere individene i grupper, dersom en ønsker det (de første 10, de neste 10 o.s.v.). Egenskapen til en nominalvariabel, nemlig at en kan klassifisere i grupper, gjelder altså også ordinal- og intervallskalaer. Klassifiseringen i fire målenivåer er med andre ord hierarkisk.

Hva slags målenivå en variabel har, er helt avgjørende for hva slags statistiske analyser vi kan bruke. Til og med den enkle (univariate), beskrivende statistikken avhenger av målenivå. Det gir for eksempel ingen mening å regne ut det aritmetiske gjennomsnittet på en nominalvariabel (der kategorienes rekkefølge ikke har noen bestemt mening) dersom disse variablene har tre eller flere kategorier.

1.7 Utforming av spørreskjemaspørsmål og svarkategorier

I en survey vil konstruksjon av måleinstrumenter først og fremst ha å gjøre med hvordan en formulerer spørsmål og svarkategorier som kan inngå i et intervjuskjema eller et spørreskjema. De fleste innføringsbøker i survey-metode inneholder råd for hvordan en teknisk kan konstruere spørsmål og svarformater som fungerer godt både praktisk og metodologisk (Bradburn & Sudman, 1979; Converse & Presser, 1986; Tourangeau, Rips & Rasinski, 2000). En liste over prinsipper for konstruksjon av spørreskjemaspørsmål kan f.eks. omfatte følgende:

1) Bruk helst spørsmål som kan besvares med tall eller kryss. Dette fordi det letter utfyllingen og fører til færre manglende svar og mer sammenlignbare svar.
2) Åpne spørsmål kan benyttes til innhenting av supplerende informasjon og kan øke forståelsen av resultatene.
3) Gjør spørsmålsformuleringene enkle og unngå grammatikalsk kompleksitet.
4) Hvert enkelt spørsmål bør være endimensjonalt. Det samme gjelder svarkategoriseringen.
5) Unngå ledende spørsmål
6) Unngå ord og uttrykk som er vage, ukjente eller har en uklar mening for informantene

7) Unngå doble nektinger. Slike doble nektinger oppstår lett når du formulerer selve spørsmålet negativt, og deretter opererer med svarkategorier som innebærer benektelse (Foddy, 1993).

8) Unngå overflødige ord. Molenaar (1982) har på grunnlag av en litteraturstudie konkludert med at jo flere ord (substantiver eller informative ord) som blir brukt når en formulerer et spørsmål, desto større er sannsynligheten for at spørsmålet vil bli galt fortolket.

9) Dersom et spørsmål bare skal besvares av en undergruppe av informanter, må dette forklares eksplisitt. Dette kalles å bruke et filter.

10) Spørsmål som senere skal analyseres mot (korreleres med) andre spørsmål, bør ikke gi for skjeve svarfordelinger

11) Det bør normalt bare være lovlig å sette ett kryss for hvert spørsmål eller hvert ledd en skal ta stilling til

12) Dersom det skal være tillatt med mer enn ett kryss, bør antall lovlige kryss spesifiseres. Dette for å motvirke tendensen til at noen svarer svært grundig og setter mange kryss, mens andre er raskere og mer overfladiske og setter få kryss.

13) Svarkategoriene må være gjensidig utelukkende

14) Svarkategoriene må til sammen dekke alle logiske muligheter for de informantene som skal svare på spørsmålet

15) Kategoriseringen bør være "nøytral" eller balansert slik at den ikke leder svarene i en bestemt retning

16) Alle deler av et spørreskjema må være begrunnet praktisk eller teoretisk. Spørsmål som ikke kan begrunnes klart er trolig overflødige, og bør kuttes ut. Alternativt bør det arbeides videre med å utvikle rasjonalet for undersøkelsen.

Det var Payne som i 1951 først satte søkelyset på betydningen av å formulere spørsmål så kort og konsist som mulig (se punkt 8 ovenfor). Han mente at en ikke bør lage spørsmål med mer enn omtrent 20 ord. Prinsippet om at spørsmål skal formuleres med så få ord som mulig, gjelder ikke uinnskrenket. Det finnes faktisk forskning som har vist at en noen ganger får bedre svar ved å bruke mange ord. Converse & Presser (1986) viser til undersøkelser som bekrefter dette. I en slik undersøkelse ble det laget to versjoner av intervjuskjemaet. Den knappeste versjonen bestod av spørsmål som hørtes slik ut:

> ”What medicines, if any, did you take or use during the past 4 weeks?”

Den mer ordrike versjonen inneholdt spørsmål av denne typen:

> ”The next question is about medicines during the past 4 weeks. We want you to think about this. What medicines, if any, did you take or use during the past 4 weeks?”

Når den lengste formuleringen ble brukt, var svarene mer utfyllende enn når den korteste versjonen ble brukt. Det er ikke lett å si hva som var forklaringen på dette. Siden dette var spørsmål som ble brukt under intervju, kan det tenkes at ordrike spørsmål stimulerte til mer

ordrike svar. Lange spørsmål kan tenkes å stimulere informanten til å reflektere mer, og dermed gi lenger og flere svar. Det kan også tenkes at når intervjueren bruker mer tid på å stille spørsmålet, får respondenten mer tid til å tenke seg om allerede før spørsmålet er avlevert. Undersøkelser har vist at intervjuere har en tendens til å gi informantene for kort tid til å svare (Cannell et al., 1979, beskrevet i Converse & Presser, 1986). Trass i at en altså noen ganger kan oppnå mer utfyllende svar ved å stille lange spørsmål, mener Converse og Presser likevel at en bør holde seg til korte spørsmål. Problemet med at intervjuere gir informantene for kort tid til å svare må løses på andre måter, mener de. Intervjuerne må skoleres og trenes bedre.

Noen ganger stiller en spørsmål der en ønsker at informantene skal ta stilling til en rekke enkeltledd. Dersom det dreier seg om en kostholdsundersøkelse, kan det for eksempel være spørsmål av typen ”Hvor godt liker du følgende typer frukt”, og listen av frukt kan omfatte de vanlige typene frukt en finner i dagligvarebutikker. Her kan en be om svar av to ulike typer. En kan be informantene om å svare på en bedømmelsesskala som går fra ”Svært godt” til ”Svært dårlig” for hver type frukt som nevnes. Alternativt kan en be om at hver informant rangerer alle de typene frukt som blir nevnt fra den typen de liker best til den typen de liker dårligst. Fordelen med å bruke bedømmelsesskalaer er at det går raskere og svare. Ulempen er at skalaen gjerne skiller dårlig, slik at svarene blir like på mange av leddene. Skalaen differensierer med andre ord ikke særlig godt. Fordelen med rangering er at en vil skille bedre mellom svarene på de ulike leddene, at kvaliteten på svarene (både validiteten og reliabiliteten) blir bedre, men samtidig tar det gjerne lang tid å rangere leddene (særlig dersom det er mange ledd). Dessuten er det mer tungvint å analysere dataene statistisk når en ber informantene rangere ledd (Visser et al., 2000). En bør derfor tenke seg nøye om før en velger rangering framfor bedømmelsesskalaer.

Når en velger å bruke bedømmelsesskalaer, må en ta stilling til hvor mange svarkategorier en skal bruke. Her må en skille mellom bipolare skalaer (skalaer med to motsatte poler, ofte med kategorier som er symmetriske om midtkategorien) og unipolare skalaer (skalaer som går fra mye av en egenskap til lite eller ingenting av den samme egenskapen). Undersøkelser har vist at for bipolare skalaer fungerer det best (gir best validitet og høyest reliabilitet) med skalaer som har syv svaralternativer (Matell & Jacoby, 1971). For unipolare skalaer er det best med fem alternativer (Wikman & Warneryd, 1990).

Et annet valg en står overfor er om en skal bruke skalaer der bare ytterpunktene er angitt med tekst (og alle verdiene imellom bare er punkter på en linje eller tall langs en tallskala) eller om en skal forsøke å utstyre alle kategoriene med egen tekst. Noen studier har vist at datakvaliteten blir best dersom alle svaralternativer (kategorier) beskrives med tekst (for eksempel Krosnick & Berent, 1993). Når en velger hvordan en skal beskrive de ulike svaralternativene langs en skala eller dimensjon, er det en fordel at en velger ord som gir en mest mulig lik avstand mellom kategoriene. Det er for eksempel uheldig å bruke disse tre svarkategoriene: ”Svært god”, ”God” og ”Dårlig”. Dette fordi avstanden mellom ”God” og ”Dårlig” åpenbart er større enn mellom ”God” og ”Svært god” (Klockars & Yamagishi, 1988; Myers & Warner, 1968).

En ser ofte at forskere opererer med serier av spørsmål der svaralternativene er ”enig – uenig”, ”sant – usant” eller ”ja-nei”. Undersøkelser har imidlertid vist at en del mennesker har lett for å svare ”sant”, ”enig” eller ”ja” uavhengig av hva spørsmålet handler om. De har en tendens til å være enige eller til å føye seg (the acquiescence response bias) (Shuman & Presser, 1981). Feiltendensen er spesielt utpreget hos informanter med begrensede kognitive ferdigheter, når spørsmålene er vanskelige å forstå, og når spørsmålene kommer langt ute i spørreskjemaet, slik at informantene har begynt å bli trøtte. Dette problemet oppstår først og fremst når en skal måle holdninger, oppfatninger, personlighetstrekk eller andre psykologiske forhold. Dersom spørsmålene handler om konkret atferd, demografiske opplysninger og tilsvarende, er det ikke spesielt problematisk for eksempel å bruke svarkategoriene ”ja” og ”nei”.

Når vi designer skalaer og spørreskjemaer, ender vi ofte opp med nokså omfattende instrumenter som krever at de som skal delta i undersøkelsene er høyt motiverte til å svare på en skikkelig måte. Vi krever at de skal lese hvert spørsmål nøye, at de vurderer de ulike svaralternativene, og at de svarer samvittighetsfullt. Men ofte vil vi møte informanter som mangler den nødvendige motivasjon eller som kanskje har problemer med å forstå mange av spørsmålene. Da risikerer vi at de svarer så lettvint som mulig. Kanskje har de en tendens til å erklære seg enige i det meste, det vi ovenfor kalte en føyelighetstendens (the acquiescence response bias), eller kanskje setter de for enkelhets skyld bare kryss i den samme svarkategorien hele veien, uten å vurdere hvert enkelt spørsmål så nøye. Det at en, uten å vurdere spørsmålene så nøye, svarer det samme på hvert enkelt av en hel serie med spørsmål, er eksempel på det som kalles respons-sett. Respons-sett kan defineres som tendensen til å svare på en bestemt måte, uavhengig av innholdet i spørsmålene som stilles. En kan imidlertid ikke uten videre si at det foreligger respons-sett selv om en informant har svart det samme på en hel serie spørsmål. Det kan i prinsippet godt tenkes at svarene virkelig reflekterer vedkommendes oppfatninger, og at vedkommende har svart både veloverveid og samvittighetsfullt.

For å motvirke tendensen til respons-sett, er det noen forskere som har foreslått at en skal veksle mellom positivt formulerte og negativt formulerte påstander, slik at en tvinger informantene til å tenke over hva de skal svare på hvert enkelt spørsmål (Likert, 1932; Anastasi, 1982). Streiner & Norman (2003) hevder imidlertid at dette er en dårlig løsning. Negativt formulerte spørsmål eller påstander bør av flere grunner unngås. For det første vil det å formulere en setning negativt gjerne endre meningsinnholdet. Dersom en svarer ’enig’ på påstanden ’Jeg føler meg vel’ betyr det noen annet enn ’uenig’ på påstanden ’Jeg føler meg ikke vel’. For det andre vil det for mange være kognitivt krevende å skulle svare benektende på en negativt formulert påstand. Informantene vil lett miste oversikten over hva et slikt svar egentlig betyr. For det tredje er det lettere for informantene å gi sin tilslutning til en negativt formulert påstand enn å svare negativt på en positivt formulert påstand. Studier har dessuten vist at negativt formulerte spørsmål har lavere validitet enn positivt formulerte spørsmål. Og sist, men ikke minst har en funnet at skalaer som består av både positivt og negativt formulerte spørsmål eller påstander har lavere reliabilitet enn skalaer der alle spørsmål er positivt formulert. Og vi kan kanskje legge til at selv interesserte og høyt motiverte informanter kan la seg irritere og frustrere over spørreskjemaer som er konstruert slik at det er kognitivt krevende å svare.

Alle skjema må prøves ut på forhånd (pilottestes) på typiske representanter for de gruppene en skal ha med i undersøkelsen. En viktig retning innen sosialpsykologisk forskning er Ajzen & Fishbeins (1980) teori om overveide handlinger (The Theory of Reasoned Action) og Ajzens (1988) teori om planlagt atferd (The Theory of Planned Behaviour). Disse forskerne anbefaler at en under utviklingen av spørsmål til en undersøkelse der en gjør bruk av disse teoriene benytter fokusgrupper eller personlige intervjuer for å identifisere det de kaller sentrale oppfatninger (salient beliefs). Slik bruk av kvalitative tilnærminger er sterkt å anbefale, særlig når en ikke bare er interessert i enkle demografiske eller atferdsmessige karakteristika, men også holdninger, oppfatninger, subjektive normer og andre forhold som er abstrakte. På slike områder er det viktig å undersøke hva som er viktige aspekter for informantene, hva slags ord de bruker for å beskrive fenomenene, og hvordan en skal formulere spørsmålene for at de skal gi en klar mening.

Spørsmålene bør ordnes etter et organiserende prinsipp. Det er en god regel å begynne med enkle, ukontroversielle spørsmål og avslutte skjemaet eller intervjuet med tilsvarende.

Når et spørreskjema eller et intervjuskjema skal danne grunnlag for artikler som skal publiseres i internasjonale tidsskrift, stilles det stadig oftere krav om at en skal si noe presist om instrumentenes kvalitet. Det vil i årene framover bli stadig vanskeligere å få akseptert for publisering manuskripter som baserer seg på spørreskjemadata som er innsamlet bare ved bruk av hjemmesnekrede spørsmål som ikke er skikkelig uttestet. Det forventes blant annet at en skal kunne si noe konkret om spørsmålenes og skalaenes reliabilitet (pålitelighet).

Den motsatte ytterlighet er å ikke ta sjansen på å formulere egne spørsmål, men bare å basere seg på skalaer og instrumenter som andre har brukt tidligere. Det er imidlertid grenser for hvor langt en kan gå i retning av bare å bruke etablerte skalaer og instrumenter. Dersom slike skalaer og instrumenter begynner å leve sitt eget selvstendige liv og blir et ensidig kvalitetskriterium i forskningen, vil en tape mye. Det vil føre til mindre originalitet, dårligere tilpasning av instrumenter til spesifikke problemstillinger og kanskje også for dårlig tilpasning til lokal kultur, språk og levesett. Utviklingen i forskningen er avhengig av at det foregår en stadig nyutvikling av måleinstrumenter. Det er imidlertid viktig at slike nye instrumenter kvalitetstestes allerede fra starten av.

Vanlige framgangsmåter for å vurdere eller sikre kvaliteten på spørreskjemaspørsmål er:

- Å reformulere spørsmål (og/eller svarkategorier) der det er et høyt antall manglende svar. Høyt antall manglende svar kan tyde på at det er noe i veien med måten spørsmålene eller svaralternativene er formulert på, og at en del av respondentene finner det vanskelig å svare på en fornuftig måte.

- Formulere spørsmål og svarkategorier slik at en får god spredning i svarene. Dersom svarene hoper seg opp i en bestemt kategori på ett enkelt spørsmål, vil spørsmålet fungere dårlig når en skal analysere dette mot andre spørsmål. I noen av kategoriene kan antall svar bli så lavt at vi ikke har muligheter for å finne ut noe særlig om denne bestemte gruppen av respondenter. I det ekstreme tilfelle at alle krysser av samme svar på et spørsmål, er ikke dette lenger en variabel, men en konstant.

- Dersom en stiller spørsmål om nøyaktig samme forhold på to steder i et skjema, kan en sjekke svarene mot hverandre og finne ut om det er stor grad av indre konsistens. Dersom direkte inkonsistens (som altså viser at det ene av svarene må være feil) forekommer hyppig, tyder dette på svakheter ved spørsmålet. Inkonsistenser kan også skyldes liten interesse hos de som deltar i undersøkelsen eller at de er trøtte eller indisponerte når de svarer. Dersom skjemaene er svært omfattende, synker ofte kvaliteten på svarene etterhvert som de fyller ut.

- Dersom en har laget serier med spørsmål som er ment å måle samme underliggende fenomen eller begrep, bør disse korrelere høyt innbyrdes. Ofte undersøker en indre konsistens i en skala ved bruk av egne statistiske størrelser, f.eks. Cronbachs alpha. En redegjørelse for Cronbachs alpha er gitt i kapittel 5.

- Ofte kan det være fornuftig å sjekke respondentenes forståelse av innholdet i spørsmålene i et skjema ved å foreta en grundig intervjuing rundt spørsmålene og forsøke å danne seg et inntrykk av hvordan respondenten faktisk har oppfattet dem.

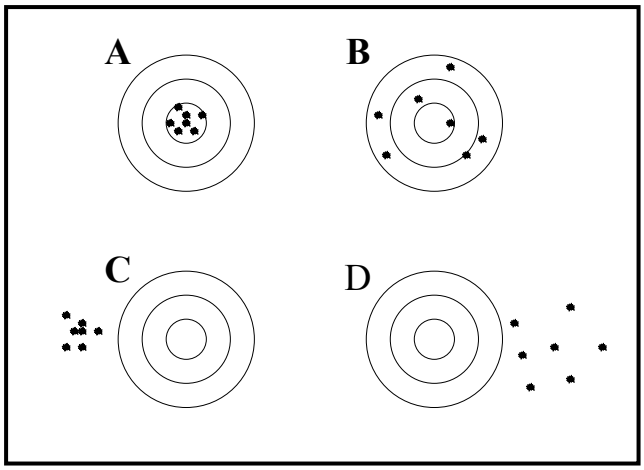
1.8 Validitet og reliabilitet

De fleste lærebøker og innføringstekster i samfunnsvitenskapelig eller atferdsvitenskapelig forskningsmetode pleier å ofre en hel del plass på begrepene validitet og reliabilitet. Validitet oversettes gjerne med gyldighet. Dersom vi måler noe på en valid måte, betyr det at vi måler det vi faktisk ønsker å måle. Validitet handler med andre ord om forholdet mellom begreper og måleinstrumenter. Noen ganger er dette ganske enkelt. Begrepet kan være kjønn, og vi måler kjønn ved at hver enkelt av de som deltar i undersøkelsen krysser av for om vedkommende er mann eller kvinne. Dersom de som deltar tar undersøkelsen seriøst og er tilstrekkelig motivert til å svare på skjemaet, vil svaret på dette spørsmålet så å si alltid bli riktig, og vi vil ikke være i tvil om at det faktisk er et gyldig mål på det vi mener med kjønn. Dersom vi forsøker å måle noe mer abstrakt, for eksempel et personlighetstrekk, kan det være langt vanskeligere å avgjøre om vi måler det vi ønsker å måle.

Reliabilitet handler om hvor nøyaktig vi måler, og oversettes gjerne med pålitelighet. La oss tenke oss at vi skal måle vekten på en gruppe personer. Vekten vi bruker har imidlertid nokså treg mekanikk, og den har en tendens til å stoppe på tall litt over eller litt under det riktige tallet. Det betyr at når vi veier samme person flere ganger etter hverandre, får vi hele tiden litt forskjellige tall. Dette er et eksempel på at reliabiliteten ikke er tilfredsstillende. Dersom vekten fungerer helt fint rent teknisk, men er feil justert, får vi et annet problem. Dersom vekten er justert slik at den viser 3% for mye, betyr det at når vi veier en person som egentlig veier 70 kilo, så viser vekten 72,1 kilo. Problemet her er ikke lav reliabilitet. Dersom vi veier den samme personen flere ganger etter hverandre, får vi det samme resultatet hver gang. Målingen er med andre ord svært reliabel. Likevel er den feil. Den er ikke valid.

Ofte illustreres forholdet mellom validitet og reliabilitet med en enkel figur som den vi har vist nedenfor (Fig. 1.3). Den viser fire blinker og vi ser av de svarte punktene hvor skuddene traff.

Fig. 1.3: Validitet og reliabilitet



A:
Høy reliabilitet
Høy validitet

B:
Lav reliabilitet
Lav validitet (!)

C:
Høy reliabilitet
Lav validitet

D:
Lav reliabilitet
Lav validitet

En kan nærme seg spørsmålet om å vurdere instrumenters reliabilitet på tre forskjellige måter:

1) Dersom vi måler en egenskap flere ganger over tid, får vi da samme (eller nesten samme) resultatet hver gang? Forutsetningen for at dette skal gi mening er at egenskapen er nokså stabil over den tidsperioden testingen foregår. Dersom vi for eksempel bruker en skala for å måle et personlighetstrekk i en gruppe, og målingen blir gjort to ganger med to ukers mellomrom, regner vi med at det bør være temmelig høyt samsvar mellom første og andre måling. Personlighetstrekk antas å være ganske stabile egenskaper, og de forventes ikke å endre seg vesentlig i løpet av to uker. De som skåret høyt første gang forventes å skåre omtrent like høyt andre gang. De som skåret lavt første gang, forventes så skåre ganske lavt andre gang. Dersom det er dårlig samsvar mellom de to målingene, tas det som et tegn på at instrumentet har lav reliabilitet.

2) Den andre tilnærmingen har å gjøre med nøyaktighet. Er resultatet av målingen virkelig i samsvar med det som er "korrekt"?

3) Reliabilitet kan også defineres som fravær av feil. Denne siste tilnærmingen gjør det mulig å operasjonalisere reliabilitet på en enkel måte. Reliabilitet defineres som andelen "sann" varians av den totale variansen på en variabel. Dersom vi var i stand til å måle de "sanne" verdiene på en egenskap, er reliabiliteten lik kvadratet av korrelasjonen mellom den variabelen vi vil måle reliabiliteten på og de sanne verdiene. Hva varians er og hva korrelasjon er vil bli nærmere beskrevet i kapittel 3.

En felles komité nedsatt av American Psychological Association, the American Educational Research Association og the National Council on Measurement used in Education nådde fram til enighet om hva som skal betraktes som de viktigste formene for validitet. Kerlinger & Lee (2000) bruker den samme inndelingen og sier at validitet grunnleggende sett kan deles i tre typer:

1) Innholdsvaliditet
2) Kriterievaliditet
3) Begrepsvaliditet.

Høy innholdsvaliditet innebærer at en måler et representativt eller adekvat spekter av del-egenskaper ved den forhold en forsøker å måle. Dersom en for eksempel skal måle daliglivsplager i en befolkning, er det viktig å ha med spørsmål om alle de dagliglivsplager som er vanlige i denne befolkningen. Dersom noen av de vanligste plagene ikke dekkes av spørsmålene som er med i skjemaet, svekkes validiteten.

Noen ganger hører en om en type validitet som er nært beslektet med innholdsvaliditet, nemlig overflate-validitet (face validity). Dette handler om hva spørsmålet eller skalaen synes å måle. Legfolk eller eksperter kan inspisere de leddene som inngår i en skala og ut fra det ser vurdere om de synes at den måler det den skal måle. I slike tilfeller kan validiteten eventuelt kvantifiseres ved bruk av koeffisienter som viser grad av enighet mellom ulike vurderere. Kerlinger & Lee hevder imidlertid at dette ikke kan kalles validitet i det hele tatt.

Høy kriterie-validitet har en når det instrumentet en har utviklet korrelerer høyt med et kriterium som en antar er et godt mål på det forhold en vil undersøke. Dersom vi for eksempel ønsker å utvikle en skala som måler depressivitet, er det et positivt tegn på validitet dersom en gruppe pasienter som har fått diagnosen depresjon jevnt over skårer høyere enn en gruppe personer som ikke har fått en slik diagnose. Vi antar da at de legene eller psykologene som har stilt diagnosen har gjort en god jobb og at en slik diagnose kan stilles med nokså stor grad av sikkerhet. Diagnosen blir dermed et kriterium på hvor god skalaen er. Dersom kriteriet en validerer utfra er en måling som blir gjort omtrent samtidig med målingen på den variabelen en ønsker å validere, kalles det samtidig validitet (concurrent validity). Dersom kriteriet er en måling som kommer senere i tid, kalles det prediktiv validitet (predictive validity).

Begrepsvaliditet kan undersøkes ved å korrelere en variabel som måler en egenskap med andre variabler som er mål på egenskaper som vi ut fra teorier eller begrepsmodeller mener bør være korrelerte med den første. I forlengelsen av dette resonnementet kan vi legge til at også mangel på sammenheng mellom variabler kan brukes som et kriterium på begrepsvaliditet. Kort sagt kan vi si at høy korrelasjon med variabler som en spesifikk variabel bør korrelere høyt med kombinert med lav eller ingen korrelasjon mellom variabler som den ikke bør korrelere med, er et tegn på tilfredsstillende validitet[4].

[4] Disse prinsippene anvendes i en svært elegant teknikk som ble lansert av Campbell & Fiske i 1959. Den baserer seg på noe som kalles multitrekk-multimetode matrisen, og poenget med teknikken er nettopp at noen sammenhenger bør gi høy korrelasjon samtidig med at andre bør gi lav korrelasjon for at validiteten skal være tilfredsstillende.

En redegjørelse for hvordan kvaliteten på målingene er sikret og en vurdering av hvor tilfredsstillende kvaliteten på målingene er, hever nivået på en vitenskapelig publikasjon betydelig. Ved å administrere et spørreskjema til samme gruppe individer på to tidspunkt etter hverandre, og ved å beregne test-retest-korrelasjonen[5], får vi en viss indikasjon på instrumentets kvalitet. Tidsavstanden mellom testingene bør være såpass stor at en ikke risikerer at den enkelte husker for mye av sine svar fra første runde. Dersom hukommelsen fungerer for godt, og dersom de som svarer er innstilt på å vise at de er konsistente i sine svar, vil en kunne overestimere reliabiliteten. Samtidig bør ikke tidsavstanden bli så stor at selve det fenomenet en studerer har endret seg vesentlig. Dersom fenomenet har forandret seg, risikerer en å underestimere reliabiliteten. En lav test-retest-korrelasjon behøver ikke alltid bety at den enkelte har svart upresist, men kan bety at reelle endringer har funnet sted.

Å måle psykologiske egenskaper og faktorer på en valid og reliabel måte, er ikke enkelt. Vi kunne liste opp en lang rekke forhold som kan komme forstyrrende inn. Men vi skal i stedet bare nevne et fascinerende eksempel. Schwarz & Clore (1983), som gjennomførte undersøkelser av livskvalitet, fikk mistanke om at informantenes sinnsstemning (mood) noen ganger påvirket vurderingene av hvordan de selv hadde det. Når en skal måle livskvalitet er en som regel ikke interessert i at informantenes sinnsstemning i øyeblikket skal påvirke svarene for sterkt. Det er mer stabile vurderinger av ulike aspekter ved eget liv og egen livssituasjon en er ute etter. En faktor som lett kan påvirke folks sinnsstemning er været. For å teste hypotesen, gjennomførte Schwarz & Clore derfor intervjuer både på dager med mye sol og på dager med overskyet vær. Det viste seg at de som ble intervjuet på solrike dager rapporterte om større tilfredshet med livet enn de som ble intervjuet på dager da det var overskyet. Men dersom de som ble intervjuet på overskyede dager forut for spørsmålet om tilfredshet med livet først måtte svare på et spørsmål om hvordan været var, endret svarene på livskvalitetsspørsmålene slik at de var like tilfredse med tilværelsen som de informantene som ble intervjuet på solrike dager. Spørsmålet om været fikk trolig informantene til å attribuere (tilskrive) noe av sin sinnsstemning til at været var dårlig, og dermed tolket de i mindre grad sin sinnsstemning som et uttrykk for sin generelle livskvalitet.

Av denne studien ser vi at en så enkel ting som været den dagen en foretar intervjuer kan påvirke informantenes svar. I tillegg illustrerer studien at ett enkelt spørsmål, som stilles før intervjuet begynner, kan endre informantenes respons på spørsmål som blir stilt under selve intervjuet. Dette viser hvor lite som skal til for å påvirke informantene når en skal måle et såpass abstrakt forhold som ”livskvalitet”.

1.9 Utvalg og populasjon

[5] Hva en korrelasjon er blir nærmere beskrevet i kapittel 3. Her er det bare nødvendig å skjønne at en korrelasjon er et uttrykk for i hvilken grad hver enkelt av informantene svarer det samme (har samme relative eller absolutte plassering på skalaen) første og andre gang. Jo mindre sammenheng det er mellom det de svarer på nøyaktig samme spørsmål når spørsmålene stilles med et passende tidsintervall i mellom, desto lavere antar vi at reliabiliteten på spørsmålet er. Det finnes ulike typer korrelasjoner å velge mellom når en skal beregne reliabilitet. Dette vil vi komme tilbake til i kapittel 5.

Før vi ser nærmere på forholdet mellom utvalg og univers, er det viktig å være klar over to begreper og hvordan de skiller seg fra hverandre. Det ene er statistisk størrelse og det andre er parameter. En statistisk størrelse er noe en regner ut på grunnlag av alle enhetene som er med i et utvalg, f.eks. et aritmetisk gjennomsnitt, et standardavvik eller et prosenttall. En parameter er den tilsvarende egenskapen for populasjonen, verdier en vanligvis ikke kjenner noe til, men som en prøver å anslå ved å basere seg på den informasjon som finnes i dataene fra utvalget. En sier gjerne at den statistiske størrelsen en har regnet ut på grunnlag av utvalget er et estimat av parameteren. Dersom vi hadde data for hele populasjonen, kunne vi regne ut parametrene direkte.

Noen av de statistiske størrelsene en kan regne ut på grunnlag av data fra et utvalg kan ikke uten videre forventes å gi et helt riktig bilde av tilsvarende parameter i populasjonen. Statistikerne har imidlertid utviklet formler som korrigerer utregningene slik at de likevel kan forventes å gi et riktig bilde. Slike korrigerte formler gir oss det som kalles forventningsrette estimat.

En annen viktig distinksjon er skillet mellom beskrivende statistikk og slutningsstatistikk (ofte brukes ordene deskriptiv- og analytisk statistikk). Den første typen statistikk handler om å regne ut statistiske størrelser som summerer opp egenskaper ved utvalget (eller estimerer de samme størrelsene med tanke på at de skal være gyldige for populasjonen). Slike egenskaper kan være de samme som er nevnt ovenfor: aritmetiske gjennomsnitt, standardavvik eller prosenttall. Eller det kan være statistiske størrelser som beskriver sammenhenger mellom variabler, for eksempel korrelasjoner eller forskjeller mellom gjennomsnitt. En del slike statistiske størrelser skal beskrives senere i denne teksten. Slutnings-statistikken er den statistikken som anvendes når en skal trekke slutninger fra utvalg til univers. Hvor presist kan en si noe om hele populasjonen når en har regnet ut en statistisk størrelse basert på et utvalg.

Universene vi generaliserer til behøver ikke alltid være virkelige univers. Noen ganger forestiller en seg at de enhetene som inngår i utvaget er trukket fra et teoretisk univers som vi tenker oss er uendelig stort. De vanligste formene for slutnings-statistikk er konfidensintervall og signifikanstester. Disse vil bli nærmere beskrevet i neste kapittel.

1.10 Tekking av utvalg

De fleste surveys gjennomføres på nokså store populasjoner. Det er derfor nesten bestandig nødvendig å trekke utvalg. Måten en trekker utvalg på, har konsekvenser for hvordan en senere kan bruke data til statistiske analyser.

Første trinn er å bestemme målpopulasjonen, det vil si å definere den befolkningen en ønsker å studere. En slik populasjon kan f.eks. være alle nålevende personer som er utdannet som psykologer i Norge. I praksis må en finne fram til et register som inneholder flest mulig av disse. Et slikt register kan være alle som er medlemmer av Norsk Psykologforening.

Imidlertid er det ikke slik at alle som har psykologisk embetseksamen i Norge er med i dette

registeret. Dermed oppstår det en forskjell mellom målpopulasjon og det som kalles survey-populasjon (sampling frame). Idealet er selvfølgelig å oppnå en best mulig overensstemmelse mellom målpopulasjon og survey-populasjon.

De vanligste måtene å trekke representative utvalg på er følgende (Cochran, 1963):

* Rent tilfeldig trekking
* Klyngeutvalg
* Stratifiserte utvalg
* Flertrinnsutvalg

Alle disse framgangsmåtene kan sikre utvalg som er av god kvalitet. Alle er eksempler på representative utvalg. Et statistisk representativt utvalg er trukket slik at alle elementer i populasjonen har en kjent sannsynlighet, forskjellig fra 0 og 1, til å komme med i utvalget[6]. Dette betyr at alle, absolutt alle som tilhører surveypopulasjonen, må ha en viss sjanse til å komme med i utvalget, selv om sannsynligheten godt kan være svært lav. Det betyr også at ingen på forhånd må være sikre på å komme med i utvalget, selv om sannsynligheten gjerne kan være svært høy. Legg merke til at definisjonen ikke sier at alle elementene i surveypopulasjonen skal ha en lik sannsynlighet for å komme med i utvalget. Den sier heller ikke at alle elementene skal trekkes ut uavhengig av hverandre. Begrepet representativt utvalg er dermed et nokså generelt begrep som dekker alle mulige former for sannsynlighetsutvalg.

Rent tilfeldige utvalg har den fordel at en uten videre kan anvende den vanlige statistikken som finnes i standard programpakker. Dersom en trekker utvalgene på andre måter, kan det lett få konsekvenser for den statistiske bearbeidelsen av data og fører ofte til at en øver vold mot statistisk teori. Kalton (1983) sier det slik:

> The regular standard error formulae found in statistics texts and incorporated in most computer programs relate only to unrestricted sampling (simple random sampling with replacement). These formulae should not be applied uncritically with other sample designs, for which they may produce overestimates or, more often, underestimates of the sampling error." (s.75)

Kalton diskuterer i sin bok hvilke konsekvenser ulike måter å trekke utvalg på har for samplingfeil. Når en trekker utvalg fra en endelig populasjon, blir for eksempel standardfeilen på ulike estimat mindre enn når en trekker fra en uendelig stor populasjon. Jo større den endelige populasjonen er, desto mindre feil gjør en imidlertid ved å basere seg på standardformler. Når populasjonene er mindre, kan en vinne noe presisjon på å basere seg på formler som tar hensyn til dette. Men det å finne fram til de aktuelle formlene og gjøre de nødvendige beregningene er som regel så arbeidskrevende at det er enklere allerede i utgangspunktet å trekke et litt større utvalg og dermed bedre sjansene for å finne de sammenhengene en er ute etter.

---

[6] Noen steder defineres representativitet på en slik måte at en godtar at noen elementer kan ha en sannsynlighet på 1,0 for å komme med i utvalget. Se for eksempel Visser et al., 2000.

Den første som lanserte idéen om å gjennomføre undersøkelser blant utvalg i stedet for i en hel befolkning var faktisk nordmannen Anders Nikolai Kiær (1838-1919). Ved et møte i ISI (the International Statistical Institute) i 1895 presenterte han en rapport med sine erfaringer om det han kalte "den representative metode". Han mente at dersom en kunne sette sammen et utvalg på en slik måte at det på sentrale variabler hadde en fordeling som tilsvarte populasjonen, så kunne en nøye seg med å beskrive dette utvalget, og med rimelig sikkerhet si at resultatene gjaldt populasjonen som helhet. Han presenterte innlegg om sin representative metode ved møter i ISI helt fram til 1903. Kiærs ideer vant litt etter litt oppslutning. I 1903 ble metoden akseptert internasjonalt.

Kiær kom imidlertid ikke så langt at han lanserte idéen om å trekke tilfeldig for å sikre representativitet. Denne ideen ble lansert av Bowley i 1906 (Bethlehem, 1999). Selv om en kan finne eksempler på at ideen med å trekke representative utvalg eksisterte så tidlig som før år 1900 (Stephan, 1948), har de moderne samplingteknikkene utviklet seg først i løpet av tiden etter 1950.

Kiærs idéer var så viktige at de er blitt karakterisert som begynnelsen på en ny epoke i statistikken. Kiær var for øvrig den første direktøren for Statistisk sentralbyrå. Ifølge den norske sosialøkonomen Tore Schweder, var det statistiske forskningsmiljøet i Skandinavia helt på høyde med Sir Francis Galton og hans samtidige. Men når Fisher kom inn på scenen i England, og med uttørkingen av statistikkskolen i København, overtok Storbritannia etter hvert føringen.

http://www.stat.fi/isi99/proceedings/arkisto/varasto/schw0844.pdf

Cochran (1963) gjennomgår metoder for å justere samplingfeilen for mer komplekse utvelgingsteknikker. Både bruken av beskrivende statistikk og bruken av slutnings-statistikk blir imidlertid enklest dersom en holder seg til rent tilfeldige utvalg. Vi skal nå se litt på andre typer utvalg og hva slags konsekvenser det har når vi ikke lenger holder oss til de rent tilfeldige utvalgene.

Klyngeutvalg trekkes ved at en velger ut hele grupper av enheter samtidig. Dersom en ønsker et representativt utvalg av norske skolebarn og trekker et rent tilfeldig utvalg av skoleklasser, får en et klyngeutvalg. Dersom vi måler forhold som kan tenkes å henge sammen med hvilken skoleklasse en går i (slik at barna innen en skoleklasse ligner hverandre mer enn på tvers av skoleklasser), vil usikkerheten bli større ved klyngeutvalg enn ved rent tilfeldige utvalg. Som en generell regel kan en si at jo større klyngene er, desto mer usikre blir de tallene en regner ut (Kalton, 1983, s.77). Det er derfor mer fordelaktig å trekke enkeltklasser enn f.eks. å trekke hele skoler. Klyngeutvalg gir normalt større usikkerhet enn rent tilfeldig utvelging. En må derfor bruke andre (og mer kompliserte) formler enn de standardformlene som brukes i de fleste statistikkbøker, og konfidensintervallene blir som regel større enn de vi får ved å benytte formler som baserer seg på rent tilfeldig trekking.

Stratifiserte utvalg trekkes ved at en først deler inn alle enhetene i surveypopulasjonen i undergrupper eller strata, f.eks. etter kjønn og alder. Deretter trekker en et rent tilfeldig utvalg innen hvert stratum. Dersom størrelsen på disse utvalgene er proporsjonal med størrelsen på strataene (f.eks. at en velger et 10% utvalg fra hvert stratum), snakker en om proporsjonal stratifisert trekking. Proporsjonal stratifisert trekking (med rent tilfeldig trekking innen hvert stratum) gir estimater som alltid er like presise eller mer presise enn de vi får ved rent tilfeldig trekking (Kalton, 1983, s.76)[7].

Dersom en har informasjon om hvordan fordelingen på en bestemt variabel ser ut innen hvert stratum i populasjonen, kan en, gitt en total størrelse på hele det utvalget som skal trekkes, beregne hvor mange enheter en må ha med fra hvert stratum for å få en mest mulig presis estimering av parameteren for hele populasjonen. Dette kalles optimal stratifisert trekking. En kan også stratifisere på andre måter. Dersom en ønsker å regne statistikk på en spesiell undergruppe, og denne undergruppen bare representerer en liten del av populasjonen, kan en velge å la denne gruppen være overrepresentert i utvalget. Mens en i alle andre strata trekker 10%-utvalg, kan en i denne spesielle gruppen for eksempel trekke et 30%-utvalg.

En kan også anvende en framgangsmåte som sikrer at en med gitte pengemidler kan få så presise estimat som mulig. Dersom det har ulik pris å innhente data fra ulike strata, kan denne prisen bygges inn i en formel, og en kan beregne nøyaktig hvordan utvalget skal være sammensatt for å gi best mulig presis informasjon innenfor en gitt budsjettramme (Kalton, 1983).

Dersom en stratifiserer og trekker rent tilfeldige utvalg innen hvert stratum, men ikke benytter proporsjonal stratifisering, må alle estimeringer av populasjonsparametre foretas med vekting. Det stratum eller de strata som er overrepresentert, må vektes ned (vektes med tallverdier under 1,0) slik at de teller like mye som i populasjonen. En kan også gi de som er underrepresentert høye vekter (vekter større enn 1,0). Vekter større enn 1,0 må imidlertid brukes med stor forsiktighet. Dersom en anvender slutningsstatistikk på data der en har vektet med tallverdier større enn 1,0, risikerer en å gi inntrykk av en presisjon det ikke er dekning for, med mindre en benytter spesiell programvare som tar hensyn til at dataene er vektet. En kan lett oppnå signifikante sammenhenger og forskjeller som ikke er reelle[8].
Til forskjell fra klyngeutvelging kan stratifisert utvelging bidra til å redusere størrelsen på standardfeilene og konfidensintervallene, gitt en bestemt utvalgsstørrelse. I sin utmerkede innføringsbok i statistikk for psykologer beskriver Guilford (1965) ved hjelp av enkle formler hvordan dette skjer.

---

[7] Visser et al (2000) har misforstått på dette punktet (s. 233). De skriver at et hvilket som helst stratifisert utvalg har en designeffekt som er lavere enn 1,0, noe som betyr at en alltid får mer presise tall enn ved rent tilfeldig trekking. Dette er ikke riktig.

[8] Noen statistikkpakker, for eksempel STATA kan behandle problemet med vekting på en langt mer elegant måte. En vekter individene i hvert stratum med et forholdstall (Nj/nj) som er den inverse av samplingfraksjonen for det aktuelle stratum. Deretter kan programmet vekte riktig samtidig som det holder styr på det egentlige n i hvert stratum og regner ut standardfeil, konfidensintervall og signifikanstester. Noen programmer kan også korrigere for designeffekten som oppstår når en trekker klyngeutvalg (cluster sampling). I SPSS inneholder modulen "Complex" statistiske prosedyrer for å vekte og for å ta hensyn til stratifisering og klyngeutvelging (cluster sampling).

Flertrinnsutvalg innebærer at en foretar trekking i flere omganger. For å trekke landsrepresentative utvalg som skal være med i intervju-undersøkelser, er en avhengig av en viss geografisk samling av intervjuobjektene av praktiske og økonomiske årsaker. Dette problemet kan løses ved at en først trekker et utvalg geografiske enheter (f.eks. kommuner), og deretter trekker utvalg av personer innen de geografiske enhetene som er trukket ut. Igjen er det slik at presisjonen reduseres. Dette kan imidlertid kompenseres ved at en trekker større utvalg.

Ofte trekkes utvalg på en slik måte at en strengt tatt ikke kan si at det er representative utvalg eller sannsynlighetsutvalg, men der en likevel kan analysere data under den forutsetning at det har samme egenskaper som et rent tilfeldig utvalg. Sekvensiell utvelging (ofte kalt systematisk utvelging) er ett slikt eksempel. Dersom en har et register som er ordnet etter kriterier som en er sikker på ikke har noen systematisk sammenheng med de forhold en ønsker å kartlegge, kan en bruke slik sekvensiell trekking. Dersom en har et register på 30.000 enheter og ønsker et utvalg på 3.000, kan en velge å starte med et hvilket som helst tall mellom 1 og 10 og deretter trekke ut hvert tiende subjekt. Oftest trekkes starttallet tilfeldig. En annen framgangsmåte består i å finne et uavhengig kriterium, og ta med i utvalget alle de som oppfyller dette kriteriet. Ett eksempel er Statens tobakkskaderåds (1999) røykevaneundersøkelser blant elever i ungdomsskolen. I disse undersøkelsene blir alle elever født den 6. uansett måned tatt med i utvalget. En går da ut fra at det å være født på denne spesielle dagen i måneden ikke skulle være avgjørende for ens røykevaner.

En annen teknikk er såkalt kvoteutvelgelse. Dette innebærer at en stratifiserer populasjonen og sørger for å fylle opp hvert stratum til et visst nivå. Dermed kan en f.eks. sikre seg at utvalget får en riktig demografisk sammensetning. Dersom de som inngår i slike utvalg ikke er trukket tilfeldig, men rekruttert på andre måter, kan slike utvalg ikke sies å være representative i statistisk forstand.

I de tilfeller en simpelthen bare tar med i utvalget en gruppe som er lett tilgjengelig, f.eks. en gruppe studenter ved en grunnfagsforelesning ved Universitetet, kaller en dette bekvemmelighetsutvalg (convenience sampling) (Kalton, 1983). Dette gir selvsagt heller ikke utvalg som er representative i statistisk forstand.

1.11 Om å øke deltakelsen i en survey

Feilkildene som oppstår i forbindelse med surveys deles gjerne inn i to kategorier; tilfeldige feil og systematiske feil (Groves, 1989). De tilfeldige kan f.eks. skyldes at en har med et forholdsvis lite utvalg, og at de tallene en beregner derfor er beheftet med stor usikkerhet. Systematiske feil kan f.eks. handle om at spørsmål er stilt på en ledende måte, og at en derfor ikke får et dekkende bilde av hva folk egentlig mener om et saksforhold.

Oppslutningen om en undersøkelse, oftest oppgitt som prosentandelen av de opprinnelig utvalgte som faktisk deltar i undersøkelsen, er en kilde til begge typer feil. Når mange lar være å delta, blir utvalget mindre, og tallene dermed mindre sikre. De feil som skyldes tilfeldigheter øker med andre ord. Dersom de som lar være å delta i en undersøkelse er ulike

de som deltar, står vi overfor et problem som oftest er langt større. Vi sitter tilbake med et utvalg som ikke er representativt. Det er systematisk forskjellig fra det opprinnelige utvalget.

Svarprosenten varierer ofte med type undersøkelse som blir gjennomført. Kerlinger (1986) hevder at i postale surveys er det vanlig med så lav svarprosent som 50-60. Heller ikke telefonsurveys pleier å gi noen oppslutning høyere enn 60%. Og i undersøkelser der en gjennomfører datainnsamlingene ved hjelp av personlige intervjuer, kommer en sjelden over 70% (Visser et al., 2000). Basert på data fra 16 europeiske land fant Hox & de Leeuw (2002) at i perioden 1970-1990 økte andel en ikke lyktes i å få kontakt med med 0,2% per år, og avslagsraten (nektet å delta) økte med 0,3% per år. Også i Norge er det en tendens til at deltakelsen i survey-undersøkelser stadig blir lavere. Dette kan ha sammenheng med at slike undersøkelser er blitt så vanlige at mange er lei av å bli intervjuet og fylle ut spørreskjema.

Et nokså dramatisk eksempel på konsekvensen av frafall dette har vi fra en undersøkelse av røyking blant norske leger. Undersøkelsen ble gjennomført i 1974, på et tidspunkt da debatten om røyking og helse var svært intens. Det var kort tid før innføringen av tobakksloven, og media hadde ofret mye spalteplass på stoff om røyking og helse. I denne situasjonen gjennomførte Statens tobakkskadråd en undersøkelse av røyking blant legene. Etter første utsendelse av skjema var det 82% som svarte. Etter to purringer hadde ytterligere 12% svart (Aarø, Bjartveit & Vellar, 1977).

Det viste seg å være en svært klar sammenheng mellom tidspunktet legene svarte på og andel som røykte daglig. Blant de menn som svarte på første henvendelse, var det 32% dagligrøykere. Blant de som svarte etter purring var tallet 58%. For de kvinnelige legenes del var tilsvarende tall 17% og 49%. Dette viser betydningen av høy deltakelse. Dersom en ikke hadde purret, hadde en beregnet et noe for lavt tall når det gjaldt andel dagligrøykere blant legene.

I alle undersøkelser der det gjennomføres purringer, bør en registrere på hvilket tidspunkt hvert svar kom inn. Denne opplysningen bør legges inn som egen variabel på datafilen. Når en bare oppnår en moderat eller lav svarprosent, og samtidig finner klar sammenheng mellom svartidspunkt og sentrale variabler i undersøkelsen, bør en utvise stor forsiktighet når en ønsker å generalisere fra det utvalget en sitter igjen med til den opprinnelige populasjonen.

Sammenhengen mellom svarvillighet og fordelinger på de spørsmålene en stiller i undersøkelsen er ikke alltid så sterke som dette. I en undersøkelse av mosjonsvaner i befolkningen fant en svært lave og bare ikke-signifikante forskjeller ved baseline mellom de som deltok i en oppfølgende undersøkelse og de som ikke deltok (Ommundsen & Aarø, 1994).

Det finnes faktisk også eksempler på at undersøkelser med lav deltakelse gir riktigere resultater enn undersøkelser med høyere deltakelse. Visser og medarbeidere (1996) sammenliknet over en 15-års periode resultatene fra to serier av undersøkelser av politiske holdninger i den voksne befolkningen i Ohio, USA. Den ene serien ble gjennomført ved bruk av postale spørreskjema, og svarprosenten var på 20 prosent, noe som er svært lavt. Den andre serien av undersøkelser var basert på telefonintervjuer og hadde en svarprosent på jevnt

over 60. Det viste seg at den postale undersøkelsen predikerte stemmegivningen ved politiske valg bedre enn undersøkelsen som var basert på telefonintervjuer. Den gjennomsnittlige feilen i undersøkelsene med 20 prosents oppslutning var bare på 1,6 prosent. I undersøkelsene med 60 prosents oppslutning var den gjennomsnittlige feilen på 5,2 prosent. En lav oppslutning er altså ikke alltid ensbetydende med dårlige data. Det som er avgjørende er om det å ikke delta i undersøkelsen versus det å delta har sammenheng med det en forsøker å måle.

En del forskere har sett nærmere på hvordan en kan oppnå høyest mulig svarprosent. Groves (1989) har undersøkt tre forhold som kan ha konsekvenser for oppslutningen om intervjuundersøkelser:

1. Utsendelse av informasjon på forhånd for å varsle om at en vil bli bedt om å delta i en undersøkelse. Slik informasjon virker imidlertid ikke alltid etter hensikten. Noen ganger går oppslutningen om en undersøkelse ned når det er gitt informasjon på forhånd. En undersøkelse fra England viste at oppslutningen økte med 3 prosentpoeng når en på forhånd sendte ut et velformulert introduksjonsbrev som orienterte om den forestående undersøkelsen (Lynn et al., 1997).

2. Aspekter ved den introduksjonen intervjueren gir når han oppsøker en informant. I en studie av de Leeuw og Hox (2004) ble det funnet at dersom en telefonintervjuer innledningsvis i samtalen sier at ”Jeg skal ikke selge noe”, økte deltakelsen med to prosentpoeng. Dette kan synes lite, men forskerne bak studien mener dette er en setning som er enkel å putte inn i en introduksjon, og at et slikt enkelt håndgrep er en kostnadseffektiv måte å øke oppslutningen på.

3. Bruk av ulike typer belønning. Ferber & Sudman (1974) viste i en undersøkelse at det først og fremst er når informantene må bruke forholdsvis mye tid på en undersøkelse at det er særlig vits i å bruke ytre belønninger som f.eks. penger.

Andre faktorer som er kjent fra litteraturen omfatter det å bruke massemedia for å øke interessen om en undersøkelse hos publikum, bruk av lokale intervjuere, trening av intervjuerne i påvirkningsteknikker, samt bruk av spesialtrenede intervjuere som tar seg av de som ikke vil delta i undersøkelsen.

Det å være svært pågående for å få folk til å delta i en undersøkelse, eller det å bruke bruk av sterke insentiver, kan selvsagt introdusere nye feilkilder. Det kan f.eks. tenkes at de som er minst velvillige til å delta, men likevel deltar, i mindre grad er opptatt av å svare ærlig og presist på spørsmålene. Å utsette noen for press for å få dem til å delta vil selvsagt også være etisk betenkelig.

1.12 Systematiske feilkilder

Ovenfor har vi dvelt en del ved den usikkerhet som knytter seg til at vi ikke undersøker hele populasjoner, men i stedet baserer oss på informasjoner innhentet fra et begrenset utvalg av informanter. Dette er bare en av et stort antall feilkilder som eksisterer i forbindelse med surveys.

En mer fullstendig liste må omfatte følgende punkter:

- Svakheter ved selve måleinstrumentene (lav validitet og reliabilitet)
- Feil som skyldes kontekstuelle faktorer ved utfylling av skjema – eller kontekst og kommunikasjon mellom intervjuer og informant når det er snakk om intervjuer
- Systematiske forskjeller mellom målpopulasjon og surveypopulasjon
- Frafall som skyldes at ikke alle som var trukket ut deltar i undersøkelsen
- Feil som gjøres under koding/innleggelse på datamaskin
- Manglende svar på enkeltspørsmål eller kombinasjoner av spørsmål
- Brudd på forutsetningene for anvendelse av statistikken.

Altfor ofte stirrer en seg blind på p-verdier og konfidensintervall, som sier noe om feilmarginene som oppstår fordi en undersøker et utvalg i stedet for hele populasjonen, og glemmer de andre kildene til feil. Det er helt nødvendig at forskeren ikke taper helheten av syne, men holder klart for seg hele spekteret av feilkilder. Samtidig er det viktig å ikke la alle de mulige feilkildene få lov til å skremme forskeren fra å utnytte de informasjoner som finnes i data. Så lenge en bruker offentlige midler til å gjennomføre forskningsprosjekter, plikter en ikke bare å ta alle mulige forbehold og være selvkritisk til egne funn. Like viktig er det å trekke ut den informasjonen som faktisk finnes i et datasett og sørge for at den blir publisert, gjort tilgjengelig for allmennheten og kan bli til nytte i praktisk samfunnsplanlegging, utforming av forebyggende tiltak, behandling, rehabilitering eller hvor kunnskapen nå enn måtte kunne komme til nytte.

1.13 Kvalitetskontroll av data

Før en setter igang arbeidet med å analysere et datasett, er det helt avgjørende at en først har forsikret seg om at data holder tilstrekkelig kvalitet. Dersom en oppdager dette først etter at en har gjort mange analyser, pådrar en seg et betydelig ekstraarbeid. Analysene må i slike tilfeller gjøres på nytt. Dersom en oppdager at en må gjøre endringer i data etter at en allerede har begynt å publisere resultater, risikerer en inkonsistenser mellom de ulike publikasjonene fra undersøkelsen.

Sjekking av datakvaliteten bør omfatte følgende trinn:

1) Kontroll-lesing av en del cases for å forsikre seg om at datakvaliteten er tilstrekkelig god. I tilfeller der datakvaliteten åpenbart er mangelfull, bør en sørge for re-innleggelse av hele datasettet. Særlig kritisk er datakvaliteten når en arbeider med panelundersøkelser og ønsker å analysere endringer i undergrupper over tid, og/eller når utvalgene er små.

2) Sjekking av alle ulovlige kodeverdier mot originalskjemaene for å korrigere feil som måtte ha oppstått under innleggingen av data.

3) Sjekking av inkonsistente svar. Dersom en person på ett spørsmål har svart at han aldri har prøvd å røyke, men på et annet har oppgitt at han røyker ukentlig, må ett av svarene være feil. I første runde må en gå tilbake til originalskjemaene for å finne ut om inkonsistensen skyldes punchefeil. Dersom svarene i skjemaet faktisk er inkonsistente, er neste trinn å lete etter holdepunkter for hvilket svar som er mest å stole på. Som en siste utvei må en enten etablere regler for hvilket svar en vil legge mest vekt på, eller omkode begge svar til "manglende data". Dersom en etablerer regler for hvilket svar en kan stole mest på, er det nødvendig at dette blir gjort eksplisitt når en rapporterer fra undersøkelsen. Når en gjennomfører parallelle undersøkelser, der en sammenlikner data på tvers av disse (for eksempel undersøkelser som er gjort i forskjellige land), er det viktig at kvalitetskontroll og "retting" (cleaning) av dataene foregår etter nøyaktig de samme prosedyrene og reglene på tvers av surveys.

4) Noen ganger vil en oppdage at kvaliteten på et skjema er så dårlig at vedkommende respondent må tas ut av datasettet. Dette ser en når antall inkonsistenser blir stort eller når manglende svar er høyt. I forbindelse med prosjektet "Health Behaviour among School Aged Children" (Wold & Aarø, 1994) har en etablert en regel om at når mer enn 25% av en del sentrale variabler ikke er besvart, skal vedkommende skjema ekskluderes fra datasettet.

Når en skal gjennomføre dataanalyser der mange variabler inngår samtidig, adderes manglende svar på enkeltvariablene opp til et betydelig antall. I slike tilfeller må en noen ganger redusere frafallet ved å sette inn erstatningstall. Det eksisterer flere prosedyrer for å gjøre dette. Det kan f.eks. settes inn gjennomsnittstall beregnet på grunnlag av alle de enheter som har gyldige verdier på denne variabelen. Andre ganger beregner en ved bruk av multippel lineær regresjon ut fra beslektede variabler hva som burde være verdien på en manglende observasjon, og setter denne inn. Innsetting av verdier for manglende observasjoner må gjøres med stor forsiktighet og grundig overveielse. Det er etter hvert utviklet svært avanserte og gode prosedyrer for beregning av erstatningstall (Switzer & Roth, 2004).

1.14 Konklusjon

Survey-metoden er svært utbredt og blir brukt i svært mange sammenhenger som f.eks. opinionsmålinger, som et pedagogisk instrument i skolen, brukerundersøkelser i næringslivet, større kartlegginger av helse og velferd og som et instrument i teoribasert forskning. Brukt som et redskap i forskningen stiller bruk av survey-metoden store krav til den som skal anvende instrumentet. Det er ikke bare data-analysen som krever innsikt og omtanke. Forberedelsene og gjennomføringen av en survey avgjør kvaliteten på de dataene vi sitter igjen med etterpå.

Det er viktig å …

(1) sikre kvaliteten på måleinstrumenter (spørreskjemaer og intervjuskjemaer),
(2) finne en mest mulig adekvat surveypopulasjon,
(3) beregne minimum utvalgsstørrelse,
(4) sikre representativitet ved trekkingen av utvalget,
(5) gjennomføre undersøkelsen på en standardisert og metodisk forsvarlig måte,
(6) sikre størst mulig oppslutning om undersøkelsen.

Sist, men ikke minst, er det nødvendig å ta utgangspunkt i klare idéer om hva som skal måles. Ofte er begrunnelsen praktisk. Innen forskningen bør en imidlertid kreve teoretisk og begrepsmessig klarhet. Alle spørsmål bør falle inn under et begrep eller en kategori av spørsmål der en har en klar mening om hva en vil måle og hvorfor. Surveys brukt i forskningssammenheng bør ikke ha preg av å være fisketurer med garn, der en sammen med større fisk får opp både småfisk, tang og søppel. Surveys bør handle om målrettet innhenting av spesifikk informasjon. De bør med andre ord ligne mer på fluefiske etter laks og ørret.

Referanser

Abramson, J.H. (1984). *Survey methods in community medicine. An introduction to epidemiological and evaluative studies.* Edinburgh: Churchill Livingstone.

Anastasi, A. (1982). *Psychological testing* (5. utgave). New York: MacMillan.

Aron, A. & Aron, E.N. (1999). *Statistics for psychology* (Second Edition). Upper Saddle River, New Jersey: Prentice Hall.

Bakketeig, L.S. & Magnus, P. (2003). *Epidemiologi*. Oslo: Gyldendal Akademisk.

Baron, R.M. & Kenny, D.A. (1986). The moderator-mediator distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182.

Bethlehem, J. (1999). Cross-sectional research (s.110-142). I Adér,H.J. & Mellenbergh,G.J. (red.). *Research methodology in the social, behavioural and life sciences*. London: Sage.

Blalock, H.M. (1972). *Social statistics*. New York: McGraw-Hill.

Born, M. (1949). *Natural philosophy of cause and chance.* Oxford: Oxford University Press.

Bradburn, N.M. & Sudman, S. (1979). *Improving interview method and questionnaire design.* San Francisco: Jossey-Bass.

Bulmer, M. (1984). *Sociological research methods. An introduction.* London: MacMillan.

Campbell, D. & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *Vol.54*, s.81-105.

Cannell, C.F., Oksenberg, L. & Converse, J.M. (1979). *Experiments in interviewing techniques.* Ann Arbor, Michigan: Institute for Social Research.

Cochran, W.G. (1963). *Sampling techniques.* New York: John Wiley & Sons.

Converse, J.M. & Presser, S. (1986). *Survey questions. Handcrafting the standardized questionnaire.* Newbury Park, California: Sage.

Dalgard, O.S. (1996). Psychiatric interventions for prevention of mental disorders. *International Journal of Technology Assessment in Health Care, 12(4)*, 604-617.

de Leeuw, E.D. & Hox, J.J. (2004). I am not selling anything: 29 experiments in telephone introductions. *International Journal of Public Opinion Research, 16(4)*, 464-473.

Duncan, O.D. (1984). *Notes on social measurement: Historical and critical.* New York: Russell Sage (Sitert etter DeVellis, 1991).

Everitt, B.S. (1996). *Making sense of statistics in psychology. A second level course.* Oxford: Oxford University Press.

Ferber, R. & Sudmann, S. (1974). Effects of compensation in consumer expenditure studies. *Annals of Economic and Social Measurement*, *Vol.3 (No.2)*, 319-331.

Foddy, W. (1993). *Constructing questions for interviews and questionnaires.* Cambridge: Cambridge University Press.

Groves, R.M. (1989). *Survey errors and survey costs.* New York: John Wiley & Sons.

Guilford, J.P. (1965). *Fundamental statistics in psychology and education.* New York: McGraw-Hill.

Hellevik, O. (2005). *Forskningsmetode i sosiologi og statsvitenskap* (7. utgave). Oslo: Universitetsforlaget.

Hernes, G. (1979). Politikk, informasjon og motivering. Noen prinsipielle synspunkter på forebyggende sosialpolitikk. Bergen: NAVF's senter for samfunnsvitenskapelig forskerutdanning, Universitetet i Bergen. Upublisert notat.

Hox, J.J. & de Leeuw, E.D. (2002). The influence of interviewers' attitude and behavior on household survey nonresponse: an international comparison. I Groves, R.M., Dillman, D.A., Eltinge, J.L. & Little, R.J.A. (red.) *Survey Nonresponse* (pp. 103-120). New York: Wiley.

Ilstad, S. (1989). *Survey-metoden. En veiledning i utvalgsundersøkelser.* Trondheim: Tapir forlag.

James, L.R. & Singh, B.H. (1978). An introduction to the logic, assumptions, and the basic analytic procedures of two-stage least squares. *Psychological Bulletin, 85,* 1104-1122.

Jøsendal, O., Aarø, L.E., Torsheim, T., & Rasbash, J. (2005). Evaluation of the school-based smoking prevention programme "BE smokeFREE". *Scandinavian Journal of Psychology, 46 (2)*, 189-199.

Kalton, G. (1983). *Introduction to survey sampling.* Beverly Hills, California: SAGE. (Quantitative Applications in the Social Sciences, nr. 35.)

Kenny, D.A. (1979). *Correlation and causality.* New York: Wiley.

Kerlinger, F.N. (1986). *Foundations of behavioral research.* New York: CBS Publishing.

Kerlinger, F.N. & Lee, H.B. (2000). *Foundations of behavioral research* (Fourth Edition). Fort Worth, Texas: Harcourt College Publishers.

Klockars, A.J. & Yamagishi, M. (1988). The influence of labels and positions in rating scales. *Journal of Eudcational Measurement, 25*, 85-96.

Kraft, P. & Svendsen, T. (1997). Tobacco use among adults in Norway 1973-95: Has the decrease levelled out? *Tobacco Control, 6*, 27-32.

Likert, R.A. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*, 44-53.

Lynn, P., Turner, R., & Smith, P. (1997). The effect of complexity and tone of an advance letter on response to an interview survey. *Survey Methods Centre Newsletter, 17*, 13-17.

Lysgaard, S. (1976). *Arbeiderkollektivet.* Oslo: Universitetsforlaget.

Matell, M.S. & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert Scale items? Study I: Reliability and validity. *Educational and Psychological Measurement, 31*, 657-674.

Molenaar, N.J. (1982). Response-effects of "Formal" characteristics of questions. I W.Dijkstra & J.Van der Zouwen (red.). *Response behaviour and the survey interview.* New York: Academic Press.

Myers, J.H. & Warner, W.G. (1968). Semantic properties of selected evaluation adjectives. *Journal of Marketing Research, 5*, 409-412.

Ommundsen, Y. & Aarø, L.E. (1994). *Folk i form til OL - kampanjen. En evaluering basert på spørreundersøkelser i voksenbefolkningen i 1990 og 1994.* Bergen: HEMIL-senteret, Universitetet i Bergen (HEMIL-rapport nr. 8/94).

Payne, S.L. (1951). *The art of asking questions.* Princeton, New Jersey: Princeton University Press.

Pierce, J.P., Gilpin, E.A., Emery, M.M., White, B.R., Berry, C.C., Farkas, A.J. (2006). Has the California Tobacco Control Program reduced smoking? In Isaacs, S.L. & Knickman, J.R., (red.), *Tobacco Control Policy* (pp. 467-85). San Francisco: Jossey-Bass.

Sherif, M. & Sherif, C.W. (1969). *Social psychology.* New York: Harper and Row.

Schuman, H. & Presser, S. (1981). *Questions and answers in attitude surveys*. San Diego, California: Academic Press.

Schwarz, N. & Clore, G.L. (1983). Mood, misattibution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology, 45*, 513-523.

Shibuya, K., Inoue, M., & Lopez, A.D. (2005). Statistical modeling and projections ofl ung cancer mortality in 4 industrialized countries. *International Journal of Cancer, 117 (3)*, 476-485.

Statens tobakkskaderåd (1999). *Tall om tobakk 1973-98.* Oslo: Statens tobakkskaderåd.

Stephan, F.F. (1948). History of the uses of modern sampling procedures. *Journal of the American Statistical Association*, *Vol.43*, s.12-39.

Stevens, S.S. (1951). Mathematics, measurements and psychophysics. I S.S.Stevens (red.): *Handbook of Experimental Psychology.* New York: John Wiley.

Streiner, D.L. & Norman, G.R. (2003). *Health measurement scales* (3. utgave). Oxford: Oxford University Press.

Strønstad, K., Aarø, L.E., Hetland, J., & Wold, B. (2002). Depressivitet og røyking – en prospektiv panelstudie blant ungdom i Hordaland. *Norsk Epidemiologi, 12 (3)*, 221-230.

Switzer, F.S. III & Roth, P.L. (2004). Coping with missing data. I Rogelberg, S.G. (red.), *Handbook of research methods in industrial and organizational psychology* (s. 310-323). Oxford, England: Blackwell.

Tourangeau, R., Rips, L.J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge: Cambridge University Press.

Visser, P.S., Krosnick, J.A., Marquette, J., & Curtin, M. (1996). Mail surveys for election forecasting? An evaluation of the Columbus Dispatch poll. *Public Opinion Quarterly, 60*, 181-227.

Visser, P.S., Krossnick, J.A., & Lavrakas, P.J. (2000). Survey research. I Reis, H.T. & Judd, C.M. (red.). *Handbook of research methods in social-and personality psychology* (s. 223-252). Cambridge: Cambridge University Press.

Weisberg, H.F. (1993). Central tendency and variability. I M.S. Lewis-Beck (red.), *Basic statistics* (s.1-88). London: Sage.

Wikman, A. & Warneryd, B. (1990). Measurement errors in survey questions: Explaining response variability. *Social Indicators Research, 22*, 199-212.

Wold, B. & Aarø, L.E. (1994). *Health Behaviour in School-Aged Children. A WHO Cross-National Survey (HBSC). Research protocol for the 1993-94 survey.* Bergen: HEMIL-senteret, Universitetet i Bergen (HEMIL-rapport nr. 4/94).

Aarø, L.E., Bjartveit, K. & Vellar, O.D. (1977). Smoking habits among Norwegian doctors 1974. *Scandinavian Journal of Social Medicine*, *Vol.5*, 127-135.

Kap 2: Univariat statistikk

2.1 Beskrivelse av fordelinger

2.1.1 Kategorielle variabler

Det første trinnet i en dataanalyse er alltid å undersøke variablene enkeltvis. Dette er viktig av flere grunner. I kapittel 1 var vi inne på dette med å undersøke at variablene ikke inneholder ulovlige verdier. Dersom en har registrert eller spurt om kjønn, kodes svarene som regel med 1 for mann og 2 for kvinne. Noen ganger ser en at rekkefølgen er omvendt, altså 1 for kvinne og 2 for mann. Når det ikke foreligger informasjon om kjønn, er denne variabelen blank, eller en kan velge å bruke en bestemt tallverdi, f.eks. -9 eller 9 for å markere at her mangler en svar. Dersom det forekommer andre kodeverdier enn de som symboliserer mann, kvnne eller manglende svar, må det være feil. Slike feil må sjekkes og rettes opp før en setter i gang dataanalysene.

Men inspeksjon av enkeltvariabler er viktig også av andre grunner. Dersom en skal regne statistikk på undergrupper, må en forsikre seg om at gruppene er store nok til at det gir mening å foreta slike beregninger. Noen ganger er retningen på en variabel av stor betydning. Dersom en for eksempel skal kombinere flere variabler til en samlet indeks, må alle enkeltvariablene snues i en slik retning at høy tallverdi reflekterer navnet på indeksen. Dersom variablene har ulike retninger, vil en indeks bli meningsløs. Den såkalte parametriske slutnings-statistikken baserer seg på antakelser om at variablene er normalfordelte, eller i det minste at de ikke avviker altfor sterkt fra en normalfordeling. Denne forutsetningen er særlig

kritisk dersom en analyserer på små grupper. En bør derfor analysere variablene en jobber med for å finne ut om fordelingene i tilstrekkelig grad samsvarer med forutsetningene for den statistikken en bruker.

En enveisfordeling kan beskrives på flere måter. En skiller gjerne mellom mål på sentral tendens (den mest typiske eller gjennomsnittlige skåren) og mål på spredning (i hvor stor grad observasjonene sprer seg ut på undergrupper eller på den skalaen en har benyttet). Det finnes mange mål på sentral tendens. Når variablene er kategorielle og på nominalnivå, oppgir en gjerne modus (hvilken kategori som inneholder det største antall respondenter) og modalprosent (hvor mange prosent av repsondentene som havnet i nettopp denne kategorien). Dersom variabelen er på ordinalnivå, er det vanlig å oppgi median. Medianen er den verdien på skalaen som deler subjektene inn i to like store grupper (halvparten i hver gruppe) (Norusis, 1993). Dersom variabelen er metrisk (intervall- eller rationivå), kan en bruke aritmetiske gjennomsnitt, som vil bli gjort rede for noe senere i dette kapittelet.

Vi har tidligere vært inne på at målenivåene er hierarkiske. Det samme er målene for sentral tendens. Modus og modalprosent er de eneste målene for sentraltendens som kan brukes når en variabel er nominell og kategoriell. Når en variabel er på ordinalnvå, kan en også bruke modus og modalprosent (dersom variabelen er kategoriell), men i tillegg kan en bruke median. Når en variabel er metrisk kan en i prinsippet bruke alle disse (modus og modalprosent forutsatt at en har delt skalaen inn i grovere kategorier), men i tillegg også aritmetiske gjennomsnitt.

La oss se på eksempelet gjengitt i Tabell 2.1 (neste side). Tabellen viser rapportert stemmegivning ved siste forutgående valg i et landsrepresentativt utvalg av personer født 1909-1985 som ble intervjuet i 2004[1]. Dette er en typisk nominalvariabel. En kan plassere hver enkelt person som deltar i undersøkelsen i en av kategoriene, men det å ordne kategoriene i en bestemt rekkefølge er ikke enkelt. En kan selvsagt forsøke å bruke en slags sosialistisk-borgerlig-dimensjon, men det er ikke uten videre enkelt å plassere alle partiene etter hverandre på denne dimensjonen. Det er for eksempel vanskelig å vite hvilket av partiene Kristelig Folkeparti og Venstre som skal plasseres nærmest Høyre. Det er også vanskelig å vite hvor en skal plassere Kystpartiet. Dersom det hadde vært enkelt og ukontroversielt å plassere partiene langs en slik dimensjon, kunne en betrakte partitilhørighet som en ordinalvariabel. Men for vårt formål passer det bedre å anta at dette ikke er mulig.

Tabellen viser hvor mange som havner i hver kategori (frekvens), prosenter av alle som har deltatt, prosenter av valide svar, og kumulative prosenter av valide svar. I dette eksempelet er det ikke særlig interessant å se på den siste kolonnen (kumulative prosenter av valide svar). Det mest interessante er nok kolonnen som inneholder prosenter av valide svar. De viser hvor mange som ville stemt på hvert av partiene i prosent av alle som har tatt stilling til spørsmålet. Det betyr at de som ikke vet, ikke vil svare eller har svart at dette ikke er et aktuelt spørsmål for dem (kanskje de ikke stemte ved siste valg) er holdt utenfor.

[1] Datene er hentet fra European Social Survey og stilt til disposisjon av Norsk Samfunnsvitenskapelig Datatjeneste (NSD) http://ess.nsd.uib.no/webview/index.jsp

Tabell 2.1: Stemmegivning ved siste forutgående valg, Norge. Frekvens- og prosentfordeling. (Data fra European Social Survey 2004, stilt til disposisjon av Norsk Samfunnsvitenskapelig Datatjeneste - NSD)

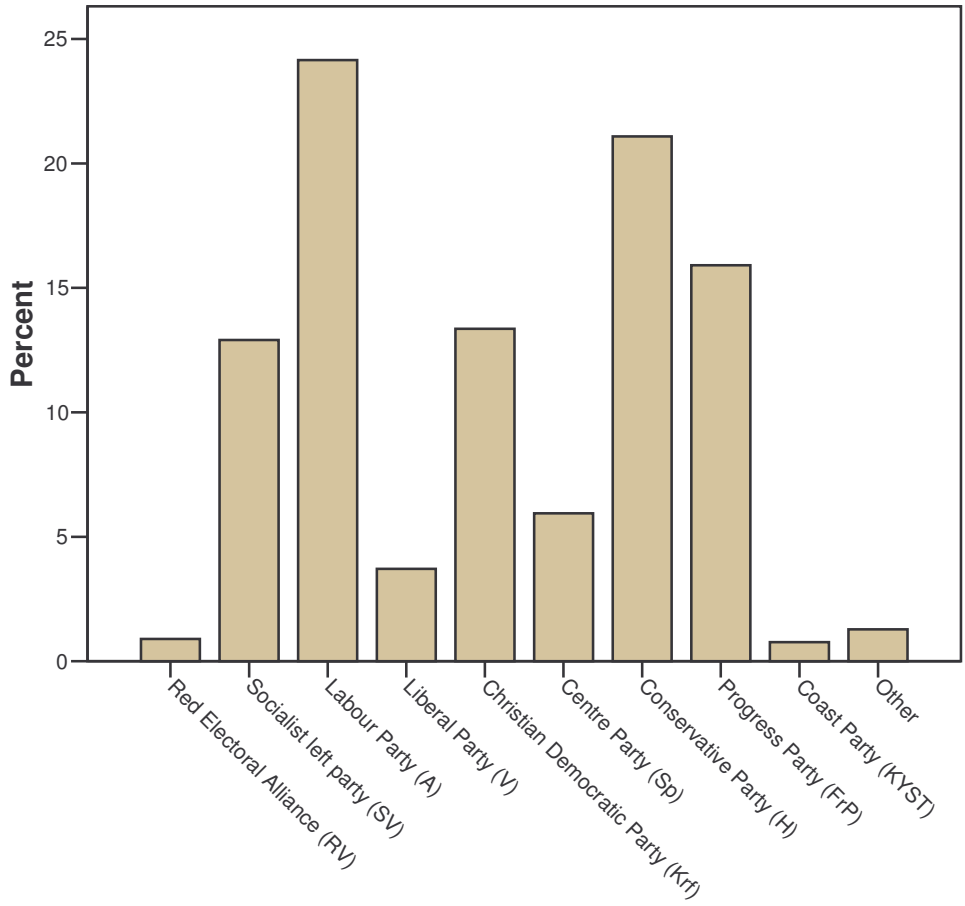
		Frekvens	Prosent	Valid prosent	Kumulativ prosent
Valide svar	1 Rød valgallianse (RV)	14	,7	,9	,9
	2 Sosialistisk venstreparti (SV)	202	9,9	12,9	13,8
	3 Arbeiderpartiet (A)	378	18,6	24,2	38,0
	4 Venstre (V)	58	2,8	3,7	41,7
	5 Kristelig folkeparti (Krf)	209	10,3	13,4	55,0
	6 Senterpartiet (Sp)	93	4,6	5,9	61,0
	7 Høyre (H)	330	16,2	21,1	82,0
	8 Fremskrittspartiet (FrP)	249	12,2	15,9	98,0
	9 Kystpartiet (KYST)	12	,6	,8	98,7
	10 Andre	20	1,0	1,3	100,0
	Total	1565	76,9	100,0	
Manglende svar	66 Ikke aktuelt	388	19,1		
	77 Ville ikke svare	59	2,9		
	88 Vet ikke	24	1,2		
	Total	471	23,1		
Total		2036	100,0		

Denne tabellen er framkommet ved at vi først valgte ut de personene som hadde oppgitt at de var norske. Vi gikk inn på *Data*, og deretter *Select Cases*. Der hentet vi inn navnet på variabelen som symboliserer land (cntry) og satte opp betingelsen *cntry = "NO"*. Grunnen til at vi satte verdien i hermetegn, er at den er en string-variabel, som kan inneholde andre tegn enn tall. Og endelig gikk vi inn på *Analyze*, *Descriptive Statistics*, og deretter *Frequencies*. Vi hentet så inn variabelen *prtvtno* og trykket deretter på *OK*.

Dersom vi skal rapportere et mål for sentral tendens i dette tilfellet, er det kun aktuelt å bruke modus og modalprosent. Modus er i dette tilfellet Arbeiderpartiet, siden det er dette partiet som har oppnådd den største oppslutningen, og modalprosenten (blant alle som har gitt valide svar) er 24,2. Dersom enda flere hadde sagt at de ville stemme på Arbeiderpartiet, ville vi

hatt en sterkere sentraltendens (en høyere modalprosent). Dersom færre hadde stemt Arbeiderpartiet (men Arbeiderpartiet likevel hadde vært større enn det nest største partiet), ville vi hatt en svakere sentraltendens (en lavere modalprosent).

Fig. 2.1: Stemmegivning ved siste forutgående valg, Norge. Stolpediagram. (Data fra European Social Survey 2004, stilt til disposisjon av NSD)

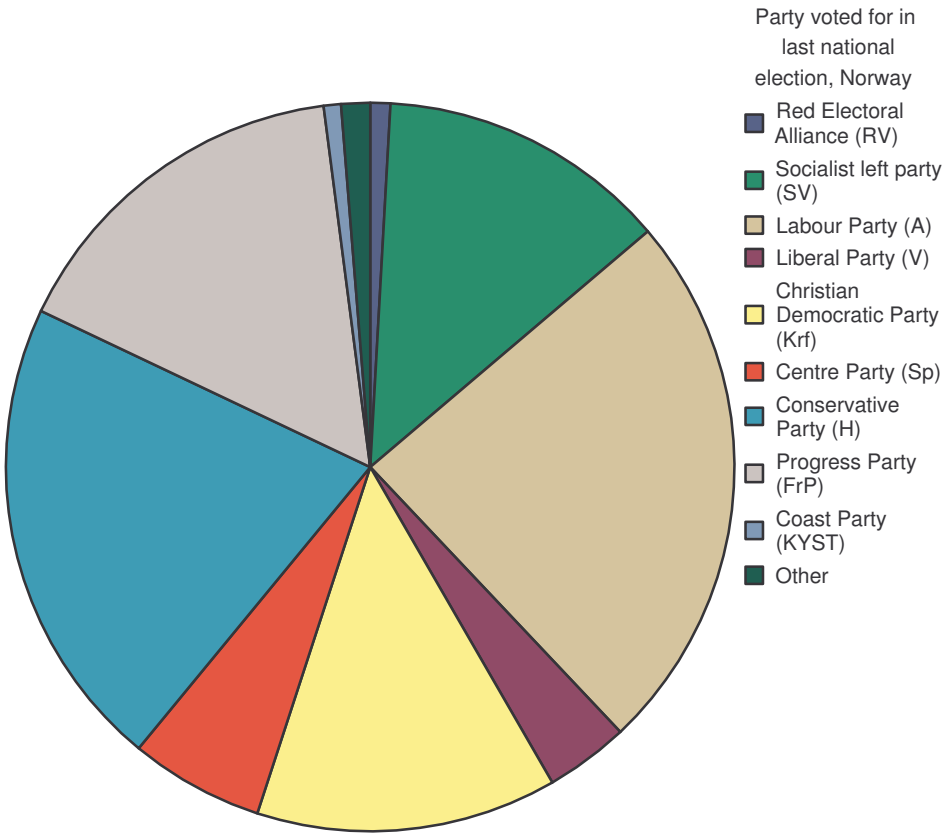


I SPSS bestiller vi et slikt stolpediagram ved å velge følgende fra menyen: *Graphs*, *Bar, Simple* og *Define.* Variabelen (*prtvtno*) legges deretter inn i Category Axis. Så merker en av for *% of cases*, og så trykker en *OK*. Omtrent samme diagram kan også bestilles gjennom å velge *Bar Charts* under *Bar* i *Frequencies*.

Jo flere kategorier svarene sprer seg på, og jo lavere andelen dermed blir i hver kategori, desto større er spredningen av svarene på en kategoriell nominalvariabel. Jo mer svarene hoper seg opp i noen få kategorier, desto mindre er spredningen. Det er likevel ikke vanlig å regne ut noen bestemte koeffisienter for å vise spredningen på kategorielle variabler, men heller å vise konkret hvordan spredningen ser ut ved bruk av stolpediagram (bar charts). Et

slikt stolpediagram er gjengitt i Fig. 2.1. Legg merke til at det er mellomrom mellom stolpene. Det finnes en type diagram der stolpene står tett sammen, uten noen mellomrom. Disse kalles histogram. Histogram gir mening når en viser fordelinger på variabler som er metriske, der det finnes en underliggende, kontinuerlig skala. Men det gir ikke mening å bruke histogram på nominalvariabler.

Fig. 2.2: Stemmegivning ved siste valg, Norge. Kakediagram. (Data fra European Social Survey 2004, stilt til disposisjon av NSD)



Noen foretrekker å beskrive fordelingen på nominalvariabler ved bruk av sektordiagram eller kakediagram. Et slikt diagram er gjengitt i Fig. 2.2. Vi har benyttet samme data som i Tabell 2.1 og Fig. 2.1. Et sektordiagram bestilles i SPSS ved å velge følgende fra menyen: *Graph*, *Pie* og *Define*. Omtrent samme diagram kan også bestilles gjennom å velge *Pie Charts* under *Bar* i *Frequencies*.

2.1.2 Metriske variabler

Det mest logiske hadde nå vært å se på hvordan vi beskriver fordelinger på ordinalvariabler. Dersom ordinalvariablene er kategorielle og med relativt få kategorier, bruker vi imidlertid stort sett samme statistiske størrelser og diagrammer som de vi har beskrevet ovenfor. Når ordinalvariablene har svært mange nivå, kan det være aktuelt å bruke statistiske størrelser som median (for sentral tendens) og kvartilavvik (for spredning). Disse statistiske størrelsene lar seg imidlertid beskrive vel så godt på metriske variabler. Vi skal derfor forenkle framstillingen ved å gjøre nettopp dette.

Dersom vi kan anta at variabelen vi undersøker er en metrisk variabel, er det flere andre statistiske størrelser som kan benyttes for å beskrive fordelingen. Artimetisk gjennomsnitt (mean) er et mål for sentral tendens. Standardavvik er et mål for spredning.

For å kunne regne ut aritmetiske gjennomsnitt og standardavvik på en variabel er det nødvendig at vi har å gjøre med en variabel som er målt på intervallnivå eller rationivå (metriske variabler). Det kan for eksempel dreie seg om et spørsmål stilt til dagligrøykere om hvor mange sigaretter de røyker per dag. La oss tenke oss at vi har spurt 20 personer. Svarene deres fordeler seg slik:

Antall sigaretter:
12 08 11 19 14 13 12 07 05 10 13 16 18 09 27 17 08 22 21 03

Disse svarene kan plasseres inn i et histogram. Dette gjør vi gjerne ved først å gruppere svarene. Vi velger følgende gruppering: 0-4, 5-9, 10-14, 15-19, 20-24 og 25-29. Dersom vi grupperer svarene slik, får vi følgende fordeling:

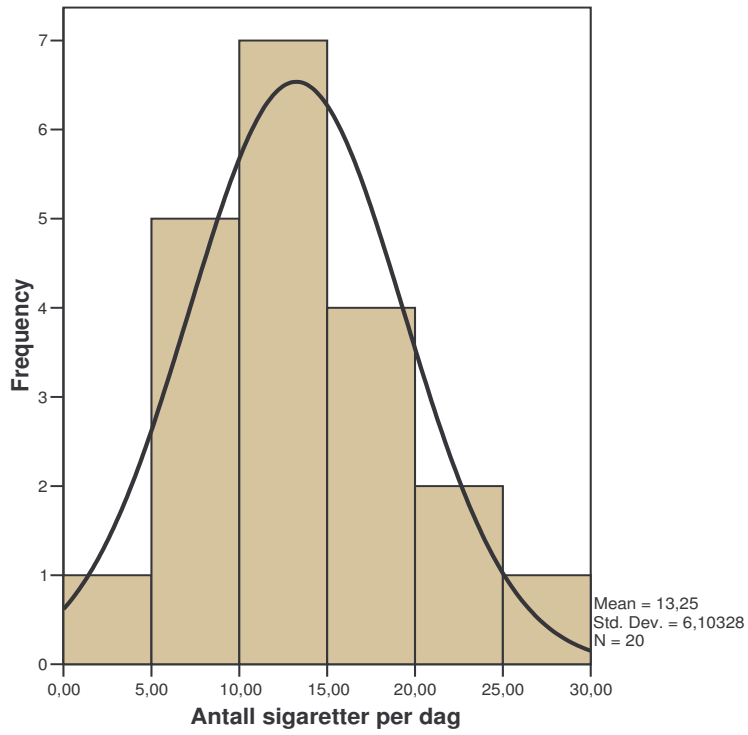
00-04: 1 personer
05-09: 5 personer
10-14: 7 personer
15-19: 4 personer
20-24: 2 personer
25-29: 1 person

Grunnen til at vi grupperer observasjonene sammen på denne måten, er at vi dermed kan tegne et histogram som på en elegant måte viser hvordan observasjonene fordeler seg. Dersom vi ikke grupperer observasjonene, vil vi få et altfor stort antall verdier på den horisontale aksen, og på enkelte verdier vil vi få null observasjoner, mens vi på andre får en eller to. Et diagram som viser en slik fordeling vil gi lite mening. Derfor deler vi skalaen inn i intervall og grupperer observasjonene. Plassert inn i et stolpediagram vil denne fordelingen se ut slik som vist nedenfor i Fig. 2.3.

Vi ser at det er forholdsvis mange som havner i de midterste kategoriene, mens det er ganske få (bare en) som rapporterer at de røyker svært lite (0-4 sigaretter per dag) eller svært mye (25 sigaretter eller flere per dag). Dette er ikke uvanlig når en skal måle mennesker på en

eller annen egenskap. En sier gjerne at fordelingen er tilnærmet normalfordelt. En slik normalfordelingskurve har en helt bestemt form, og kan beskrives ved hjelp av en matematisk formel. Den foreliggende fordelingen er ikke helt nøyaktig lik en normalfordeling. Når en har såpass få personer med i fordelingen som i dette eksempelet (n=20), vil fordelingen vanligvis (tilfeldigvis) kunne avvike en hel del fra normalfordelingen, selv om fordelingen slik den ser ut i universet er normalfordelt. Som vi skal se senere i denne teksten, viser det seg forresten at antall sigaretter dagligrøykere pleier å røyke per dag ikke er nøyaktig normalfordelt i den norske befolkningen, men fordelingen er likevel nokså lik en normalfordeling.

Fig. 2.3: Histogram over antall sigaretter per dag blant dagligrøykere. Hypotetisk eksempel.



For å få fram dette histogrammet i SPSS, legger vi først inn alle tallene (antall sigaretter slik de er listet opp på forrige side) i SPSS sitt regneark (*Data View*). Tallene legges etter hverandre vertikalt, med ett tall i hver celle. Deretter går vi inn i *Variable View* og setter navn på vatiabelen, f.eks. *v1*. Det kan også være greit å lage en forklaring (*Label*) på variabelen. Vi kan for eksempel beskrive den som *Antall sigaretter per dag*. Deretter trykker vi på *Graphs* og *Histogram* og henter inn variabelen i boksen til høyre. Og til slutt trykker vi på *OK*. Vi ser at SPSS automatisk grupperer svarene slik vi har foreslått ovenfor.

For å beskrive en slik fordeling pleier en som regel (som allerede nevnt) å regne ut to statistiske størrelser: det aritmetiske gjennomsnittet og standardavviket. Formelen for et aritmetisk gjennomsnitt ser slik ut:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (2.1)$$

$\bar{X}$ Det aritmetiske gjennomsnittet av skårene på variabelen *x*

x_i Skåren til enhet (f.eks. person) nr *i*

n Antall enheter

Dette betyr ganske enkelt at vi legger sammen alle de 20 tallene som viser hvor mange sigaretter de vanligvis røyker per dag og deler på antallet observasjoner (altså n=20). I vårt tilfelle blir summen av observasjonene 265 sigaretter. Dersom vi deler på 20 får vi 13,25. Resultatet av utregningen er med andre ord at de gjennomsnittlig røyker 13,25 sigaretter pr. dag. Dette kaller vi altså et aritmetisk gjennomsnitt. Et aritmetisk gjennomsnitt som er regnet ut på grunnlag av tall fra et rent tilfeldig utvalg personer er et forventningsrett estimat av det aritmetiske gjennomsnittet i populasjonen.

Det finnes flere andre mål for sentral tendens som også kan benyttes her, for eksempel modus, som vi har beskrevet tidligere (det antallet sigaretter som flest røyker per dag) eller median som tilsvarer det antallet som deler fordelingen i to slik at 50% faller på venstre side og 50% faller på høyre side[2]. Modus i fordelingen ovenfor er 10-14 sigaretter per dag, siden det er denne kategorien som har flest observasjoner. Siden 50% har oppgitt at de røyker 12 eller færre sigaretter per dag og 50% har oppgitt at de røyker 13 eller flere, plasserer vi medianen midt imellom 12 og 13. Medianen er med andre ord 12,5.

Tidligere har vi understreket at dikotome variabler har bestemte interessante og nyttige egenskaper. En av disse egenskapene er at det gir mening å regne ut det aritmetiske gjennomsnittet av en slik variabel. Først må vi da omkode de to verdiene på variabelen til tallene 0 og 1. Tallverdien 1 bør stå for det kjennetegnet vi er spesielt interessert i, mens 0 står for det å ikke ha dette kjennetegnet[3]. Når vi da legger sammen verdiene på alle enhetene og deler på antall enheter (n), får vi ut proporsjonen som har denne bestemte egenskapen. Ganger vi med hundre får vi prosent som har denne egenskapen. For dikotome variabler

---

[2] Weisberg (1993) presenterer enda flere mål for sentral tendens, blant annet trimmete gjennomsnitt, geometriske gjennomsnitt og harmoniske gjennomsnitt.

[3] Dersom vi tar en kategoriell variabel og lager en slik koding for hver verdi (0 for det å ikke tilhøre den bestemte kategorien og 1 for det å tilhøre kategorien), kaller vi hver slik variabel for en dummy-variabel og slik koding kalles dummy-koding. For å representere en flerkategoriell variabel ved bruk av dummy-koder behøver vi et antall dummy-variabler som er lik antall kategorier på den opprinnelige variabelen minus én.

fungerer med andre ord utregning av gjennomsnitt som en måte å beregne proporsjoner eller prosenter på. Dette viser at prosenter kan betraktes som en type aritmetiske gjennomsnitt for dikotomier.

For å beskrive en fordeling trenger vi også et mål for spredning. Igjen finnes det flere å velge mellom. Her skal vi beskrive et mål som kalles standardavvik. Formelen for å beregne standardavviket i et utvalg ser slik ut:

$$SD_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n}} \qquad (2.2)$$

SD_x Standardavviket til variabelen x

$\bar{X}$ Det aritmetiske gjennomsnittet til variabelen x

x_i Skåren til enhet nr. i

n Antall enheter

Vi skal igjen bruke eksempelet med antall sigaretter røykt per dag. Først trekker tar vi hver enkelt skår og trekker fra gjennomsnittet. Denne differansen kvadrerer vi. For de fem første observasjonene blir tallene slik:

$(12 - 13{,}25)^2 = \ \ 1{,}5625$
$(08 - 13{,}25)^2 = 27{,}5625$
$(11 - 13{,}25)^2 = \ \ 5{,}0625$
$(19 - 13{,}25)^2 = 33{,}0625$
$(14 - 13{,}25)^2 = \ \ 0{,}5625$

Dersom vi gjør dette for hele rekka på 20 tall og legger sammen alle tallene vi får, blir resultatet 707,75. Legg merke til at både de verdiene som er større enn det aritmetiske gjennomsnittet og de som er mindre blir positive tall før de legges sammen. Dette fordi kvadratet av et negativt tall blir et positivt tall. Summen av alle de kvadrerte differansene kalles kvadratsummen. Denne deler vi med antall observasjoner (altså 20) og får 35,3875. Dette tallet kalles variansen til fordelingen eller variansen på den aktuelle variabelen. Ved å regne ut kvadratroten av dette tallet får vi standardavviket i utvalget. Resultatet blir 5,9487. Ved å runde av til to desimaler etter kommaet får vi 5,95. Vi har med andre ord regnet ut at standardavviket til fordelingen er på 5,95 sigaretter.

Ovenfor har vi sett at det aritmetiske gjennomsnittet på en dummy-kodet variabel kan fortolkes som en proporsjon (eller en prosent når proporsjonen ganges med hundre). Også varians og standardavvik kan beregnes på slike dummyvariabler. Dersom sannsynligheten for å ha et bestemt kjennetegn er p, mens sannsynligheten for ikke å ha dette kjennetegnet følgelig er 1 – p, kaller vi følgende uttrykk for variabelens varians:

$$SD^2 = p(1-p) \qquad (2.3)$$

SD^2 Varians (standardavviket kvadrert)

p Proporsjonen som har et bestemt kjennetegn

Variabelens standardavvik er kvadratroten av variansen, med andre ord SD (standard deviation) (Blalock, 1972). Det er verdt å legge merke til at en dikotomi har størst varians når fordelingen mellom de som har kjennetegnet og de som ikke har det ligger nær 50-50. Når fordelingen er nøyaktig 50 – 50 blir variansen 0,25. Dersom fordelingen er 90-10 blir variansen mye lavere, nemlig 0,09.

Ovenfor viste vi hvordan vi regnet ut variansen og standardavviket på en intervallvariabel (daglig sigarettforbruk blant dagligrøykere) for et utvalg på 20 personer. Det er imidlertid sjelden vi er interessert i å beskrive bare det utvalget av personer vi har trukket. Som regel vil vi helst komme fram til en beskrivelse som gjelder hele populasjonen. Her støter vi på et lite problem. Den formelen vi har benyttet gir oss ikke det som i statistikken kalles et forventningsrett estimat av standardavviket i populasjonen. Et forventningsrett estimat har vi dersom vi benytter en formel som med utgangspunkt i de data vi har for utvalget gir et riktigst mulig tall når vi skal forsøke å si noe om egenskaper ved populasjonen. For å regne ut et slikt forventningsrett estimat må vi gjøre en liten endring i formelen vi bruker. Den blir da seende ut slik som i formel 2.4 nedenfor.

$$SD_{x(POP)} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}} \qquad (2.4)$$

$SD_{x(POP)}$ Estimat av standardavviket til variabelen x i populasjonen

$\bar{X}$ Det aritmetiske gjennomsnittet til variabelen x i utvalget

x_i Skåren til enhet nr. i i utvalget

n Antall enheter i utvalget

Vi ser at nevneren i brøken ikke lenger bare er n, men er blitt til n-1. Uttrykket n-1 sier vi er antall frihetsgrader. I denne sammenhengen skal vi ikke gå nærmere inn på begrepet frihetsgrader, men bare slå fast at ved å dele på antall frihetsgrader i stedet for antall observasjoner i utvalget, får vi et forventningsrett estimat av standardavviket i populasjonen eller universet. Anvendt på eksempelet ovenfor finner vi da at vi deler kvadratsummen (707,75) med 19 (20 minus 1) og får 37,25. Dette tallet er et estimat av variansen på denne variabelen i populasjonen. Ved å regne ut kvadratroten av dette tallet får vi et estimat av

standardavviket i populasjonen, som viser seg å være lik 6,10. Estimatet av standardavviket i populasjonen er altså litt større enn standardavviket i utvalget. Når utvalgene er store, betyr det i praksis svært lite om vi deler på antall observasjoner (n) eller om vi deler på antall frihetsgrader (n-1).

Tabell 2.2: Beskrivende statistikk på en metrisk variabel fra SPSS

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
Antall sigaretter per dag	20	3,00	27,00	13,2500	6,10328	37,250
Valid N (listwise)	20					

Tabell 2.2 viser en utskrift som er hentet ut fra SPSS og der datene er de samme som for Fig. 2.3. Analysen er bestilt fra menyen ved å gå inn på følgende stikkord: *Analyse*, *Descriptive statistics*, *Descriptives*, og *Options*. Options gir muligheter for å bestille noe mer statistikk enn det som er standard, for eksempel spisshet (kurtosis) og skjevhet (skewness). Skevheten er i dette tilfellet 0,465 og spissheten er -0,090. Fordelingen er med andre ord litt høyreskjev (halen mot høyre), men temmelig lik en normalfordelingskurve når det gjelder spisshet. Legg merke til at SPSS har regnet ut standardavviket til å være 6,10. Det betyr at SPSS automatisk antar at tallene våre stammer fra et utvalg som rent tilfeldig er trukket fra en større (i prinsippet uendelig stor) populasjon. SPSS antar at vi er interessert i å estimere standardavviket i denne populasjonen.

Spisshet (kurtosis) og skjevhet (skewness) måler avvik fra normalfordelingen. Den såkalte parametriske statistikken, det vil si statistikk som baserer seg på analyser der variablene (i det minste den avhengige variabelen) forutsettes å være målt på intervallnivå (Sheskin, 1997), bygger på bestemte forutsetninger[4]. En av disse forutsetningene er at den avhengige variabelen skal være normalfordelt. Det er derfor svært nyttig å kjenne til de to statistiske størrelsene kurtosis og skjevhet, samt å kunne teste om en variabel avviker signifikant fra å være normalfordelt. En perfekt normalfordelt variabel har en spisshet på 0,0 og en skjevhet på 0,0. Dersom spissheten er større enn 0,0 (altså positiv), er fordelingen spissere (med en høyere topp på midten) enn en normalfordeling. Dersom spissheten er under 0,0 (altså negativ), er fordelingen flatere enn en normalfordeling. Dersom skjevheten er større enn 0,0 (positiv), betyr det at tyngdepunktet av observasjonene hoper seg opp mot venstre i fordelingen, mens vi mot høyre (høyere verdier) vil observere en lengre hale enn den vi ser

[4] Den parametriske statistikken bygger på tre forutsetninger: (1) At variablene er målt på intervallnivå, (2) at variablene er normalfordelte, og (3) at variansen er homogen på tvers av subgrupper eller over alle nivå på den variabelen en analyserer mot. Ofte, men ikke alltid, må også en fjerde forutsetning være oppfylt, nemlig at det er uavhengighet mellom observasjonene på tvers av enheter eller subjekter.

mot venstre. Vi sier gjerne at en slik fordeling er høyreskjev. Dersom variabelen er venstreskjev, vil tyngdepunktet ligge mot høyre og halen mot venstre.

Som allerede beskrevet ovenfor, defineres medianen i en fordeling som den verdien som deler subjektene i to grupper. Dersom vi har spurt røykere om hvor mange sigaretter de røyker pr. dag, og det viser seg at akkurat halvparten røyker 12 sigaretter eller mindre per dag mens den andre halvparten røyker 13 sigaretter eller mer, sier vi at medianen ligger mellom 12 og 13. Det er i et slikt tilfelle konvensjon å si at medianen er 12,5 sigaretter per dag. En høyreskjev fordeling kan defineres som en fordeling der det aritmetiske gjennomsnittet er høyere enn medianen. En venstreskev fordeling vil ha et aritmetisk gjennomsnitt som ligger lavere enn medianen. Dersom aritmetisk gjennomsnitt og median er akkurat like, betyr det at skjevheten (skewness) er nøyaktig 0,0, og fordelingen er verken venstreskjev eller høyreskjev. Som regel vil en slik fordeling være nokså symmetrisk. Vi kan likevel ikke uten videre si at den er helt symmetrisk, for det kan foreligge avvik fra symmetri som ikke fanges opp av formelen for skjevhet.

I eksempelet vi har brukt ovenfor, var medianen 12,5 og det aritmetiske gjennomsnittet 13,25. Det aritmetiske gjennomsnittet er altså litt høyere en medianen. Vi kan dermed fastslå at fordelingen er litt høyreskev. Dette kan vi slå fast også ved å studere fig. 2.3, som har mer hale mot høyre enn mot venstre, eller ved å se på tallet for skjevhet som er regnet ut av SPSS (0,465).

Formelen for å regne ut skjevheten til en variabel er ganske enkel (Lewis-Beck, 1993, s.21):

$$Skjevhet = \frac{\sum \frac{(x_i - \bar{X})^3}{SD_x}}{n} \qquad (2.5)$$

x_i Skåren til observasjon nr. *i*

$\bar{X}$ Det aritmetiske gjennomsnittet til variabelen *x*

SD_x Standardavviket til variabelen *x*

n Antall observasjoner

2.2 Sannsynlighet og sannsynlighetsfordelinger

Til nå har vi nøyd oss med å bruke statistikken rent beskrivende. I det første eksempelet beskrev vi politisk stemmegivning blant et representativt utvalg på 2036 personer trukket fra hele den voksne norske befolkningen. Det andre eksempelet dreiet seg om å beskrive forbruket av sigaretter per dag blant 20 dagligrøykere, som vi antar også stammet fra et representativt utvalg av befolkningen. De statistiske størrelsene vi har regnet ut er alle sammen eksempler på beskrivende (eller deskriptiv) statistikk. Vi skal etter hvert se på det

som kalles slutningsstatistikk (eller analytisk statistikk) anvendt på enveis-fordelinger. Men først er det nødvendig å si noe om sannsynlighet og sannsynlighetsfordelinger.

2.2.1 Enkel sannsynlighetsregning

Et sentralt begrep i slutningsstatistikken er "sannsynlighet". Vi skal ikke her gi noen grundig og fullstendig redegjørelse for sannsynlighetsteori, men det er nødvendig å si litt om dette temaet.

De fleste lærebøker som tar for seg temaet sannsynlighet starter med å beskrive hva som skjer når en kaster mynt og krone. Når en kaster en mynt opp i luften slik at den snurrer raskt rundt, er det ikke lett å spå om hvilken side som viser opp når den lander. Det er to mulige utfall, nemlig mynt og kron. I teorien kan en kanskje tenke seg at mynten blir stående på høykant, men det skjer så sjelden at vi i prinsippet kan se helt bort fra denne tredje utfallsmuligheten.
Dersom vi kaster mange ganger, vil vi se at den noen ganger havner med myntsiden opp, andre ganger med kronsiden opp. Dersom mynten er rund, av ensartet metall og med mønstre som ikke skaper noen vesentlig forskjell i vektfordeling på de to sidene, vil vi i det lange løp (etter mange nok forsøk) finne at omtrent 50% av utfallene er mynt og 50% er kron. Jo flere ganger vi forsøker, desto nærmere vil vi komme en 50-50-fordeling (prosentvis like mange av begge utfall). Vi sier da at sannsynligheten for det ene utfallet (mynt) er 0,5. Og sannsynligheten for det andre utfallet (kron) er også 0,5. Til sammen blir dette 1,0. Sannsynligheten for at det enten skal bli mynt eller kron er med andre ord 1,0 (eller 100%).

Et annet eksempel som stadig går igjen handler om det å kaste terning. Terninger har normalt seks sider nummererte fra 1 til 6. Dersom det er en "rettferdig" terning, vil vil finne at i omtrent 1/6 av tilfellene får vi en ener. I 1/6 av tilfellene får vi en toer etc. Slik blir det i hvert fall dersom vi kaster mange nok ganger. Vi sier da at sannsynligheten for utfallet 1 er 1/6, noe som tilsvarer en sannynlighet på p=0,167 (avrundet ved 3. desimal). Tallet 0,167 er altså en p-verdi. Bokstaven p står for probabilitet (probability). Sannsynligheter uttrykkes gjerne med en gerenell formel som ser slik ut:

$$p(g) = \frac{\text{Antall utfall av typen g}}{\text{Antall mulige utfall}} \qquad (2.6)$$

$p(g)$ - sannsynligheten for et utfall av typen g

For at en skal kunne regne sannsynligheter slik som vist her, må de ulike utfallene være gjensidig utelukkende. Og det stemmer da også godt med eksemplene ovenfor. Når en kaster mynt og kron, kan en ikke få begge deler samtidig. Ett bestemt kast gir enten det ene eller det

andre. De to utfallene er gjensidig utelukkende. Det samme gjelder kast med terning. En kan ikke både få en toer og en treer samtidig når en kaster bare en terning. Når alle mulige utfall er gjensidig utelukkende, er summen av sannsynlighetene for alle utfall alltid lik 1,0. Når det bare finnes to mulige utfall, er det dessuten slik at sannsynligheten for det ene utfallet alltid er lik 1,0 minus sannsynligheten for det andre. Dersom andelen gutter som blir født i en befolkning er lik 51,8%, så sier vi at sannsynligheten for at en tilfeldig valgt fødsel skal være en gutteføds el $p_{gutt} = 0{,}518$. Sannsynligheten for jenteføds el er da $p_{jente} = 1{,}0 - p_{gutt} = 1{,}000 - 0{,}518 = 0{,}482$.

Det at sannsynligheten vi regner ut på grunnlag av et bestemt eksperiment (for eksempel når vi kaster mynt og kron eller når vi kaster en terning) nærmer seg et bestemt tall (0,5 eller 0,167) kalles de store talls lov. Når en kaster en mynt bare noen få ganger, kan en lett få p-verdier som avviker en hel del fra 0,5, men etter hvert som en øker antallet eksperiment (kast), vil p-verdien som en regner ut nærme seg 0,5 mer og mer (men med stadig mindre, tilfeldige avvik).

Vi så ovenfor at når en har to utfall, kan de adderes, og blir til sammen 1,0. Også når en har flere enn to utfall kan sannsynligheter adderes. Sannsynligheten for å få enten en ener eller en toer når en kaster en rettferdig terning er 1/6 + 1/6 =2/6 eller p = 0,333. Slik addisjon benyttes når en skal regne ut summen av flere mulige utfall.

Noen ganger bruker en multiplikasjon. Det vil for eksempel være tilfelle dersom en kaster mynt to ganger etter hverandre og skal regne ut sannsynligheten for å få samme utfall to ganger. Dersom sannsynligheten for å få mynt begge ganger er 0,5, er sannsynligheten for at en skal få mynt begge gangene 0,5 * 0,5 = 0,25. Sannsynligheten for å få 0,5 seks ganger på rad kan en regne ut ved å gange 0,5 med seg selv slik: $0{,}5^6 = 0{,}5 * 0{,}5 * 0{,}5 * 0{,}5 * 0{,}5 * 0{,}5 = 0{,}016$. Sannsynligheten for å få mynt seks ganger på rad er med andre ord p = 0,016. Eller regnet om til prosent er sannsynligheten for å få mynt seks ganger på rad lik 1,6%.

2.2.2 Sannsynlighetsfordelinger

Så langt har vi bare interessert oss for sannsynligheten for ett bestemt utfall, for flere mulige utfall og for det å få samme utfall flere ganger på rad. Men dersom vi for eksempel kaster en mynt mange ganger, kan vi systematisere alle de mulige utfallene for alle gangene vi kaster samlet. Vi kan for eksempel kaste mynt og kron fem ganger på rad og så stille spørsmålet om hva sannsynligheten er for å få mynt null ganger, en gang, to ganger, tre ganger, fire ganger eller fem ganger.

Dette kan regnes ut ved hjelp av enkel kombinatorikk. Vi bruker følgende formel:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \qquad (2.7)$$

n - antall mulige utfall

k - antall utfall av en bestemt type

Utropstegnet heter i matematikkspråket fakultet. Fakultet av tallet fire vil si at en ganger tallene 1 * 2 * 3 * 4, med andre ord tallrekken fra 1 til tallet selv. Fakultet av null er definert som 1. Dersom en har en litt avansert kalkulator, eller stiller kalkulatoren som følger med Windows på "vitenskapelig", kan en eksperimentere og se hva fakultet av ulike tall blir. Da vil en for eksempel se at 4! = 24 og at 7! = 5040.

Formelen for sannsynligheten for k mynt av n kast er vist nedenfor (ligning 2.8).

$$p \text{ (k bestemte utfall på n forsøk)} = \binom{n}{k} p^k q^{n-k} \qquad (2.8)$$

p - sannsynligheten for et bestemt utfall (f.eks. mynt)

q - sannsynligheten for annet utfall (1-p)

n - antall forsøk

k - antall ganger det bestemte utfallet (i dette eksempelet mynt)

Johann Friedrich Carl Gauss var født i Brunswick, Tyskland i 1777. Allerede tidlig i barneskolen så lærerne at han var utrolig begavet. Han oppdaget for eksempel på egen hånd at han kunne summere alle tall fra 1 til 100 ved å gange 101 (summen av to og to tall) med 50.

Gauss er kjent for å ha beskrevet binomialfordelingen, normalfordelingen, samt aritmetiske og geometriske gjennomsnitt. Han engasjerte seg i forskning på tallteori, differensiallikninger og hypergeometriske funksjoner. Men han interesserte seg også for tema utenom matematikken og statistikken, for eksempel forskning på jordmagnetisme og astronomi. Han tilbrakte det meste av sitt yrkesaktive liv ved Universitetet i Göttingen.

Ett av de sitatene Gauss er kjent for, illustrerer kanskje hans grunnleggende holdning til vitenskap: "When a philosopher says something that is true, then it is trivial. When he says something that is not trivial, then it is false."

http://www-groups.dcs.st-and.ac.uk/~history/Mathematicians/Gauss.html

Ved å sette tallene inn i disse ligningene, kan vi lett regne ut følgende rekke med sannsynligheter:

0,03125 (sannsynligheten for å få null ganger mynt i fem kast)
0,15625 (sannsynligheten for å få en gang mynt i fem kast)
0,31250 .
0,31250 .
0,15625 .
0,03125 (sannsynligheten for å få fem ganger mynt i fem kast)

Summen av disse tallene blir nøyaktig 1,00000.

Dette er et eksempel på en svært enkel sannsynlighetsfordeling. Nærmere bestemt er dette en binomisk fordeling. En kan regne ut slike fordelinger for et hvilket som helst antall forsøk, og sannsynlighetene for de to mulige utfallene behøver ikke være 0,5 og 0,5 slik som her. Sannsynligheten for ett utfall kan være hvilket som helst tall mellom 0,0 og 1,0, og sannsynligheten for det andre utfallet blir da 1,0 minus sannsynligheten for det første.

Johann Friedrich Carl Gauss (1777-1855)

Vi har flere ganger tidligere i denne teksten nevnt noe som kalles en normalfordelingskurve. Denne fordelingen ble første gang beskrevet av den tyske matematikeren Johann Gauss. Den blir derfor ofte kalt gausskurven. Vi nevnte at mange egenskaper i befolkningen er tilnærmet normalfordelte. Det gjelder for eksempel folks høyde, ulike karaktertrekk, og det gjelder tilnærmet også for antall sigaretter dagligrøykere røyker per dag. Men det viser seg at det er mye annet som er normalfordelt. Det kan blant annet vises matematisk at etter hvert som en øker antall forsøk, vil binomialfordelingen ligne mer og mer på en normalfordelingskurve. Når p = 0,5, og dermed også q = 0,5, vil normalfordelingen kunne brukes som en god tilnærming til binomialfordelingen allerede ved 10 forsøk.

Det er ikke bare binomialfordelingen som nærmer seg normalfordelingen når en øker antall forsøk eller observasjoner. En lang rekke typer statistiske størrelser fordeler seg slik. Dette er svært gunstig, for det gir oss et nyttig redskap i forbindelse med slutnings-statistikken. Nå må vi straks føye til at det finnes mange typer fordelinger som kan brukes til samme formål. Vi skal etter hvert støte på t-fordelingen, F-fordelingen og χ^2-fordelingen (kji-kvadrat-

fordelingen). Men grunnprinsippene når en skal bruke disse sannsynlighetsfordelingene til å trekke slutninger fra utvalg til populasjon er hele veien de samme.

2.3 Samplingfordelinger og standardfeil

Vi har nå sett litt på begrepet sannsynlighet og vi har lært litt om sannsynlighetsfordelinger. Før vi kan vise hvordan en bruker slutningsstatistikk, er det dessuten nødvendig å kjenne til begrepene samplingfordeling (sampling distribution) og standardfeil (standard error). Den første som brukte begrepet standardfeil (standard error) var George Udny Yule i 1897. Det å forstå hva en samplingfordeling er, er helt grunnleggende dersom en skal forstå noe av slutningsstatistikken, altså den delen av statistikken som handler om å generalisere fra utvalg til univers. Mohr (1993) skriver om dette:

> The idea of the sampling distribution is fundamental to an understanding of classical inference. It is the keystone of the whole process; without clarity in regard to the sampling distribution, only a muddy and confused notion of significance testing and confidence intervals is possible (s.95) [mine understrekninger].

Mohr (1993) har dessuten presentert en svært enkel og pedagogisk god beskrivelse av hva en samplingfordeling er, og hvordan den kan brukes til å regne ut statistisk signifikans og konfidensintervall.

Før vi går nærmere inn på beskrivelsen av standardfeil, konfidensintervall og signifikanstesting, er det viktig å gjøre rede for litt mer av begrepsbruken i forbindelse med statistikk som regnes på populasjoner og utvalg. Når vi beregner en statistisk størrelse som er ment å beskrive egenskaper ved variabler (prosenter, aritmetiske gjennomsnitt, standardavvik eller lignende) eller sammenhenger mellom variabler basert på data fra en hel populasjon, kaller vi denne størrelsen for en parameter[5]. Når en beregner tilsvarende tall for ut utvalg, bruker en betegnelsen statistiske størrelser (statistics). Når vi forsøker å beskrive populasjonen på bakgrunn av data fra et representativt utvalg, sier vi at vi estimerer, og de statistiske størrelsene er estimater. Et forventningsrett estimat er en statistisk størrelse (beregnet på et utvalg) som i det lange løp gir den aller riktigste antakelsen om parameteren en forsøker å estimere. En statistiker ville sannsynligvis formulert denne siste setningen på en litt annen måte, men for våre formål skulle dette gi en tilstrekkelig god forståelse.

Både beregning av konfidensintervall og signifikanstesting er eksempler på slutningsstatistikk (eller analytisk statistikk). Dette er former for statistikk som benyttes når vi med utgangspunkt i et utvalg skal forsøke å si noe om populasjonen. Dersom vi for eksempel skal regne ut konfidensintervallet til et aritmetisk gjennomsnitt, regner vi først ut gjennomsnittet for det utvalget vi har undersøkt. Deretter beregner vi et intervall rundt dette gjennomsnittet. Intervallet angir det området vi med en viss sikkerhet (for eksempel 95%) kan si at den tilsvarende parameteren (det aritmetiske gjennomsnittet) for populasjonen befinner seg.

---

[5] Ordet parameter blir også brukt om konstanter som inngår i likninger som beskriver sannsynlighetsfordelinger. Sir Ronald Fisher brukte begrepet på denne måten første gang i 1932.

GEORGE UDNY YULE

George Udny Yule (1871-1951), født på en gard i Skottland, introduserte begrepet standardfeil i 1897. Etter studier i England og Tyskland arbeidet han en tid som assistent for Carl Pearsons. Mens Pearson var mest opptatt av å anvende statistisk teori innen biologien, var Yule mer opptatt av samfunnsforskning og epidemiologi. De ble senere bitre motstandere, blant annet fordi Pearson mislikte en koeffisient til beskrivelse av sammenhenger i 2x2-tabeller som Yule hadde introdusert. Yule's viktigste arbeider var innen områdene korrelasjon (blant annet spuriøsitet), regresjon, tidsserier, mendelske arvelover, epidemiologi og forskning om statistiske framgangsmåter for å vurdere virkningene av vaksiner. Yule's Introduction to the Theory of Statistics (1910) fikk stor utbredelse og innflytelse, ikke minst etter at den ble revidert av Maurice Kendall i 1937.

http://www.morris.umn.edu/~sungurea/introstat/history/w98/Yule.html
http://www.economics.soton.ac.uk/staff/aldrich/Figures.htm#yu

Signifikanstesting skjer på en litt annen måte. Der starter vi ut med en antagelse om parameteren. Deretter formulerer vi det som kalles nullhypotesen. Deretter beregner vi den statistiske størrelsen basert på utvalget, og så tester vi hvor sannsynlig det er å få et slikt resultat dersom nullhypotesen var riktig. Dersom avviket mellom nullhypotesen og det resultatet vi har fått er tilstrekkelig stort og antall observasjoner i utvalget tilstrekkelig høyt, kan vi forkaste hypotesen. La oss for eksempel si at vi har gjennomført en undersøkelse blant ungdom, og i undersøkelsen inngår en skala som er ment å måle depressivitet. Med bakgrunn i tidligere undersøkelser antar vi at jenter skårer noe høyere på depressivitet enn gutter. Dette er vår hypotese. For å gjennomføre en signifikanstesting, må vi imidlertid lansere en nullhypotese. I dette tilfellet er nullhypotesen at det ikke er noen forskjell mellom gutter og jenter i gjennomsnittlig skår på depressivitet. Dersom forskjellen mellom gutter og jenter er tilstrekkelig stor, og dersom antallet gutter og jenter som er med i undersøkelsen er

tilstrekkelig stort, vil vi kunne forkaste nullhypotesen. Vi har ikke dermed bevist den alternative hypotesen (altså at det er en forskjell på gutter og jenter i populasjonen), men vi har funnet at det er liten sannsynlighet (for eksempel mindre enn 5 % sannsynlighet) for at nullhypotesen er korrekt.

Uansett om vi beregner konfidensintervall eller tester hypoteser, er samplingfordelingen og standardfeilen viktige statistiske begreper å kjenne til. La oss derfor se nærmere på disse.

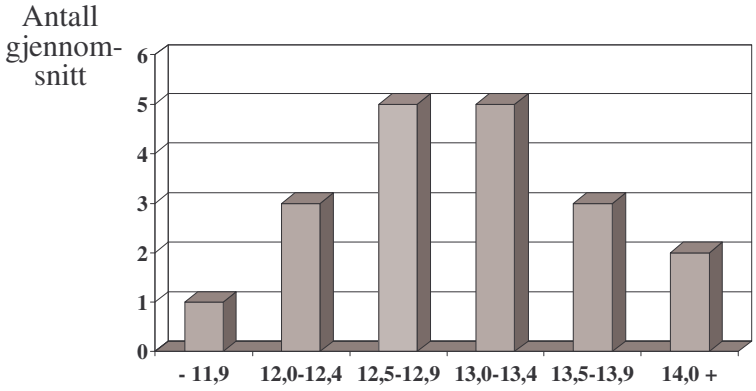
Når vi på grunnlag av et utvalg trukket fra en større populasjon regner ut en statistisk størrelse, f.eks. prosent som har en bestemt egenskap, vil resultatet være beheftet med en viss usikkerhet. La oss si at egenskapen vi vil måle er andel som bekrefter at de vil stemme ja til norsk medlemskap i EU. Utvalget vi har trukket er på 500 personer, og andel som svarer bekreftende er 43%. For å få et mål på hvor usikkert dette tallet er, tenker vi oss at vi trekker et nytt utvalg på samme måte og med samme utvalgsstørrelse for å se om dette avviker fra det første. Andelen som bekrefter at de vil stemme ja i dette andre utvalget kan for eksempel vise seg å være 39%. Slik kan vi fortsette med stadig nye utvalg, og hver gang regner vi ut prosenten som vil stemme ja. Etter hvert vil det framkomme et mønster. Andel som vil stemme ja varierer hele tiden, men hoper seg opp rundt en gjennomsnittsverdi. Det aritmetiske gjennomsnittet i denne fordelingen vil med økende antall nye utvalg (i samsvar med de store talls lov) ha en tendens til å nærme seg den tilsvarende tallverdien for hele populasjonen (parameteren). Standardavviket til den fordelingen som framkommer er den størrelsen vi kaller standardfeil. Selve fordelingen kalles en samplingfordeling (Kalton, 1983). Standardfeilen er med andre ord standardavviket i samplingfordelingen.

Dersom vi gjør samme forsøket på nytt, men med en lavere utvalgsstørrelse, blir standardfeilen større. Dersom vi øker utvalgsstørrelsen, blir standardfeilen mindre. Standardfeilen er et uttrykk for hvor presist vi kan uttale oss om en parameter (som altså gjelder hele populasjonen) på basis av en statistisk størrelse (som er beregnet med utgangspunkt i et utvalg).

Ovenfor har vi brukt standardfeilen til en proporsjon eller andel (prosent) som eksempel. Vi kunne på tilsvarende måte regnet ut standardfeilen til hvilken som helst statistisk størrelse som beskriver et utvalg, for eksempel et aritmetisk gjennomsnitt. La oss på nytt bruke eksempelet med antall sigaretter røykt daglig blant dagligrøykere. Vi trekker et utvalg på 20 personer og regner ut gjennomsnittlig antall sigaretter de røyker per dag. Det første tallet vi får er 13.25. Så trekker vi et nytt utvalg på 20, og denne ganger blir resultatet 12,9. Slik fortsetter vi inntil vi har fått 20 slike gjennomsnitt. Vi vil hele tiden se at tallene blir litt forskjellige, men at de har en tendens til å samle seg rundt et bestemt punkt.

La oss tenke oss at vi ikke begrenset oss til å gjøre dette 20 ganger slik som i dette eksempelet, men i stedet flere tusen ganger. Hele tiden måtte utvalget ha samme størrelse, nemlig 20 personer. Hvis vi grupperte disse flere tusen observasjonene av aritmetiske gjennomsnitt og laget et tilsvarende histogram, ville vi få det vi ovenfor har definert som en samplingfordeling. Siden antall observasjoner var svært stort, kunne vi ha svært små intervall, og i stedet for stolper kunne vi trekke linjer mellom topp-punktene på stolpene. Resultatet ville bli en fordeling som praktisk talt er identisk med en normalfordelingskurve.

Fig. 2.4: Gjennomsnittlig antall sigaretter i 20 utvalg dagligrøykere med 100 personer i hvert. Frekvensfordeling (hypotetisk eksempel)



Gjennomsnittlig antall sigaretter pr. dag (gruppert)

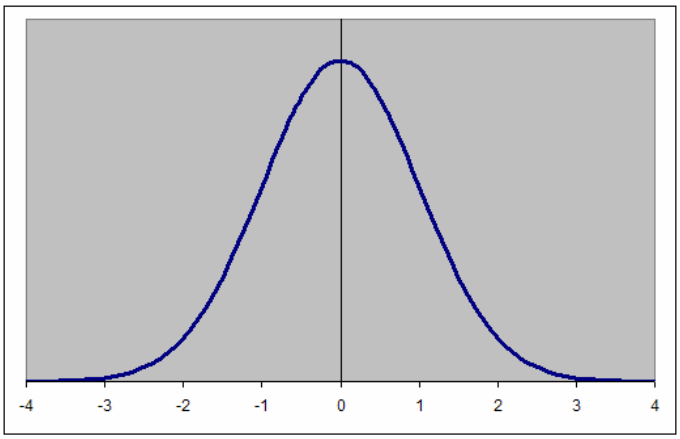
Matematikerne har nemlig vist at samplingfordelingen til et aritmetisk gjennomsnitt er tilnærmet normalfordelt. Dette gjelder i hvert fall for variabler som er omtrent normalfordelte, slik vi antar at sigarettforbruket blant dagligrøykerne er. Matematikerne har dessuten vist at selv om fordelingen på den variabelen vi interesserer oss for ikke er normalfordelt, er samplingfordelingen tilnærmet normalfordelt når bare utvalgene blir store nok.

Dette kalles sentralgrenseteoremet. Konklusjonen er altså at når utvalgene ikke er for små, kan vi anta at samplingfordelingen til et aritmetisk gjennomsnitt er tilnærmet normalfordelt. Dette er svært praktisk, for normalfordelingskurven har en del kjente egenskaper som vi kan gjøre oss nytte av. Det er dessuten ikke bare samplingfordelingen til det aritmetiske gjennomsnittet som er normalfordelt. Det finnes en lang rekke samplingfordelinger som er tilnærmet normalfordelte. La oss derfor se litt nærmere på normalfordelingskurven (Fig. 2.5).

Normalfordelingskurven er klokkeformet og symmetrisk om gjennomsnittet. Den er laget slik at den har et gjennomsnitt på 0,0 og et standardavvik på 1,0. Det er vanlig å prosentuere arealet i fordelingen. Hele arealet under kurven er på 100 prosent. Det er selvsagt slik at 50 prosent av arealet ligger til venstre for gjennomsnittet og 50 prosent til høyre. Videre vet vi at 68,26 prosent av arealet ligger innenfor pluss/minus ett standardavvik. Punktene langs den horisontale aksen (x-aksen) i et diagram som viser normalfordelingen angis som regel i form av antall standardavvik. Dette kalles vanligvis for en z-skår. Dersom vi slår opp i en statistisk tabell som viser sammenhengen mellom hvor vi er på x-aksen og arealet til venstre (eller til høyre) for dette punktet, kan vi for eksempel finne ut at 2,5% av arealet ligger til høyre for z-verdien 1,96 og 2,5% av arealet ligger til venstre for z-verdien -1,96. Dette betyr at 5% av arealet i en normalfordeling ligger utenfor en avstand på pluss eller minus 1,96 fra

midtpunktet (gjennomsnittet). Videre kan vi finne ut at det er 1% av arealet som faller utenfor z-verdien pluss eller minus 2,58. Endelig kan vi se at en promille av observasjonene faller utenfor z-verdien pluss eller minus 3,30. Det er nettopp disse tallverdiene vi gjør bruk av når vi bruker normalfordelingskurven som et hjelpemiddel til å regne ut konfidensintervall og sannsynligheter i forbindelse med signifikanstesting.

Fig. 2.5: Normalfordelingskurven



Ovenfor har vi vist hvordan en kan konstruere en samplingfordeling ved å trekke tusenvis av utvalg av en bestemt størrelse og hver gang estimere en bestemt parameter[6]. Heldigvis finnes det enklere måter å beregne standardfeil på enn ved å trekke et stort antall utvalg av en viss størrelse. Vi kan regne ut standardfeilen ut fra antagelser om den parameteren vi interesserer oss for. Vi kan også anslå størrelsen på standardfeilen ved å bruke den informasjonen som finnes i bare ett enkelt utvalg. Det er dette siste en som regel gjør når en regner ut konfidensintervall eller gjennomfører en signifikanstesting.

2.4 Konfidensintervall og signifikanstesting av enkeltvariabler

2.4.1 Konfidensintervall for proporsjoner

Dersom en har trukket et rent tilfeldig utvalg på 500 personer fra hele den voksne, norske befolkningen, fått alle disse til å svare på et spørreskjema, og regnet ut at det er 30% som har bestemt seg for å stemme ja til norsk medlemskap i EU, vet vi at denne statistiske størrelsen er et forventningsrett estimat av tilsvarende parameter for hele befolkningen. Dersom vi ønsker å regne ut hvor stor usikkerheten er, hvor stort intervall vi må ta med for å ha en 95% sannsynlighet for at parameteren faller innenfor, bruker vi en svært enkel framgangsmåte.

---

[6] På internett finnes det flere hjemmesider der en kan eksperimentere med samplingfordelinger for å se hva som skjer når en endrer formen på den fordelingen en sampler fra eller øker utvalgsstørrelsen. Bruk en nettleser og søk på "sampling distributions", så vil du finne noen slike.

Først regner vi ut den såkalte standardfeilen til denne prosenten (SE_p) ved hjelp av følgende formel:

$$SE_p = \sqrt{\frac{p(100-p)}{n-1}} \qquad (2.9)$$

SE_p Standardfeilen (standard error) til prosenten p

p Prosenten som har et bestemt kjennetegn

n Antall enheter

Ved å sette inn tallene fra eksempelet ovenfor kan vi regne ut standardfeilen.

$$SE_p = \sqrt{\frac{30 * (100 - 30)}{500 - 1}} = 2{,}05$$

SE_p Standardfeilen til prosent som vil stemme ja

Prosenten som sier de vil stemme ja: 30

Antall personer som har svart: 500

Fra statistisk teori vet vi at vi kan bruke normalfordelingen som en god tilnærming når vi skal regne ut konfidensintervallet til p[7]. I en vanlig tabell over normalfordelingskurven (z-verdier) vil vi finne at for å regne ut et 95% konfidensintervall må vi multiplisere standardfeilen med 1,96. Dermed kan vi regne ut at konfidensintervallet dekker området 30% ± (1,96 * 2,05%) = 26%-34%. Vi har altså regnet ut at det er 95% sannsynlighet for at prosent som har bestemt seg for å stemme ja til EU i hele den voksne befolkningen ligger et sted mellom 26% og 34%.

Så enkelt kan en altså beregne konfidensintervallet til et prosenttall dersom en har benyttet rent tilfeldig trekking. Eksempelet viser hvordan vi kan benytte normalfordelingskurven som et hjelpemiddel til å regne ut et konfidensintervall.

Så langt har vi ikke tatt hensyn til at populasjonene som utvalgene trekkes fra varierer i størrelse. Når populasjonen er så stor som hele den voksne, norske befolkningen, spiller heller ikke dette noen praktisk rolle. En kan like gjerne anta at populasjonen er uendelig stor, og dermed unngå det kompliserende elementet som ligger i å ta hensyn til størrelsen på populasjonen.

[7] Den eksakte fordelingen som skal brukes er binomialfordelingen. Dette er bare nødvendig når tallene er små, eller mer presist, når p*n eller (1-p)*n er mindre enn 10. Når den minste av disse to verdiene er 10 eller større, er det forsvarlig å benytte normalfordelingskurven som tilnærming (Guilford, 1965).

Tabell 2.3: Beskrivende statistikk (bl.a. konfidensintervall) for en dikotom variabel.

Descriptives

			Statistic	Std. Error
v1 For norsk EU-medlemskap?	Mean		,3000	,02051
	95% Confidence Interval for Mean	Lower Bound	,2597	
		Upper Bound	,3403	
	5% Trimmed Mean		,2778	
	Median		,0000	
	Variance		,210	
	Std. Deviation		,45872	
	Minimum		,00	
	Maximum		1,00	
	Range		1,00	
	Interquartile Range		1,00	
	Skewness		,876	,109
	Kurtosis		-1,238	,218

Merkelig nok gir ikke SPSS noen muligheter til å få ut konfidensintervall for prosenter. Men ved å kode tallene som dummy-variabler (mot norsk medlemskap i EU = 0; for norsk medlemskap = 1), kan en regne ut konfidensintervallet til sannsynligheten. En kan for eksempel gå inn i *Analyze*, *Descriptive Statistics*, *Explore* og *Statistics* og markere for *Descriptives*. Her kan en også bestemme hva slags konfidensintervall en vil ha, f.eks. 95% eller 99%. I Tabell 2.3 ser vi at mean = 0,30. Dette tilsvarer at 30% sa de ville stemme ja. Deretter ser vi at konfidenstintervallet (95%) går fra 0,2597 til 0,3403. I prosenter blir dette 25,97 – 34,03, eller nokså nøyaktig 26% - 34%, som er det samme som vi har regnet ut for hånd. Legg merke til at også standardfeilen er med på utskriften. Den er beregnet til 0,02051, eller om lag 2,05%, slik vi også har regnet ut for hånd.

$$SE_p = \sqrt{\frac{p(100-p)}{n-1}(1-\frac{n}{N})} \qquad (2.10)$$

SE_p Standardfeilen (standard error) til prosenten p

p Prosenten som har et bestemt kjennetegn

n Antall enheter

N Antall personer i populasjonen (universet)

Når populasjonene er små, kan en vinne presisjon ved å ta hensyn til størrelsen. Dersom vårt utvalg på 500 personer som tok stilling til norsk medlemskap i EU hadde vært trukket fra et begrenset univers (på for eksempel 1000 personer), ville vi kunne bruke formel 2.10 for å beregne standardfeilen på et prosentestimat.

Uttrykket inne i den siste parentesen $(1-\frac{n}{N})$ kalles som tidligere nevnt korreksjon for at vi har en endelig populasjon (finite population correction). Vi ser at jo større universet er i forhold til utvalget, desto mer nærmer uttrykket seg 1,0, og desto mer vil standardfeilen nærme seg den standardfeilen vi får ved å bruke formelen uten en slik korreksjon. Videre ser vi at jo større del populasjonen vi tar med i utvalget, desto mer vil korreksjonen nærme seg 0,0. Når vi tar med hele populasjonen i utvalget, blir uttrykket lik 0,0. Når dette ganges med det uttrykket som står foran, blir resultatet en standardfeil på null. Dette virker logisk og rimelig. Dersom en har med hele populasjonen i utvalget, og en er interessert i å trekke slutninger fra utvalg til populasjon, kan en, når hele populasjonen inngår i utvalget, gjøre dette med null feil, altså med perfekt presisjon.

I psykologisk forskning er vi ikke alltid interessert i å trekke slutninger til bestemte populasjoner, men ønsker kanskje å finne fram til sammenhenger som har noe mer allmenn gyldighet. Derfor vil det som regel være mest meningsfylt å bruke den vanlige slutningsstatistikken, som forutsetter at en har en uendelig stor populasjon, selv om slike populasjoner alltid bare er hypotetiske.

Dersom vi tar tallene i eksempelet ovenfor og setter inn i formelen får vi følgende:

$$SE_p = \sqrt{\frac{30(100-30)}{500-1}\left(1-\frac{500}{1000}\right)} = 1{,}45$$

SE Standardfeilen til prosenten som sier de vil stemme ja

Prosenten som sier de vil stemme ja: 30

Antall personer i utvalget: 500

Antall personer i universet (populasjonen): 1000

Et 95% konfidensintervall ville i dette tilfellet blitt 30% ± 2,8% = 27,2% - 32,8%. Mens konfidensintervallet under forutsetning av en uendelig stor populasjon dekker et område på 8 prosentpoeng, dekker konfidensintervallet i dette siste tilfellet 5,6%. Og som nevnt ovenfor; jo mer størrelsen på utvalget nærmer seg størrelsen på populasjonen, desto mindre blir konfidensintervallet, og blir til slutt 30% ± 0,0%).

2.4.2 Konfidensintervall for aritmetiske gjennomsnitt

Tidligere har vi sett hvordan en kan regne ut gjennomsnittet og standardavviket på en metrisk variabel (for eksempel antall sigaretter dagligrøykere pleier å røyke per dag). På samme måte

som for prosenter, er vi også her interessert i å beregne et konfidensintervall for gjennomsnittet. For å greie dette må vi kunne beregne standardfeilen til gjennomsnittet.

Det gjør vi ved hjelp av følgende formel:

$$SE_{\bar{X}} = \sqrt{\frac{SD_X^{\,2}}{n}} \qquad (2.11)$$

$SE_{\bar{X}}$ Standardfeilen til det aritmetiske gjennomsnittet til variabelen x

SD_X Det estimerte standardavviket til variabelen x i populasjonen

n Antall enheter

Legg merke til at det standardavviket som skal brukes er det estimerte standardavviket til variabelen x i populasjonen. Dersom vi i stedet tar utgangspunkt i standardavviket i utvalget (der vi har delt med antall observasjoner i stedet for med antall frihetsgrader), ville formelen se slik ut:

$$SE_{\bar{X}} = \sqrt{\frac{SD_X^{\,2}}{n-1}} \qquad (2.12)$$

$SE_{\bar{X}}$ Standardfeilen til det aritmetiske gjennomsnittet til variabelen x

SD_X Standardavviket til variabelen x i utvalget (ikke populasjonsestimat)

n Antall enheter

Vi ser at dette uttrykket er slående likt formelen for standardfeilen på et prosentestimat.

Dersom vi igjen benytter oss av eksempelet med antall sigaretter og bruker formel 2.11, får vi at

$$SE_{\bar{X}} = \sqrt{\frac{6{,}10^2}{20}} = 1{,}36$$

$SE_{\bar{X}}$ Standardfeilen til det aritmetiske gjennomsnittet til variabelen x

Det estimerte standardavviket til variabelen x i populasjonen: 6,10

Antall enheter: 20

Standardfeilen er altså på 1,36 sigaretter. Dersom vi ønsker å regne ut et 95% konfidensintervall, gjør vi det omtrent slik som vist for prosenttall ovenfor. Men fordi samplingfordelingen til et artimetisk gjennomsnitt er t-fordelt og ikke normalfordelt, må vi undersøke hvilken t-verdi som skal brukes når vi regner ut et 95% konfidensintervall. Vi husker at for normalfordelingen var den kritiske verdien for et 95% konfidensintervall 1,96. Når en skal slå opp i tabellene over t-fordelinger finner vi at det eksisterer en t-fordeling for hvert antall frihetsgrader. I vårt tilfelle var antall frihetsgrader 19 (n-1). Den kritiske verdien for et 95% konfidensintervall på en t-fordeling med 19 frihetsgrader tilsvarer tallet 2,09. Vi multipliserer standardfeilen (1,36) med den kritiske verdien (2,09). Konfidensintervallet er lik gjennomsnittet for utvalget pluss/minus tallet vi da har fått. Anvendt på eksempelet med røykevaner får vi at 1,36 multiplisert med 2,09 blir 2,84. Konfidensintervallet blir med andre ord 13,25 sigaretter +/- 2,84 sigaretter, hvilket blir 10,41 – 16,09. Vi har med andre ord beregnet at det er 95 prosent sikkert at det gjennomsnittlige antall sigaretter røykt per dag i universet (populasjonen) ligger et sted mellom 10,41 og 16,09 sigaretter per dag. Dette konfidensintervallet dekker et ganske stort område. Dette kommer av at utvalget tross alt er ganske lite, bare 20 personer.

Når antall frihetsgrader vokser, nærmer t-fordelingene mer og mer z-fordelingen. Når antall frihetsgrader passerer 30, er det for enkelhets skyld like greit å bruke z-fordelingen i stedet for t-fordelingen.

Tabell 2.4: Deskriptiv statistikk over sigarettforbruk (sigaretter per dag) blant dagligrøykere (Tallene for konfidensintervall som vi har regnet ut ovenfor blir ikke nøyaktig de samme som i tabellen. Dette skyldes avrundingsfeil)

Descriptives

			Statistic	Std. Error
v1 Antall sigaretter per dag	Mean		13,2500	1,36473
	95% Confidence Interval for Mean	Lower Bound	10,3936	
		Upper Bound	16,1064	
	5% Trimmed Mean		13,0556	
	Median		12,5000	
	Variance		37,250	
	Std. Deviation		6,10328	
	Minimum		3,00	
	Maximum		27,00	
	Range		24,00	
	Interquartile Range		9,50	
	Skewness		,465	,512
	Kurtosis		-,090	,992

Ved å øke størrelsen på utvalget vil vi redusere størrelsen på konfidensintervallet. Vårt estimat blir med andre ord mer presist når vi øker utvalgsstørrelsen. Dersom vi i eksempelet ovenfor øker antall observasjoner til 100 (med nøyaktig samme gjennomsnitt og standardavvik på fordelingen), synker standardfeilen fra 1,36 til 0,60. Dersom vi øker n til 500, synker standardfeilen til 0,27. Med økende antall observasjoner får vi altså et stadig mer presist estimat av gjennomsnittsforbruket av sigaretter blant dagligrøykere i hele befolkningen. Med et utvalg på 20 kunne vi med 95% sikkerhet slå fast at det virkelige tallet (for befolkningen som helhet) lå mellom 10,41 og 16,09 (konfidensintervallet). Med et utvalg på 100 ville konfidensintervallet omfatte området fra 12,07 til 14,43. Med et utvalg på 500 vil konfidensintervallet omfatte området fra til 12,72 til 13,78. Når utregningene ovenfor ikke stemmer helt med utskriften fra SPSS, skyldes dette, som allerede nevnt, avrundingsfeil.

Ovenfor var vi inne på at det er forskjell på å generalisere statistisk til en uendelig stor populasjon versus det å generalisere statistisk til en begrenset populasjon. Usikkerheten minsker med minskende størrelse på populasjonen. Og vi viste hvordan dette slår ut når en skal beregne konfidensintervallet til en proporsjon.

$$SE_x = \sqrt{\frac{SDx^2}{n}(1-\frac{n}{N})} \qquad (2.13)$$

SE_x Standardfeilen til det aritmetiske gjennomstnittet av variabelen x

SD_x Estimert standardavviket til variabelen x i populasjonen

n Antall enheter i utvalget

N Antall enheter i populasjonen (universet)

Helt tilsvarende regner vi dersom det handler om konfidensintervallet til det aritmetiske gjennomsnittet på en intervallvariabel beregnet for et utvalg fra en endelig populasjon. Korreksjonen for at vi har en endelig populasjon er, som vi ser av formelen nedenfor, nøyaktig den samme som det vi benyttet for prosenter (se formel 2.10).

Vi ser altså at presisjonen, når vi skal estimere parametre i populasjonen på bakgrunn av et utvalg, er avhengig av størrelsen på populasjonen utvalget er trukket fra. Enda viktigere enn størrelsen på populasjonen er størrelsen på utvalget. Forbausende ofte ser vi at størrelsen på et utvalg fastsettes uten noen form for beregninger av hvor stor grad av presisjon en gjerne vil oppnå, og dette til tross for at det er en enkel sak å foreta slike beregninger.

2.4.3 Hypotesetesting

Ovenfor har vi konsentrert oss mest om å beskrive hvordan samplingfordelinger kan brukes til å beregne konfidensintervall. Beregning av konfidensintervall angir hvor stor usikkerhet

det er omkring et estimat når vi med utgangspunkt i et utvalg skal si noe om populasjonen. En beslektet, men likevel noe annerledes form for slutningsstatistikk er hypotesetesting. Når en skal teste en hypotese, handler det også om forholdet mellom utvalg og populasjon. Populasjonen kan være endelig, som når en survey gjennomføres på et representativt utvalg trukket fra en bestemt befolkning, eller en kan anta at den er uendelig stor. Uansett er hypotesetestingen basert på at en generaliserer til en populasjon som er uendelig stor. Hypotesen handler om hva en tror om populasjonen.

Hypoteser blir, som vi allerede har nevnt tidligere, presentert parvis. Den ene kalles nullhypotesen og den andre kalles den alternative hypotesen. Nullhypotesen postulerer som regel at det ikke er forskjeller eller sammenhenger. For eksempel at det ikke eksisterer noen forskjell mellom to eller flere gjennomsnitt (eller prosenter) eller at det ikke er noen sammenheng mellom to variabler. Den alternative hypotesen sier det motsatte, med andre ord at det finnes forskjeller eller sammenhenger: at det er en forskjell mellom gjennomsnitt eller at det er en sammenheng mellom to variabler.[8]

La oss bruke forskjeller mellom to prosenter som eksempel. Når en skal gjennomføre en hypotesetesting, tenker en seg at nullhypotesen er riktig, altså at der i populasjonen ikke er noen forskjell mellom prosent som har et bestemt kjennetegn i to undergrupper, for eksempel blant menn og kvinner. Dersom en har funnet ut at det er en viss forskjell mellom menn og kvinner i utvalget, er spørsmålet om denne forskjellen er så stor at en kan forkaste nullhypotesen, altså antagelsen om at det i populasjonen ikke er noen forskjell mellom menn og kvinner. Før en gjennomfører en slik hypotesetesting, skal en bestemme seg for et signifikansnivå. Det vil si at en skal bestemme hvor stor sannsynligheten skal være for at en rent tilfeldig oppnår en så stor forskjell i utvalget gitt at det ikke eksisterer en slik forskjell i populasjonen. Det har utviklet seg en tradisjon for at en sier noe er signifikant dersom denne sannsynligheten er mindre enn 0,05. Men i tillegg pleier en å rapportere om p er mindre enn 0,01 eller 0,001. En kaller disse tre signifikansnivåene for 5-prosent nivået, 1-prosent nivået og 0,1-prosent nivået.

Når en leser forskningsrapporter og artikler der det er brukt slutningsstatistikk, ser en at det ofte spesifiseres om en har brukt enhalet eller tohalet test. La oss si at vi har benyttet en skala til måling av depressive tendenser i en landsrepresentativ undersøkelse blant ungdom i en bestemt aldersgruppe. Vi interesserer oss særlig for kjønnforskjeller, og vi vil teste om det er en signifikant forskjell mellom gutter og jenter når det gjelder deres gjennomsnittlige skår på skalaen. I denne situasjonen kan vi velge mellom to alternative måter å teste på. Nullhypotesen er i begge tilfeller at det ikke er noen forskjell mellom gutter og jenter når det gjelder gjennomsnittlig depresjonsskår. Dersom det er snakk om å bruke en enhalet test, må vi formulere den alternative hypotesen slik at den har en bestemt retning. Vi kan for eksempel på grunnlag av tidligere undersøkelser trekke den konklusjon at dersom det eksisterer noen forskjell, må den gå i retning av at jentene skårer høyere enn guttene. Da blir dette den alternative hypotesen. Men dersom tidligere forskning ikke med sikkerhet kan si noe om hvilken retning en eventuell forskjell har, at det like gjerne kan være guttene som

[8] I den multivariate statistikken kan en ha hypoteser om sammenhenger mellom mer enn to variabler samtidig. Dette er for eksempel tilfelle i multippel regresjonsanalyse der en kan teste sammenhengen mellom et sett prediktorvariabler og en kriterievariabel.

jentene i populasjonen som har høyest skår, må vi bruke en tohalet test. Fordelen med å bruke en enhalet test er at det er lettere å påvise signifikante forskjeller. Men bruken av enhalete tester kan lett kritiseres. En kan bli beskyldt for å ha benyttet enhalet test nettopp for å oppnå signifikans. En garderer seg mot slik kritikk ved å bruke tohalete tester.

Hypotesetesting av sammenhenger mellom variabler (herunder forskjeller mellom grupper) skal vi komme tilbake til i kapittel 3. Her skal vi se nærmere på hvordan vi kan utføre hypotesetesting på enkeltvariabler.

2.4.4 Signifikanstesting av enveisfordeling med bare to kategorier (celler)

Som vi har beskrevet tidligere i dette kapittelet, kan en kategoriell variabel beskrives best ved å prosentuere. En slik prosentfordeling er gjengitt i Tabell 2.5. Dataene stammer fra en undersøkelse av norske legers røykevaner som ble gjennomført i 1974. På det tidspunktet var det sterk debatt i media og blant politikere om innføringen av en lov der en blant annet ville forby tobakksreklame. De som var for forbudet mot tobakksreklame påberopte seg støtte fra opinionen og fra ekspertene. I den forbindelse var det svært interessant å finne ut hva legene mente om dette spørsmålet. Undersøkelsen viste at blant de legene som hadde tatt stilling til spørsmålet, var det et flertall for. Blant mannlige leger var det 64% som var for, og blant kvinnelige leger var det 81% som var for. I prosent av alle legene (de som ikke hadde tatt standpunkt medregnet) var det 57% av de mannlige legene som var for og 72% av de kvinnelige legene var for. Tabell 2.5 viser en enveis frekvens- og prosentfordeling av svarene fra de mannlige legene.

For være sikrere på at det virkelig var et flertall av legene som var for et reklameforbud, var det nødvendig å bruke slutningsstatistikk. Det vil i dette tilfellet si enten signifikanstesting eller konfidensintervall.

Tabell 2.5: Norske (mannlige) legers syn på forbudet mot tobakksreklame. En undersøkelse fra 1974. Eksempel på en enveis frekvens- og prosentfordeling. Utskrift fra SPSS versjon 11.0.0.

V46 Holdning til reklameforbudet[a]

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Enig	418	56,5	56,8	56,8
	2 Uenig	232	31,4	31,5	88,3
	3 Vet ikke	86	11,6	11,7	100,0
	Total	736	99,5	100,0	
Missing	System	4	,5		
Total		740	100,0		

a. V11 Kjønn = 1 mann

Vi skal nå se på de mannlige legene som hadde tatt standpunkt til spørsmålet. Dersom vi setter tallene for mannlige leger (418 mannlige leger er for og 232 er imot, det vil si at 64% av 650 er for) inn i formel 2.9, får vi en standardfeil på 1,88%. Dersom vi multipliserer dette tallet med 1,96 får vi det tallet vi skal trekke fra og legge til prosentestimatet for å få vite hvor stort område et konfidensintervall på 95% dekker. Resultatet blir 3,7%. Konfidensintervallet går dermed fra 60,3 – 67,7%. Vi kan med 95% sikkerhet si at den virkelige verdien faller innenfor dette området. Siden konfidensintervallet ikke omfatter tallet 50 prosent, kan vi si at det er signifikant flere enn 50 prosent av de mannlige legene som hadde tatt standpunkt som støttet forbudet mot tobakksreklame. Dersom vi gjør tilsvarende utregninger for kvinnelige leger, får vi et like overbevisende resultat. Det var 72 prosent av de kvinnelige legene som hadde tatt standpunkt som var for reklameforbudet, og antall som hadde tatt standpunkt var 346. Dermed skulle det være en enkel sak selv å regne ut standardfeil og konfidensintervall og sjekke om konfidensintervallet omfatter tallet 50 prosent.

Det finnes en mer direkte måte å undersøke signifikans på når en ønsker å teste en empirisk mot en hypotetisk fordeling. Den hypotetiske fordelingen angir hvordan vi tenker oss at fordelingen ser ut i populasjonen. Dette kalles, slik vi allrede har fortalt om ovenfor, en null-hypotese. Vi antar at fordelingen i populasjonen er 50 prosent for reklameforbudet og 50 prosent mot. Den alternative hypotesen er at andelen som er for reklameforbudet er forskjellig fra 50%. I utvalget er fordelingen den at 418 mannlige leger er for reklameforbudet mens 232 er imot. Det betyr at 64 prosent er for.

For å signifikansteste her kan en bruke en χ^2-test for frekvenstabeller. Tidligere har vi lært om normalfordelingskurven (eller z-fordelingen), og vi har brukt t-fordelingen. En χ^2-fordeling er egentlig en hel familie av fordelinger, og hvilken av disse som skal benyttes i det enkelte tilfelle avhenger av hvor mange frihetsgrader en opererer med. Disse fordelingenen brukes omtrent som en z-fordeling. Når en har regnet ut en χ^2 –verdi, kan en slå opp i en tabell og finne ut om den verdien en har regnet ut er høyere enn de kritiske verdiene i tabellen.

Den aktuelle signifikanstestingen kan nokså enkelt utføres av SPSS, men er også svært enkel å utføre for hånd. Vi støtter oss til Guilfords gamle lærebok i psykologisk statistikk (Guilford, 1965, se også Henkel, 1976). Vi tar utgangspunkt i den generelle formelen for å regne ut en kji-kvadrat-verdi for frekvenstabeller:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} \qquad (2.14)$$

f_o Den observerte frekvensen i en celle i en frekvenstabell

f_e Den forventede frekvensen i den samme cellen

Antall frihetsgrader i en enveis freksvensfordeling er lik antall celler minus én. Under forutsetning av at vi skal teste en observert enveis fordeling med bare to kategorier mot en hypotetisk fordeling, kan vi anvende en svært enkel formel. Legg merke til at summasjonstegnet er borte. Det er med andre ord nok å bruke en celle til beregningen, og det er det samme hvilken en velger:

$$\chi^2 = \frac{2(|f_o - f_e| - 0,5)^2}{f_e} \qquad (2.15)$$

f_o Den observerte frekvensen i den ene cellen i en tofelts-tabell

f_e Den forventede frekvensen i den samme cellen

De to loddrette strekene som står i telleren betyr at dersom uttrykket i mellom blir negativt, skal en erstatte minusen med pluss. Eller sagt på en annen måte, en bruker den absolutte verdien av differansen mellom f_o og f_e. Uttrykket inne i parentesen avsluttes med –0,5. Å trekke fra tallet 0,5 kalles Yates kontinuitetskorreksjon. En slik korreksjon er bare nødvendig når en regner ut χ^2-verdien på tabeller med bare en frihetsgrad[9]. Korreksjonen betyr mest når de forventede frekvensene er lave. For tabeller basert på et stort antall observasjoner er en slik korreksjon mindre viktig.

Den observerte fordelingen hentet fra Tabell 2.5 blir følgende: Enig i reklameforbudet: 418 personer, uenige i reklameforbudet: 232 personer, til sammen 650 personer. Dersom fordelingen blir antatt å være 50-50 i populasjonen, blir de forventede frekvensene 325 i begge cellene. La oss sette inn tallene og ta utgangspunkt i cellen med flest frekvenser.

$$\chi^2 = \frac{2(|418 - 325| - 0,5)^2}{325} = 52,65$$

Den observerte frekvensen i den ene cellen: 418

Den forventende frekvensen i den samme cellen: 325

I dette tilfellet skal vi slå opp i en tabell med en frihetsgrad (to celler minus én), og vi finner da at den verdien vi har regnet ut er større enn den verdien som tilsvarer en signifikans på p<0,001-nivået (10,827). Dette betyr at sannsynligheten for at så mange skulle svare at de

[9] Yates kontinuitetskorreksjon er noe omstridt blant fagstatistikere. Det er uenighet om hvor nødvendig det er å bruke denne korreksjonen, og hvor nøyaktig den er. Denne diskusjonen er gjort rede for i Howell, 1997. Selv om det i mange år har vært vanlig å bruke Yates kontinuitetskorreksjon, mener for eksempel Field (2000) at en like godt kan la være.

støttet reklameforbudet, dersom det virkelige tallet i universet var 50%, er mindre enn 0,001. Signifikanstestingen stemmer med andre ord godt med det resultatet vi fikk når vi beregnet konfidensintervallet. Dersom vi regner ut χ^2-verdien uten å bruke Yates kontinuitetskorreksjon blir resultatet 53,225.

Vi skal nå se på hvordan denne signifikanstestingen kan utføres i SPSS. Vi kan definere en variabel som heter "Weight" som variabel nr. 1. Deretter en som vi kaller "Holdning til reklameforbudet". I "Weight" legger vi inn 418 i første rad og 232 i andre rad. I holdningsvariabelen legger vi inn tallet 1 (positiv til reklameforbudet) i første rad og tallet 2 (negativ til reklameforbudet) i rad 2. Deretter trykker vi på *Data*, *Weight Cases*, og legger så inn "Weight"-variabelen. Trykk deretter *OK*. Deretter velger du *Analyze*, *Nonparametric Tests* og *Chi-square*. Legg så inn variabelen "Holdning til reklameforbudet" i *Test Variable List* og trykk *OK*. Uskriften vil da gi de to tabellene som er vist i Tabell 2.6. Ved å endre på de to tallene under "Weight", kan vi variere antallet i de to cellene og se hva resultatene blir.

Som tidligere nevnt har det vært uenighet om bruken av Yates kontinuitetskorreksjon (som bare skal anvendes når antall frihetsgrader er lik en), og vi ser at SPSS ikke anvender denne for testing av enveis frekvenstabeller. Egentlig er det temmelig merkelig at ikke Yates kontinuitetskorreksjon blir brukt her, for i de fleste statistikkbøker blir dette anbefalt. Dette gjelder også statistikktekster som er laget spesielt for å brukes sammen med SPSS (se for eksempel Weinberg & Abramowitz, 2002, side 534).

Første tabellen i utskriften viser de to frekvensene (antall som var for og antall som var mot reklameforbudet), forventede frekvenser under null-hypotesen (like mange i hver celle) og differansen mellom disse tallene (her kalt residualer). Den andre tabellen viser χ^2-verdien, antall frihetsgrader og den tilsvarende p-verdien. Vi ser at p-verdien er satt til p = 0,000. Egentlig er dette misvisende, for p-verdien blir aldri så liten som nøyaktig null. Men når det står tre nuller etter desimalen, betyr det at den er så liten at den er mindre enn p = 0,001. Dette er som regel det strengeste kravet en stiller til signifikans, så det har ikke noen hensikt å rapportere med flere desimaler enn dette. Når p-verdien er så lav, så betyr det at sannsynligheten for å få et så stort avvik fra en 50-50 fordeling i utvalget, gitt at null-hypotesen er riktig (at fordelingen i populasjonen er like mange for som mot) er mindre enn p = 0,001, altså mindre enn en promille. Vi forkaster altså null-hypotesen og har sannsynliggjort at det er et flertall av mannlige leger som var for reklameforbudet.

I sin argumentasjon for reklameforbudet hevdet Statens tobakkskaderåd at det ikke bare var et flertall av legene som støttet reklameforbudet, men at det til og med blant de røykende legene var et flertall som støttet det. Blant de kvinnelige røykende legene var dette åpenbart riktig. Av de kvinnelige røykende legene som hadde tatt et standpunkt, var det 80 prosent som var for reklameforbudet. Blant de mannlige legene var det imidlertid bare 56% (182 av til sammen 323) som var for. Det var derfor usikkert om Statens tobakkskaderåd hadde sine ord i behold.når det gjaldt de mannlige legene. For å teste dette ble det utført en χ^2-test.

Tabell 2.6: Chi-kvadrat-testing av enveis frekvensfordeling. Tabeller fra SPSS, versjon 12.

v1 Holdning til forbudet mot tobakksreklame

	Observed N	Expected N	Residual
1,00 For	418	325,0	93,0
2,00 Mot	232	325,0	-93,0
Total	650		

Test Statistics

	v1 Holdning til forbudet mot tobakksreklame
Chi-Square[a]	53,225
df	1
Asymp. Sig.	,000

a. 0 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is 325,0.

Tallene går fram av utregningen nedenfor.

$$\chi^2 = \frac{2(|182 - 161{,}5| - 0{,}5)^2}{161{,}5} = 4{,}95$$

Den observerte frekvensen i den ene cellen: 182

Den forventende frekvensen i den samme cellen: 161,5

Dette tallet er større enn 3,84, som for en frihetsgrad tilsvarer et signifikansnivå på 5%. Men det er mindre enn 5,41, som tilsvarer et signifikansnivå på 1%. Dette ser vi ved å slå opp i en tabell over χ^2-fordelingen. Vi kan derfor slå fast at sjansen for å at så mange leger skulle være mot tobakksreklame i utvalget er mindre enn 5%, dersom det virkelige tallet i populasjonen (parameteren) skulle være halvparten.

Utregningene ovenfor er basert på at vi har trukket et rent tilfeldig utvalg av leger. Det er imidlertid ikke tilfelle. Blant både mannlige og blant kvinnelige leger ble utvalget trukket ved bruk av proposjonal stratifisering. En slik prosedyre gir imidlertid minst like god presisjon som et rent tilfeldig utvalg, så på dette punktet kan ikke framgangsmåten kritiseres. Videre er utregningen basert på at vi har et utvalg fra en uendelig stor populasjon. Det er selvsagt feil. Antall leger i Norge i 1974 var omtrent 6000. Ved å korrigere for dette ville vi få redusert størrelsen på konfidensintervallene, og vi ville enda lettere oppnå signifikans. Så heller ikke

på dette punktet har vi vært for radikale. En annen mulig feilkilde er frafallet. Imidlertid var svarprosenten i denne undersøkelsen meget høy (94,2 %), så det er lite sannsynlig at bildet ville endret seg vesentlig selv om vi hadde fått alle legene til å svare.

Vi har ovenfor benyttet både konfidensintervall og signifikanstesting for å undersøke det samme, nemlig om andelen leger som var positive til reklameforbudet var forskjellig fra 50%. Om forholdet mellom konfidensintervall og signifikanstesting sier Howitt & Cramer (2000):

> ... confidence intervals contain enough information to judge statistical significance. However, statistical significance alone does not contain enough information to calculate confidence intervals. (p. 410)

2.4.5 Forutsetninger for bruk av χ^2-testen på enveis frekvensfordelinger

Mange signifikanstestser for metriske data baserer seg på at fordelingene på de variablene en tester skal ha en bestemt form (i hverfall når antall observasjoner ikke er høyt). De skal være tilnærmet normalfordelte. De kalles parametriske eller fordelings-avhengige tester (distribution-tied tests). χ^2-testen for enveis frekvensfordelinger er et eksempel på en ikke-parametrisk signifikanstest[10]. Disse kalles også fordelings-uavhengige tester (distribution-free tests).

Likevel er ikke χ^2-testen for enveis frekvensfordelinger helt uten krav til de data en bruker. Følgende forutsetninger må være oppfylt for at en skal kunne bruke denne testen:

1. Kategoriene (eller cellene) må være gjensidig utelukkende. En og samme person kan ikke gi mer enn ett svar, og dette svaret må entydig kunne plasseres i bare en av cellene.
2. Observasjonene må være uavhengige av hverandre. Det betyr vi med utgangspunkt i en bestemt observasjon (for eksempel en leges svar) ikke må kunne si noe om en annen observasjon (en annen leges svar).
3. Antallet i cellene må ikke være for lavt, det må være minst så stort som 5 i hver enkelt celle. Og det dreier seg da om det forventede antallet, ikke om det observerte.
4. Vi må dessuten passe på at summen av forventede frekvenser alltid er lik summen av observerte frekvenser.

---

[10] For en oversikt over de ikke-parametriske statistiske testene viser vi til en klassisk tekst av Siegel (1956) eller til en nyere utgave av Siegel & Castellan (1988). En noe mer omfattende innføring i ikke-parametrisk statistikk finner vi i Sheskin (1997). I de fleste større statistikkprogrammer finner vi de ikke-parametriske testene plassert i en egen avdeling som kalles "nonparametric tests". Slik er det også i SPSS. De ikke-parametriske testene er som regel mindre kontroversielle enn de parametriske testene fordi de ikke bygger på forutsetninger om intervallnivå og normalfordelte variabler. En ulempe med de parametriske testene er at de ofte har noe lavere styrke (power) enn de parametriske (de har vanskeligere for å gi signifikante sammenhenger eller forskjeller). Noen av dem har likevel en teststyrke som ligger nokså nær opp til de tilsvarende parametriske testene.

2.4.6 Testing ved bruk av binomialfordelingen

Men hva skal en så gjøre dersom en skal teste en fordeling som inneholder et så lavt antall observasjoner at en får mindre enn fem observasjoner i minst en av cellene? I dette tilfellet bruker vi binomialfordelingen, som vi har gjort rede for tidligere i dette kapittelet. Ved å bruke binomialfordelingen kan vi få ut sannsynligheter som er presist riktige, og ikke bare tilnærmet riktige slik som når vi bruker χ^2-testen. For å teste dette kan vi se hva som skjer dersom vi reduserer antall observasjoner i fordelingen vi har prøvd ut ovenfor til 7 og 2 observasjoner. Antall forventede frekvenser i hver celle (når vi tester mot en 50-50-fordeling skal da bli 4,5. Dersom vi på nytt bruker χ^2-testen for enveis frekvensfordelinger, vil det komme en fotnote til tabellen som sier at "2 cells (100,0%) have expected frequencies less than 5. The minimum expected cell frequency is 4,5." Vi bør da kjøre analysen på nytt, men passe på å be om å be om en eksakt test ("Exact"). Dersom vi gjør dette, får vi den utskriften som er vist i Tabell 2.7.

Vi ser nå at det er kommet ut to nye p-verdier. Den første tilsvarer den p-verdien som χ^2-testen gir, ("Asymptotic Significance"). Men siden de forventede frekvensene i cellene er så lave, er det riktigere å bruke "Exact Significance". Vi ser at p-verdien er lik 0,180. Dette er et tall som er høyere enn den mest liberale signifikansgrensen vi etter vanlige konvensjoner kan tillate oss å bruke, nemlig 0,05. Vi har med andre ord ikke fått signifikans. Med et utvalg på bare 9 leger der 7 av disse var for reklameforbudet mens 2 var imot, ville vi ikke kunne forkaste nullhypotesen om en 50-50-fordeling i populasjonen.

Tabell 2.7: Testing av enveis frekvensfordeling ved bruk av binomialfordelingen

v1 Holdning til forbudet mot tobakksreklame

	Observed N	Expected N	Residual
1,00 For	7	4,5	2,5
2,00 Mot	2	4,5	-2,5
Total	9		

Test Statistics

	v1 Holdning til forbudet mot tobakksreklame
Chi-Square[a]	2,778
df	1
Asymp. Sig.	,096
Exact Sig.	,180
Point Probability	,141

a. 2 cells (100,0%) have expected frequencies less than 5. The minimum expected cell frequency is 4,5.

”Point probability” er sannsynligheten for å få en fordeling som er nøyaktig så forskjellig fra 50-50-fordelingen (det vil si sannsynligheten for å få en 7-2-fordeling pluss sannsynligheten for å få en 2-7-fordeling).

Den eksakte testen (som altså er basert på binomialfordelingen) kan bare benyttes nå vi har to kategorier på variabelen (med andre ord to celler i tabellen). Når vi ønsker å teste enveisfordelinger med mer enn to kategorier, er vi henvist til å bruke χ^2-testen. Alternativt kan vi slå sammen celler slik at vi bare sitter igjen med to. Sammenslåing av celler kan imidlertid lett føre til at en bryter med forutsetningene for signifikanstestingen. Dersom vi systematisk forsøker å kombinere celler slik at vi skal få fram et signifikant resultat, er det stor fare for at vi ”kapitaliserer” på tilfeldig variasjon i dataene. Sammenslåing av celler må derfor gjøres etter klare kriterier som vi kommer fram til på forhånd, altså uavhengig av de resultatene vi kan lese ut av tabellene.

2.4.7 Signifikanstesting av endringer på dikotomier når en har paneldata (McNemars test)

Noen ganger følger en samme gruppe av personer over tid, og en er interessert i å se om fordelingen på en kategoriell variabel endrer seg. Det enkleste tilfellet er selvsagt endringer i en dikotomi. Et eksempel kan være andel røykere i samme gruppe på to tidspunkt. Slike data finnes i et HEMIL-senter-prosjekt som ble gjennomført etter oppdrag fra Den Norske Kreftforening. Oppdraget gikk ut på å evaluere virkningene av et antirøykeprogram som ble gjennomført i tre intervensjonsgrupper og en kontrollgruppe. For-undersøkelsen (baseline) ble gjennomført i november 1994 og første oppfølgingsundersøkelse (post-test) ble gjennomført allerede i mai 1995. På disse tidspunktene var de som deltok alle elever i 7. klasse (det som i dag er 8. klasse) i grunnskolen, med andre ord gjennomsnittlig 13,5 år gamle midt i skoleåret. Spørsmålet forskerne stilte seg var om det allerede etter så kort tid hadde funnet sted en viss rekruttering av røykere i kontrollgruppen. Jo større rekruttering av røykere det hadde vært i kontrollgruppen, desto mer sannsynlig var det at en kunne vente å finne noen effekt ved å sammenlikne med intervensjonsgruppene. For å finne ut av dette ble det først laget en enkel krysstabell som er vist i tabell 2.8.

Tabell 2.8: Endringer i røyking (røyker/røyker ikke) fra forundersøkelse (T_1) til første etterundersøkelse (T_2) i kontrollgruppen.

		Røyker ved T_2?	
		ja	nei
Røyker ved T_1?	ja	49	**20**
	nei	**93**	852

Tallet nederst til høyre viser antall ungdommer som ikke røykte ved forundersøkelsen og heller ikke røykte ved oppfølgingsundersøkelsen (852). Tallet øverst til venstre viser antall ungdommer som røykte på begge tidspunkt (49). Ingen av disse har endret røykevaner slik vi har målt røyking her.

De to andre tallene (vist med uthevet skrift) viser de som endret atferd. Øverst til høyre ser vi antall ungdommer som røykte ved forundersøkelsen, men har sluttet ved etterundersøkelsen. Det er 20 personer. Nederst til venstre ser vi antallet som har begynt å røyke. Det er 93 personer. Vi ser med andre ord at det er langt flere som har begynt å røyke enn som har sluttet. Det er et uttrykk for den rekruttering av røykere som gjerne finner sted i begynnelsen av ten-årene. Spørsmålet er om dette er en signifikant endring.

For å teste dette, sammenlikner vi ganske enkelt størrelsen på de to gruppene som har endret seg. Nullhypotesen er at gruppene er like store. Dermed kan vi ganske enkelt bruke den testen vi har lært ovenfor der vi sammenlikner to frekvenser. Vi setter de aktuelle tallene inn i formel 2.2 og får følgende:

$$\chi^2 = \frac{2(|93 - 56{,}5| - 0{,}5)^2}{56{,}5} = 45{,}88$$

Den observerte frekvensen i den ene cellen: 93

Den forventende frekvensen i den samme cellen: 56,5

Denne testen kalles McNemars test for endringer (McNemar's test for the significance of changes, McNemar, 1969). Testen har en frihetsgrad. Tallverdien vi har regnet ut (45,88) er langt høyere enn de kritiske verdiene for ulike signifikansnivå i en χ^2-tabell. Vi kan derfor med stor sikkerhet trekke den konklusjon at det har foregått en økning i andel røykere blant ungdommene i kontrollgruppen[11].

Analyser som ble gjort av korttids-effekter av intervensjonene viste da også store forskjeller mellom gruppene. Økningen i andel som røykte, uansett hvordan vi definerte røyking, var større i kontrollgruppen enn i noen av de andre gruppene. Intervensjonen hadde en betydelig effekt (Jøsendal, Aarø & Bergh, 1998).

Dersom en ønsker å analysere endringer over tid på kategorielle variabler som er målt på ordinalnivå, og som har mer enn to kategorier, brukes vanligvis Wilcoxon's test (Siegel & Castellan, 1988).

[11] Det må likevel tas ett forbehold. I denne undersøkelsen er utvalget trukket på skolenivå. En har trukket hele skoler, og tatt med alle elevene på det aktuelle klassetrinnet, i stedet for å trekke enkeltelever. Dette bryter med forutsetningen for å anvende McNemars test. Siden resultatet er så soleklart, er det likevel ganske usannsynlig at det ville blitt noe annerledes selv om en hadde tatt hensyn til designeffekten.

La oss så gjøre denne analysen med SPSS. For å få fram den aktuelle tabellen må vi legge inn tre variabler:
V1: Røyker ved tidspunkt 1 (1=ja; 2=nei)
V2: Røyker ved tidspunkt 2 (1=ja; 2=nei)
V3: Weight – størrelsen på de fire gruppene som framkommer ved å kombinere V1 og V2

Dataene vil se slik ut når vi legger dem inn i "Data view":
1,00 1,00 49,00
1,00 2,00 20,00
2,00 1,00 93,00
2,00 2,00 852,00

Først trykker vi på *Data*, deretter på *Weight Cases*, og så *Weight Cases By*. Og som variabel som skal brukes til å vekte med benytter vi V3. Deretter går vi til *Analyze*, så til *Nonparametric Tests*, og så til *2 Related Samples*. Til slutt legger vi inn V1 og V2 under *Test Pair(s) List*, klikker på *McNemar's* og deretter *OK*. Dersom alt dette er gjort riktig, får vi ut de tabellene som er vist nedenfor (Tabell 2.9).

Tabell 2.9: Testing av endring i røykestatus ved bruk av McNemars test

v1 Røyker T1 & v2 Røyker T2

	v2 Røyker T2	
v1 Røyker T1	1	2
1	49	20
2	93	852

Test Statistics[b]

	v1 Røyker T1 & v2 Røyker T2
N	1014
Chi-Square[a]	45,876
Asymp. Sig.	,000

a. Continuity Corrected

b. McNemar Test

Vi ser at χ^2-verdien er lik den vi har regnet ut ovenfor. Her har SPSS benyttet Yates kontinuitetskorreksjon, slik det skal gjøres. SPSS har, som vi ser, også gitt oss en p-verdi som

er satt lik 0,000. Som vi tidligere har vært inne på er p-verdier som er nøyaktig lik null sjelden vare i den virkelige verden. Men vi må tolke tallet dit hen at p-verdien i hvert fall er mindre enn 0,001. Dermed kan vi trekke den konklusjon at testen viser signifikans. Nullhypotesen, som sier at det ikke har skjedd noen endring i røykevaner (like mange har begynt som de som har sluttet i populasjonen), kan forkastes. Gitt at nullhypotesen var riktig, ville sannsynligheten for å få et slikt resultat i utvalget (at langt flere begynte enn sluttet) være mindre enn en promille ($p<0,001$).

2.4.8 Testing av et empirisk gjennomsnitt mot et hypotetisk

I avsnittene 2.4.4 til 2.4.7 har vi sett hvordan vi tester enveisfordelinger på kategorielle variabler mot hypotetiske fordelinger. Vi har i samme slengen tatt med oss McNemars test, selv om den egentlig er en test der vi sammenlikner to fordelinger, nemlig fordelingen på en dikotom variabel på ett tidspunkt mot den samme variabelen på et senere tidspunkt. Imidlertid er testen laget slik at det en egentlig undersøker er en enveis-fordeling: antall som har endret i den ene retningen mot antall som har gått i motsatt retning. Det er derfor nokså naturlig å omtale McNemars test i forbindelse med testing av empiriske enveis frekvensfordelinger mot hypotetiske.

Vi skal nå se på en annen situasjon, nemlig der vi har regnet ut det aritmetiske gjennomsnittet på en metrisk variabel og ønsker å teste denne mot en hypotese om den tilsvarende parameteren for populasjonen.

Vi forestiller oss at vi har administrert en rekke tester, deriblant en IQ-test til ti barn som får spesialundervisning på grunn av lærevansker. Resultatene av testingen ser slik ut:

Barn nr.	IQ
1	96
2	89
3	97
4	91
5	102
6	96
7	93
8	88
9	104
10	90

Vi har kjennskap til at gjennomsnittlig skår på denne IQ-testen for norske skolebarn er 100. Dette kommer av at testen er utviklet og laget slikat den skal ha en gjennomsnitt på akkurat 100. Vi antar at testen nylig er utviklet, slik at vi kan regne med at gjennomsnittstallet fremdeles er gyldig. Vi er interessert i å finne ut om de ti barna vi har testet har en

gjennomsnitts-skår som er signifikant lavere enn gjennomsnittstallene for alle norske barn i samme aldersgruppe[12].

For å finne ut av dette bruker vi en t-test for gjennomsnitt i ett utvalg.

$$t = \frac{\bar{X} - \mu}{SE_{\bar{x}}} \qquad (2.16)$$

t t-skår

$\bar{X}$ Gjennomsnitt i utvalget

$SE_{\bar{x}}$ Standardfeilen til gjennomsnittet i utvalget

Standardfeilen regner vi ut ved hjelp av formel 2.11. Deretter bruker vi formel 2.16 for å regne ut t-verdien. Svaret vi får ved å bruke disse formlene er at t-verdien er lik -3,151. Siden t-fordelingen ikke er en bestemt fordeling, men et helt sett av fordelinger, må vi vite hvilken fordeling vi skal bruke. Hvilken fordeling vi skal benytte er avhengig av antall frihetsgrader. Antall frihetsgrader er i dette tilfellet lik antall observasjoner minus tallet 1. Med ti observasjoner får vi ni frihetsgrader.

Det finnes egne tabeller som viser hvor stor t-verdien må være for å oppnå signifikans på t-testen. Det viser seg at med 9 frihetsgrader er den kritiske verdien for å oppnå signifikans på p<0,05-nivået 2,262. Det betyr at for å oppnå signifikans må t-verdien vi regner ut være større enn pluss 2,262 eller mindre enn minus 2,262. I vårt tilfelle har vi fått tallet -3,151, noe som gir signifikans på p<0,05-nivået. Dersom vi skulle oppnådd signifikans på p<0,01-nivået, måtte t-verdien være større enn 3,250 eller mindre enn minus 3,250. Det er den ikke. Ergo har vi ikke oppnådd signifikans på dette nivået.

I SPSS kan vi utføre denne testen ved å legge inn de ti observasjonene i en kolonne i regnearket (Data View). Så går vi inn på Variable View og kaller den IQ. Deretter går vi til *Analyze*, *Compare Means*, og *One-Sample T Test*. Deretter legger vi variabelen inn i ruten *Test Variable(s)*. Så legger vi tallet 100 inn som *Test Value*. Vi behøver ikke å endre på noe under *Options*. Se tabell 2.10.

[12] Bruken av IQ-testing er blitt sterkt kritisert både blant psykologer og pedagoger, så dette eksempelet er kanskje det minst politisk korrekte vi kunne velge. Imidlertid har IQ-tester en egenskap som passer glimrende i dette tilfellet. De er laget slik at de har et kjent gjennomsnitt og en kjent fordeling. IQ-skalaer skal ha et gjennomsnitt på 100 og et standardavvik på 15. Dette vil egentlig bare være riktig kort tid etter at testen er standardisert. Dette fordi den gjennomsnittlige skåren endrer seg over tid. La oss for eksemplets skyld her likevel anta at gjennomsnittet for befolkningen er 100.

Tabell 2.10: Testing av aritmetisk gjennomsnitt for empirisk fordeling mot kjent parameter

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
IQ	10	94,6000	5,42013	1,71399

One-Sample Test

	Test Value = 100					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
IQ	-3,151	9	,012	-5,40000	-9,2773	-1,5227

Resultatene vi får ut er gjengitt i Tabell 2.10.Vi ser at gjennomsnittet for utvalget er beregnet til 94,6. Standardavviket er beregnet til 5,52. Standardfeilen er 1,71. Selve signifikanstestingen er vist i den andre del-tabellen. T-verdien er beregnet til – 3,151, og med 9 frihetsgrader gir dette en p-verdi på 0,012. Dette tallet er lavere enn 0,05, men det er høyere enn 0,01. Det betyr at den beregnede gjennomsnittlige IQ i dette utvalget er signifikant lavere enn 100 dersom vi bruker et signifikansnivå på 0,05. Men dersom vi stilte krav om signifikans på p<0,01-nivået, hadde vi altså ikke oppnådd signifikans.

Dersom antall observasjoner hadde vært større, nærmere bestemt minst n = 30, kunne vi i stedet for å bruke en t-fordeling ha testet ved hjelp av z-fordelingen. Dette fordi t-fordelingene, etter hvert som en øker n (og dermed også øker antall frihetsgrader), mer og mer nærmer seg en z-fordeling.

2.4.9 Avvik fra normalfordeling

For å teste om en emprisk fordeling avviker fra normalitet, har vi to statistiske tester til disposisjon: Kolmogorov-Smirnov-testen[13] og Shapiro-Wilks-testen. Dersom vi ved bruk av disse testene får som resultat at p-verdien er lavere enn 0,05 (alternativt 0,01 eller 0,001), betyr det at fordelingen på variabelen er signifikant forskjellig fra en normalfordelt variabel med samme gjennomsnitt og standardavvik. Dersom en har muligheter til å velge mellom de to testene, bør en velge Shapiro-Wilks-testen, som gir noe mer nøyaktige resultater enn Kolmogorov-Smirnov-testen (Field, 2000).

[13] Kolmogorov-Smirnov-testen brukes ofte til å teste avvik fra andre fordelinger enn normalfordelinger, for eksempel avvik fra uniforme fordelinger, Poisson-fordelinger og eksponensielle fordelinger.

Kolmogorov-Smirnov-testen finner en i SPSS under *Analysis*, *Nonparametric Tests*, 1-S*ample K-S*. Det regnes ut en z-verdi og en tilsvarende p-verdi (Asymptotic Sig. (2-tailed)). Dersom p-verdien er lavere enn 0,05, kan vi forkaste null-hypotesen om at variabelen for eksempel er normalfordelt.

Dersom variabler avviker for sterkt fra normalfordelingen til at det er forsvarlig å benytte parametrisk statistikk, finnes det i prinsippet to løsninger. Vi kan transformere skalaen slik at variabelen får en fordeling som avviker mindre fra normalitet, eller vi kan velge statistiske tester som ikke bygger på en forutsetning om normalfordelte variabler på metrisk nivå. Slike tester kalles (som nevnt tidligere) ikke-parametriske.

Dersom antall observasjoner er stort, behøver vi imidlertid ikke nødvendigvis å forkaste bruken av parametrisk statistikk, selv om avviket fra normalitet skulle være signifikant. Det som i slike tilfeller er avgjørende, er størrelsen på avviket uttrykt ved kurtosis og skjevhet. Når antall observasjoner øker, avtar dessuten den feilen en gjør ved å bruke parametrisk statistikk på variabler som avviker fra normalfordeling. Waternaux (1976) har vist at når en analyserer på mer enn 100 subjekter, forsvinner problemet med at en underestimerer varians på variabler med negativ kurtosis (flat fordeling) og når en analyserer på mer enn 200 subjekter, forsvinner problemet med at en underestimerer varians på variabler med positiv kurtosis (spisse fordelinger) (se også Tabachnick & Fidell, 1996, s. 71-78).

Referanser

Field, A. (2000). *Discovering statistics. Using SPSS for Windows*. London: Sage.

Guilford, J.P. (1965): *Fundamental statistics in psychology and education.* New York: McGraw-Hill.

Henkel, R.E. (1976). *Tests of significance.* Beverly Hills: Sage.

Howell, D.C. (1997). *Statistical methods for psychology* (4. Utgave). Belmont, California: Duxbury.

Howitt, D. & Cramer, D. (2000). *An introduction to statistics in psychology* 2. utgave. Harlow, Essex, England: Pearsons.

Jøsendal, O., Aarø, L.E. & Bergh, I.H. (1998). Effects of a school-based smoking prevention programme among sub-groups of adolescents. *Health Education Research*, *13(2)*, 215-224.

Lewis-Beck, M.S. (1993). Applied regression: An introduction. I M.S. Lewis-Beck (red.), *Regression analysis*. (s.1-68). London: Sage.

McNemar, Q. (1969): *Psychological statistics.* New York: John Wiley & Sons.

Norusis, M.J. (1993). *SPSS for windows. Base system user's guide. Release 6.0.* Chicago, Illinois: SPSS Inc.

Sheskin, D.J. (1997). *Handbook of parametric and nonparametric statistical procedures*. Boca Raton, Florida: CRC Press.

Siegel, S. (1956): *Nonparametric Statistics for the Behavioral Sciences.* Tokyo: McGraw Hill.

Siegel, S. & Castellan, N.J. (1988). *Nonparametric Statistics for the Behavioral Sciences.* New York: McGraw Hill.

Tabachnick, B.G. & Fidell, L.S. (1997). *Using multivariate statistics* (Tredje utgave). New York: Harper & Collins.

Waternaux, C.M. (1976). Asymptotic distribution of the sample roots for a nonnormal population. *Biometrica, 63(3)*, 639-645.

Weinberg, S.L. & Abramowitz, S.K. (2002). *Data analysis for the behavioral sciences using SPSS.* Cambridge: Cambridge University Press.

Kap 3: Bivariat statisikk

Bivariat statisikk dreier seg om å beskrive sammenhenger mellom to variabler samt å trekke slutninger om slike sammenhenger fra utvalg til populasjon. Hvordan slike sammenhenger skal beskrives og hvordan en skal regne ut konfidensintervall eller signifikansteste avhenger først og fremst av variablenes målenivå. Når en skal analysere sammenhenger mellom kategorielle variabler (med et lite antall kategorier), bruker en som regel krysstabeller med prosentuering og chi-kvadrat-testen for uavhengighet i krysstabeller. Når en på en enkel måte skal beskrive styrken på slike sammenhenger, finnes det en lang rekke assosiasjonsmål en kan velge mellom. Når en skal analysere sammenhengen mellom to metriske variabler, bruker en gjerne Pearsons produkt-moment korrelasjon og en signifikanstest for å se om en kan forkaste hypotesen om at korrelasjonen er lik null. Denne korrelasjonen baserer seg imidlertid på en antakelse om at sammenhengen er lineær, at den kan beskrives langs en rett linje. Dersom sammenhengen ikke er lineær, finnes det forskjellige alternative måter å beskrive sammenhengen på. Når en har to ordinalvariabler, finnes det flere koeffisienter (og signifikanstester) å velge mellom, blant annet Spearmans rangkorrelasjonskoeffisient og Goodman-Kruskals Gamma. Når den ene variabelen er kategoriell og den andre er metrisk, benytter en som regel eta-koeffisienten og t-test for forskjeller mellom gjennomsnitter eller enveis variansanalyse for å signifikansteste. Det er alle disse statistiske størrelsene dette kapittelet skal handle om. Mot slutten av kapittelet skal vi dessuten ta for oss ulike former for kontroll for tredjevariabler.

3.1 Bivariat analyse av kategorielle variabler

Det er sjelden forskere er fornøyde med bare å se på enveis frekvensfordelinger. Det er likevel viktig at denne første delen av den statistiske bearbeidelsen blir gjort skikkelig. Forskeren bør ha et inngående kjennskap til hver enkelt variabel før han eller hun begynner å se på relasjoner mellom variabler. Det en særlig bør være på vakt mot, er skjeve fordelinger, små grupper og ekstreme verdier. Hvis noe av dette er problematisk, bør en vurdere å transformere variablene slik at de blir mindre skjeve, eller en bør vurdere å slå sammen nabokategorier der dette er logisk mulig.

Som vi allerede har nevnt ovenfor, foregår analysen av relasjoner mellom to kategorielle variabler vanligvis ved bruk av krysstabeller og prosentueringer. I tillegg foretas gjerne en signifikanstesting som sier noe om det er en statistisk sikker eller signifikant sammenheng mellom de to variablene. Videre regnes det ofte ut forskjellige assosiasjonsmål.

3.1.1 Firefelts-krysstabeller

Et eksempel på en svært enkel krysstabell er vist i Tabell 3.1. Tabellen er hentet fra en undersøkelse gjennomført blant et representativt utvalg elever i ungdomsskolen av Statens tobakkskaderåd (nå: Sosial- og helsedirektoratet, Avdeling tobakk) i år 2000. Det er blant annet stilt spørsmål om hvilket klassetrinn elevene går på, deres røykevaner, og om de har vært med på et intervensjonsprogram som heter "VÆR røykFRI"[1], og som tilbys alle ungdomsskoler i Norge. Programmet ble i år 2000 brukt av omtrent halvparten av alle ungdomsskolene i Norge. I tabellen ser vi andel som røyker daglig etter om de har deltatt i programmet. Tabellen omfatter bare elever i 10. klasse. Blant de som har deltatt i programmet ser vi at andel som røyker daglig er 12,4 prosent. Blant de som ikke har deltatt i programmet ser vi at andelen som røyker er 19,6 prosent. Spørsmålet er om denne forskjellen er så stor at vi med rimelig grad av sikkerhet kan forkaste nullhypotesen, med andre ord forkaste hypotesen om at det i populasjonen av norske tidendeklassinger ikke er noen forskjell i andel dagligrøykere mellom de som er med i VÆR røykFRI og de som ikke er med.

For å teste dette må vi gjøre en kji-kvadrat-testing, altså en signifikanstesting som er basert på kji-kvadrat-fordelingen. Som vi har vært inne på tidligere, er dette i likhet med z-fordelingen en sannsynlighetsfordeling. Når en tester sammenhenger i krysstabeller er det denne fordelingen som er riktigst å benytte.

På samme måte som når vi ovenfor testet enveis-fordelinger, dreier det seg også her om å sammenlikne en observert og en hypotetisk fordeling. Den observerte fordelingen er vist i Tabell 3.1. Der ser vi at det blant de som har deltatt i VÆR røykFRI er 55 ungdommer som røyker daglig, mens 387 ungdommer ikke røyker daglig. Blant de som ikke har deltatt i VÆR røykFRI er fordelingen 85 og 349 ungdommer.

Men før vi regner ut resultatet, skal vi se hvordan vi lager en tabell med forventede frekvenser, med andre ord en tabell som viser hvordan tallene ville fordelt seg dersom det ikke var noen sammenheng mellom deltakelse i VÆR røykFRI-programmet og røykevaner

[1] Programmet er nå revidert og har endret navn til "FRI".

(røyke daglig/ikke røyke daglig). For å få fram det forventede antallet i den cellen som ligger i første rad og første kolonne, ganger vi summen av observasjonene i første rad (442) med summen av observasjonene i første kolonne (140) og deler på totalt antall observasjoner i tabellen (876). Svaret blir i dette tilfellet 70,64. Formelen vi bruker er vist i 3.1.

Tabell 3.1: Røyker daglig etter deltakelse i VÆR røykFRI. Elever i 10. klasse.

v1 Med i VÆR røykFRI? * v2 Røyker daglig? Crosstabulation

			v2 Røyker daglig?		
			1,00 Ja	2,00 Nei	Total
v1 Med i VÆR røykFRI?	1,00 Ja	Count	55	387	442
		% within v1 Med i VÆR røykFRI?	12,4%	87,6%	100,0%
	2,00 Nei	Count	85	349	434
		% within v1 Med i VÆR røykFRI?	19,6%	80,4%	100,0%
Total		Count	140	736	876
		% within v1 Med i VÆR røykFRI?	16,0%	84,0%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	8,318[b]	1	,004		
Continuity Correction[a]	7,795	1	,005		
Likelihood Ratio	8,368	1	,004		
Fisher's Exact Test				,004	,003
Linear-by-Linear Association	8,309	1	,004		
N of Valid Cases	876				

a. Computed only for a 2x2 table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 69,36.

Dersom vi regner ut de forventede frekvensene for hver av cellene, vil vi se at marginalfordelingene (summen av hver kolonne og summen av hver rad) er nøyaktig de samme som før. Vi vil dessuten se at dersom vi prosentuerer horisontalt (slik at prosentene adderer seg opp til 100 for hver rad), får vi de samme prosentene i begge radene, nemlig 15,98. Dette betyr at dersom vi ikke fant noen sammenheng mellom deltakelse i VÆR røykFRI og røykevaner, ville det vært 15,98 prosent røykere både blant de som hadde deltatt og blant de som ikke hadde deltatt. Dette tilsvarer nullhypotesen. Nullhypotesen sier at det ikke er noen sammenheng, eller sagt på en annen måte, det er ingen forskjell i røyking mellom de som har deltatt og de som ikke har deltatt i VÆR røykFRI.

$$f_{e\,jk} = \frac{f_j * f_k}{n} \qquad (3.1)$$

$f_{e\,jk}$ Forventet antall observasjoner i cellen i rad j og kolonne k

f_j Antall observasjoner i rad j

f_k Antall observasjoneri kolonne k

n Totalt antall observasjoner i tabellen

Tabell 3.2: Observerte og forventede frekvenser

		Røyker daglig	Røyker ikke daglig	Sum
Deltatt i VÆR røykFRI	Observerte frekvenser	55	387	442
	Forventede frekvenser	70,64	371,36	442
Ikke deltatt i VÆR røykFRI	Observerte frekvenser	85	349	434
	Forventede frekvenser	69,36	364,64	434
Sum	Observerte frekvenser	140	736	876
	Forventede frekvenser	140	736	876

Dersom vi nå setter tallene fra tabellen inn i formel 3.2, blir resultatet 7,795. Dette ser vi stemmer med utskriften fra SPSS, der det står at Chi-kvadrat-verdien, når en anvender Yates' kontinuitetskorreksjon, blir 7,795. I toveis krysstabeller er antall frihetsgrader lik antall rader minus én multipliert med antall kolonner minus én. I firefeltstabeller (2x2-tabeller) er derfor antall frihetsgreader lik 1. Det er derfor vi bruker kontinuitetskorreksjonen. Når vi slår opp i en Kji-kvadrat-tabell finner vi at tallet 7,795 er mindre enn den kritiske verdien for signifikans på p<0,001- nivået (10,84), men den er større enn den kritiske verdien for signifikans på p<0,01- nivået (6,64). Det er med andre ord mindre enn 1 prosent sannynlig at vi ville finne en så stor forskjell i røykevaner mellom de som har deltatt og de som ikke har deltatt i VÆR røykFRI dersom det ikke var noen forskjell i populasjonen. Vi kan derfor forkaste

nullhypotesen. Det ser med andre ord ut til at deltakelse i VÆR røykFRI henger sammen med mindre dagligrøyking.

$$\chi^2 = \sum \frac{(|f_o - f_e| - 0,5)^2}{f_e} \qquad (3.2)$$

χ^2 Chi-kvadrat-verdien

f_o Den observerte frekvensen i en celle i en frekvenstabell

f_e Den forventede frekvensen i den samme cellen

De som ønsker å gjøre den analysen som er vist i tabell 3.1 på egen hånd, kan gå fram på følgende måte. Først legger en inn følgende tall i data-arket (*Data View*) i SPSS:

1	1	55
1	2	387
2	1	85
2	2	349

Deretter går en inn i *Variable View* og kaller variabelen i første kolonne for V1 og setter inn labelen "Med i VÆR røykFRI?". Deretter kaller en variabelen i den andre kolonnen for V2 og setter inn labelen "Røyker daglig?". Deretter setter en under *Values* inn at 1 betyr ja og 2 betyr nei på begge disse variablene. Og så kaller en den siste variabelen *Weight*, men uten å sette inn noen tekst under *Values* på denne.

Dernest går en inn på *Data*, *Weight Cases*, klikker på *Weight cases by*, legger inn variabelen Weight og klikker så *OK*. Deretter går en inn på *Analyze*, *Descriptive Statistics* og *Crosstabs*. Så legger en inn V1 – "Med i VÆR røykFRI?" som rekkevariabel (Row) og "Røyker daglig?" som kolonnevariabel (Column). Gå så inn på *Statistics* og trykk på *Chi Square* og *Continue*. Gå deretter inn på *Cells* og trykk på *Row* og *Continue*. Når en så til slutt trykker på *OK*, får en ut de tabellene som er vist under Tabell 3.1.

I tabell 3.1 vises også en del annen statistikk. Pearsons Chi-square er den vanlige chi-kvadrat-testen uten at en har anvendt Yates kontinuitetskorreksjon. Likelihood ratio er en alternativ test som kan brukes. Fishers eksakte test (Fisher's exact test) er en test som må brukes når antall forventede observasjoner i minst en av de fire cellene i tabellen er lavere enn 5,00. Den siste statistiske størrelsen "Linear-by-Linear Association" gir bare mening dersom en har krysstabeller med flere enn to rader eller flere enn to kolonner. Bokstavene d.f. står for "degrees of freedom" eller frihetsgrader. Siden det er den samme sammenhengen som testes hele tiden, er antall frihetsgrader alltid 1. Tallene i kolonnene til høyre for d.f.-kolonnen er alle sammen p-verdier. Det betyr at SPSS har slått opp i tabellen for oss, slik at vi slipper. Vi

ser at testene gir temmelig sammenfallende resultat. Alle de tohalede testene gir p-verdier på 0,004 eller 0,005. Som vi allerede har slått fast ovenfor, er dermed resultatet at sammenhengen er signifikant på p<0,01-nivået, men ikke på p<0,001-nivået.

Det er viktig å være klar over at vi dermed ikke har bevist at VÆR røykFRI-programmet har hatt en effekt på elevenes røyking. Alt vi har vist er at det blant de elevene som har deltatt er signifikant mindre røyking. Hva som ville være tilfelle dersom disse elevene ikke hadde deltatt i programmet, vet vi ikke sikkert. Kanskje hadde andelen som røykte blant disse vært like lav uansett. Kanskje er det skoler med få elever som røyker som melder seg på og vil delta i programmet. For å finne ut av dette måtte det gjennomføres en felteksperimentell undersøkelse der vi randomiserte skolene slik at vi rent tilfeldig plasserte skoler i en intervensjonsgruppe og lot de øvrige utgjøre en kontrollgruppe. En slik studie er faktisk gjennomført, og viste at det også med et slikt design ble mindre røyking blant elevene ved skoler som deltok i programmet (Jøsendal, Aarø & Bergh, 1998; Jøsendal & Aarø, 2005). Dette er en langt sikrere indikasjon på kausalitet.

3.1.2 Større krysstabeller

Vi skal nå se på testing av sammenhenger i tabeller som er større enn en firefeltstabell. Det betyr tabeller med minst tre rader eller minst tre kolonner. Denne gangen skal vi bruke data fra prosjektet Health Behaviour in School Aged Children – A Cross National Study (HBSC). Vi skal ta for oss resultater fra Hordaland fylke, og vi skal se på sammenhengen mellom klassetrinn og det å oppleve å bli mobbet. Vi tenker oss her at dataene er samlet inn ved rent tilfeldig trekkeing av enkeltelever. Dersom en har trukket utvalget på andre måter, er det feil å bruke de signifikanstestene som benyttes nedenfor. Vi skal senere se at dataene ikke er trukket slik som forutsatt, og vi skal komme tilbake til hvordan en da skal analysere disse dataene.

Det er tre klassetrinn som deltar i denne undersøkelsen, nemlig 6., 8. og 10. klasse i grunnskolen. Mobbing er målt på en skala som har fem kategorier:

1 – ikke mobbet
2 – en eller to ganger
3 – av og til
4 – ukentlig
5 – flere ganger i uken

Vi har slått sammen kategoriene 3, 4 og 5, gitt samlekategorien tallverdien 1 og sagt at de elevene som har krysset av for ett av disse svarene mobbes av og til eller ofte. Og så har vi slått sammen kategoriene 1 og 2 og gitt disse tallverdien 2. Disse elevene mobbes sjelden eller aldri. Resultatet av analysen vises i tabell 3.3.

Vi ser at i denne utskriften mangler Chi-kvadrat-verdien beregnet med bruk av Yates' kontinuitetskorreksjon. Det er fordi denne tabellen er større enn en firefelts-tabell. Den har tre rader og to kolonner, og antall frihetsgrader blir dermed (3-1)*(2-1) = 2. For tabeller som har to eller flere frihetsgrader skal korreksjonen ikke benyttes.

Vi ser at Chi-kvadrat-verdien er 6,071. Dersom vi slår opp i en tabell, vil vi finne at den kritiske verdien for å få signifikans på p<0,05-nivået ved to frihetsgrader er 5,99. Forskjellen i andel som mobbes over klassetrinn er med andre ord signifikant. Vi kan forkaste nullhypotesen om at det er like mye mobbing på de tre klassetrinnene.

Tabell 3.3: Mobbing etter klassetrinn. Elever fra Hordaland (HBSC-studien 1997)

GRADE * Mobbes? Crosstabulation

			Mobbes?		
			av og til eller ofte	sjelden eller aldri	Total
GRADE	Grade 6	Count	20	116	136
		% within GRADE	14,7%	85,3%	100,0%
	Grade 8	Count	11	125	136
		% within GRADE	8,1%	91,9%	100,0%
	Grade 10	Count	10	143	153
		% within GRADE	6,5%	93,5%	100,0%
Total		Count	41	384	425
		% within GRADE	9,6%	90,4%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6,071[a]	2	,048
Likelihood Ratio	5,791	2	,055
Linear-by-Linear Association	5,379	1	,020
N of Valid Cases	425		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 13,12.

Vi ser av tabellen at vi ikke ville oppnådd signifikans dersom vi hadde basert oss på den signifikanstestingen som kalles ”Likelihood ratio”. Noen ganger vil denne være mindre sensitiv enn Pearsons, og noen ganger vil den være mer sensitiv. Den ene er ikke mer korrekt enn den andre. Den vanligste testen i forbindelse med krysstabeller er Pearsons, og vi holder oss derfor til den. For en forsker som oppdager at en signifikanstest viser signifikans mens en annen ikke gjør det, kan det være fristende i ettertid å velge den som gir signifikans (dersom en da ønsket å oppnå signifikans). Dette er imidlertid en framgangsmåte som bryter med hele logikken bak signifikanstesting. En må på forhånd bestemme seg for hvilken test en skal benytte og holde seg til den. Og dersom en benytter en annen test enn det som er standard i forskningslitteraturen, bør en ha klare og overbevisende argumenter for det valget en har gjort.

Dersom vi ser nøyere på tabellen, oppdager vi at forskjellen mellom 6. og 8. klassetrinn er på 6,6 prosentpoeng, mens forskjellen mellom 8.og 10. klassetrinn er på 1,6 prosentpoeng. Vi kan derfor stille følgende spørsmål. Er det forskjellen mellom 6. og 8.klasse som er utslagsgivende? Kan det tenkes at den er signifikant, mens forskjellen fra 8. til 10. klasse ikke er signifikant? Dette er det mulig å teste. Vi legger inn et ekstra filter som gjør at vi bare får med elever fra to klassetrinn om gangen, og kjører ut tabellen på nytt. Vi får da følgende resultat:

6. klasse mot 8. klasse: $\chi^2 = 2{,}949$; 1 frihetsgrad; $p = 0{,}086$
8. klasse mot 10. klasse: $\chi^2 = 0{,}257$; 1 frihetsgrad; $p = 0{,}612$
6. klasse mot 10. klasse: $\chi^2 = 5{,}166$; 1 frihetsgrad; $p = 0{,}023$

Bare den siste av de tre testene gir signifikans. Det viser at det er forskjellen mellom laveste og høyeste klassetrinn som gjør utslaget. Ut fra tallene er det rimelig å anta at mobbingen er mest utbredt i 6. klasse og at den så blir lavere i 8. klasse og aller lavest i 10. klasse, men vi trenger et større datamateriale for å fastslå med sikkerhet hvordan forskjellen over klassetrinn egentlig ser ut.

For å finne ut hvordan mønsteret ser ut i det øvrige Norge, tar vi ut en krysstabell der vi utelukker alle elever fra Hordaland (se Tabell 3.4). Vi ser her at sammenhengen er litt svakere. Forskjellen målt i prosentpoeng mellom 6. og 10. klasse er 5,7 for hele landet mot 8,2 for Hordaland. Likevel er mønsteret det samme. Andelen som mobbes av og til eller ofte avtar over klassetrinn. Når vi her får langt klarere signifikans ($p<0{,}001$), skyldes det at antall observasjoner er så mye høyere. Jo høyere antall observasjoner, desto lettere er det å oppnå signifikans gitt at styrken på sammenhengen (eller forskjellen mellom gruppene) er den samme. Når vi tester forskjellen mellom 8. og 10. klasse for alle skoler i landet (utenom Hordaland), viser det seg imidlertid at den er så liten (to prosentpoeng) at den ikke oppnår signifikans ($\chi^2 = 2{,}820$; 1 frihetsgrad; $p = 0{,}093$).

Som vi tidligere har vært inne på, baserer den vanlige statistikken som en finner i lærebøker og programpakker seg på at utvalgene er trukket ved bruk av rent tilfeldig trekking. De dataene vi har benyttet i eksemplene over (fra HBSC-prosjektet), er imidlertid samlet inn klassevis. En har ikke trukket enkeltelever, men skoleklasser. Dermed har vi sannsynligvis gjort en feil. Dersom det viser seg at forekomsten av mobbing er en variabel som henger sammen med skoleklasse (med andre ord at det i noen skoleklasser foregår systematisk mer mobbing enn i andre skoleklasser), ville vi ganske sikkert ikke fått samme resultater. Blant annet er det svært tvilsomt om vi hadde oppnådd en signifikans når vi testet forskjeller over de tre klassetrinnene for Hordaland.

Heldigvis er det utviklet statistisk programvare som kan brukes i slike tilfeller (Bryk & Raudenbush, 1992; Goldstein, 1995). I SPSS finnes det en modul som kalles ”Complex”. I denne finnes det en egen prosedyre for testing av sammenhenger i krysstabeller når en har trukket utvalget på andre måter enn ved rent tilfeldig trekking. Dersom ulike grupper er trukket med ulik sannsynlighet kan en vekte observasjonene for at de skal gi et riktig bilde av populasjonen. En kan også legge inn informasjon om stratifisering og om trekking av klynger (clusters). Alt dette kan krysstabellanalysen i Complex korrigere for.

Tabell 3.4: Mobbing etter klassetrinn. Elever fra hele landet unntatt Hordaland.

GRADE * Mobbes? Crosstabulation

			Mobbes?		
			av og til eller ofte	sjelden eller aldri	Total
GRADE	Grade 6	Count	260	1291	1551
		% within GRADE	16,8%	83,2%	100,0%
	Grade 8	Count	189	1255	1444
		% within GRADE	13,1%	86,9%	100,0%
	Grade 10	Count	164	1318	1482
		% within GRADE	11,1%	88,9%	100,0%
Total		Count	613	3864	4477
		% within GRADE	13,7%	86,3%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	21,473[a]	2	,000
Likelihood Ratio	21,296	2	,000
Linear-by-Linear Association	20,903	1	,000
N of Valid Cases	4477		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 197,72.

3.1.3 Dekomponering av sammenhenger i større krysstabeller

Når en har chi-kvadrat-testet en krysstabell som er større enn 2x2 og funnet at sammenhengen er statistisk sikker (signifikant), lurer en ofte på hvilke deler av tabellen det var som gav signifikans. En krysstabell som er større enn 2x2 kan jo deles inn i flere undertabeller, og hver slik tabell gir informasjon om deler av den sammenhengen som finnes i den samlede tabellen.

Det finnes flere strategier for å dekomponere en større tabell og dekomponere en slik samlet chi-kvadrat-test i del-tabeller og del-chi-kvadrat-verdier. Maxwell (1961) (se også Kimball, 1954) gjør rede for to slike metoder. Uansett metode handler det om å dele inn tabellen i undertabeller, og beregne en chi-kvadrat-verdi for hver slik tabell. Antall undertabeller en bør analysere tilsvarer antall frihetsgrader i den store tabellen. Antall frihetsgrader i en toveis krysstabell er som kjent antall kolonner (K) minus en multiplisert med antall rader (R) minus en, med andre ord: (K-1)*(R-1). Dette betyr altså at en tabell med 3 rader og 4 kolonner har (3-1)*(4-1) = 6 frihetsgrader. En trenger altså seks undertabeller for å dekomponere chi-kvadrat-verdien i en 3 x 4-tabell.

Maxwell (1961) gjør oppmerksom på at en kan finne signifikante del-tabeller selv om hovedtabellen ikke viser signifikans. Han mener dette bidrar til å gjøre signifikanstestingen mer sensitiv. Vi kan føye til at dersom vi ikke finner en samlet signifikans når vi tester hele

tabellen under ett, vil det å tolke på signifikans vi finner i under-tabeller, være spekulativt. Det vil lett føre til at vi kapitaliserer på tilfeldig variasjon eller tilfeldige feil i data. I all signifikanstesting er det et overordnet prinsipp at dersom testingen av en totaleffekt ikke gir signifikans, skal en heller ikke legge vekt på signifikante del-effekter.

Dekomponering av en 3x3-tabell kan f.eks. foregår slik (a1 er antall observasjoner i cellen der rad 1 og kolonne 1 krysser hverandre, a2 er antall observasjoner i cellen der rad 1 og kolonne 2 krysser hverandre etc.):

Fig. 3.1: Dekomponering av en 3 x 3-tabell etter Maxwell (1961)

Hovedtabell:

a1	a2	a3
b1	b2	b3
c1	c2	c3

Deltabeller:

a1	a2
b1	b2

a1	a2	a3
b1	b2	b3

a1	a2
b1	b2
c1	c2

a1	a2	a3
b1	b2	b3
c1	c2	c3

Vi ser at antall deltabeller er fire. Dette tilsvarer antall frihetsgrader i tabellen (3-1)*(3-1) = 4. En behøver ikke følge akkurat et slikt mønster som det dette eksempelet viser. Hovedpoenget er å følge en framgangsmåte som er statistisk korrekt og som samtidig synliggjør de kontrastene som er særlig interessante.

3.1.4 Valgmuligheter i utskriftene fra krysstabellanalyser

I forbindelse med analyse av krysstabeller varierer det en del hvilke statistiske størrelser fra utskriftene som er meningsfullt å bruke. Blant annet varierer dette med hvor mange rader og kolonner tabellen har og hvor mange forventede observasjoner en har i cellene. Det er særlig viktig å være oppmerksom på følgende:

- Når antall observasjoner i cellene blir svært lavt i en 2x2-tabell, skal en i stedet for χ^2testen benytte Fishers nøyaktige sannsynlighetstest (Fishers's exact probability test). Dette fordi chi-kvadrat-fordelingen blir en for unøyaktig tilnærming ved et lavt antall observasjoner.

- Når en beregner χ^2-verdier for 2x2-tabeller, benyttes helst Yates kontinuitetskorreksjon for å få en bedre tilpasning til χ^2-fordelingen. Denne kontinuitetskorreksjonen er beskrevet ovenfor, og som vi allerede har nevnt, er det noe uenighet blant fagstatistikere om den bør anvendes eller ikke. De fleste lærebokforfattere vil nok anbefale at den brukes.

3.2 Bivariat analyse av metriske variabler

Når en har med metriske variabler å gjøre, er det vanlig å beskrive bivariate sammenhenger ved bruk av Pearsons produkt-moment korrelasjon. Når det snakkes om korrelasjoner, uten noen nærmere forklaring, er det som regel denne koeffisienten det siktes til.

Karl Pearson (1857-1956) drev forskning om arvelighet og evolusjon. Hans mest produktive periode var mellom 1893 og 1912. Han utviklet statistiske redskaper som korrelasjonskoeffisienten (Pearsons produkt-moment korrelasjonskoeffisient) og regresjonsanalyse. Han lanserte begrepet "standardavvik" i 1893. Pearson kom dårlig overens med en annen berømt engelsk statistiker, Ronald Fisher. Mens Fisher spesialiserte seg på statistikk for små utvalg, var Pearson opptatt av å anvende statistikk på store utvalg. Konflikten mellom Pearson og Fisher var såpass bitter at Fisher lot være å ta en stilling han hadde søkt på, og fått tilbud om, fordi det ville bety at han måtte jobbe under Pearson. Dette var en stilling som sjefsstatistiker for Galton-laboratoriet i London.

Pearson var fritenker og sosialist og likte dårlig alt som kunne minne om underkastelse under autoriteter. Dette kom til uttrykk da han var student ved Cambridge. Det var obligatorisk frammøte til andakt (chapel), og studentene måtte dessuten være til stede ved teologiske forelesninger. Pearson var svært interessert i religion, men han hatet tanken på obligatoritet. Han engasjerte seg sterkt i en argumentasjon mot både forelesningene og andaktene, og vant til slutt fram. Bestemmelsene om obligatoritet ble opphevet. Til universitetsledelsens store overraskelse fortsatte han likevel å møte opp til andaktene. For Pearson var det stor forskjell på tvungen og frivillig deltakelse. Pearson avslo å motta ordener og hedersbevisninger. Han holdt forelesninger om kvinnespørsmål og marxisme.

Karl Pearson (1857-1956)

3.2.1 Hva er en metrisk variabel?

Variabler som er målt på ekte metrisk nivå (intervall eller rationivå) har en ikke mange av i atferdsforskningen og samfunnsforskningen. Eksempler på slike variabler er høyde, vekt og temperatur. Ofte arbeider en med variabler som en for enkelhets skyld antar er metriske, selv om de strengt tatt ikke er det. I forbindelse med holdningsmåling kan en for eksempel komme med påstander og be om at informantene svarer ved å merke av på en linje som går fra ”Helt enig” til ”Helt uenig”. Kanskje kan en da anta at en har målt på metrisk nivå. Informasjonen en registerer er avstanden fra det ene ytterpunktet på skalaen til stedet der repsondenten har markert. Avstanden kan måles i centimeter eller i antall (like store) intervaller som er merket av på linjen. Enda vanligere er det å måle holdninger ved at respondentene krysser av i bokser som har egne verditekster, for eksempel:

- ❑ Helt enig
- ❑ Ganske enig
- ❑ Litt enig
- ❑ Verken enig eller uenig
- ❑ Litt uenig
- ❑ Ganske uenig
- ❑ Helt uenig

En nummererer boksene fra 1 til 7, og lar disse tallverdiene symbolisere de ulike svarene. Det er nokså vanlig å analysere slike variabler som om de skulle være intervallvariabler. En antar da at avstanden mellom svaralternativene er lik over hele skalaen, med andre ord at avstanden mellom for eksempel ”Helt enig” og ”Ganske enig” er like stor som avstanden mellom ”Ganske enig” og ”Litt enig”. At informantene kognitivt fungerer slik at avstandene er like, er imidlertid en tvilsom antakelse. Det kan likevel argumenteres med at den feilen en gjør i så fall har liten praktisk betydning.

3.2.2 Korrelasjonsdiagrammet (punktdiagrammet) og produkt-moment-korrelasjonen

La oss likevel starte ut med variabler som alle lett kan være enige om er målt på metrisk nivå, nemlig høyde og vekt. La oss tenke oss at vi måler høyde og vekt på en gruppe rekrutter som er på sesjon. Gruppen består av 15 personer, og tallene er slik:

Person nr.	Høyde i cm.	Vekt i kg.
01	179	81
02	167	72
03	191	85
04	186	93
05	178	78
06	189	91
07	172	67
08	177	78
09	184	89
10	182	90
11	175	85
12	185	81
13	175	72
14	181	71
15	170	77

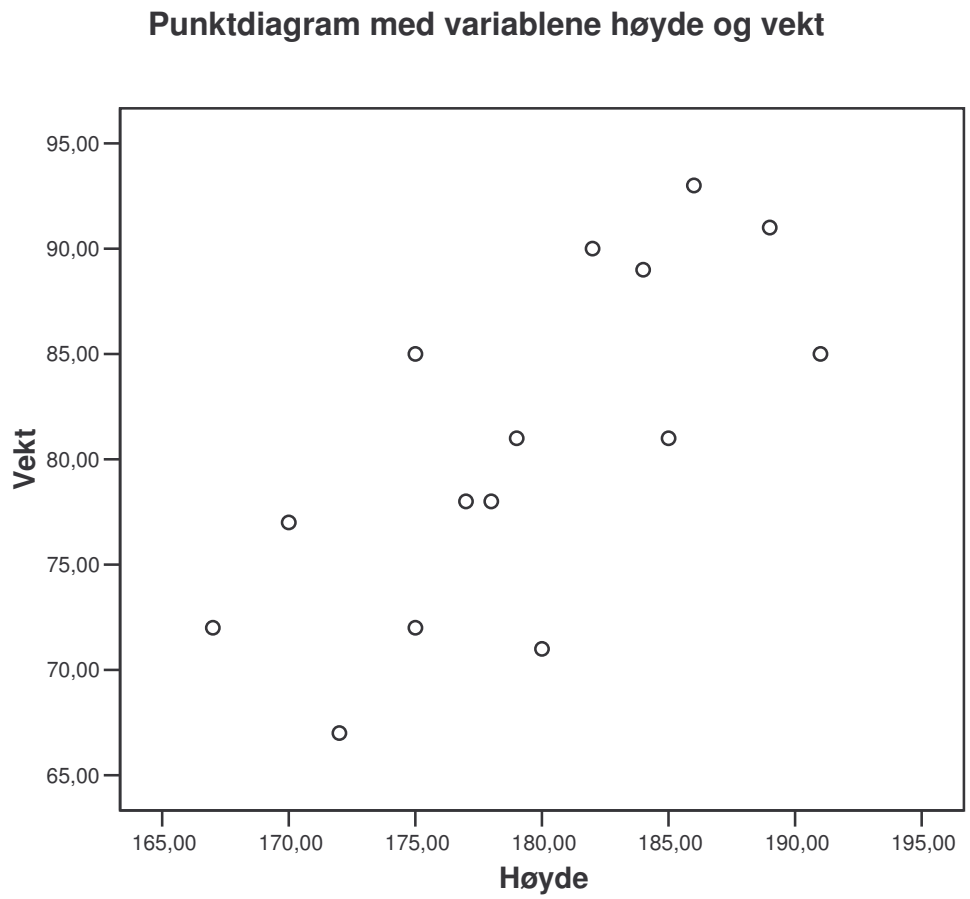
For å lage et punktdiagram legger vi først tallene inn i SPSS *Data Editor*. En må da være påpasselig med å legge inn tallene som tilhører samme person på samme linje. Dersom en legger inn tallene i en annen rekkefølge, vil ikke programmet forstå at opplysningene hører sammen tre og tre (person nr., høyde og vekt). Husk å sette navn på variablene i *Variable View* og kall dem for ”Personnr”, ”Høyde” og ”Vekt”. Når tallene og variabelnavnene er lagt inn riktig, går en inn i *Graphs*, *Scatter* og *Design* og legger inn ”Høyde” som X-akse-variabel og ”Vekt” som Y-akse variabel. Og dersom en vil, kan en legge inn en overskrift på diagrammet i *Title*. Når en så trykker *OK*, får en fram et punktdiagram. Diagrammet er gjengitt i Fig. 3.2.

Hvert punkt i planet representerer en av rekruttene. Ved å lese av på x-aksen og y-aksen, er det lett å gjenfinne hver enkelt person. Den personen som ligger aller lengst til venstre i diagrammet ser vi er 167 cm. Høy og veier 72 kilo. Denne personen gjenfinner vi tabellen over som nr. 02.

Når vi inspiserer dette diagrammet, legger vi merke til at punktene ikke fordeler seg helt tilfeldig utover i rommet. De danner en slags figur, en avlang ”sky” som strekker seg fra nederst til venstre mot øverste høyre hjørne i firkanten som omgir punktene. Vi får inntrykk

av at det er en viss sammenheng mellom høyde og vekt. Men foreløpig har vi ikke noen måte å uttrykke denne sammenhengen på. Vi trenger en koeffisient som kan brukes til dette formålet. En slik koeffisient finnes, og det er altså Pearsons produkt-moment korrelasjon, eller bare Pearsons r.

Fig. 3.2: Punktdiagram med variablene høyde og vekt (n=15)



Aron & Aron (1999) gir en pedagogisk sett meget god forklaring på hva en produkt-moment-korrelasjon er. De starter med å si at du først skal standardisere begge de variablene du skal korrelere. Det vil si at du fra hver observasjon trekker gjennomsnittsverdien og deler på standardavviket. Dermed blir gjennomsnittet på begge variablene lik 0,0 og standardavviket blir lik 1,0. Dette kalles z-skårer eller standardskårer. For hver enhet (for eksempel for hver person) i datasettet, skal du gange standardskåren vedkommende har på den ene variabelen med standardskåren vedkommende har på den andre variabelen. Dette kalles kryssprodukter. Deretter legger du sammen kryssproduktene for alle enhetene (personene) og deler på n (antall enheter eller personer). Resultatet av denne utregningen er Pearsons produkt-moment korrelasjon. En slik korrelasjon er med andre ord gjennomsnittet av kryssproduktene når du har z-transformert (eller standardisert) begge variablene du skal korrelere. Denne statistiske størrelsen har altså svært gunstige egenskaper, for eksempel at den blir null når det ikke er noen lineær sammenheng mellom variablene, pluss 1,0 når den lineære sammenhengen er perfekt positiv og -1,0 når den lineære sammenhengen er perfekt negativ.

Men vi kan også vise hva en produkt-moment-korrelasjon er med utgangspunkt i den formelen en vanligvis bruker. Formelen er gjengitt nedenfor (formel 3.3).

$$r = \frac{\sum_{1}^{n}((x_i - \bar{X})(y_i - \bar{Y}))}{n * sd_x * sd_y} \qquad (3.3)$$

x_i En enhets verdi på variabel x

$\bar{X}$ Det aritmetiske gjennomsnittet på variabelen x

y_i En enhets verdi på variabel y

$\bar{Y}$ Det aritmetiske gjennomsnittet på variabelen y

sd_x Standardavviket på variabelen x

sd_y Standardavviket på variabelen y

n Antall observasjoner

Vi skal nå regne ut korrelasjonen mellom de to variablene høyde og vekt ved å bruke de dataene som er gjengitt ovenfor. For å regne ut korrelasjoner mellom variabler i SPSS går vi inn i *Analyze*, *Correlate* og *Bivariate*. Dedretter merker vi variablene "Høyde" og "Vekt" og legger begge inn i feltet *Variables*. Vi kan også godt forsikre oss om at det er satt en hake utenfor ordet *Pearson*. Til slutt klikker vi på *OK*. Vi får da ut det resultatet som er vist i Fig. 3.3.

Fig. 3.3: Korrelasjonsanalyse fra SPSS

Correlations

		Høyde	Vekt
Høyde	Pearson Correlation	1	,703**
	Sig. (2-tailed)		,003
	N	15	15
Vekt	Pearson Correlation	,703**	1
	Sig. (2-tailed)	,003	
	N	15	15

**. Correlation is significant at the 0.01 level

Vi ser at resultatet er gjengitt to ganger, både i cellen øverst til høyre og i cellen nederst til venstre. Vi ser videre at korrelasjonen er 0,703. Det er med andre ord en ganske høy korrelasjon mellom høyde og vekt. Og vi ser at korrelasjonen er positiv. Det betyr at en høy person har en tendens til å veie mer enn en lav person, et resultat som ikke er særlig

overraskende. I de to andre cellene ser vi tallet 1. Ett-tallet symboliserer at når en korrelerer en variabel med seg selv, blir korrelasjonen 1,0. Sammenhengen er med andre ord perfekt. I alle cellene står tallet 15. Det viser antall observasjoner hver korrelasjon er basert på. I denne tabellen er n den samme i alle cellene. Dersom vi hadde hatt med mange varibler i analysen hadde vi fått ut mange korrelasjoner, og dersom vi hadde manglet opplysninger for enkelte av personene på noen av variablene, ville n variert fra celle til celle.

Vi har tidligerte vært inne på at en korrelasjonskoeffisient varierer mellom -1,00 og +1,00. Dersom vi har en perfekt positiv korrelasjon, ville punktene i korrelasjonsdiagrammet ligge på en rett linje som starter ned til venstre og peker oppover mot høyre. Dersom vi har en perfekt negativ korrelasjon, ville punktene ligge langs en rett linje som starter øverst til venstre og peker nedover mot høyre. Dersom korrelasjonen var null, ville vi ikke finne noe slikt mønster som kan beskrives langs en rett linje. Dersom det ikke var noen sammenheng mellom variablene, ville punktene fordele seg nokså tilfeldig utover i planet.

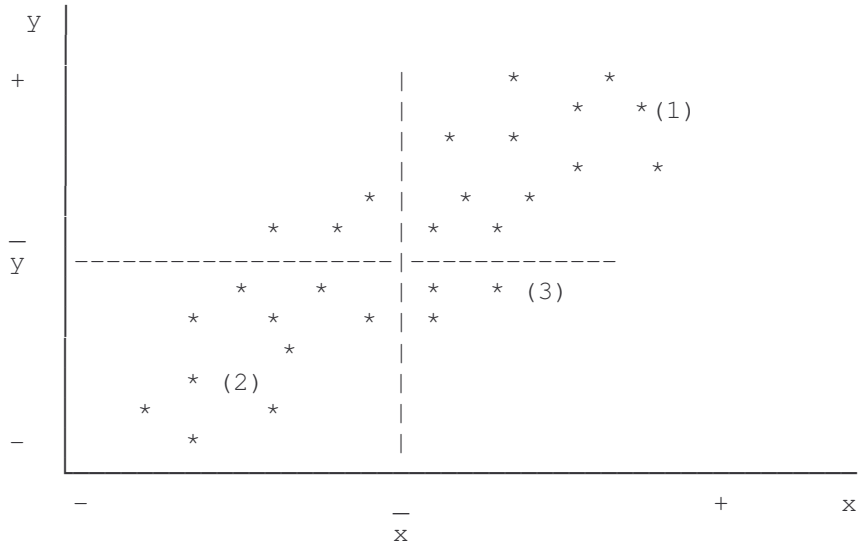
Dersom en kvadrerer en korrelasjonskoeffisient (ganger den med seg selv), får vi vite hvor mye av variansen i den ene av variablene som forklares av den andre variabelen. Dersom vi kvadrerer 0,703, får vi tallet 0,494. Det betyr at vi her har en forklart varians på 49,4 %.

Det er viktig å være klar over at vi noen ganger har sammenhenger som ikke kan beskrives langs en rett linje. Når vi printer ut et korrelasjonsdiagram, kan vi tenke oss at vi likevel kan se et tydelig mønster. Det kan for eksempel være formet som en U. Punktene starter øverst til venstre, sprer seg ned mot midten av diagrammet, og stiger deretter oppover mot høyre. Selv om et slikt diagram viser en tydelig sammenheng mellom to variabler, vil denne sammenhengen ikke komme til uttrykk når vi regner ut en produkt-moment-korrelasjon. Den kan lett bli nokså nær 0,00. Korrelasjonen fanger altså ikke opp den sammenhengen en tydelig kan se er der. I så fall må en finne andre måter å beskrive sammenhengen på. Hvordan dette kan gjøres, skal vi komme tilbake til senere.

La oss se litt nærmere på hva som skjer når vi regner ut en korrelasjonskoeffisient ved bruk av formel 3.3 Vi skal ikke bry oss noe særlig om den delen av formelen som faller under streken. I forbindelse med definisjonsformler for korrelasjoner kan vi som regel tenke på uttrykket under streken som et virkemiddel til å gi koeffisienten de riktige egenskaper (variasjon mellom -1,0 og +1,0, verdien 0,0 ved uavhengighet etc.). La oss konsentrere oss om det som står over streken: $\sum((x_i - \bar{X})(y_i - \bar{Y}))$.

La oss ta utgangspunkt i en observasjon, den som er merket med (1) i Fig. 3.4. Denne observasjonen ligger høyt på x-aksen, slik at verdien (x- $\bar{x}$) blir positiv og nokså stor. Observasjonen skårer også høyt på y-aksen, slik at verdien (y- $\bar{y}$) også blir positiv og forholdsvis høy. Dermed blir også produktet av de to parentesene også et forholdsvis stort, positivt tall. Observasjonen bidrar dermed til en høy positiv korrelasjon. Dersom vi tar for oss observasjonen (2), vil vi se at de to parentesene gir store negative verdier. Når vi multipliserer et negativt tall med et negativt tall, får produktet positivt fortegn. Dermed har også denne observasjonen bidratt til en høy, positiv korrelasjon. Når vi vurderer observasjonen (3) på tilsvarende måte, oppdager vi at den vil bidra negativt. Verdien på den ene aksen (x-aksen) ligger over gjennomsnittet og gir dermed et positivt tall. Verdien på den andre aksen (y-aksen) ligger lavere enn gjennomsnittsverdien, og gir dermed et negativt tall. Produktet blir dermed et negativt tall.

Fig. 3.4: Korrelasjonsdiagrammets fire kvadrater



Dette kan generaliseres til følgende: Alle observasjonene i nederste venstre og øverste høyre kvadrat bidrar til en positiv korrelasjon. Alle observasjonene i nederste høyre og øverste venstre kvadrat bidrar til en negativ korrelasjon. Flertallet av observasjoner ligger her i nederste venstre og øverste høyre kvadrat. Derfor blir korrelasjonen i dette tilfellet positiv og høy. Dersom observasjonene hovedsakelig hadde ligget i de andre kvadratene, ville korrelasjonen blitt negativ. Vi skal senere se at denne logikken er ganske lik den som benyttes for de fleste assosiasjonsmål for firefeltstabeller. De fire kvadratene i Fig. 3.4 tilsvarer i en viss forstand de fire cellene i en 2x2-tabell.

3.2.3 Konfidensintervall og signifikanstesting av korrelasjoner

I Fig. 3.3 ser vi at korrelasjonen er utstyrt med to stjerner. Og under tabellen ser vi forklaringen. Der står det at "Correlation is significant at 0.01 level". Korrelasjonen er altså signifikant på p<0,01-nivået. Nullhypotesen er at korrelasjonen i universet som vårt utvalg er trukket fra er på 0,00. Den alternative hypotesen er at korrelasjonen er forskjellig fra 0,00. Men den kan like gjerne være negativ som positiv. Vi skal nå se nærmere på hvordan denne testen utføres, og vi skal se hvordan vi regner ut et konfidentintervall rundt en korrelasjonsverdi.

For å regne ut konfidensintervallet til en korrelasjon må en først transformere korrelasjonen til en z-verdi. Dette kalles Fishers z-transformasjon etter statistikeren Ronald Fisher som utviklet denne metoden. Formelen for en slik z-transformasjon er gitt i 3.4.

$$z_r = 0,5 * \ln \frac{1+r}{1-r} \qquad (3.4)$$

z_r z-verdien som svarer til en bestemt korrelasjon r

r korrelasjonskoeffisienten som skal transformeres

For en korrelasjon på 0,60 regner vi ut den tilsvarende z-verdien slik:

$$z_{r=0,60} = 0,5 * \ln \frac{1+0,60}{1-0,60} = 0,5 * \ln \frac{1,60}{0,40} = 0,5 * \ln 4 = 0,5 * 1,386 = 0,693$$

For korrelasjoner mellom -0,50 og +0,50 er den tilsvarende z-verien ikke så veldig forskjellig fra korrelasjonen selv. Sammenhengen mellom korrelasjonen og z-verdien er omtrent lineær. Z-verdien som svarer til en korrelsjon på 0,20 er lik 0,203. For en korrelasjon på 0,50 er den tilsvarende z-verdien 0,549. Men etter hvert som korrelasjonen nærmer seg -1,00 eller +1,00, blir avviket mellom korrelasjonen og z-verdien større og større. Når korrelasjonen er 0,95 er z-verdien 1,832. Når korrelasjonen er 0,99 er z-verdien 2,647.

For å kunne regne ut et konfidentintervall rundt en korrelasjon, må en anta at de to variablene som inngår i utregningen av korrelasjonen er bivariat normalfordelte. Det betyr at dersom en plotter alle observasjonene inn i et tredimensjonalt diagram, bør fordelingen se ut omtrent som en mexikansk hatt. Profilen ser ut som en normalfordelingskurve uansett hvilken vinkel en betrakter den fra. Weinberg & Abramowitz (2002) summerer opp alle forutsetningene som må være til stede på følgende systematiske måte:

1) Hver av de to variablene må være normalfordelte
2) Dersom en velger ut alle observasjonene som har en bestemt verdi på variabelen x (for eksempel x=a), skal fordelingen av disse observasjonene på y være normalfordelt. Dette kalles en betinget fordeling av y gitt at x=a.
3) Dersom en velger ut alle observasjonene som har en bestemt verdi på variabelen y (for eksempel y=b), skal fordelingen av disse observasjonene på y være normalfordelt. Dette kalles en betinget fordeling av x gitt at y=b.
4) Alle de betingede fordelingene av y gitt x skal ha samme standardavvik.
5) Alle de betingede fordelingene av x gitt y skal ha samme standardavvik.
6) Alle gjennomsnittene til de betingede fordelingene av y gitt x skal falle langs en rett linje.
7) Alle gjennomsnittene til de betingede fordelingene av y gitt x skal falle langs en rett linje.

Vi kan skille mellom tre kategorier av forutsetninger:

Normalitet – forutsetningene 1,2 og 3
Homoskedastisitet – forutsetningene 4 og 5
Linearitet – forutsetningene 6 og 7

Når vi skal regne ut konfidentintervallet til en korrelsjon, må vi først transformere r til z_r. Konfidensintervallet til z_r kan regnes ut på en ganske enkel måte. Først beregner vi standardfeilen til z_r som vises inne i parentesen i formel 3.5. Deretter ganger vi denne med den kritiske z-verdien. Det tallet vi da får ut kan legges til eller trekkes fra z_r, og den laveste verdien vi får er den nederste grensen for konfidensintervallet og den høyeste verdien angir den øverste grensen for konfidensintervallet. Deretter kan vi transformere resultatet tilbake fra z_r – verdier til r.

$$CI_{z_r} = z_r \pm z_{\text{kritisk verdi}} \left(\frac{1}{\sqrt{n-3}} \right) \qquad (3.5)$$

CI_{z_r} Konfidensintervallet til z_r

z_r Den z-verdien som tilsvarer korrelasjonen *r*

$z_{\text{kritisk verdi}}$ z-verdien som tilsvarer konfidensintervallet en beregner

n Antall observasjoner

La oss ta et eksempel. Korrelasjonen mellom høyde og vekt som er gjengitt i Fig. 3.3 ble 0,703. For enkelhets skyld sier vi at den ble 0,70. Antall observasjoner var 15. Ved hjelp av formlene ovenfor skulle det være greit å regne ut et 95 prosents konfidensintervall, det vil si øvre og nedre grense for det området der vi med 95% sannsynlighet kan si at korrelasjonen i populasjonen befinner seg.

Først transformerer vi korrelasjonen til z_r ved hjelp av formel 3.4.

$$z_{r=0,70} = 0,5 * \ln \frac{1+0,70}{1-0,70} = 0,5 * \ln \frac{1,70}{0,30} = 0,5 * \ln 5,67 = 0,5 * 1,73 = 0,87$$

Deretter bruker vi formel 3.5 for å regne ut konfidensintervallet til z_r.

$$CI_{z_r} = z_r \pm z_{\text{kritisk verdi}} \left(\frac{1}{\sqrt{n-3}} \right) = 0,87 \pm 1,96 \left(\frac{1}{\sqrt{15-3}} \right) = 0,87 \pm 1,96 \left(\frac{1}{3,46} \right) = 0,87 \pm 0,57$$

Den laveste verdien blir med andre ord 0,30 og den høyeste blir 1,44.

For å transformere verdien til bake til r, må vi bruke enda en formel, som av matematikk-kyndige lett kan utledes av formelen vi brukte når vi regnet motsatt vei (formel 3.4). Den nye formelen er gjengitt i 3.6.

$$r = \frac{e^{2z} - 1}{e^{2z} + 1} \qquad (3.6)$$

r korrelasjonskoeffisienten

e Det naturlige logaritmetallet (2,718)

z z-verdien som skal transformeres til korrelasjon

Dersom vi setter inn de to z-verdiene vi regnet ut ovenfor, og som angir nedre og øvre grense for konfidensintervallet uttrykt ved z-tall, får vi følgende utregninger:

$$r = \frac{e^{2z} - 1}{e^{2z} + 1} = \frac{2,718^{2z} - 1}{2,718^{2z} + 1} = \frac{2,718^{2(0,30)} - 1}{2,718^{2(0,30)} + 1} = \frac{1,82 - 1}{1,82 + 1} = 0,29$$

$$r = \frac{e^{2z} - 1}{e^{2z} + 1} = \frac{2,718^{2z} - 1}{2,718^{2z} + 1} = \frac{2,718^{2(1,44)} - 1}{2,718^{2(1,44)} + 1} = \frac{17,81 - 1}{17,81 + 1} = 0,89$$

Vi kan med 95% sikkerhet si at korrelasjonen mellom høyde og vekt i den populasjonen utvalget er trukket fra ligger mellom 0,29 og 0,89. Dette er et nokså bredt område, og det skyldes at antall observasjoner er så lavt som 15. Hadde antall observasjoner for eksempel vært 10 ganger så stort, ville konfidensintervallet blitt langt mindre.

Statistikkpakken SPSS gir dessverre ingen muligheter til å regne ut noe så enkelt som konfidentintervall for korrelasjoner.

Signifikanstesting av korrelasjoner kan derimot lett gjøres ved bruk av SPSS. For å signifikansteste en korrelasjon bruker en vanligvis ikke en test som baserer seg på z-fordelingen, men derimot en nært beslektet test som baserer seg på en t-fordeling. Til forskjell fra z-fordelingen finnes det mange forskjellige t-fordelinger, avhengig av hvor mange frihetsgrader en opererer med. Antall frihetsgrader er i dette tilfellet lik antall observasjoner minus 2 (altså n-2).

For å regne ut t-verdien bruker en formelen som er gjengitt i 3.7.

$$t = \frac{r}{\sqrt{\frac{(1 - r^2)}{(n - 2)}}} \qquad (3.7)$$

r korrelasjonskoeffisienten

n antall observasjoner

La oss ta eksempelet fra Fig. 3.3. Vi runder av korrelasjonen til 0,70 og setter inn i formel 3.7.

$$t = \frac{r}{\sqrt{\frac{(1-r^2)}{(n-2)}}} = \frac{0,70}{\sqrt{\frac{(1-0,70^2)}{(15-2)}}} = \frac{0,70}{\sqrt{\frac{0,51}{13}}} = \frac{0,70}{0,20} = 3,50$$

Dersom vi slår opp i en t-tabell og ser på raden med tretten frihetsgrader, finner vi følgende kritiske nivåer for de mest brukte signifikansnivåene:

p<0,05 – 2,160
p<0,01 – 3,012
p<0,001 – 4,221

Den verdien vi har regnet ut (t = 3,50) ligger mellom de to siste. Det betyr at korrelasjonen er signifikant forskjellig fra null på p<0,01-nivået. Dersom vi ser tilbake på fotnoten under Fig. 3.3, ser vi at dette stemmer.

Så sent som i 1995 publiserte Olkin & Finn en relativt enkel metode for å regne ut konfidensintervallet til forskjellen mellom to korrelasjoner. For store utvalg er standardfeilen til forskjellen r_{yx} mellom korrelasjonene fra forskjellige utvalg slik som vist i formel 3.7B. For å regne ut et konfidentintervall kan vi bruke z-fordelingen som tilnærming. Dersom konfidentintervallet ikke omfatter verdien null, betyr det at korrelasjonene er signifikant forskjellige.

Vi må imidlertid vise stor forsiktighet når vi skal tolke på forskjeller mellom korrelasjoner. Korrelasjoner lar seg påvirke av forskjeller i spredning på variablene mellom de to populasjonene. Testing av forskjeller i ustandardiserte regresjonskoeffisienter er i slike tilfeller en bedre tilnærming (Cohen et al., 2003, s. 47).

$$SE_{r_V - r_W} = \sqrt{\frac{1-r_V^2}{n_V} + \frac{1-r_W^2}{n_W}} \qquad (3.7.b)$$

$SE_{r_V - r_W}$ Standardfeilen til forskjellen mellom to korrelasjoner fra to forskjellige utvalg

r_V Korrelasjonen mellom to variabler fra utvalg V

r_W Korrelasjonen mellom de to samme variablene fra utvalg W

n_V Antall enheter fra utvalg V

n_W Antall enheter fra utvalg W

3.3 Sammenhengen mellom en metrisk variabel og en dikotomi

3.3.1 Forskjell i gjennomsnitt mellom to uavhengige grupper på en metrisk variabel

Ovenfor har vi gjort rede for hvordan en analyserer sammenhengen mellom to kategorielle variabler og hvordan en analyserer sammenhengen mellom to metriske variabler (når en kan anta at sammenhengen mellom de metriske variablene er lineær). Nå er vi kommet dit at vi skal se hvordan en analyserer sammenhengen mellom to variabler der den ene er dikotom og den andre er metrisk.

Den aller enkleste situasjonen består i at vi har målt en egenskap på en metrisk variabel i to forskjellige grupper, og ønsker å sammenlikne det aritmetiske gjennomsnittet i de to gruppene.

Når vi signfikanstester forskjellen mellom gjennomsnitts-skårene i to uavhengige grupper, baserer vi oss på en test som benytter t-fordelingen. Formelen for å regne ut t-verdien er gitt nedenfor (formel 3.8).

$$t = \frac{\bar{X}_1 - \bar{X}_2}{sd\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (3.8)$$

$\bar{X}_1$ Det aritmetiske gjennomsnittet i gruppe 1

$\bar{X}_2$ Det aritmetiske gjennomsnittet i gruppe 2

sd Estimat av det felles standardavviket for de to gruppene (se formel 3.9)

n_1 Antall observasjoner i gruppe 1

n_2 Antall observasjoner i gruppe 2

Vi har tidligere nevnt at t-statistikken ikke er en bestemt fordeling, men en familie av fordelinger. Antall frihetsgrader avgjør hvilken vi skal bruke i hvert enkelt tilfelle. Antall frihetsgrader er i dette tilfellet definert som $n_1 + n_2 - 2$.

Siden det er forskjellen mellom de to aritmetiske gjennomsnittene vi skal teste, er det vel ganske naturlig at det er denne gjennomsnitten vi opererer med over brøkstreken. Under streken beregner vi standardfeilen til forskjellen mellom gjennomsnittene i de to gruppene.

For å regne ut det felles standardavviket for de to gruppene trenger vi en ny formel som er gitt nedenfor (formel 3.9).

$$sd = \sqrt{\frac{sd_1^{\,2}(n_1 - 1) + sd_2^{\,2}(n_2 - 1)}{n_1 + n_2 - 2}} \qquad (3.9)$$

sd Det felles standardavviket for de to gruppene

sd_1 Standardavviket i gruppe 1

sd_2 Standardavviket i gruppe 2

n_1 Antall observasjoner i gruppe 1

n_2 Antall observasjoner i gruppe 2

Men hva er det en slik t-test for forskjeller mellom to gruppegjennomsnitt egentlig viser? Som i all annen signifikanstesting handler det om en null-hypotese og en alternativ hypotese. Nullhypotesen postulerer at det i populasjonen som utvalget er trukket fra ikke er noen forskjell mellom gjennomsnittene i de to gruppene. Den alternative hypotesen sier at det er en forskjell. Dersom vi for eksempel oppnår signifikans på p<0,01-nivået, kan vi trekke den konklusjon at sannsynligheten for at nullypotesen er riktig er mindre enn 0,01. Vi kan med andre ord forkaste nullhypotesen og vi har sannsynliggjort at det faktisk er en forskjell mellom de to gjennomsnittene.

La oss ta et eksempel. Blant 20 psykiatriske pasienter ved en sykehusavdeling har vi rekruttert de ti mest motiverte til å delta i en intensiv musikkterapi. De har noe varierende diagnoser, men har alle sammen vist seg å profitere lite på verbal terapi. Den ene gruppen (tilfeldig valgt blant de 20) gjennomgår i løpet av 6 uker en intensiv musikkterapi med en sesjon hver dag (behandlingsgruppen). Den andre gruppen (kontrollgruppen) får ikke noe alternativt tilbud. Begge gruppene testes med en skala som måler forekomsten av depressivitet. De som gikk i terapi ble testet umiddelbart etter at terapien var avsluttet mens pasientene i kontrollgruppen ble testet samtidig. Skalaen går fra 0 til 30. De som har fått skåren null har ikke rapportert noen plager i det hele tatt. Dersom en får en skår på 30, betyr det at en har rapportert maksimal hyppighet av alle plagene som inngikk i skalaen.

De to gruppene får følgende resultater:

De som har gått i musikkterapi:

13, 16, 18, 9, 11, 17, 12, 22, 21, 10

De som ikke har gått i musikkterapi:

23, 18, 22, 19, 12, 21, 17, 24, 16, 20

Dersom vi bruker en av de formlene vi lærte i kapittel 2, kan vi regne ut gjennomsnittet i de to gruppene. Gjennomsnittlig skåre for psykiske plager blant de som ikke hadde gått i terapi viste seg å være 19,2 (standardavvik 3,61). Blant de som hadde gått i terapi var skåren lavere, nærmere bestemt 14,9 (standardavvik 4,58). Det ser altså ut til at de som har gått i musikkterapi gjennomsnittlig har en lavere forekomst av psykiske plager enn de som ikke har

deltatt. Men er denne forskjellen så stor at vi kan feste lit til dette resultatet? En måte å finne ut av dette på er å signifikansteste forskjellen. Nullhypotesen er at det i populasjonen ikke er noen forskjell mellom gruppene. Den alternative hypotesen sier at det er en forskjell.

Ved hjelp av tallene ovenfor er det nokså enkelt å foreta en t-test for forskjell i gjennomsnitt mellom de to gruppene. Vi setter først tallene inn i formel 3.9 for å regne ut det felles standardavviket.

$$sd = \sqrt{\frac{3,61^2(10-1)+4,58^2(10-1)}{10+10-2}} = \sqrt{\frac{117,2889+188,7876}{18}} = 4,12$$

Deretter setter vi tallene inn i formel 3.8 for å regne ut t-verdien.

$$t = \frac{19,2-14,9}{4,12\sqrt{\frac{1}{10}+\frac{1}{10}}} = \frac{4,3}{4,12*0,4472} = 2,33$$

Antall frihetsgrader er lik summen av observasjoner i de to gruppene minus to, altså 10 + 10 – 2 = 18. Dersom vi slår opp i en tabell over kritiske verdier for t-fordelingen gitt 18 frihetsgrader finner vi følgende verdier:

p<0,05 – 2,101
p<0,01 – 2,878
p<0,001 – 3,922

Den verdien vi har regnet ut (t = 2,334) ligger mellom de to første tallene. Det betyr at vi har oppnådd signifikans på p<0,05-nivået. Det er med andre ord mindre enn fem prosents sannsynlighet for at det ikke er noen forskjell mellom de to gruppene i populasjonen.

En kan kanskje spørre seg hvilken populasjon vi sikter til. Vi har faktisk ikke trukket disse pasienene fra noen virkelig populasjon. I slike tilfeller forestiller vi oss en teoretisk populasjon og trekker slutninger til denne. Dersom de 20 pasientene som inngår i undersøkelsen var trukket fra en slik pasientpopulasjon, hvilke slutninger kunne vi da trekke om denne populasjonen? Det er dette spørsmålet en slik signifikanstesting skal besvare.

Hva vi egentlig har vist ved å gjennomføre denne testen kan en diskutere. Har vi vist at musikkterapien har hatt effekt? Hva slags forklaringer kan vi tenke oss på den forskjellen som er funnet? Det kan for eksempel godt tenkes at det ikke er selve musikkterapien som har virket positivt, men at en kunne oppnådd det samme gjennom andre aktiviteter der en involverte pasientene. For å finne ut av dette, kunne en for eksempel gjennomføre en studie der en hadde med en tredje gruppe av pasienter. Disse pasientene kunne engasjeres i andre aktiviteter. Dersom også disse pasientene fikk signifikant høyere gjennomsnittsskåre på depressvitet enn pasientene i musikkterapigruppen, kunne en trekke noe sikrere konklusjoner fra studien. En ville da ha fått en sterkere bekreftelse på at det var selve musikkterapien som var utslagsgivende.

En slik analyse som vi har regnet ut ovenfor kan vi også lage ved bruk av SPSS. Vi må først legge inn alle tallene for psykiske plageskårer i en kolonne og opplysninger om hvilken gruppe hver pasient tilhører i en annen kolonne. Deretter går vi inn i *Analyse*, *Compare Means*, og *Independent Samples T-Test*. Så legger vi inn variabelen som inneholder informasjon om plageskåren som *Test Variable* og variabelen som inneholder informasjon om hvilken gruppe pasienter det dreiet seg om som *Grouping Variable*. Når du deretter klikker på *OK*, får du ut den utskriften som er gjengitt i Fig. 3.5.

Fig. 3.5: Gjennomsnittlig plageskår i to grupper av pasienter og t-test for uavhengige grupper

Group Statistics

	Gått i musikkterapi?	N	Mean	Std. Deviation	Std. Error Mean
Plageskår	ja	10	14,9000	4,58136	1,44875
	nei	10	19,2000	3,61478	1,14310

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
Plageskår	Equal variances assumed	1,433	,247	-2,330	18	,032	-4,30000	1,84541	-8,17707	-,42293
	Equal variances not assumed			-2,330	17,076	,032	-4,30000	1,84541	-8,19217	-,40783

Vi kjenner igjen antall observasjoner (10 i hver gruppe), gjennomsnittstallene og standardavvikene i den øverste delen av tabellen. Vi kjenner også igjen t-verdien vi har regnet ut (første linje i nederste tabelldel – equal variances assumed). Vi ser at den har fått negativt fortegn i utregningen som SPSS har gjort. Fortegnet avhenger av hvilken gruppe som får høyest kodeverdi. Dersom vi hadde kodet de pasientene som ikke hadde gjennomgått musikkterapi med et ett-tall og de som hadde gjennomgått musikkterapi med et to-tall, ville t-verdien fått et positivt fortegn. Dette har ingen betydning for den absolutte tallverdien. Vi ser også at SPSS har slått opp i tabellen for oss, og funnet ut at med 18 frihetsgrader tilsvarer en t-verdi på 2,33 en p-verdi på .032. Dette tallet er lavere enn 0,05 og høyere enn 0,01. Med andre ord er gruppeforskjellen signifikant på $p<0,05$-nivået, men ikke på $p<0,01$-nivået. Dette stemmer altså med konklusjonen ovenfor.

Bruken av en t-test på forskjeller i gjennomsnitt mellom to grupper bygger på bestemte forutsetninger. For det første må den avhengige variabelen være metrisk. For det andre må observasjonene i de to gruppene være uavhengige. Det kan for eksempel ikke være slik at det er de samme individene som er testet to ganger. Disse to første forutsetningene er nokså

grunnleggende. For det tredje må observasjonene i de to populasjonene som gruppene representerer være tilnærmet normalfordelte. For det fjerde må de to populasjonene ha samme varians. Denne siste forutsetningen blir i SPSS automatisk testet ved bruk av Levenes test (se Fig. 3.5). Siden p-verdien på Levenes test i dette tilfellet var så høy som 0,247, har vi ikke vist at variansen i de to gruppene er signifikant forskjellig, og vi kan derfor anvende t-testen slik vi har gjort. Dersom Levene's test hadde vist signifikans, måtte vi ha benyttet en litt annen variant av t-testen for forskjeller mellom to gruppegjennomsnitt. I Fig. 3.4 er den gjengitt på siste linje (equal variances not assumed).

Det er forsket en hel del på hva som skjer når en bryter med forutsetningen om normalitet. En har funnet at t-testen for to uavhengige grupper er ganske robust. Dersom en har et antall observasjoner på minst 30 i hver gruppe, kan en være nokså trygg på at testen fungerer som den skal, selv om variablene avviker nokså sterkt fra normalitet. Dersom utvalgene er små og fordelingene i de to gruppene er sterkt skjeve i hver sin retning, risikerer en imidlertid at testen fungerer dårlig og at resultatene blir upålitelige. Dersom en er i tvil om en kan bruke t-testen for sammenlikning av gjennomsnitt i to uavhengige grupper, kan en i stedet benytte en ikke-parametrisk test som tester omtrent det samme. Den ikke-parametriske testen baserer seg ikke på noen forutsetning om normalitet eller lik varians. For å teste forutsetningen om normalitet kan en benytte Kolmogorov-Smirnov-testen eller Shapiro-Wilk-testen. Shapiro-Wilk-testen regnes for å være noe mer nøyaktig enn Kolmogorov-Smirnov-testen. Et godt ikke-parametrisk alternativ til t-testen som er beskrevet ovenfor er Mann-Whitney-testen (Siegel, 1956; Siegel & Castellan, 1988; Sheskin, 1997).

Tidligere i dette kapittelet har vi lært om korrelasjoner. De uttrykker styrken på sammenhengen mellom to variabler, forutsatt at sammenhengen er lineær (at observasjonene i punktdiagrammet best kan beskrives langs en rett linje). Signifikanstestingen av forskjellen mellom to gjennomsnittene i de to uavhengige gruppene ovenfor sier bare at det er lite sannsynlig at en så stor forskjell skyldes tilfeldigheter. Men den sier ikke noe om hvor sterk sammenhengen er. Jacob Cohen (1988), som er kjent for å ha utviklet en hel rekke forskjellige koeffisienter til å måle styrken på sammenhengen mellom variabler, har foreslått en koeffisient han har kalt *d*, med andre ord Cohens *d* (Se formel 3.10).

$$d = \frac{\bar{X}_1 - \bar{X}_2}{sd} \qquad (3.10)$$

d Cohens d

$\bar{X}_1$ Gjennomsnittet i gruppe nr. 1

$\bar{X}_2$ Gjennomsnittet i gruppe nr. 2

sd Det felles standardavviket for de to gruppene (jfr. formel 3.9)

Setter vi tallene fra eksempelet ovenfor inn i denne formelen får vi følgende:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{sd} = \frac{19,2 - 14,9}{4,12} = 1,04$$

Cohen har sagt at en *d* på 0,20 er liten, en *d* på 0,50 kan regnes som middels effekt, mens en *d* på 0,80 eller større må regnes som en høy effektstørrelse. Siden tallet under streken er et estimat av gruppenes felles standardavvik, kan vi si at Cohens *d* måler styrken på sammenhengen (eller forskjellen mellom gruppene) i antall standardavvik. En *d* på 1,04 er med andre ord en effektstyrke på så vidt mer enn ett standardavvik.

Et annet assosiasjonsmål som kan brukes i denne situasjonen (en metrisk variabel og en dikotomi) er eta (ή). Eta i eksempelet ovenfor viser seg å bli 0,481. Dersom en kvadrerer eta, får en vite hvor mye av variansen i den metriske variabelen som forklares av den dikotome variabelen. Eta kvadrert er i dette tilfellet 0,232. Den forklarte variansen er med andre ord 23.2 prosent.

Eta kan regnes ut ved å gå inn i *Analyze*, *Compare Means*, *Means*, og deretter legger en inn den metriske variabelen i *Dependent List* og dikotomien i *Independent List*. Trykk deretter *Option* og klikk sett inn hake i *Anova Table and Eta*. Trykk så på *Continue* og *OK*. Den delen av utskriften som viser eta er gjengitt i Fig. 3.6.

Fig. 3.6: Eta-koeffisienten fra en SPSS-utskrift.

Measures of Association

	Eta	Eta Squared
Plageskår * Gått i musikkterapi?	,481	,232

Blant de øvrige tabellene som vises i utskriften (ikke vist her), er det verdt å legge merke til signifikanstestingen av eta. Den er basert på en famile av fordelinger som kalles F-fordelinger. Egentlig er denne testen noe som kalles en enveis variansanalyse, som vi kommer tilbake til senere i denne teksten. Det viser seg at resultatet av den testen er identisk med resultatet av den testen vi har brukt for å teste forskjellen mellom to gjennomsnitt. P-verdiene er nøyaktig like. Det kan vises matematisk at testene egentlig er identiske.

Tidligere har vi lært om produkt-moment-korrelasjonen. Dersom vi beregner denne korrelasjonen mellom de to variablene i eksempelet ovenfor (gruppe og skåre på psykiske plager), får vi resultatet som er vist i Fig. 3.7.

Det er interessant å legge merke til at korrelasjonen blir nøyaktig den samme som når vi regner ut eta. Dessuten ser vi at p-verdien når en signifikanstester korrelasjonen er identisk med p-verdien vi fikk når vi t-testet forskjellen mellom de to gjennomsnittene. Når en bruker Pearons r på sammenhengen mellom en metrisk variabel og en dikotomi, kalles den koeffisienten en regner ut en punkt biseriell korrelasjon (Point biserial correlation).

Fig. 3.7: Produkt-moment-korrelasjonen mellom en metrisk variabel og en dikotomi.

Correlations

		Plageskår	Gått i musikkterapi?
Plageskår	Pearson Correlation	1	,481*
	Sig. (2-tailed)		,032
	N	20	20
Gått i musikkterapi?	Pearson Correlation	,481*	1
	Sig. (2-tailed)	,032	
	N	20	20

*. Correlation is significant at the 0.05 level (2-tailed).

3.3.2 Forskjell i gjennomsnitt mellom to korrelerte grupper på en metrisk variabel

Noen ganger ønsker en å se på forskjeller mellom to gjennomsnitt der de samme personene er målt to ganger, eller der de to settene av observasjoner på den metriske variabelen på annen måte henger sammen parvis. Dersom en har testet en gruppe ektefeller på en og samme test, er en kanskje interessert i å se om det er kvinnene eller mennene som har skåret høyest. I slike situasjoner bruker en t-testing av forskjeller i gjennomsnitt for korrelerte data (noen ganger kalt avhengige data, parrede data eller matchede data).

Dersom en har gjort målinger for par av observasjoner, slik som beskrevet ovenfor, kan en regne ut en differanse for hvert par. Videre kan en beregne standardavviket til denne tallrekken av differanser. Ved hjelp av disse tallene kan en regne ut en t-verdi slik som vist i formel 3.11. Antall frihetsgrader er lik antall par minus en (n-1).

$$t = \frac{\overline{D}}{\frac{sd_D}{\sqrt{n}}} \qquad (3.11)$$

$\overline{D}$ Gjennomsnittet av alle differansene

sd_D Standardavviket til alle differansene

n Antall par av observasjoner

Vi skal på nytt ta utgangspunkt i pasienter som deltar i musikkterapi. Denne gangen har vi undersøkt bare en enkelt gruppe. Til gjengjeld har vi testet pasientene på en depresjonsskala både før og etter seks uker med terapisesjoner. Vi skal derfor gjøre en analyse der vi

sammenlikner deres gjennomsnittlige skåre før de fikk terapi med skåren de fikk etter å ha fullført terapien. Skårene så slik ut:

Før terapien:	Etter terapien:	Differanse:
15	13	2
21	16	5
24	18	6
14	09	6
16	11	5
15	17	-2
09	12	-3
22	22	0
20	21	-1
17	10	7

Ved å bruke formler vi lærte i kapittel 2, kan vi regne ut gjennomsnittet og standardavviket til differansene. Vi får da 2,40 og 3,6575. Ved å sette tallene inn i formel 3.11, kan vi regne ut at t-verdien er

$$t = \frac{\overline{D}}{\frac{sd_D}{\sqrt{n}}} = \frac{2,40}{\frac{3,6575}{\sqrt{10}}} = 2,075$$

Så slår vi opp i en t-tabell på den rekken som viser kritiske verdier ved 9 frihetsgrader. De kritiske verdiene er da som følger:

p<0,05 – 2,262
p<0,01 – 3,250
p<0,001 – 4,781

Vi ser at t-verdien vi har regnet ut er mindre enn det kritiske tallet for p<0,05-nivået. Vi må derfor konkludere med at vi ikke har oppnådd statistisk signifikans.

La oss så utføre den samme analysen ved bruk av SPSS. Vi går da inn i *Analyze*, *Compare Means*, *Paired Samples T-Test* og legger inn de to variablene som inneholder skårene fra pretesten og posttesten i feltet under *Paired Variables*. Når vi deretter trykker på *OK*, får vi ut den utskriften som vises i Fig. 3.8.

Den øverste tabellen viser gjennomsnitt, antall observasjoner, standardavvik og standardfeilen til gjennomsnittet for pretest og posttest. Den andre tabellen viser korrelasjonen mellom skåren på pretest og posttest, antall par av observasjoner og signifikansen til korrelasjonen. Den tredje tabellen viser gjennomsnittet av differansenen mellom pretest og posttest, standardavviket til differansene, standardfeilen til gjennomsnittet av differansene samt et 95

prosents konfidensintervall for gjennomsnittet av differansene. Vi ser at konfidensintervallet omfatter tallet 0,0, med andre ord tallet som symboliserer ingen endring i gjennomsnittet. Vi kan derfor forvente at endringen ikke er signifikant. Og det bekreftes av p-verdien som vises helt til høyre i nederste tabell (p=0,068). Analysen bekrefter altså igjen at forskjellen mellom gjennomsnittsskår før terapien og etter terapien ikke var signifikant.

Fig. 3.8: T-test for forskjeller mellom to gjennomsnitt ved parrede observasjoner. SPSS.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Plageskår før terapien	17,3000	10	4,47338	1,41461
	Plageskår etter terapien	14,9000	10	4,58136	1,44875

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Plageskår før terapien & Plageskår etter terapien	10	,674	,033

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	Plageskår før terapien - Plageskår etter terapien	2,40000	3,65756	1,15662	-,21646	5,01646	2,075	9	,068

Et mål på effektstørrelse er vist i formel 3.12. Effektstørrelsen er lik gjennomsnittet av alle differansene dividert med det estimerte standardavviket av differanseskårene i populasjonen.

$$d = \frac{\overline{D}}{sd_D} \qquad (3.12)$$

$\overline{D}$ Gjennomsnittet av alle differansene

sd_D Standardavviket til differansene (estimert for populasjonen)

Reglene for tolkning av dette effektmålet er de samme som for sammenlikning av to uavhengige grupper. Liten effektstørrelse er definert som 0,20, middels som 0,50 og stor som 0,80 (Cohen, 1988).

I dette delkapittelet har vi begrenset oss til å se på sammenhengen mellom metriske variabler og kategorielle variabler når de metriske er dikotome. Sammenhengen mellom metriske variabler og polytomier (mangekategorielle variabler) har vi ikke kommet inn på. Dette fordi slike analyser åpner opp en ny verden av statistiske teknikker, nemlig variansanalyser. Variansanalyse er en famlie av statistiske teknikker som blant annet er svært nyttig i forbindelse med randomiserte, kontrollerte forsøk der en ser på virkningene av behandling. Temaet variansanalyse skal derfor tas opp i et eget kapittel.

3.4 Assosiasjonsmål for kategorielle variabler

Vi har i dette kapittelet sett at det finnes assosiasjonsmål for to metriske variabler (Pearsons produkt-moment korrelasjon) og for sammenhengen mellom en metrisk variabel og en dikotomi (Cohens *d* og eta). Når det gjaldt sammenhengen mellom to kategorielle variabler, som vi behandlet aller først i dette kapittelet, sa vi ingen ting om assosiasjonsmål. Dette temaet unngikk vi fordi det er en nokså komplisert sak. Og vi kunne med god samvittighet vente fordi vi klarte oss rimelig godt ved å beskrive sammenhengene ved bruk av forskjeller i prosenttall.

Men nå er det på tide å gjøre rede for assosiasjonsmål for kategorielle variabler. Det finnes et stort antall slike. Ulike assosiasjonsmål for krysstabeller gir ofte høyst ulike tallverdier. En må ha god kjennskap til egenskapene til assosiasjonsmålene for å bruke disse på en fornuftig måte.

Galtung (1967) summerer opp hvilke krav som bør stilles til et assosiasjonsmål for krysstabeller. Han nevner følgende:

- Det bør bli 0 (null) når variablene er uavhengige

- Tallet bør bli maksimalt stort når det foreligger maksimal avhengighet

- For ordinalt målenivå og høyere bør det angis retning

- Størrelsen bør (forutsatt at variablene har retning) variere fra -1 via 0 til +1. Sammenhenger mellom variabler der minst den ene er uten retning, bør variere mellom 0 og +1.

- Tallverdien bør være enkel å fortolke.

- Størrelsen på koeffisienten bør være uavhengig av antall observasjoner som inngår i analysen.

- Målet bør ha en kjent samplingfordeling, slik at en kan regne ut konfidensintervall og utføre signifikanstesting.

Noen av de viktigste assosiasjonsmålene for krysstabeller er følgende:

- **Phi**
  en Kji-kvadrat-basert koeffisient beregnet på 2x2-tabeller

- **Cramers V**
  en Kji-kvadrat-basert koeffisient beregnet på større tabeller
- **Yules Q**
  et "reduksjon i feil ved prediksjon"-mål for 2x2-tabeller
- **Goodman & Kruskals Gamma**
  et "reduksjon i feil ved prediksjon"-mål for større tabeller på ordinalnivå

Johan Galtung (født i Oslo i 1930) er i dag mest kjent som fredsforsker og for å ha grunnlagt Fredsforskningsinstituttet i Oslo (PRIO). Det som kanskje er mindre kjent er hans interesse for matematikk og statistikk. Hans bok om *Theory and method of social research* inneholder blant annet en interessant og innsiktsfull gjennomgang av assosiasjonsmål for krysstabeller. Fra 1969 til 1977 var han professor i fredsforskning ved Universitetet i Oslo. Han har senere hatt professorater ved mange universiteter i mange land, blant annet Santiago, Chile, der han jobbet med sin metodebok, i Geneve, og ved de amerikanske universitetene Columbia, Princeton og University of Hawaii. I sin biografi (Johan uten land) sier han at det bare er en ting han angrer på her i livet, nemlig at han ikke tidligere dro fra Norge. Han har bevart sin interesse for matematikk. Så sent som i august 2007 arrangerte han et seminar om matematikk og fred (Mathematics of, by and for peace) i Jondal, Hardanger.

3.4.1 Fra Pearsons r til assosiasjonsmål for 2x2-tabeller

Det er egentlig et ganske nært slektskap mellom produkt-moment-korrelasjonen og assosiasjonsmålene som blir brukt når en skal si noe om sammenhengen mellom to dikotome variabler. Dersom vi ser nøyere på assosiasjonsmål for 2x2-tabeller, vil vi oppdage at formlene ofte ser slik ut:

$$Koeff = \frac{bc - ad}{z} \qquad (3.13)$$

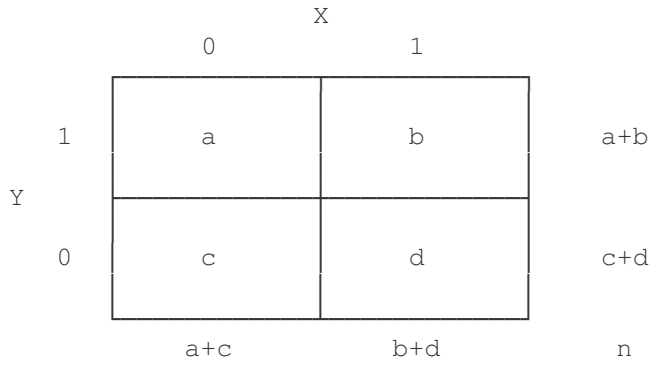
Koeff Et generelt uttrykk for assosiasjonsmål for 2x2-tabeller

a, b, c, og d Antall observasjoner i cellene a, b, c og d

z Et uttrykk som er konstruert med tanke på å gi koeffisienten gode egenskaper

Bokstavene a,b,c og d refererer til antall observasjoner i hver av cellene i en tabell som ser slik ut:

Fig. 3.9: Notasjon for firefelts-tabeller



Størrelsen z varierer fra assosiasjonsmål til assosiasjonsmål, og har til hensikt (på samme måte som nevneren i produkt-moment-korrelasjonen) å gi koeffisienten mest mulig fornuftige egenskaper.

På samme måte som for Pearsons produkt-moment-korrelasjon (beskrevet tidligere i dette kapittelet) bidrar alle observasjoner i nederste venstre og øverste høyre kvadrat til en positiv korrelasjon (uttrykket bc). Alle observasjoner i nederste høyre og øverste venstre kvadrat bidrar negativt (ad). Uttrykkene bc og ad kalles kryssproduktene i firefeltstabellen. Telleren i formelen for mange av assosiasjonsmålene for firefeltstabeller inneholder med andre ord kryssprodukter. Det samme var tilfelle med produkt-moment-korrelasjonen.

Det er imidlertid en viktig forskjell mellom de to formlene (formelen for Pearsons r og den generelle formelen for assosiasjonen i en firefeltstabell). Når vi skal regne ut en produkt-moment-korrelasjon, betyr en observasjons plassering innen hver kvadrat en hel del. Når vi har laget en krysstabell, teller alle observasjonene i hver celle (kvadrat) likt. Ved å forenkle til en 2x2-tabell (eller sagt på en annen måte; ved å dikotomisere eller dele observasjonene på hver variabel inn i to grupper), mister vi med andre ord en hel del detaljer og nyanser. Eksempelet tjener likevel til å vise den sterke likheten mellom produkt-moment-korrelasjonene og assosiasjonsmål for 2x2-tabeller.

Det er vanlig å skille mellom to hovedgrupper av assosiasjonsmål for krysstabeller. En gruppe av disse er basert på chi-kvadrat-verdier og har nært slektskap med Pearsons produkt-moment-korrelasjon. Disse assosiasjonsmålene egner seg godt for situasjoner der den ene eller begge variablene er nominalvariabler. Den andre gruppen baserer seg på en tankegang om sannsynligheten for å gjette riktig dersom en tar utgangspunkt i det en vet om den ene variabelen. Disse kan ofte passe for ordinalvariabler. Vi skal her først ta for oss de chi-kvadrat-baserte assosiasjonsmålene.

3.4.2 Fra dikotomier til polytomier (kategorielle variabler med flere enn to verdier)

Chi-kvadrat-baserte assosiasjonsmål[2]

Tidligere har vi lært noe om å signifikansteste ved bruk av Chi-kvadrat-fordelingen. Vi lærte at vi regner ut chi-kvadrat-verdien ved å addere sammen de kvadrerte differansene mellom observerte og forventede frekvenser i en frekvenstabell etter å ha delt på forventet frekvens. Et eksempel på en tabell med observerte frekvenser og deretter samme tabell med forventede frekvenser er gjengitt i Fig. 3.10. Tabellen over forventede frekvenser er beregnet på grunnlag av tabellen over observerte frekvenser med utgangspunkt i formel 3.1. Hvis vi beregner Cramers V eller et annet rimelig bra assosiasjonsmål for krysstabeller med utgangspunkt i tabellen over forventede frekvenser, vil vi finne at resultatet blir en sammenheng på nøyaktig 0,00. Tabellen representerer med andre ord nullhypotesen. Spørsmålet vi undersøker ved χ^2 -testingen er om de observerte frekvensene er forskjellige nok fra de forventede frekvensene til at null-hypotesen kan forkastes.

Med utgangspunkt i differansene mellom de to tabellene regner vi ut tabellens χ^2 -verdi. For hver celle i tabellen kvadrerer vi avvikene mellom observerte og forventede frekvenser og deler på forventet frekvens. Deretter adderer vi alle disse tallstørrelsene. Resultatet av denne utregningen er, som vi har lært tidligere, χ^2 -verdien. For tabellen som er vist i Fig. 3.10 blir χ^2-verdien 24,603.

En del av assosiasjonsmålene for krysstabeller baserer seg på chi-kvadrat-verdien. Ett av de mest brukte assosiasjonsmålene er, som tidligere nevnt, phi. Formelen for phi ser slik ut:

$$\varphi = \sqrt{\frac{\chi^2}{n}} \qquad (3.14)$$

φ Phi-koeffisienten (mål for assosiasjon i en 2x2-tabell)

n Antall observasjoner i tabellen

[2] Det finnes flere chi-kvadrat-baserte assosiasjonsmål for krysstabeller enn phi og Cramers V som vi altså presenterer her. Ett av disse er Kontingens-koeffisienten, et annet er Tschuprow's T-koeffisient. For spesielt interesserte kan vi henvise til Galtung (1967) eller Liebtrau (1983).

Fig. 3.10: Observerte og forventede frekvenser i en krysstabell

f_o

10	20	30	60
25	10	5	40
35	30	35	100

f_e

21	18	21	60
14	12	14	40
35	30	35	100

Med utgangspunkt i en firefeltstabell der antall observasjoner i cellene er symbolisert med bokstavene a, b, c og d, slik som vist ovenfor, kan vi regne ut phi ved formelen:

$$\varphi = \frac{(bc - ad)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \qquad (3.15)$$

φ Phi-koeffsienten (mål for assosiasjon i en 2x2-tabell)

a, b, c, d Antall observasjoner i hver av feltene i tabellen (se Fig. 3.11)

Her kjenner vi straks igjen uttrykket i telleren, som er det samme som ble brukt ovenfor for å illustrere likheten mellom en produkt-moment-korrelasjon og et assosiasjonsmål for en 2x2-tabell.

Det forunderlige er at dersom vi beregner en vanlig produkt-moment-korrelasjon for en 2x2-tabell, får vi alltid den samme tallverdien som når vi regner ut phi. Det er nokså lett å vise matematisk at dette alltid er tilfelle. Når vi regner ut assosiasjonsmål for en 2x2-tabell er altså phi identisk med Pearsons r. Dette understreker ytterligere likheten mellom assosiasjonsmålene på tvers av variablenes målenivå. Phi-koeffisienten har den uheldige egenskap at når tabellene har flere enn to rader eller flere enn to kolonner blir maksimumsverdien større enn 1,0.

Når vi har rent kategorielle variabler på nominalnivå, men antall rader og/eller kolonner er større enn 2, kan vi, som tidligere nevnt, benytte et assosiasjonsmål som er nært beslektet med phi, nemlig Cramers V (noen ganger kalt Cramers phi). Formelen for Cramers V ser slik ut:

$$V = \sqrt{\frac{\chi^2}{(n)(df_{\min})}} \qquad (3.16)$$

V Cramers V

χ^2 Chi-kvadrat-verdien fra krysstabellen

n Antall observasjoner i tabellen

$df_{\min}$ Det minste tallet av antall rader og antall kolonner i tabellen

Dersom vi har en tabell med bare to kolonner eller bare to rader, ser vi at formelen blir identisk med formelen for phi. Uttrykket df_{min} blir i slike tilfeller nemlig alltid lik 1. Phi er en Cramers V for en tabell med enten to kolonner eller to rader eller 2x2-tabeller. For tabellen vist i Fig. 3.10 blir Cramers V = 0,496 eller omtrent 0,50. Cramers V har en maksimums-størrelse på 1,00 også når tabellene har flere enn to rader og kolonner.

3.4.3 Fra dikotomier til ordinalvariabler

Redusert sannsynlighet for å gjette feil ved prediksjon

Youles Q er en av de eldste målene for assosiasjon mellom to dikotome variabler. G. Udny Yule publiserte en artikkel om dette assosiasjonsmålet i 1912 (Yule, 1912). Bokstaven Q er brukt til ære for Quételet, en av de store statistikerne i det 19 århundre. Vi tar utgangspunkt i en 2x2-tabell og betegner cellene med de vanlige symbolene.

Vi tenker oss at tabellen er arrangert slik at høy verdi på variabelen X er til høyre og høy verdi på variabelen Y er oppover. Dette er symbolisert ved verdiene 0 og 1. Dette betyr at alle observasjonene som havner i a og d er inkonsistente (høy verdi på den ene variabelen er kombinert med lav på den andre) mens b og c er konsistente (høy på begge eller lav på begge). Tallene på ytterkantene av tabellen (a+b, c+d, a+c og b+d) kalles marginalfordelingene.

Formelen for Yules Q kan da uttrykkes slik:

$$Q = \frac{bc - ad}{bc + ad} \qquad (3.17)$$

Q Yules Q

a, b, c, d Antall observasjoner i cellene i en firefeltstabell (se Fig. 3.9)

I formlene for andre assosiasjonsmål vil vi ofte kjenne igjen uttrykkene (bc) og (ad). Disse kalles, som vi har vært inne på tidligere, kryssprodukter. Ved bruk av dette ordet kan vi faktisk formulere formelen verbalt på en enkel måte. "Yules Q er lik differansen mellom kryssproduktene delt med summen av kryssproduktene." Når kryssproduktene er like, blir korrelasjonen 0 (null). Når det første kryssproduktet er størst, blir korrelasjonen positiv. Når det andre kryssproduktet er størst, blir korrelasjonen negativ.

Vi ser at Yules Q blir pluss 1,0 når a, d eller begge er lik 0. I et slikt tilfelle blir det siste kryssproduktet lik 0 (null), og alt som blir igjen av formelen er (bc)/(bc) = 1,0. Tilsvarende blir Yules Q lik minus 1,0 når b, c eller begge disse er lik 0 (null). Når det første kryssproduktet blir lik 0 (null), er alt som blir igjen av formelen uttrykket -(ad)/(ad) = -1,0.

En viktig egenskap ved Yules Q er at den ikke er avhengig av marginalfordelingene. Dette betyr at dersom vi sammenlikner en egenskap hos to grupper, f.eks. andel røykere blant et representativt utvalg fra hele befolkningen med andel røykere fra en gruppe innsatte i fenglser, er sammenhengen uavhengig av hvor mange vi har i de to utvalgene.

James Davis (1971) har sett nærmere på hvor meningsfylt Yules Q er som assosiasjonsmål. Han hevder at i 40 år forble Yules Q en statistisk størrelse som ikke hadde en klar fortolkning. Med Goodman & Kruskals artikkel fra 1954 om assosiasjonsmål for krysstabeller, fikk imidlertid Yules Q en klar mening. En Yules Q-verdi på z betyr at vi gjør det z bedre enn rent tilfeldig når vi skal predikere den ene variabelen fra den andre og anvender en regel om at 0 på den ene skal bety 0 på den andre og 1 på den ene betyr 1 på den andre.

Formelen for utregning av standardfeilen til samplingfordelingen til Q er gitt nedenfor (formel 3.18).:

$$SE_Q = \sqrt{\frac{(1,00 - Q^2)^2 * \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)}{4}} \qquad (3.18)$$

SE_Q Standardfeilen til Yules Q

Q Yules Q

a, b, c, d Antall observasjoner i cellene i tabellen

Dersom vi ønsker å regne ut et 95-prosents konfidensintervall for Yules Q, bruker vi samme formelen som vi har brukt flere ganger tidligere. Vi multipliserer standardfeilen med 1,96 og trekker dette tallet ifra samt adderer det til Q-verdien.

$$CI_{Q95} = Q \pm 1,96 * SE_Q$$

Yules Q er laget for 2x2-tabeller. Ofte er tabellene våre mye større, de har flere enn to rader og/eller flere enn to kolonner. For slike større tabeller, forutsatt at begge variablene er på

minst ordinalnivå, finnes det et assosiasjonsmål som er beslektet med Yules Q, nemlig Goodman - Kruskals Gamma. For å regne ut Gamma må vi først tenke oss at vi har laget en lang rekke firefelts-tabeller, firefelts-tabeller som er mulig å konstruere på grunnlag av en større tabell ved å slå sammen celler. Dette skal gjøres etter bestemte regler. På grunnlag av tabellene teller vi opp konsistente og inkonsistente observasjoner og setter disse inn i formelen for Yules Q. Resultatet er Gamma. Goodman & Kruskals Gamma er med andre ord en Yules Q som er generalisert til større tabeller. Yules Q er Gamma for en 2x2-tabell.

Formelen for Gamma kan skrives på følgende generelle måte (3.19):

$$\lambda = \frac{KP - IP}{KP + IP} \qquad (3.19)$$

λ Goodman - Kruskals gamma-koeffisient

KP Kryssproduktet for konsistente par

IP Kryssproduktet for inkonsistente par

Yules Q og Goodman-Kruskals Gamma er enkle, nært beslektede og lett fortolkbare assosiasjonsmål, og Gamma tar altså hensyn til ordinal-informasjonen i data. Tradisjonelt har imidlertid to andre assosiasjonsmål for ordinalvariabler vært mye benyttet, nemlig Spearmans Rho og Kendalls Tau.

Fig. 3.11: Yngre og eldres syn på countrymusikk.

Alder * Liker countrymusikk? Crosstabulation

			Liker countrymusikk?		
			Ja	Nei	Total
Alder	-49 år	Count	10	90	100
		% within Alder	10,0%	90,0%	100,0%
	50 år+	Count	30	70	100
		% within Alder	30,0%	70,0%	100,0%
Total		Count	40	160	200
		% within Alder	20,0%	80,0%	100,0%

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Phi	-,250			,000
	Cramer's V	,250			,000
Ordinal by Ordinal	Gamma	-,588	,130	-3,651	,000
N of Valid Cases		200			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

La oss se på phi og Yules Q for en 2x2-tabell fra SPSS. Vi tenker oss at vi har intervjuet 100 personer som er 49 år eller yngre og 100 personer som er 50 år eller eldre om hva de synes om country-musikk. Vi har gjort det så enkelt at vi har bedt alle sammen om å svare ja eller nei. Resultatet vises i Fig. 3.11.

Første del av tabellen viser at blant de som er 49 år eller yngre er det 10% som liker countrymusikk. Blant de som er 50 eller eldre er tallet 30%. Dette er en relativt kraftig sammenheng. Oppslutningen om countrymusikk er tre ganger så stor i den eldste av de to gruppene. Hva så med de to koeffisientene vi er interessert i? Det første vi legger merke til er at vi ikke har fått ut noen Yules Q i det hele tatt. Derimot har vi fått ut Goodman-Kruskals Gamma, som egentlig er en koeffisient for tabeller som er større enn en 2x2-tabell. Dersom vi kontrollregner, slik det er gjort nedenfor, vil vi se at tallet vi har fått er identisk med det vi får når vi bruker formelen for Yules Q. Og siden Yules Q er å betrakte som et spesialtilfelle av gamma, kan vi se på gamma-verdien fra Fig. 3.11 som en Yules Q. I tabellen vises også en standardfeil og en signifikanstest av gamma.

$$Q = \frac{bc - ad}{bc + ad} = \frac{90*30 - 10*70}{90*30 + 10*70} = \frac{2700 - 700}{2700 + 700} = \frac{2000}{3400} = 0,588$$

Selv om det ikke går fram av tabellene som er vist i Fig. 3.11, kan vi opplyse om at χ^2-verdien til denne tabellen er lik 12,50. Dette gir oss en mulighet til å regne ut phi-koeffisienten på en enkel måte. Vi setter bare tallene inn i formel 3.14 slik som vist nedenfor:

$$\varphi = \sqrt{\frac{\chi^2}{n}} = \sqrt{\frac{12,5}{200}} = \sqrt{0,0625} = 0,25$$

Vi ser at dette stemmer med det som er vist i siste tabellen i Fig. 3.11. Tabellen viser her både en phi-koeffisient og en Cramers V. Det hadde vært mer korrekt om tabellen bare hadde vist en phi-koeffisient, for den er egentlig det samme som en Cramers V for en 2x2-tabell.

Også for 2x3 og 3x2-tabeller blir phi og Cramers V identiske. For større tabeller blir de forskjellige. For tabeller som er større enn en 2x2-tabell gir det bare mening å bruke Goodman-Kruskals gamma dersom relasjonen mellom de to variablene er monoton over kategorier.

Spearmans ρ (rho)

Spearmans ρ har flere svakheter som gjør den dårlig egnet som assosiasjonsmål. Den er avledet av formelen for produkt-moment-korrelasjonen og forutsetter at en er villig til å behandle en rangvariabel som en intervallvariabel. Galtung (1967) påpeker flere andre svakheter ved ρ, f.eks. at den ikke blir null i situasjoner der den burde bli lik null. ρ er likevel ett av de mest brukte assosiasjonsmålene for rangvariabler, kanskje nettopp fordi den likner så sterkt på en produkt-momentkorrelasjon.

En kan beregne en Spearmans ρ ved å først rangere alle verdiene på hver variabel som skal inngå i analysen, og deretter regne ut den vanlige produkt-moment-korrelasjonen mellom variablene. Det finnes også en egen formel for bereging av ρ. Først rangerer en alle verdiene på begge variablene og kaller differansen mellom rangtallene for par nr. i d_i. Deretter regner en ut slik:

$$\rho_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} \qquad (3.20)$$

ρ_S Spearmans rho

d_i Differansen mellom rangtallene en enhet har på de to variablene

n Antall enheter (antall par av observasjoner)

ρ varierer mellom -1 og +1.

Dersom to eller flere verdier på en av variablene er like, har vi det som kalles "ties". I slike tilfeller gir en gjennomsnittlig rangverdi på de observasjonene som har samme verdi. Dersom vi f.eks. rangerer 10 observasjoner og vi etter å ha rangert de 5 første kommer til tre observasjoner som har samme verdi, får alle disse tre rangtallet 7. Når vi har slike "ties" i data, blir ikke lenger maksimums- og minimumsverdiene av ρ +1 og -1. For å justere for dette brukes en noe mer komplisert formel, som f.eks. er gjengitt i Gibbons (1993).

For å signifikansteste eller regne ut konfidensintervall beregnes standardfeilen til ρ ved følgende enkle formel:

$$SE_\rho = \rho\sqrt{n-1} \qquad (3.21)$$

SE_ρ Standardfeilen til Spearmans rho

ρ Spearmans rho

n Antall enheter (antall par av observasjoner)

Når antall observasjoner er høyt (>30), kan en anta at samplingfordelingen til Rho er tilnærmet normalfordelt. Når antall observasjoner er 30 eller lavere, må en anvende egne tabeller (f.eks. Gibbons, 1993).

Kendalls τ (tau)

Kendalls τ er basert på idéen om konkordans og diskonkordans. Vi kan tenke oss at gitt et visst antall observasjoner, kan disse sammenliknes parvis. Dersom vi har n elementer, er antall par av elementer lik

$$n_{par} = \frac{1}{2} n(n-1) \qquad (3.22)$$

n_{par} Antall par som kan dannes av n elementer

n Antall elementer

Noen ganger skriver en n_{par} som: $\binom{n}{2}$

At et par er konkordant betyr at de er ordnet i samme rekkefølge på de to variablene (f. eks. at case 1 har høyere verdier enn case 2 på begge variablene). At et par er diskonkordant betyr at de er ordnet i motsatt rekkefølge på de to variablene, f.eks. at case 1 er høyere enn case 2 på den ene variabelen og lavere enn case 2 på den andre variabelen. Kendalls τ (tau) er gitt av følgende formel:

$$\tau = \frac{2(C-D)}{n_{par}(n_{par}-1)} = \frac{2C-2D}{{n_{par}}^2 - n_{par}} \qquad (3.23)$$

τ Kendalls tau

C Antall konkordante par

D Antall diskonkordante par

n_{par} Totalt antall par

τ er med andre ord basert på andelen konkordante par minus andelen diskonkordante par eller andel overensstemmende minus andel uoverensstemmende par.

τ har flere uheldige egenskaper, f.eks. at den bare kan bli lik 1,0 under svært spesielle forhold. Den er blitt forsøkt modifisert og er kommet i tre utgaver (τ_a, τ_b og τ_c). De ulike utgavene av τ behandler ties på ulike måter. τ_c har imidlertid oppnådd en slik kompleksitet at den ikke lenger tilfredsstiller kravet om å være enkelt fortolkbar og forståelig. τ_b viser seg i en 2x2-tabell å bli identisk med phi (og dermed også Pearsons produkt-moment-korrelasjon).

3.4.4 Når dikotome variabler representerer metriske variabler

Når den ene variabelen er på intervallnivå og den andre er en dikotomi, er det legitimt å beregne korrelasjonen ved å anvende formelen for en produkt-moment-korrelasjon. Dette kalles, som vi har vært inne på tidligere i dette kapittelet, en punkt-biseriell korrelasjon. Dersom den dikotome variabelen antas å reflektere en underliggende kontinuerlig og normalfordelt variabel, kan en estimere korrelasjonen mellom denne underliggende variabelen og intervallvariabelen. Dette kalles en biseriell korrelasjon (McNemar, 1969).

Fra før vet vi at produkt-moment-korrelasjonen mellom to (genuine) dikotomier er det samme som en phi-koeffisient (og at denne svarer til en Kendalls Tau_b og en Cramers V for 2x2-tabeller. Dersom begge variablene er dikotome og antas å representere underliggende kontinuerlige, normalfordelte variabler, kan vi estimere sammenhengen mellom disse underliggende variablene. Dette kalles tetrakorisk korrelasjon (McNemar, 1969).

Slektskapet mellom de mange ulike assosiasjonsmålene kommer tydelig fram når vi bruker formlene på sammenhengen mellom to dikotomier. I dette tilfellet finner vi at følgende assosiasjonsmål blir identiske:

Cramers V
φ (phi)
Kendalls τ_b (tau-b)
η (eta)
Pearsons r

Dette illustrerer hvor nært slektskap det er mellom en rekke av de mest kjente målene for statistisk assosiasjon. Vi kan også føye til at for sammenhengen mellom to dikotomier kjenner vi igjen uttrykket i telleren (bc-ad) i enda flere assosiasjonsmål: Kendalls Tau_a, Kendalls Tau_c, Goodman-Kruskals lambda, Somers d_{yx}, samt alle de Chi-kvadrat-baserte assosiasjonsmålene.

En nyttig oversikt med angitt slektskap mellom de mest kjente assosiasjonsmålene er presentert i Galtung (1967).

3.5 Kontroll for tredjevariabler

Når vi kommer til den multivariate statistikken, er det mange fenoméner fra den mer elementære statistikken som er viktig å ha klart for seg. Blant disse er kontroll for tredjevariabler og statistisk interaksjon.

Å kontrollere for en tredjevariabel vil si å undersøke hvordan sammenhengen mellom to variabler forandrer seg dersom en tar hensyn til en tredje. Ofte kalles den tredje variabelen for en kovarians-variabel. Dette begrepet stammer fra kovariansanalyse. Vi velger i denne teksten konsekvent å snakke om tredjevariabler. Tredjevariabler kan modifisere sammenhengen mellom to variabler på flere forskjellige måter. Den kan mediere sammenhengen. I noen tilfeller vil vi si at x virker på y gjennom den tredje variabelen (altså z). I så fall kalles den en mediator. Alternativt kan sammenhengen mellom x og y variere i retning og styrke avhengig

av hvor en befinner seg på den tredje variabelen. Dette kalles å moderere sammenhengen, og tredjevariabelen kalles i dette tilfellet for en moderator.

Noen ganger kontrollerer vi for en tredjevariabel fordi vi ønsker å finne ut om en sammenheng mellom to variabler er spuriøs, med andre ord at den skyldes en utenforliggende variabel og at den derfor er triviell eller misvisende og ikke verdt å legge vekt på. I så fall kaller vi denne tredjevariabelen for en konfunderende variabel (confounder). Et ofte brukt eksempel er sammenhengen mellom det å spise is og drukning. Det viser seg at i perioder der mange spiser is, er det også mange som drukner. Det er ikke dermed vist at spising av is øker risikoen for å drukne. Forklaringen er vel heller at i perioder med fint vær er det både mange som spiser is og det er mange som bader. Jo flere som bader, desto større er risikoen for drukningsulykker. La oss ta et annet (tenkt) eksempel. I en større epidemiologisk undersøkelse finner en at de som skårer høyt på depressivitet har en økt risiko for å dø av lungekreft. Det viser seg imidlertid at begge disse variablene er korrelert med sigarettrøyking. De som røyker skårer gjennomsnittlig høyere på depressivitet, og de har en mye høyere risiko for å dø av lungekreft. Når en kontrollerer for sigarettrøyking, forsvinner kanskje hele sammenhengen.

Den enkleste og ofte beste måten å foreta kontroll for en tredjevariabel på når den er kategoriell, er å se på en sammenheng for hver subgruppe på tredjevariabelen.

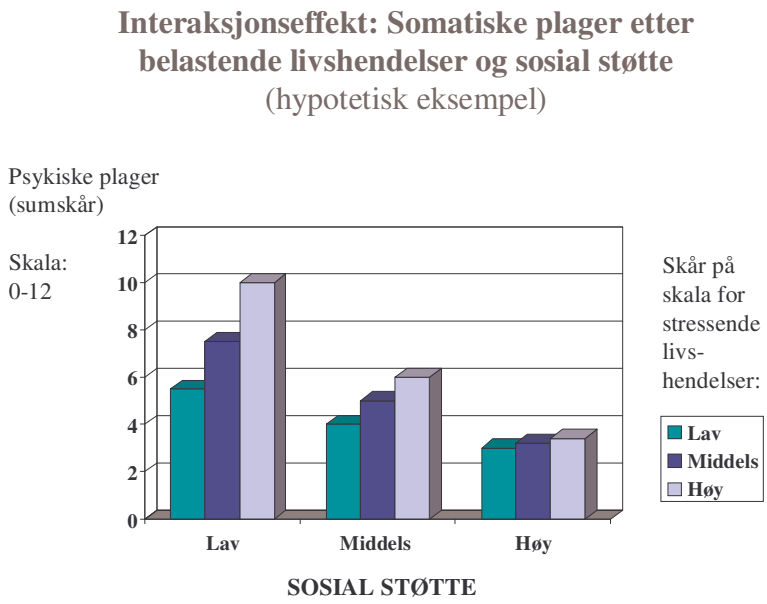
Figurene 3.12 og 3.13 viser et hypotetisk eksempel på en interaksjonseffekt. Vi tar utgangspunkt i den velkjente bufferhypotesen. Den går kort fortalt ut på at personer som opplever mye sosial støtte fra andre mennesker tåler større belastninger eller mer stress enn andre. Sammenhengen mellom belastninger og helse er med andre ord svakere hos de som opplever sitt sosiale nettverk som støttende.

I eksempelet antar vi at vi har målt somatiske hvrdagsplager som for eksempel hodepine, vondt i magen og muskelsmerter forskjellige steder på kroppen. Vi har laget oss en såkalt sumskår (forklares grundigere i neste kapittel). Denne sumskåren er en skala som går fra 0 til 12. En person som har fått skåren 0 (null) har ikke rapportert om noen plager i det hele tatt. En person som har fått tallverdien 12 har rapportert maksimalt med plager. Vi antar at jo høyere skår en person har fått, desto mer plager har vedkommende.

Vi har også målt den enkeltes opplevelse av sosial støtte fra andre. Her har vi også gått veien om å stille en rekke spørsmål. Disse er så addert sammen til en sumskår. Skårene på denne er deretter delt i tre grupper; lav, middels og høy sosial støtte, slik at vi har omtrent like mange i hver gruppe. Når vi analyserer sosial støtte mot plageskåren finner vi at gjennomsnittlig plageskår er lavest blant de som opplever mye sosial støtte, noe høyere blant de som opplever middels sosial støtte, og høyest blant de med lavest grad av sosial støtte.

Endelig har vi målt det som kalles stressende livshendelser. Stressende livshendelser er slikt som å oppleve dødsfall i familien, å bli skilt, å miste jobben etc. Her har vi talt opp antall slike hendelser som er rapportert og delt inn i tre grupper; lav middels og høy. Når vi analyserer stressende livshendelser mot plageskåren, finner vi at gjennomsnittlig plageskår er lavest blant de som opplever færrest stressende livshendelser, høyere blant de som opplever noe flere stressende livshendelser og høyest blant de som har opplevd flest stressende livshendelser.

Fig. 3.12: Kontroll for tredjevariabel - med interaksjon.



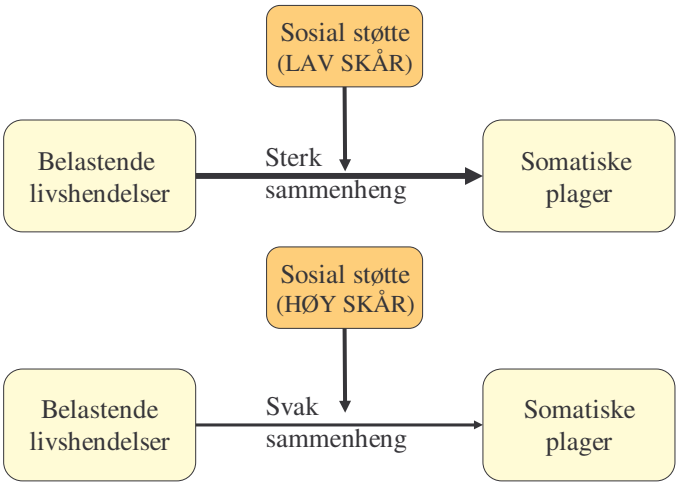
Det er likevel ikke disse to sammenhengene hver for seg vi er mest interessert i. Mest spennende er det å se om sammenhengen mellom stressende livshendelser og plager er avhengig av hvordan en skårer på sosial støtte. Dersom vi ser nøyere på Fig. 3.12 oppdager vi at blant de med lite sosial støtte er det en sterk sammenheng mellom belastende livshendelser og somatiske plager. Blant de med mye sosial støtte er sammenhengen nesten ikke til stede. De med middels sosial støtte kommer i en mellomposisjon. Vi har med andre ord fått bekreftet antakelsen om at der finnes en statistisk interaksjonseffekt. I vårt tilfelle kan den tolkes som en støtte til bufferhypotesen.

I virkelighetens verden har det vist seg vanskelig å finne så klare bekreftelser på buffer-hypotesen. De gangene en finner statistiske interaksjonseffekter viser det seg gjerne at de er ganske svake, eller at de skyldes en litt for enkel og naiv bruk av statistikk.

Statistiske interaksjonseffekter kan framkomme på så mange måter. Her har vi sett på forskjeller i gjennomsnitts-skår mellom ulike grupper av personer. Noen ganger får en fram interaksjonseffekter ved å se på prosenter. Andre ganger ved å bruke korrelasjoner eller regresjonsanalyse.

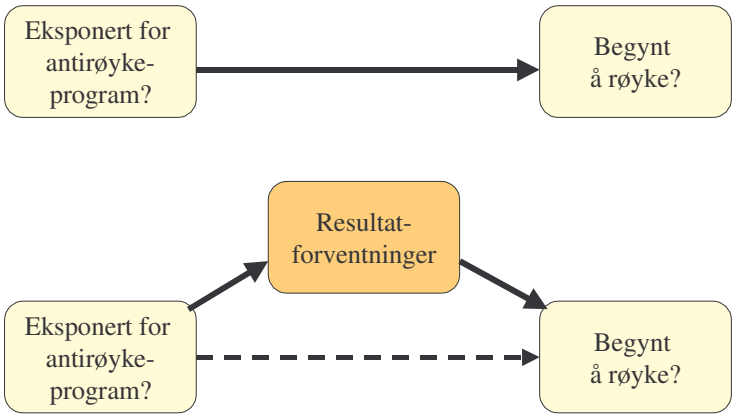
I eksempelet referert ovenfor har vi ikke gjort noe forsøk på å signifikansteste interaksjonseffekten. Vi vet med andre ord ikke om den er sterk nok til at vi med rimelig sikkerhet kan forkaste hypotesen om at det ikke er noen slik interaksjonseffekt i populasjonen (null-hypotesen). Hvordan dette kan gjøres skal vi komme tilbake til i kapittelet om variansanalyse (kapittel 4).

Fig. 3.13: Interaksjonseffekt: Styrken, retningen eller formen på sammenhengen mellom en prediktor (belastende livshendelser) og en kriterievariabel (somatiske plager) avhenger av skår på en tredjevariabel (sosial støtte)



Dette var et eksempel på en interaksjonseffekt. Noen ganger gjennomfører en noe som kalles en additiv kontroll for en tredjevariabel (se Fig. 3.14). Et eksempel på en slik additiv kontroll har vi i et materiale som er samlet inn ved HEMIL-senteret som en del av evalueringen av en skolebasert antirøykekampanje. Kampanjen var ganske vellykket, og det viste seg at blant de elevene som hadde blitt eksponert for tiltaket, var det færre som begynte å røyke. Spørsmålet var om dette kunne forklares på noen måte ved hjelp av alle de faktorene som var undersøkt i datainnsamlingene. En mulig forklaring var at en hadde lyktes i å påvirke elevenes forventninger til røykingen. Hadde en kanskje lyktes i å få elevene til å forstå at røyking har få positive konsekvenser og mange negative? Hva slags forventninger en person har til konsekvensene av egen atferd kalles resultatforventninger. Begrepet stammer fra Banduras sosiale læringsteori (som i dag kalles sosial kognitiv teori). For å finne ut om det er forskjeller i resultatforventninger mellom intervensjonsgruppen og kontrollgruppen som forklarer forskjellen i røykevaner, må en gjennomføre en såkalt additiv kontroll for tredjevariabler. Dersom det er slik at mindre positive eller mer negative forventninger til konsekvensene av røyking har ført til at færre røyker, vil sammenhengen mellom det å ha blitt eksponert for tiltaket og røykevaner bli redusert eller kanskje helt borte når en kontrollerer for resultatforventninger. Slike analyser av dataene er foreløpig ikke gjennomført, så det er foreløpig et åpent spørsmål om resultaforventningene har fungert som mediator.

Fig. 3.14: Additiv kontroll for tredjevariabel. Sammenhengen mellom en prediktor (eksposisjon for et antirøykeprogram) og en kriterievariabel (begynt å røyke) endrer seg når en kontrollerer statistisk for en tredjevariabel (resultatforventninger).



Hvordan en skal regne ut styrken på en sammenheng mellom to variabler der en har kontrollert for en tredjevariabel avhenger av hva slags målenivå det er på variablene. Dersom alle de tre variablene er målt på intervallnivå, bruker en det som kalles en partiell korrelasjonskoeffisient. Formelen for den partielle korrelasjonen er vist nedenfor (formel 3.24).

$$r_{xy.z} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{1 - r^2_{xz}} \sqrt{1 - r^2_{yz}}} \qquad (3.24)$$

$r_{xy.z}$ - Den partielle korrelasjonen mellom variablene x og y når det kontrolleres for z

r_{xy} - Korrelasjonen mellom x og y

r_{xz} - Korrelasjonen mellom x og z

r_{yz} - Korrelasjonen mellom y og z

La oss ta et tenkt eksempel, hentet fra Aron, Aron & Coups (2006) lærebok i statistikk. I en studie blant personer i alderen 20-35 år har vi målt graden av stress de opplever i hverdagen. Vi har funnet at grad av stress øker med alderen. Korrelasjonen er så sterk som 0,40. Vi har imidlertid også spurt hvor mange barn de har, og det viser seg at graden av opplevd stress også henger sammen med antall barn. Korrelasjonen er her enda sterkere, nemlig 0,50.

Dessuten henger antall barn sammen med alder, og korrelasjonen er her så høy som 0,60. Vi får derfor en mistanke om at det ikke er alderen som er avgjørende for hvor stor grad av stress de opplever, men at det er antall barn som er viktigst. For å se hvor sterk sammenhengen er mellom alder og stress når vi har kontrollert for antall barn, regner vi ut den partielle korrelasjonen slik den er vist i formel 3.24. Variabelen x viser alder, variabelen y er grad av opplevd stress og variabelen z er antall barn. Vi får da følgende regnestykke:

$$r_{xy.z} = \frac{0,40 - 0,50 * 0,60}{\sqrt{1 - 0,50^2}\sqrt{1 - 0,60^2}} = \frac{0,40 - 0,30}{\sqrt{0,75}\sqrt{0,64}} = \frac{0,10}{0,866*0,800} = \frac{0,10}{0,693} = 0,14$$

r_{xy} - Korrelasjonen mellom x og y = 0,40

r_{xz} - Korrelasjonen mellom x og z = 0,50

r_{yz} - Korrelasjonen mellom y og z = 0,60

I SPSS finner en partielle korrelasjoner under *Analyze*, *Correlate* og *Partial*. En legger de to primære variablene (x og y) inn i ruten som heter *Variables*. Variabelen (eller variablene) en ønsker å kontrollere for legges inn i ruten *Controlling for*. Vanligvis vil en velge to-halet signifikanstesting (*Two-tailed*). En kan få ut den enkle korrelasjonen mellom x og y ved å gå inn i *Options* og velge *Zero order correlations*.

Den bivariate korrelasjonen var 0,40, altså en middels sterk korrelasjon. Den partielle korrelasjonen viser seg imidlertid å være så lav som 0,14. Vi har dermed vist at sammenhengen mellom alder og opplevd stress er temmelig lav når vi kontrollerer for hvor mange barn de har. Konklusjonen blir med andre ord en helt annen.

Davis (1971) forteller i sin innføringsbok i surveyanalyse om en notasjon som gjelder resultatet av kontroll for tredjevariabler der det ikke er snakk om interaksjonseffekter. Notasjonen kan føres tilbake til Kendall & Lazarsfeld (1950). De skiller mellom følgende situasjoner:

Ingen effekt:	Sammenhengen i undergruppene (på tredjevariabelen) har samme fortegn og er like sterk som i totalmaterialet.
Forklaring:	En sammenheng som var til stede i totalmaterialet (positiv eller negativ) forsvinner.
Suppressor-effekter:	En sammenheng blir sterkere eller den får motsatt fortegn.

Den kanskje mest vanlige situasjonen i psykologisk forskning og samfunnsforskning, at en sammenheng blir noe svakere, har ikke noe spesielt navn. Det samme er tilfelle med den temmelig trivielle situasjonen at en sammenheng på omtrent null etter kontroll for en

tredjevariabel fremdeles er nokså nær null. En sammenheng kan også skifte retning dersom en kontrollerer for en tredjevariabel.

Det finnes andre framgangsmåter for å undersøke effekten av additiv kontroll (kontroll uten interaksjon) for tredjevariabler. En statistisk størrelse som Yules Q kan regnes ut som en partiell Yules Q. At den er partiell betyr at den er regnet ut for hver underkategori på tredjevariabelen, og at det er regnet ut et slags gjennomsnitt av alle disse sammenehengene. Dette er en parallell til partielle korrelasjoner for intervallvariabler.

Dersom det foreligger interaksjonseffekter, vil en partiell koeffisient kunne skjule vesentlig informasjon. For å finne ut om det foreligger en statistisk sikker interaksjon, kan en bruke spesielle signifikanstester. Dersom en analyserer metriske variabler kan interaksjonseffekter undersøkes ved bruk av multippel lineær regresjon. Dersom en analyserer kategorielle variabler er det mest aktuelt med logistisk regresjon og Goodmans log-lineære analyse av flerveis krysstabeller.

I dette kapittelet har vi bare så vidt kommet inn på de mange ulike typer statistiske teknikker en bruker for å analysere mediator- og moderator-situasjoner. Dette skal vi komme tilbake til i kapitlene om variansanalyse og regresjonsanalyse.

Referanser

Aron, A. & Aron, E.N. (1999). *Statistics for psychology* (Second Edition). Upper Saddle River, New Jersey: Prentice Hall.

Aron, A., Aron, E.N. & Coups, E.J. (2006). *Statistics for psychology* (Fourth Edition). Upper Saddle River, New Jersey: Prentice Hall.

Bryk, A.S. & Raudenbush, S.W. (1992). *Hierachical linear models*. Newbury Park, California: Sage.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, New Jersey: Lawrence Erlbaum.

Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (Third Edition). Mahwah, New Jersey: Lawrence Erlbaum.

Davis, J. (1971): *Elementary survey analysis.* Englewood Cliffs, New Jersey: Prentice Hall.

Galtung, J. (1967). *Theory and method of social research*. New York: Columbia University Press.

Gibbons, J.D. (1993). *Nonparametric measures of association.* Newbury Park: Sage (Quantitative Applications in the Social Sciences, No.91).

Goldstein, H. (1995). *Multilevel statistical models* (second edition). London: Arnold.

Goodman, L. & Kruskal, W. (1954): Measures of association for cross-classifications. *Journal of the American Statistical Association*, *Vol.49*, 732-764. (Referert i Davis, 1971)

Jøsendal, O., Aarø, L.E. & Bergh, I.H. (1998). Effects of a school-based smoking prevention programme among sub-groups of adolescents. *Health Education Research*, *13(2)*, 215-224.

Jøsendal, O. & Aarø, L.E. (2002). VÆR røykFRI – evaluering av et tiltak for røykfrie skoler. *Tidsskrift for Den norske lægeforening, 122 (4)*, 403-407.

Kendall, P.L. & Lazarsfeld, P.F.: (1950): Problems of survey analysis. I Robert K. Merton & Paul F. Lazarsfeld (red.): *Continuities in Social Research.* New York: Free Press, 135-167.

Kimball, A.W. (1954). Short-cut formulas for the exact partition of χ^2 in contingency tables. *Biometrics*, *10*, 452-458.

Liebtrau, A.M. (1983): *Measures of association.* Beverly Hills: SAGE.

Maxwell, A.E. (1961). *Analysing qualitative data.* London: Chapman & Hall.

McNemar, Q. (1969): *Psychological statistics.* New York: John Wiley & Sons.

Olkin, I. & Finn, J.D. (1995). Correlations redux.*Psychological Bulletin, 118*, 155-164.

Sheskin, D.J. (1997). *Handbook of parametric and nonparametric statistical procedures*. Boca Raton, Florida: CRC Press.

Siegel, S. (1956): *Nonparametric Statistics for the Behavioral Sciences.* Tokyo: McGraw Hill.

Siegel, S. & Castellan, N.J. (1988). *Nonparametric Statistics for the Behavioral Sciences.* New York: McGraw Hill.

Weinberg, S.L. & Abramowitz, S.K. (2002). *Data analysis for the behavioral sciences using SPSS.* Cambridge: Cambridge University Press.

Yule, G.U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, *Vol.75*, 579-642. (Referert etter Davis, 1971)

Kap 4: Variansanalyse

4.1 Innledning

Variansanalyse (Analysis of Variance – ANOVA) er en familie av statistiske teknikker som brukes når en skal analysere en avhengig metrisk variabel mot kategorielle uavhengige variabler. De kategorielle variablene bør ha et relativt lite antall kategorier (eller nivåer, som kategoriene gjerne kalles i variansanalyse). Den metriske avhengige variabelen skal helst være normalfordelt (eller i hvert fall ikke for sterkt avvikende fra normalfordeling).

En kan se på variansanalyse som en teknikk for å analysere forskjeller i aritmetiske gjennomsnitt mellom grupper. I forrige kapittel så vi på analyse av forskjeller i gjennomsnitter mellom to grupper ved bruk av t-test. De analysene vi presenterte der kan betraktes som spesialtilfeller av variansanalyse. Men til forskjell fra t-testing av forskjeller i gjennomsnitt mellom grupper, som er begrenset til situasjoner der en sammenlikner bare to grupper, kan en i variansanalyse ha flere enn to grupper, og en kan ha mer enn en uavhengig variabel. Variansanalysen er spesielt godt egnet når en på en pedagogisk enkel måte vil beskrive interaksjonseffekter.

Som noen kanskje husker fra kapittel 3, skilte vi der mellom t-test for urelaterte (uavhengige) grupper og t-test for relaterte (korrelerte) grupper. Urelaterte grupper har en for eksempel når en sammenlikner personer med tre inntektsnivå (for eksempel høy, middels og lav inntekt). Hver enkelt person kan bare plasseres i en av gruppene, og vi antar at de som inngår i en av gruppene ikke har noen bestemt relasjon til de som inngår i de andre to. Et eksempel på relaterte grupper har en dersom en gjennomfører en undersøkelse blant ektepar, og sammenlikner disse par for par på en eller annen variabel, for eksempel en skala for måling av holdninger til ett eller annet politisk spørsmål. Relaterte grupper har en også når en sammenlikner de samme personenes skår på en test på to eller flere tidspunkt.

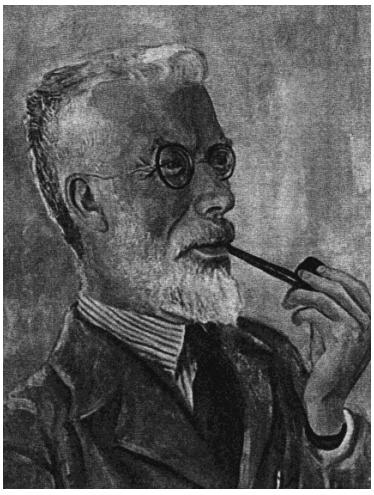
På helt tilsvarende måte som ved t-testing, skiller vi mellom variansanalyse for urelaterte grupper og variansanalyse for relaterte grupper. Ofte er det snakk om variansanalyse med repeterte målinger, og variansanalyse av relaterte grupper kalles da også variansanalyse med repeterte målinger. Når antall ganger er tre eller flere, bruker vi altså variansanalyse for repeterte målinger i stedet for t-test for relaterte grupper.

Kovariansanalyse (Analysis of Covariance – ANCOVA) er en statistisk teknikk som bygger på variansanalyse, men der en i tillegg tar inn en tredje kategori av variabler. Denne tredje gruppen av variabler skal være metriske, og en må anta at deres sammenhenger med den

avhengige variabelen er lineære. I tillegg til at en kan undersøke hvilken sammenheng kovariansvariablene har med den avhengige variabelen, kan en kontrollere for dem når en undersøker sammenhengen mellom de kategorielle, uavhengige variablene og den avhengige variabelen. På samme måte som i ANOVA kan en teste hypoteser om interaksjoner mellom de kategorielle prediktorene, men da etter kontroll for kovariansvariablene. Kovariansanalyse skal ikke beskrives nærmere i denne teksten.

Sir Ronald Aylmer Fisher er ett av de virkelig store navnene i statistikkens historie. Han har æren for å ha utviklet noen av de mest sentrale begrepene og teknikkene i moderne statistikk, for eksempel varians, variansanalyse, statistiske størrelser til beskrivelse av utvalg (til forskjell fra populasjonsparametre), signifikansnivå, nullhypotese og randomisering. Han viste seg tidlig å ha usedvanlige matematiske evner. Som voksen var han slett ingen enkel person å ha med å gjøre. Han var stadig involvert i feider med kolleger, blant annet med Karl Pearson. Han var likevel ikke helt uten humoristisk sans. En gang han skulle krysse gaten sammen med en annen kjent statistiker, William G. Cochran, og Cochran nølte på grunn av mange biler og høy risiko, sa han følgende: ”Oh, come on, a spot of natural selection won’t hurt us.” Fishers publikasjoner var ofte krevende lesning, og han var ikke alltid så nøye med å presentere alle de forutsetningene og bevisene som matematikere legger så stor vekt på. Og når han begynte en setning med ”Det er åpenbart at …”, kunne det i følge statistikeren Gosset ta både en og to timer med hard innsats før en forstod hva han mente. Ett av høydepunktene i Fishers karriere var to opphold ved Iowa State College i 1931 og 1936. To amerikanske statistikere (George Snedecor og E.F. Lindquist) skrev hver sine lærebøker i statistikk som begge var basert på Fishers idéer. Den ene ble en bestselger og solgte i mer enn 100 000 eksemplar. Den andre har hatt stor betydning for pedagogisk og psykologisk forskning helt fram til i dag. Ratioen mellom to uavhengige chi-kvadrat-fordelte størrelser (mean squares between over mean squares within) har en bestemt fordeling, F-fordelingen, som har fått sitt navn etter Fisher. Det var Snedecor som tildelte Fisher denne æren, en ære Fisher heller ville vært foruten. Fisher skal etter sigende aldri ha tilgitt Snedecor for dette.

(Fritt etter Aron & Aron, 1999, s. 326-327 og Serlin, 2005, s. 55)



Sir Ronald Aylmer Fisher (1890-1962)

Multivariat variansanalyse (MANOVA) er en utvidelse av variansanalysen til en situasjon der en i stedet for å ha en enkelt avhengig variabel har flere. Manova skal ikke presenteres nærmere her, men det finnes en rekke gode innføringstekster (Weinfurt, 1995; Tabachnik & Fidell, 2001). For at bildet skal bli komplett må vi også nevne Multivariat kovariansanalyse (Multivariate Covariance Analysis – MANCOVA) som utvider analysen både til flere samtidige avhengige variabler og til kontroll for et tredje sett av variabler (kovarians-variabler).

Det finnes to former for variansanalyse, nemlig det som kalles fikserte modeller (fixed effects models) og det som kalles tilfeldige modeller (random effects models). De fikserte modellene tar utgangspunkt i at kategoriene på de uavhengige variablene er faste, og en analyserer variasjoner i gjennomsnittsverdier på normalfordelte variabler innen de ulike subgruppene. I variansanalyse som baserer seg på tilfeldige modeller ser en på kategoriene på de uavhengige variablene som et utvalg, trukket fra et større univers av slike kategorier. Dermed blir det aktuelt å regne statistikken på en slik måte at en tar i betraktning den usikkerheten som ligger i det å bare ha et utvalg av kategorier og ikke et fullstendig sett (Tabachnik & Fidell, 2001). Vi skal i denne teksten bare presentere variansanalyse med fikserte modeller.

Det var den engelske statistikeren og genetikeren Ronald Fisher som først utviklet variansanalytiske teknikker. Begrepet variansanalyse brukte han første gang i en overskrift i en artikkel fra 1918 (Fisher, 1918). Selve teknikken ble imidlertid utviklet noe senere. Hans statistikktekst fra 1925 var det som gjorde variansanalysen kjent for et noe større publikum av fagfolk (Fisher, 1925). Denne boken kom i tretten senere utgaver og ble sist utgitt i 1970. Fisher aksepterte i 1919 en stilling som statistiker ved Rothamsted Agricultural Experiment Station i Hertfordshire, et lite stykke nord for London. Dette instituttet var ett av de eldste for jordbruksforskning i England. Det var etablert så tidlig som i 1837 og hadde som formål å forske på effekten av ulike former for jord og gjødsling på nyttevektster. Variansanalyse viste seg å være en glimrende statistisk teknikk i denne forskningen.

4.2 Enveis variansanalyse

Den enkleste formen for variansanalyse består i å undersøke sammenhengen mellom en avhengig variabel (metrisk) og en uavhengig variabel (kategoriell). Dette kalles enveis variansanalyse fordi en har bare en uavhengig variabel. Dersom en har to uavhengige variabler kalles det en toveis variansanalyse, dersom en har tre kalles det en treveis osv. Flerveis variansanalyse kalles med en fellesbetegnelse for faktoriell variansanalyse.

La oss ta et eksempel. Det er gjennomført et pedagogisk forsøk med matematikkundervisning på tre forskjellige måter. På forhånd delte en elevene tilfeldig inn i tre grupper med ti elever i hver og fem gutter og fem jenter i hver av gruppene. I den ene gruppen ble elevene undervist på tradisjonell måte, altså med kateterundervisning der læreren gjennomgikk stoffet på tavlen. I en annen gruppe lot en elevene benytte en pc-basert interaktiv opplæring. I den tredje gruppen benyttet man en variant av problembasert læring. Etter to måneder fikk alle elevene gjennomgå en matematikktest som bestod av til sammen 20 oppgaver.

Tabell 4.1: Antall rette svar på matematikkoppgavene etter pedagogisk opplegg

Elev nr.	Gruppe (trad. underv.)	Kjønn	Test-resultat	Elev nr.	Gruppe (pc-basert)	Kjønn	Test-resultat	Elev nr.	Gruppe (probl.-Basert)	Kjønn	Test-resultat
01	1	1	10	11	2	1	18	21	3	1	16
02	1	1	8	12	2	1	15	22	3	1	13
03	1	1	11	13	2	1	14	23	3	1	15
04	1	1	9	14	2	1	19	24	3	1	16
05	1	1	14	15	2	1	16	25	3	1	15
06	1	2	7	16	2	2	17	26	3	2	17
07	1	2	13	17	2	2	11	27	3	2	20
08	1	2	12	18	2	2	13	28	3	2	19
09	1	2	10	19	2	2	13	29	3	2	17
10	1	2	16	20	2	2	10	30	3	2	17

Resultatene av denne testen (antall riktige svar) er gjengitt i Tabell 4.1. Variabelen kjønn skal vi ikke benytte her, men komme tilbake til nedenfor når vi beskriver toveis variansanalyse.

Vi legger først dataene inn i dataarket i SPSS. Vi legger inn gruppe (nummerert fra 1 til 3) som v1, kjønn som v2 og skår på testen som v3. Deretter kjører vi et Boxplot ved å velge følgende fra menyen: *Graphs*, *Boxplot, Simple* og *Define*. Vi legger v1 inn i feltet som heter *Category Axis* og v3 inn in feltet som heter *Variable*. Når vi deretter klikker *OK*, får vi ut det diagrammet som er vist i Fig. 4.1

For hver gruppe vises median (svart, kraftig horisontal strek), området mellom første og tredje kvartil (grått felt) og minimums- samt maksimums-verdier. Vi får straks et inntrykk av at den første gruppen har gjort det noe dårligere enn de to andre og at det kanskje er den gruppen som har fulgt det problembaserte opplegget som har gjort det aller best.

Et neste naturlig trinn er å regne ut gjennomsnittlig skår på matematikktesten i de tre gruppene. Enklest kan vi ta ut denne informasjonen ved å velge følgende fra menyen i SPSS: *Analyze*, *Compare Means*, *Means* og så legger vi inn v1 under *Indpendent List* og v2 under *Dependent List*. Deretter klikker vi på *OK*. Vi får da ut de resultatene som vises i Fig. 4.2.

Vi ser at gruppen som har fulgt det problembaserte opplegget har oppnådd aller høyest skår med 16,5 riktige svar i gjennomsnitt. De som har fulgt det pc-baserte opplegget har fått en gjennomsnittsskår på 14,6. De som har fulgt den tradisjonelle kateterundervisningen har bare oppnådd en gjennomsnittlig skår på 11,0. Men vi vet fremdeles ikke om disse forskjellene er så store at de er statistisk sikre. Det er her variansanalysen kan hjelpe oss.

Fig. 4.1: Testresultater etter gruppe. Boxplot fra SPSS

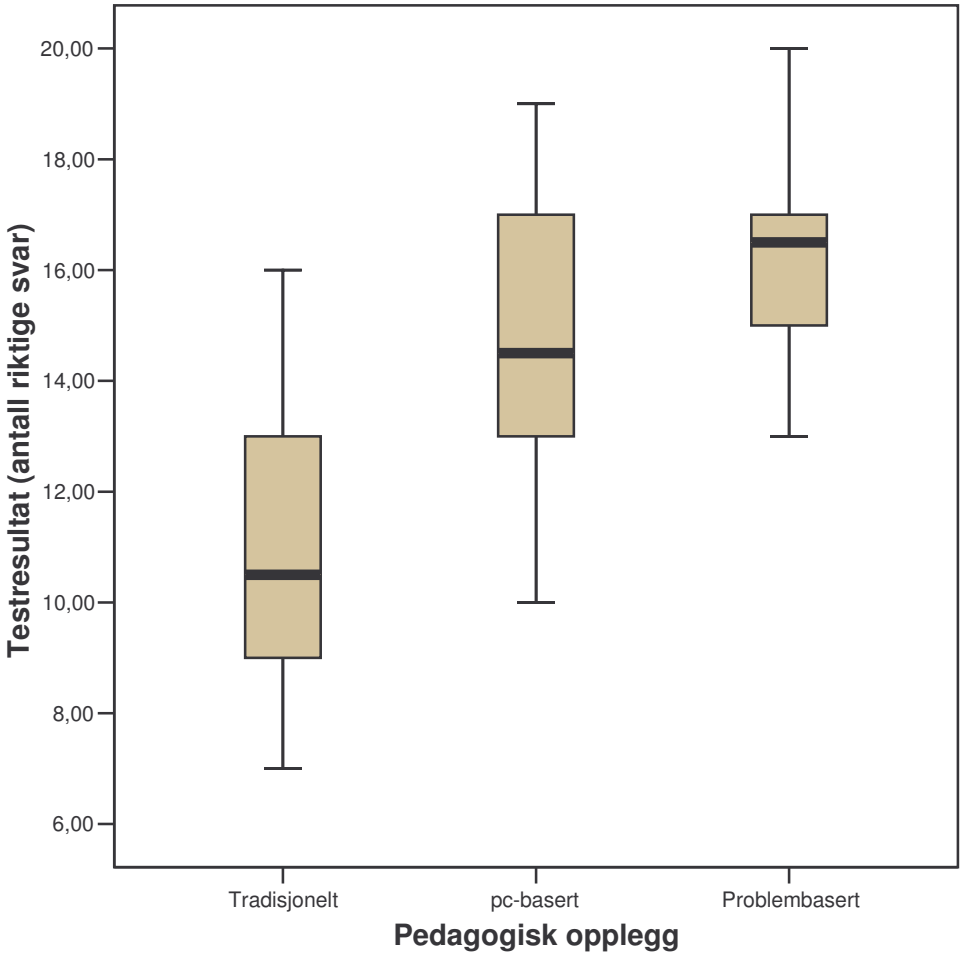


Fig. 4.2: Gjennomsnittlig matematikkskår etter gruppe.

Report

v3 Testresultat (antall riktige svar)

v1 Pedagogisk opplegg	Mean	N	Std. Deviation
1,00 Tradisjonelt	11,0000	10	2,78887
2,00 pc-basert	14,6000	10	2,95146
3,00 Problembasert	16,5000	10	2,01384
Total	14,0333	30	3,42892

I variansanalyse skilles det mellom gjennomsnittet av kvadratene mellom grupper (mean of squares between) og gjennomsnittet av kvadratene innen grupper (mean of squares within). Gjennomsnittet av kvadratene mellom grupper regnes ut slik som vist i formel 4.1.

$$MS_B = \frac{SS_B}{df} = \frac{\sum_{j=1}^{K} n_j (\bar{x}_j - \bar{\bar{x}})^2}{k-1} \qquad (4.1)$$

MS_B Gjennomsnittet av kvadratene mellom grupper (Mean of squares between groups)
SS_B Kvadratsummen mellom gruppene (Sum of squares between groups)
df Antall frihetsgrader (Degrees of freedom)
k Antall grupper
n_j Antall observasjoner i gruppe nr. j
$\bar{x}_j$ Gjennomsnittet i gruppe nr. j
$\bar{\bar{x}}$ Totalgjennomsnittet (Grand mean)

Dersom vi setter inn tallene som er gjengitt i Fig. 4.2 (regnet ut på basis av Tabell 4.1) i denne formelen og regner ut, får vi resultatet som er vist nedenfor.

$$MS_B = \frac{\sum_{j=1}^{K} n_j (\bar{x}_j - \bar{\bar{x}})^2}{k-1} =$$

$$\frac{(10(11-14,03)^2) + (10(14,6-14,03)^2) + (10(16,50-14,03)^2)}{3-1} =$$

$$\frac{91,81+3,25+61,01}{2} = 78,04$$

Gjennomsnittet av kvadratene innen grupper regnes ut slik som vist i formel 4.2.

$$MS_W = \frac{SS_W}{df} = \frac{\sum_{i,j} (x_{i,j} - \bar{x}_j)^2}{n-k} \qquad (4.2)$$

MS_W Gjennomsnittet av kvadratene innen grupper (Mean of squares within)
SS_W Kvadratsummen innen gruppene (Sum of squares within)
df Antall frihetsgrader
$x_{i,j}$ Skåren til observasjon nr. i innen gruppe j
$\bar{x}_j$ Gjennomsnittet av skårene i gruppe j
n Antall observasjoner totalt
k Antall grupper

Deretter kan vi sette de tre kvadratsummene inn i formel 4.2, og får da følgende:

$$MS_W = \frac{\sum_{i,j}(x_{i,j} - \bar{x}_j)^2}{n-k} = \frac{70,00 + 78,38 + 36,50}{30-3} = 6,847$$

Dette tallet kan regnes ut på en litt annen måte. Dersom vi først regner ut variansen innen hver gruppe og deretter regner ut det vektede gjennomsnittet av de tre variansene, får vi samme tallet. Til å vekte bruker vi antall observasjoner i gruppene. Siden antallet er like stort i de tre gruppene i dette eksempelet, behøver vi ikke å vekte her, men bare ta gjennomsnittet av de tre variansene. MS_W blir ofte omtalt som et estimat av innengruppe-variansen i populasjonen.

Dersom vi setter inn tallene fra tabell 4.1 i tabell 4.2, kan vi lett regne ut kvadratsummene innen de tre gruppene.

Tabell 4.2: Utregning av kvadratsummer innen grupper (Within groups sum of squares)

X	$X - \bar{X}$	$(X - \bar{X})^2$	X	$X - \bar{X}$	$(X - \bar{X})^2$	X	$X - \bar{X}$	$(X - \bar{X})^2$
10	-1	1	18	3,4	11,56	17	0,5	0,25
8	-3	9	15	0,4	0,14	20	3,5	12,25
11	0	0	17	2,4	5,76	16	-0,5	0,25
9	-2	4	11	-3,6	12,96	19	2,5	6,25
14	3	9	13	-1,6	2,56	13	-3,5	12,25
7	-4	16	13	-1,6	2,56	15	-1,5	2,25
13	2	4	14	-0,6	0,36	17	0,5	0,25
12	1	1	19	4,4	19,36	16	-0,5	0,25
10	-1	1	10	-4,6	21,16	17	0,5	0,25
16	5	25	16	1,4	1,96	15	-1,5	2,25
		70			78,38			36,5

Tallene for gjennomsnittet av kvadratene mellom og innen gruppene kan nå settes inn i en enkel formel som gir oss en F-verdi (formel 4.3). Dette er en statistisk størrelse som kan brukes til å signifikansteste med, på samme måte som z-fordelingen, χ^2 – fordelingen og t-fordelingen. Når det gjelder F, er det egentlig snakk om et stort antall fordelinger avhengig av antall frihetsgrader for gjennomsnittet av kvadratene mellom grupper og antall frihetsgrader for gjennomsnittet av kvadratene innen grupper.

$$F_{k-1,n-k} = \frac{MS_B}{MS_W} \qquad (4.3)$$

$F_{k-1,n-k}$ F-verdien med tilhørende frihetsgrader

$k-1$ Antall grupper minus 1 (frihetsgrader mellom grupper)

$n-k$ Antall observasjoner minus antall grupper (frihetsgrader innen grupper)

MS_B Gjennomsnittet av kvadratene mellom grupper (Mean of sum of squares between)

MS_W Gjennomsnittet av kvadratene innen grupper (Mean of sum of squares within)

Vi ser av formelen at F-verdien egentlig bare er et uttrykk for hvor stor variasjonen er mellom gruppene sett i forhold til hvor stor variasjonen er innen gruppene. Jo større forskjeller mellom gjennomsnittene og jo mindre variasjon innen gruppene, desto høyere blir F-verdien. Setter vi inn tallene som er regnet ut for gjennomsnittet av kvadratene mellom grupper og tilsvarende innen grupper, får vi følgende:

$$F_{k-1,n-k} = \frac{MS_B}{MS_W} = \frac{78,04}{6,847} = 11,398$$

$$F_{2,17} = 11,398$$

Siden F-verdien vi har regnet ut har 2 og 17 frihetsgrader, må vi slå opp i en tabell som viser kritiske verdier for akkurat denne kombinasjonen. Vi finner da følgende:

p<0,05 – 3,59
p<0,01 – 6,11
p<0,001 – 10,66

Vi ser at den F-verdien vi har regnet ut ovenfor er høyere enn noen av disse. Det betyr at forskjellen mellom de tre gruppene når det gjelder gjennomsnittlig skår på matematikktesten er signifikant på p<0,001-nivået. Vi kan forkaste nullhypotesen som sier at det i populasjonen er samme gjennomsnitts-skår i alle gruppene. I et matematisk språk formuleres gjerne nullhypotesen som en ligning. H_0: $\mu 1 = \mu 2 = \mu 3$. Tegnet μ står for gjennomsnittsskåren i populasjonen. Den alternative hypotesen er at ikke alle gjennomsnittsverdiene er like. Når vi har forkastet nullhypotesen har vi sannsynliggjort den alternative hypotesen (H_1), nemlig at gjennomsnittene i gruppene i populasjonen ikke er like.

Vi kan kjøre den samme analysen i SPSS. En går da inn i *Analysis*, *Compare Means*, og *One-Way ANOVA*, legger v1 inn in feltet *Factor* og v3 inn i *Dependent List*. Når en så trykker på *OK*, får en ut tabellen som er gjengitt i Fig. 4.3.

Vi ser at tallene stemmer ganske bra med det vi har regnet ut manuelt ovenfor. De små forskjellene i tall skyldes avrundingsfeil. Vi ser at p-verdien som er oppgitt her er lik 0,000[1]. Dette er konsistent med at p<0,001 slik vi allerede har konkludert med ovenfor.

Fig. 4.3: Enveis variansanalyse fra SPSS

ANOVA

v3 Testresultat (antall riktige svar)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	156,067	2	78,033	11,395	,000
Within Groups	184,900	27	6,848		
Total	340,967	29			

Dersom vi legger til grunn at nullhypotesen er korrekt, hvilket vi gjør i all signifikanstesting, betyr det at det ikke er noen forskjell på de tre gjennomsnittene i de tilsvarende gruppene i populasjonen. Testen forutsetter dessuten at det ikke er noen forskjell i varians i de tre gruppene. I en viss forstand kan vi si at vi under forutsetning av at nullhypotesen er korrekt kan se på de tre gruppene som utvalg fra den samme populasjonen. Vi kan derfor betrakte gjennomsnittet av kvadratene innen grupper (MS_W) som et estimat av variansen i denne populasjonen. Dette er kanskje ikke så overraskende. Derimot kan det være vanskeligere å forstå at også gjennomsnittet av kvadratene mellom gruppene (MS_B) er et estimat av den samme variansen. Men slik er det. Dersom nullhypotesen er korrekt kan vi forvente at de to tallene blir omtrent like. Selv om nullhypotesen skulle være korrekt, blir de likevel sjelden helt like, for siden vi baserer våre utregninger på tall fra et utvalg, vil tilfeldige variasjoner medføre at de to tallene blir litt forskjellige. Men poenget er at dersom de blir usannsynlig forskjellige, tyder det på at nullhypotesen er feil.

Når vi har funnet at det er signifikant forskjell mellom de tre gruppene, vet vi ikke uten videre at det er signifikante forskjeller mellom alle kombinasjoner av to og to grupper. Men siden forskjellen er størst mellom gruppe 1 (de som har hatt tradisjonell matematikkundervisning) og gruppe 3 (de som har hatt problembasert læring), kan vi regne med at denne forskjellen har bidratt til at forskjellene totalt sett er blitt signifikante. Vi vet derimot ikke om for eksempel forskjellen mellom gruppene 2 (pc-basert læring) og 3 (problembasert læring) er signifikant. Dette må vi eventuelt undersøke ved bruk av t-tester for forskjeller mellom ukorrelerte grupper slik vi lærte i kapittel 3.

Det å systematisk teste par av grupper mot hverandre kalles multippel sammenlikning (multiple comparison). Slike sammenlikninger er imidlertid både en kontroversiell og komplisert sak. Det å foreta en serie med slike tester for å se hvilke forskjeller som blir signifikante, bryter nemlig med logikken bak signifikanstestingen. Dersom en gjennomfører flere tester med et bestemt signifikansnivå, vil sjansen for at minst en av disse skal vise signifikans være betydelig høyere enn det signifikansnivået en bruker. Sannsynligheten for at en test skal vise signifikans når signifikansnivået settes til p<0,05 er 0,143 dersom en foretar tre tester og 0,265 dersom en gjør seks tester (Aron & Aron, 1999).

[1] Egentlig blir p-verdien ikke nøyaktig null, men når den blir tilstrekkelig lav, avrundes den likevel nedover til null.

En vanlig løsning på dette problemet er å benytte det som kalles Bonferroni-prosedyren (også kalt Dunns test). Poenget er å operere med et strengere signifikansnivå, slik at muligheten for tilfeldige signifikanser reduseres. Dersom en gjør to tester, bør p-verdien halveres. Dersom en gjør tre tester bør den settes til en tredjedel etc.

Hvis en gjør såkalte planlagte sammenlikninger, vil en ofte være interessert i å sammenlikne et begrenset antall par av grupper. I slike situasjoner vil Bonferroni-prosedyren være tilfredsstillende. Dersom en i stedet gjør post hoc-sammenlikninger (altså sammenlikner alle slags kombinasjoner, etter å ha påvist en signifikant forskjell mellom alle gruppene sett under ett), vil Bonferroni-prosedyren som regel bli for streng. Ved å følge denne prosedyren stiller en så strenge krav til p-verdi at det blir umulig å påvise signifikante forskjeller. Det er derfor blitt utviklet en rekke mer liberale prosedyrer, som alle har fått navn etter den statistikeren som utviklet metoden. Eksempler er prosedyrer utviklet av Duncan, Neuman-Keuls, Scheffé og Tukey. Blant statistikere er det uenighet om hvilken prosedyre som er riktigst å bruke.

Bruken av enveis variansanalyse bygger på en del forutsetninger:

1. Hver enkelt observasjon skal være uavhengig av alle de andre (dersom det da ikke er snakk om variansanalyse mellom relaterte grupper). I undersøkelser der personer utgjør enhetene, betyr det at det kun skal foreligge en skår per person. Dersom en person er registrert med to eller flere skårer, blir det feil.
2. I populasjonen skal skårene på den avhengige variabelen være normalfordelte innen hver av gruppene.
3. I populasjonen skal skårene på den avhengige variabelen ha lik varians i alle undergrupper.

Den første av disse forutsetningene kan ikke brytes uten at en risikerer å gjøre alvorlige feil. Forutsetningen om normalfordelinger innen gruppene er viktig dersom antall observasjoner i hver gruppe er lavt. Når antallet er minst 30 observasjoner i hver gruppe, kan en benytte enveis variansanalyse selv om fordelingen avviker fra normalitet. Når antallet observasjoner er mindre enn 30, bør en imidlertid være påpasselig med å sjekke om det foreligger sterke avvik fra normalitet eller om det finnes "uteliggere" (enkeltobservasjoner med svært høye eller svært lave verdier, og som dermed avviker sterkt fra de øvrige). Forutsetningen om lik varians i gruppene kan godt brytes dersom antallet observasjoner i hver gruppe er relativt høyt og dersom gruppene er omtrent like store. Men dersom antallet i gruppene er forskjellig, vil det være problematisk å bruke variansanalyse dersom variansen i en gruppe er minst 1,5 ganger så stor som i en annen gruppe.

Fig. 4.4: Levenes test av varianshomogenitet

Test of Homogeneity of Variances

v3 Testresultat (antall riktige svar)

Levene Statistic	df1	df2	Sig.
1,069	2	27	,357

Dersom vi er usikre på forutsetningen om lik varians i gruppene, kan vi benytte Levenes test, som også ble nevnt i kapittel 3 under t-test for forskjeller i gjennomsnitt mellom to uavhengige grupper. Dersom vi kjører Levenes test på eksempelet vi har brukt ovenfor (skårer på matematikktest i tre grupper), får vi det resultatet som er vist i Fig. 4.4. Vi ser at p-verdien er så høy som 0,357. Vi kan med andre ord ikke forkaste nullhypotesen om lik gjennomsnitt i de tre gruppene (i populasjonen). Det er derfor grunn til å feste lit til den variansanalysen som er gjennomført ovenfor.

Ikke alle statistikere er like begeistret over bruken av variansanalyse når en sammenlikner et større antall grupper. Rosnow og Rosenthal (1989) hevder at når vi tester forskjeller mellom tre eller flere grupper, kan vi være sikre på å ha testet noe som er ganske uinteressant. De argumenterer for at det er mer hensiktsmessig å teste forskjeller mellom grupper parvis, og at dette bør gjøres på en planlagt måte.

4.3 Effektstørrelse

Under t-testing for forskjeller i gjennomsnitt mellom to grupper så vi at effektmålet (*Cohens d*) var definert som forskjellen mellom gruppene dividert med den felles variansen for de to gruppene. Fullt så enkelt er det ikke når vi har flere enn to grupper, slik tilfellet er for enveis variansanalyse. Cohen har imidlertid utviklet et effektmål også for variansanalyse. Dette målet kalles Cohens f, og er gjengitt i formel 4.4.

$$f = \frac{sd_B}{sd_W} = \frac{\sqrt{\frac{MS_B}{n}}}{\sqrt{MS_W}} \qquad (4.4)$$

f Cohens f
sd_B Standardavviket mellom grupper
sd_W Standardavviket innen grupper
MS_B Gjennomsnittet av kvadratene mellom grupper (Mean of squares between)
MS_W Gjennomsnittet av kvadratene innen grupper (Mean of squares within)
n Antallet i hver gruppe (forutsatt at gruppene er like store)

Cohen har definert en f på 0,10 som en svak effekt, 0,25 som en middels sterk effekt og 0,40 som en sterk effekt. Dersom vi setter inn tallene fra eksempelet som er gjengitt i Fig. 4.3, får vi følgende utregning:

$$f = \frac{sd_B}{sd_W} = \frac{\sqrt{\frac{MS_B}{n}}}{\sqrt{MS_W}} = \frac{\sqrt{\frac{78,033}{10}}}{\sqrt{6,848}} = 1,07$$

Dette er en effekt som overstiger normtallene fra Cohen. Konklusjonen er med andre ord at effekten er meget sterk.

Det finnes en alternativ formel til å beregne Cohens f (se formel 4.5).

$$f = \sqrt{\frac{F}{n}} \qquad (4.5)$$

f Cohens f

F F-verdien

n Antall observasjoner i hver gruppe (når alle gruppene er like store)

Dette er en svært nyttig formel når en ut fra rene variansanalyser skal beregne effektstørrelser fra studier der effektstørrelser ikke er rapportert. Dersom vi setter inn tallene fra eksempelet gjengitt i Fig. 4.3, får vi følgende:

$$f = \sqrt{\frac{F}{n}} = \sqrt{\frac{11,395}{10}} = 1,07$$

Et mer tradisjonelt mål for sammenhengen mellom en kategoriell variabel og en metrisk variabel er forklart varians. En koeffisient som er variansbasert er eta. Eta kvadrert forteller hvor mye forklart varians i den avhengige (metriske) variabelen som forklares av den uavhengige (gruppevariabelen). Dersom vi bestiller en eta-koeffisient fra prosedyren *Analyze*, *Compare Means* og *Means* i SPSS, får vi den tabellen som er vist i Fig. 4.5. Der ser vi at eta er 0,667 og at forklart varians er 0,458. Eta er definert som kvadratsummen mellom grupper dividert med total kvadratsum. Dersom vi henter de relevante kvadratsummene ut av Fig. 4.3, og dividerer 156,067 med 340,967, får vi tallet 0,458. Cohen (1988) har ikke bare sagt noe om hvordan en skal vurdere Cohens f, men også hvordan en skal vurdere forklart varians. Han sier at en eta kvadrert på 0,01 tilsvarer en svak effekt. En eta kvadrert på 0,06 tilsvarer en middels sterk effekt. En eta kvadrert på 0,14 tilsvarer en sterk effekt[2].

Fig. 4.5: Mål for assosiasjon mellom en metrisk og kategoriell variabel (eta-koeffisienten)

Measures of Association

	Eta	Eta Squared
v3 Testresultat (antall riktige svar) * v1 Pedagogisk opplegg	,677	,458

[2] Det eksisterer en egen formel der en kan regne ut eta kvadrert på grunnlag av Cohens f (eller en kan regne motsatt vei). Ifølge Aron & Aron (1999) vil imidlertid ikke formelen alltid stemme helt med det statistikkprogrammene regner ut for oss. Dette fordi Cohens f er regnet ut på grunnlag av estimerte standardavvik i populasjonen mens den forklarte variansen oftest regnes ut som en ren beskrivelse av utvalget.

4.4 Faktoriell variansanalyse

I tabell 4.1 finnes det to uavhengige variabler. I tillegg til tre grupper definert ut fra hvilket pedagogisk opplegg elevene har fulgt, har vi også variabelen kjønn. Når vi har to uavhengige (kategorielle) variabler og en avhengig (metrisk) variabel, er det aktuelt å benytte toveis variansanalyse. Dermed er vi kommet tilbake til et tema som vi berørte på slutten av kapittel 3, nemlig statistiske interaksjonseffekter. Vi skal forsøke å forklare nærmere hva en statistisk interaksjonseffekt er med utgangspunkt i tallene som er vist i Fig. 4.6.

For å lage en toveis variansanalyse med en avhengig variabel i SPSS går vi først inn i *Analyze*, deretter *General Linear Model*, og til slutt *Univariate*. Dersom vi bruker dataene fra Tabell 4.2, legger vi v3 inn som *Dependent Variable* og v1 og v2 legges inn som *Fixed Factors*. Velg så *Options*, legg inn v1*v2 under *Display Means for* og sett hake i *Descriptive Statistics*. Trykk til slutt på *OK*. Da vil en blant annet få ut de tallene som vises i Fig. 4.6.

Fig. 4.6: Gjennomsnittlig skår på matematikktesten etter pedagogisk opplegg og kjønn

Descriptive Statistics

Dependent Variable: v3 Testresultat (antall riktige svar)

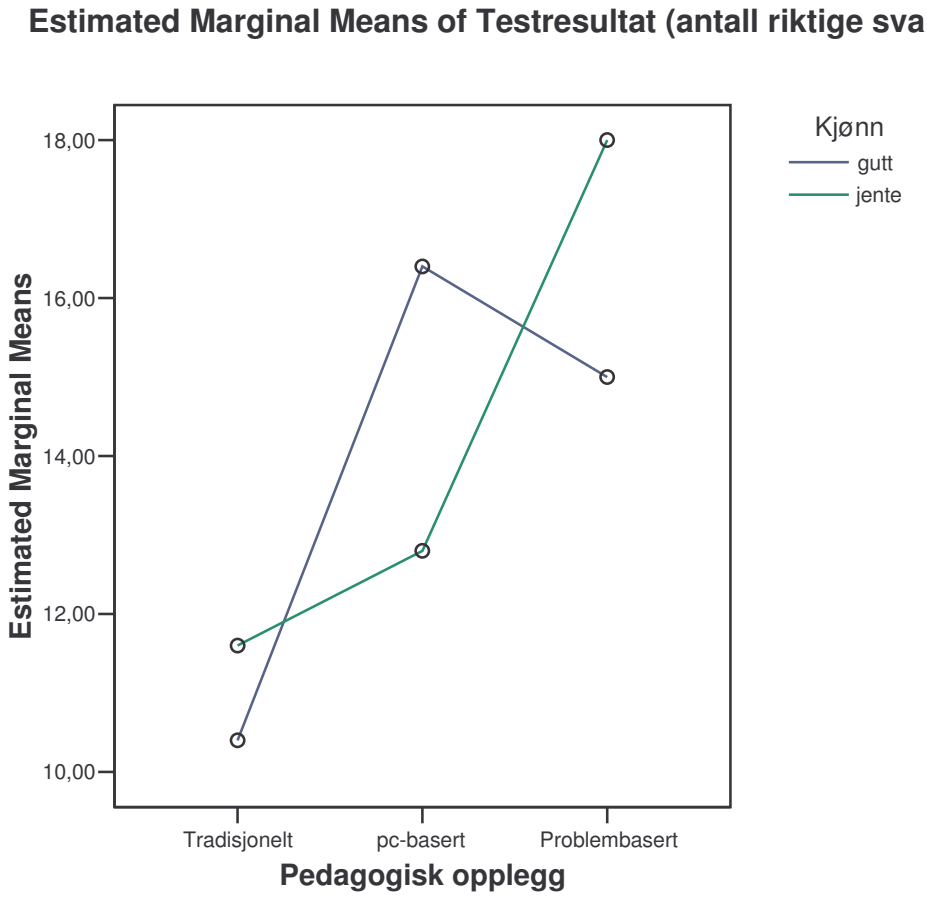
v1 Pedagogisk opplegg	v2 Kjønn	Mean	Std. Deviation	N
1,00 Tradisjonelt	1,00 gutt	10,4000	2,30217	5
	2,00 jente	11,6000	3,36155	5
	Total	11,0000	2,78887	10
2,00 pc-basert	1,00 gutt	16,4000	2,07364	5
	2,00 jente	12,8000	2,68328	5
	Total	14,6000	2,95146	10
3,00 Problembasert	1,00 gutt	15,0000	1,22474	5
	2,00 jente	18,0000	1,41421	5
	Total	16,5000	2,01384	10
Total	1,00 gutt	13,9333	3,19523	15
	2,00 jente	14,1333	3,75817	15
	Total	14,0333	3,42892	30

De tallene som er regnet ut for gutter og jenter samlet kjenner vi igjen fra Fig. 4.2. Gjennomsnittlig skår på matematikktesten er lavest for de elevene som fulgte det tradisjonelle opplegget med gjennomsnittlig 11 riktige svar i gruppen. Blant de som fulgte det pc-baserte opplegget ble gjennomsnittlig skår noe høyere, nemlig 14,6. Best gjennomsnittlig resultat fikk de elevene som fulgte den problembaserte undervisningen. Her ble gjennomsnitts-skåren 16,5. Vi har tidligere funnet ut at forskjellen mellom de tre gruppene var statistisk signifikant (Fig. 4.3).

Det nye med tallene som er vist i Fig. 4.6, er at vi her ser tall for gutter og jenter separat. Det er akkurat fem gutter og fem jenter i hver av de tre gruppene. Dette gjør det enkelt å benytte toveis variansanalyse. En nærmere inspeksjon av tallene i Fig. 4.6 viser at både blant gutter og jenter gikk det dårligst når de hadde fulgt det tradisjonelle opplegget (gjennomsnitt 10,4 og 11,6). Men når vi ser på de to andre gruppene får vi oss en overraskelse. Det viser seg at for guttenes del er de som har fulgt den pc-baserte undervisningen som får best resultat (gjennomsnittlig 16,4 riktige svar blant guttene mot bare 12,8 blant jentene). Blant de som har fulgt den problembaserte undervisningen, er det akkurat motsatt. Her der det jentene som oppnår best resultat med en gjennomsnittlig skår på 18,00 mot guttenes 15,00. La oss se hvordan dette ser ut når vi tegner et strekdiagram.

For å lage det aktuelle strekdiagrammet, gjør en det samme som ble beskrevet ovenfor (andre avsnittet i del 4.4). I tillegg trykker en på *Plots* og legger inn v1 under *Separate Lines* og v2 under *Horizontal Axis*. Så trykker en på *Add* og deretter *Continue* og *OK*. Resultatene som framkommer er vist i Fig. 4.7 og 4.8.

Fig. 4.7 Gjennomsnittlig skår på matematikktesten etter pedagogisk opplegg og kjønn. Strekdiagram.



Det framkommer et mønster som kanskje kan virke litt forvirrende. Men egentlig forteller diagrammet akkurat det samme som tallene vi gikk gjennom ovenfor. Begge grupper kommer dårligst ut med det tradisjonelle undervisningsopplegget (lavest gjennomsnittlig skår). Guttene kommer best ut når de følger det pc-baserte, og jentene kommer best ut når de følger det problembaserte opplegget. At det her er snakk om en statistisk interaksjonseffekt ser vi best når vi sjekker om linjene er parallelle. Parallelle linjer betyr at det ikke foreligger noen statistisk interaksjonseffekt. Når linjene viser svært forskjellig retning fra gruppe til gruppe, tyder det på at det foreligger en interaksjon. At det foreligger en statistisk interaksjon, betyr i dette tilfellet at sammenhengen mellom pedagogisk opplegg og matematikkresultater er forskjellig for gutter og jenter. Mer generelt kan vi formulere det slik vi har gjort tidligere. En statistisk interaksjon foreligger når sammenhengen mellom to variabler (retning, form eller styrke) er avhengig av en tredje variabel.

Men så kommer vi til et viktig og avgjørende spørsmål: Hvordan kan vi vite at den statistiske interaksjonseffekten er signifikant? At den ikke er et utslag av tilfeldige variasjoner i data. For å finne ut mer om dette, må vi se på tallene som er gjengitt i Fig. 4.8.

Fig. 4.8: Skår på matematikktesten etter pedagogisk opplegg og kjønn. Toveis variansanalyse.

Tests of Between-Subjects Effects

Dependent Variable: v3 Testresultat (antall riktige svar)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	214,567[a]	5	42,913	8,148	,000
Intercept	5908,033	1	5908,033	1121,778	,000
v1	156,067	2	78,033	14,816	,000
v2	,300	1	,300	,057	,813
v1 * v2	58,200	2	29,100	5,525	,011
Error	126,400	24	5,267		
Total	6249,000	30			
Corrected Total	340,967	29			

a. R Squared = ,629 (Adjusted R Squared = ,552)

Tabellen fra SPSS i Fig. 4.8 viser flere forskjellige F-verdier. På linjen som begynner med ”v1” vises en F-verdi (14,861) med 2 og 24 frihetsgrader og forteller at sammenhengen mellom gruppe (definert ved pedagogisk opplegg) og testresultat er signifikant ($p<0,001$). På linjen som begynner med ”v2” vises en F-verdi (0,057) med 1 og 24 frihetsgrader som forteller at sammenhengen mellom kjønn og testresultater ikke er statistisk signifikant ($p=0,813$). På linjen som begynner med ”v1*v2” gjengis en F-verdi (5,525) med 2 og 24 frihetsgrader som forteller at interaksjonen mellom pedagogisk opplegg og kjønn er statistisk signifikant på $p<0,05$-nivået ($p = 0,011$ er mindre enn $p = 0,050$). Dette er den testen vi var ute etter. Nullhypotesen er at sammenhengen mellom pedagogisk opplegg og testskår er lik for gutter og jenter. Den alternative hypotesen sier at sammenhengen ikke er lik for gutter og jenter. Siden vi har oppnådd signifikans, kan vi forkaste nullhypotesen. Vi har sannsynliggjort at det i populasjonen er forskjell mellom gutter og jenter når det gjelder sammenhengen

mellom pedagogisk opplegg og testresultat. Gutter ser ut til å profitere mest på et pc-basert undervisningsopplegg, mens jentene ser ut til å profitere mest på problembasert læring.

Når alle testene har en frihetsgrad nr 2 på 24, er det fordi vi bruker det antall frihetsgrader som er assosiert med feilvarians (error). For å få fram F-verdiene i Tabell 4.8, kan vi bare dele ”mean square” med ”error”, og antall frihetsgrader leser vi ut av tabellen under df (degrees of freedom).

I variansanalyse finnes det flere valgmuligheter med hensyn til hvordan en regner ut kvadratsummene og de tilsvarende variansestimatene. Vi har ovenfor brukt en metode som automatisk benyttes i SPSS, dersom en ikke eksplisitt gir beskjed om noe annet. Den kalles type III kvadratsummer (Type III Sum of Squares). Metoden har også mange andre navn, og kalles blant annet Metode 1, regresjonsmetoden, uveide gjennomsnitters metode eller den unike metoden. Hver enkelt effekt (for eksempel effekten av kjønn) er justert for alle andre effekter (i vårt eksempel: pedagogisk opplegg og interaksjonen mellom pedagogisk opplegg og kjønn). Tabachnik & Fidell (2001) anbefaler denne metoden når en har eksperimentelle data (slik tilfellet er her). Cramer (2003) mener denne metoden også kan anbefales når en bruker flerveis variansanalyse på ikke-eksperimentelle data, så lenge en er interessert i det unike bidraget av hver enkelt variabel og det unike bidraget til eventuelle interaksjonseffekter.

Vi skal ikke her vise hvordan en regner ut kvadratsummer, kvadratgjennomsnitt og F-verdier når en gjennomfører faktoriell varianananalyse. Men hele tiden handler det om å se hvor mye av variansen i den avhengige variabelen som kan forklares ut fra hver av de uavhengige variablene og ut fra interaksjonseffektene. I SPSS kan en spesifisere hvor mange nivå av interaksjonseffekter en ønsker å teste ut og en kan et stykke på vei bestemme hvordan hovedeffekter og interaksjonseffekter skal kontrolleres for øvrige hovedeffekter og interaksjonseffekter. Også i faktoriell variansanalyse kan en teste om variansene i de ulike gruppene er signifikant forskjellige, og det er den samme Levenes test som vi har nevnt tidligere som blir benyttet. For å få denne med i utskriften må en huke av på *Homogeneity Test* under *Options*.

I eksemplene som er brukt ovenfor har vi benyttet data som har hatt like stort antall i hver celle eller gruppe. I den enveis variansanalysen var det ti i hver gruppe. I den toveis variansanalysen var det fem i hver gruppe. Det betyr at de to uavhengige variablene er uavhengige. De korrelerer ikke med hverandre. Dersom en bruker ett av de vanligste assosiasjonsmålene for krysstabeller, vil en finne at sammenhengen er nøyaktig null. Dermed er det helt entydig hvordan de henger sammen med den avhengige variabelen. Den variansen som forklares av den ene variabelen overlapper ikke med den variansen som forklares av den andre.

Dersom de to uavhengige variablene (faktorene) er korrelerte med hverandre, oppstår det et problem. En del av den variansen den ene forklarer i den avhengige variabelen kan også forklares ut fra den andre. De overlapper med andre ord. Så spørsmålet er hvordan en skal dele den forklarte variansen mellom dem på en fornuftig måte. Her vil det eksistere ulike løsninger alt etter hva slags problemstilling en forsøker å belyse. Det finnes tre ulike måter å fordele variansen mellom faktorene (Overall & Spiegel, 1969). I en toveis variansanalyse blir resultatet for interaksjonseffekten det samme uansett metode, men resultatene for hovedeffektene vil variere (Jaccard, 1998).

Kanskje er den løsningen som i SPSS er valgt som standardløsningen (Type III) oftest den beste, nemlig at en justerer hver effekt for alle de andre effektene (også eventuelle interaksjonseffekter), slik at det en sitter tilbake med hele veien er unike sammenhenger. Den klassiske eksperimentelle løsningen (Type II) er en prosedyre der hovedeffektene blir justert for alle andre hovedeffekter, mens interaksjonene blir kontrollert for alle andre effekter, bort sett fra høyere ordens interaksjoner. Tabachnik og Fidell (2001) anbefaler denne metoden når en analyserer data fra ikke-eksperimentelle undersøkelser der en særlig er interessert i hovedeffektene.

4.5 Toveis variansanalyse med repeterte målinger

I kapittel 3 lærte vi om to ulike typer t-tester for signifikanstesting av forskjeller mellom to grupper. Den ene testen skal benyttes når en har to uavhengige (korrelerte) grupper. Et eksempel på en slik situasjon har en dersom en ønsker å teste forskjellen mellom menn og kvinner på en eller annen metrisk variabel. Men dersom en har målt en gruppe individer to ganger på en metrisk variabel, for eksempel før og etter en eller annen intervensjon, og ønsker å se om gjennomsnitts-skåren har endret seg, må en benytte t-test for avhengige (korrelerte) grupper.[3]

Dersom en har målt de samme personene tre eller flere ganger på den samme metriske variabelen og ønsker å undersøke om gjennomsnittsskåren har endret seg over tid, kan en imidlertid ikke bruke t-testing. I stedet bruker en enveis variansanalyse med repeterte målinger.

La oss bruke et tenkt eksempel. En gruppe på 16 musikkskoleelever har i løpet av ett skoleår blitt testet tre ganger. Første gangen var like etter sommerferien. Andre gang var rett før juleferien og tredje gang var etter påske. Hver gang ble de vurdert av en liten gruppe musikere som etter å ha hørt hver enkelt elev spille på sitt instrument ble enige om en skår på en poengskala fra null til tolv der null stod for svært dårlige ferdigheter og tolv stod for svært gode ferdigheter. Resultatene ble slik som gjengitt i tabell 4.3.

Disse variablene kan legges inn på regnearket I SPSS slik vi har lært å gjøre tidligere. Vi kan legge inn kjønn som v1 og skårene på de tre tidspunktene som v2, v3 og v4. Deretter går en inn på *Analysis*, *General Linear Model*, *Repeated Measures*, kaller *Within Subject Factor* for Skår, setter *Number of Levels* til 3 (fordi vi har målinger på tre tidspunkt), og trykker deretter på *Define*. Legg deretter inn v1, v2 og v3 under *Within Subjects Variables*, og v1 under Between Subjects Factor. Gå deretter inn på *Plots*, legg v1 inn på *Separate Lines* og Skår inn på *Horizontal Axis*. Trykk deretter på *Add* og så *Continue*. Trykk så på *Options* og legg *v1*Skår* inn i *Display Means for*. Sett til slutt haker i *Descriptive Statistics* og *Estimates of Effect Size* og trykk på *Continue* og *OK*.

[3] Dersom en hadde målt en del ektepar på en eller annen metrisk variabel og ønsket å teste forskjeller mellom menn og kvinner, ville vi også ha avhengighet i data (korrelerte data). Også i denne situasjonen bør en benytte t-test for korrelerte data.

Tabell 4.3: Gjennomsnittlig skår for ferdigheter på musikkinstrument på tre tidspunkt

Person nr.	Kjønn	Skår1	Skår 2	Skår 3
1	1	10	11	11
2	1	8	10	13
3	1	6	9	10
4	1	7	11	14
5	1	7	10	12
6	1	9	9	12
7	1	8	12	15
8	1	11	13	12
9	2	9	12	16
10	2	10	12	17
11	2	9	13	15
12	2	5	8	9
13	2	7	11	15
14	2	6	12	18
15	2	8	10	13
16	2	8	11	14

Dersom du har klart å følge instruksjonene i boksen ovenfor, vil du få ut den tabellen som er gjengitt i Fig 4.9. Den viser at gjennomsnittsskåren for alle (gutter og jenter samlet) er lik 8,00, 10,88 og 15,50 (dersom vi avrunder til to desimaler). Musikkferdighetene har med andre ord bedret seg markant. For guttenes del øker gjennomsnittlig skår fra 8,25 via 10,63 til 12,38. Jentene oppnår gjennomsnittstallene 7,75, 11,13 og 14,63. Jentenes instrumentferdigheter øker med andre ord litt kraftigere enn guttenes. Dette kan vi kanskje se enda tydeligere i Fig. 4.10, der de samme resultatene er gjengitt grafisk.

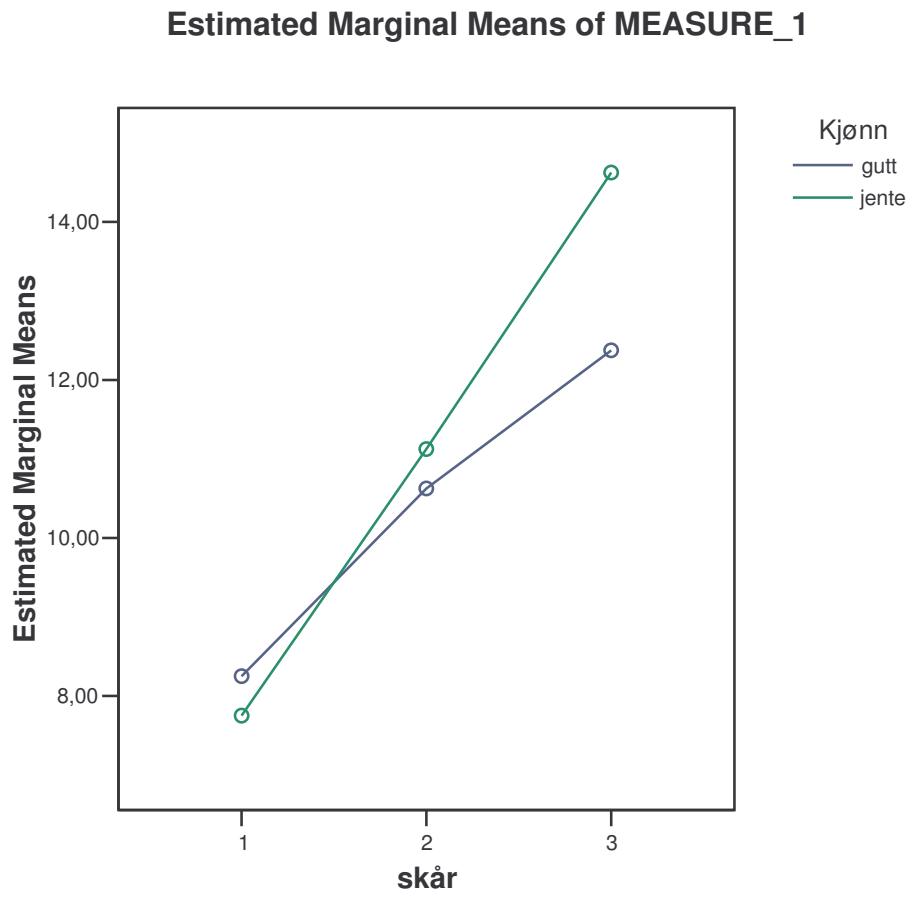
Fig. 4.9: Skår på instrumentferdigheter etter tidspunkt og kjønn

Descriptive Statistics

	v1 Kjønn	Mean	Std. Deviation	N
v2 Skår tidspunkt 1	1,00 gutt	8,2500	1,66905	8
	2,00 jente	7,7500	1,66905	8
	Total	8,0000	1,63299	16
v3 Skår tidspunkt 2	1,00 gutt	10,6250	1,40789	8
	2,00 jente	11,1250	1,55265	8
	Total	10,8750	1,45488	16
v4 Skår tidspunkt 3	1,00 gutt	12,3750	1,59799	8
	2,00 jente	14,6250	2,77424	8
	Total	13,5000	2,47656	16

Men når vi ser disse resultatene, er det to spørsmål som er naturlig å stille. Er økningen over de tre tidspunktene for alle sammen (både guttene og jentene sett under ett) så kraftig at den er statistisk signifikant? Nullhypotesen er i dette tilfellet at det ikke er forskjell i gjennomsnitt mellom de tre tidspunktene. Det er dessuten naturlig å spørre om endringen har vært signifikant forskjellig for gutter og jenter. Nullhypotesen er selvsagt at endringene har vært helt parallelle for gutter og jenter.

Fig. 4.10: Gjennomsnittlig instrumentferdighetsskår etter tidspunkt og kjønn.



Forutsatt at vi har gjort alt slik som beskrevet ovenfor, vil vi også ha fått ut den tabellen som vises i Fig. 4.11. På første linje (skår – sphericity assumed) vises en test av endringer i skår på testen. F-verdien er 72,135 og antall frihetsgrader er 2 og 28. Det gir en p-verdi som er lavere enn 0,001 (altså $p<0,001$). Det betyr at det er en signifikant forskjell i musikkferdigheter mellom de tre tidspunktene, og fra tidligere vet vi at det er snakk om en sterk økning fra tidspunkt 1 via tidspunkt 2 til tidspunkt tre. Økningen er så sterk at det tilsvarer en kvadrert partiell eta på 0,837.

På linjen som begynner med skår*v1 – Sphericity assumed finner vi en annen F-test. Her er det snakk om en test av interaksjonseffekten mellom tidspunkt og kjønn. Det er en test for å se om vi kan forkaste antagelsen om at endringen i de to gruppene (gutter og jenter) har vært parallell. Vi ser at F-verdien er lik 4,617 og at antall frihetsgrader er 2 og 28. Dette gir en p-verdi på 0,018. Dermed kan vi trekke den konklusjon at resultatet er signifikant på $p<0,05$-nivået. Det betyr at vi kan forkaste nullhypotesen om at endringen over tid har vært parallell for gutter og jenter. Det ser ut til at vi kan konkludere med at endringen har vært signifikant mer positiv blant jenter enn blant gutter. Interaksjonseffekten mellom kjønn og tidspunkt forklarer 24,8% av variansen i skår på musikkferdigheter (partiell eta kvadrert er på 0,248).

Når en skal undersøke interaksjonseffekter er det oftest best å ta ut et linjediagram slik som det vi har gjengitt i Fig. 4.10. Her ser vi at stigningen i skår både fra første til andre tidspunkt og fra andre til tredje tidspunkt er sterkere blant jenter enn blant gutter. Jo mer parallelle linjene er, desto mindre er interaksjonseffekten. Jo mindre parallelle de er, desto sterkere er interaksjonseffekten. Endringsmønsteret behøver ikke være så enkelt som i dette tilfellet. Det kunne for eksempel vært en interaksjonseffekt mellom første og andre tidspunkt og deretter parallell endring. Signifikanstestingen er ikke avhengig av hvordan mønsteret ser ut, og er like sensitiv overfor ulikheter i endring uansett hvor i tidsforløpet de måtte finne sted.

Fig. 4.11: Musikkferdighetsskår etter tidspunkt og kjønn. Toveis univariat variansanalyse med repeterte målinger.

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
skår	Sphericity Assumed	242,167	2	121,083	72,135	,000	,837
	Greenhouse-Geisser	242,167	1,333	181,611	72,135	,000	,837
	Huynh-Feldt	242,167	1,526	158,646	72,135	,000	,837
	Lower-bound	242,167	1,000	242,167	72,135	,000	,837
skår * v1	Sphericity Assumed	15,500	2	7,750	4,617	,018	,248
	Greenhouse-Geisser	15,500	1,333	11,624	4,617	,036	,248
	Huynh-Feldt	15,500	1,526	10,154	4,617	,029	,248
	Lower-bound	15,500	1,000	15,500	4,617	,050	,248
Error(skår)	Sphericity Assumed	47,000	28	1,679			
	Greenhouse-Geisser	47,000	18,668	2,518			
	Huynh-Feldt	47,000	21,370	2,199			
	Lower-bound	47,000	14,000	3,357			

Utskriften fra SPSS inneholder imidlertid en hel del informasjon ut over det vi har presentert her. Blant annet får vi resultatet av en test som kalles Mauchly's Test of Sphericity. Dersom den viser signifikans, er ikke den F-testen vi nettopp har utført helt presis. Da kan en i stedet bruke en av de tre andre testene som er presentert i Fig. 4.11 (Greenhouse-Geisser, Huynh-Feldt eller Lower-bound). Vi skal ikke her gå nærmere inn på de tre testene, men overlater de som har spesielle interesser eller behov i retning av å lære seg disse testene til andre tekster (For eksempel Field, 2000).

Det er selvsagt også mulig å utføre en enveis variansanalyse med repeterte målinger, det vil si en analyse der en ikke har noen annen faktor enn f. eks. tidspunkt for målingen. Dette kan en

gjøre ved å unnlate å legge inn en ”Between Subjects Faktor” (i vårt eksempel ovenfor benyttet vi variabelen v1 - kjønn). Utskriften vil i dette tilfellet mangle en del av den informasjonen som er knyttet til ”between subjects”-faktoren. Den øvrige informasjonen skulle være lett forståelig ut fra det vi har forklart ovenfor.

Eksempelet vi har gjennomgått i dette avsnittet viser en toveis variansanalyse med repeterte målinger der vi har to grupper og tre tidspunkt. En kan ha et hvilket som helst antall grupper fra to og oppover, og et hvilket som helst antall måletidspunkt fra to og oppover. En kan dessuten ha flere gruppevariabler (faktorer). Når modellene blir svært omfattende og komplekse kreves det imidlertid mange enheter (subjekter), og tolkningen av resultatene kan bli vanskeligere.

Referanser

Aron, A. & Aron, E.N. (1999). *Statistics for psychologists* (Second Edition). Upper Saddle River, New Jersey: Prentice Hall.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, New Jersey: Lawrence Erlbaum.

Cramer, D. (2003). *Advanced quantitative data analysis.* Maidenhead, England: Open University Press.

Field, A. (2000). *Discovering statistics using SPSS for Windows*. London: Sage.

Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian Inheritance. *Transaction of the Royal Society of Edinburgh, 52,* 399-433.

Fisher, R.A. (1925). *Statistical methods for research workers.* Edinburgh: Oliver and Boyd.

Jaccard, J. (1998). *Interaction effects in factorial analysis of variance*. Thousand Oaks, California: Sage.

Overall, J.E. & Spiegel, D.K. (1969). Concerning least squares analysis of experimental data. *Psychological Bulletin, 72*, 311-322.

Rosnow, R.L. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276-1284.

Serlin, R.C. (2005). Analysis of variance. I Everitt, B.S. & Howell, D.C. (red.), *Encyclopedia of statistics in behavioural science, Vol. 1* (pp. 52-56). London: Wiley.

Tabachnik, B.G. & Fidell, L.S. (2001). *Using multivariate statistics* (Fourth Edition). Boston/London: Allyn & Bacon.

Weinfurt, K.P. (1995). Multivariate analysis of variance. I Grimm, L.G. & Yarnold, P.R. (Red.), *Reading and understanding multivariate statistics* (s. 245-276). Washington D.C.: American Psychological Association.

Kap 5: Prinsipal komponentanalyse, faktoranalyse og reliabilitet

De første trinnene i en data-analyse er vanligvis (a) å undersøke enveis-fordelinger og gjøre seg opp en mening om hvordan enkeltvariablene fungerer og (b) å se på de bivariate sammenhengene mellom variabler. Disse første to trinnene er begge viktige. For det første er det nødvendig å kjenne de mer enkle og grunnleggende egenskapene til de variablene en skal analysere på og eventuelt gjennomføre nødvendige omkodinger og transformasjoner for å gjøre variablene mer anvendbare. For det andre er det nødvendig å vite om variablene har de egenskapene som kreves for å ta i bruk mer avanserte statistiske teknikker, dersom problemstillingene inviterer til å ta i bruk slike.

Forskere som ikke har særlig god kjennskap til den multivariate statistikken, nøyer seg gjerne med å foreta slike enklere analyser, og kan dermed gå glipp av viktig informasjon som finnes i data. De mer rutinerte og statistikk-glade forskerne beveger seg ofte for raskt over de mer elementære analysene og risikerer derfor å gå glipp av den ofte viktige informasjonen som er tilgjengelig på dette nivået. De risikerer dessuten å misbruke statistiske metoder fordi de ikke nøye nok undersøker om de bryter noen av forutsetningene for bruk av mer avanserte statistiske teknikker.

Før vi beveger oss inn i den multivariate statistikkens verden, er det viktig å få sagt at en ikke skal velge mer avanserte statistiske teknikker enn høyst nødvendig. Dersom en skal anvende multivariate statistiske teknikker, bør dette være begrunnet i at en vil ha svar på spesifikke forskningsspørsmål som bare kan besvares ved bruk av disse mer avanserte teknikkene. Etter denne advarselen skal vi se nærmere på ett viktig område; analyse av grupper av variabler der relasjonene er symmetriske (Rosenberg, 1968) med tanke på datareduksjon. Datareduksjon vil her si at en prøver å redusere et større antall variabler til et mindre antall ved å lage sumskårer eller indekser.

5.1 Målemodeller og ulike typer indikatorer

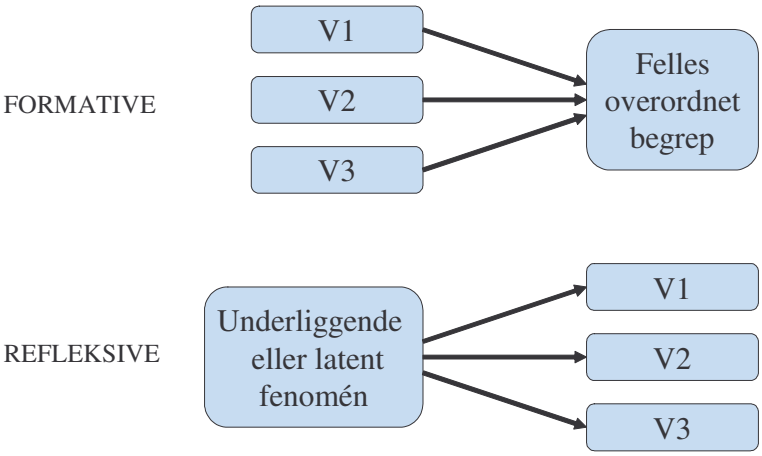
Sumskårer og indekser er redskaper til å redusere en ofte omfattende og noen ganger uoversiktlig mengde enkeltvariabler til et mindre og mer håndterbart antall. Ved å gi avkall på mindre vesentlige detaljer, kan en slik skape enkelhet og oversikt.

Et vanlig kriterium på at en gruppe variabler kan kombineres til en sumskår, er at de er høyt korrelerte innbyrdes. Disse høye interkorrelasjonene tas gjerne som et tegn på at de variablene som inngår har noe til felles. Noen ganger sier vi at de er uttrykk for et underliggende fenomén eller begrep. Indikatorer som bygger på en antakelse om at de reflekterer en underliggende latent faktor, kalles ofte refleksive indikatorer (Mastekaasa, 1987). DeVellis (1991) kaller et sett av ledd som reflekterer en underliggende latent variabel for en skala.

Høye interkorrelasjoner er imidlertid ikke det eneste kriteriet på hvilke variabler som bør inngå i en sumskår eller en skala. Det er ikke engang det viktigste kriteriet. Noen ganger kan en konstruere indekser på rent teoretisk eller begrepsmessig grunnlag og behøver ikke å bry seg så mye om å studere de innbyrdes sammenhengene mellom de enkeltvariabler og ledd som skal brukes. Ett eksempel er indekser for bruk av alkohol basert på spørsmål om inntak av henholdsvis øl, vin og brennevin. Dersom en er interessert i å beregne det rapporterte inntaket av ren alkohol i en befolkning, spiller det liten rolle om bruk av en type alkoholholdige drikker er positivt eller negativt korrelert med bruken av andre alkoholholdige drikker. Vi kan i prinsippet tenke oss at bruk av en type alkohol er negativt korrelert med bruk av andre typer alkohol, at det å ha et høyt forbruk av en type alkoholiske drikker henger sammen med redusert forbruk av andre. Likevel er et samlet mål på inntak per tidsenhet av ren alkohol eller såkalte alkoholenheter i mange sammenhenger meningsfylt. Hvis intern konsistens i slike tilfeller likevel brukes som kriterium på hvilke variabler som kan kombineres til en indeks, kan det føre til at en unnlater å konstruere indekser som det ellers kan være godt begrepsmessig eller teoretisk grunnlag for.

På samme måte kan vi tenke oss at vi lager indekser som måler inntaket av fett i kosten ved å bygge på opplysninger om bruk av ulike matslag. En slik indeks kan være nyttig med tanke på å predikere forekomsten av hjerte- karsykdom i en befolkning, uansett den innbyrdes sammenhengen mellom forbruket av ulike matprodukter. De enkeltvariablene som inngår i en slik indeks kalles formative indikatorer (Mastekaasa, 1987) (se Fig. 5.1). Mens refleksive indikatorer ses på som en effekt av eller et produkt av en underliggende eller latent faktor, betraktes de formative indikatorene i en viss forstand som årsaker til det en forsøker å summere opp gjennom en indeks (Bagozzi & Fornell, 1982).

Fig. 5.1: Formative og refleksive indikatorer



Det finnes flere statistiske modeller som handler om hvordan variabler reflekterer noe underliggende eller latent. Innen disse modellene skilles det mellom den delen av en variabels varians som reflekterer den underliggende latente faktoren og den øvrige variansen som kalles feilvarians. Den klassiske målemodellen (DeVellis, 1991) bygger på følgende forutsetninger:

1) Mengden av feilvarians som assosieres med individuelle ledd varierer tilfeldig
2) Ett ledds feilvarians er ukorrelert med alle andre ledds feilvarians
3) Feiltermene ikke er korrelert med den sanne skåren på den latente variabelen.

De to første forutsetningene er vanlige antakelser som en rekke statistiske analyser er basert på. En modell som stiller enda strengere krav til data er den som kalles parallelle tester. Her forutsettes det

a) at den latente variabelen virker like sterkt inn på hver av variablene en har målt, og at
b) mengden av feilvarians er like stor for alle variabler.

Alle disse strenge kravene til forholdet mellom latent variabel og de variablene en har målt er imidlertid ikke nødvendige for å kunne si noe om forholdet mellom sanne skårer og observerte skårer. En modell som kalles essensielt tau-ekvivalente tester (essentially tau equivalent tests – eller randomly parallel tests) bygger på at mengden med feilvarians på en målt variabel ikke behøver å være lik feilvariansen på de andre variablene. Det finnes enda mer liberale modeller, f.eks. Den kongeneriske modellen. Under denne modellen er det ikke

engang nødvendig at den latente variabelen virker like sterkt inn på hver av variablene som blir målt. Den forutsetter bare (i tillegg til de klassiske forutsetningene nevnt ovenfor) at variablene deler en felles underliggende variabel. En mer utfyllende presentasjon og drøfting av de ulike målemodellene er gitt i DeVellis (1991). En enda mer liberal modell er den generelle faktormodellen som tillater at variablene som inngår i en skala er relatert til flere forskjellige latente variabler.

De forskjellige målemodellene står i en spesiell relasjon til hverandre. Den kongeneriske modellen er et spesialtilfelle av den generelle faktormodellen. Essensielt tau-ekvivalente tester er et spesialtilfelle av den kongeneriske modellen. Parallelle tester er på sin side et spesialtilfelle av de essensielt tau-ekvivalente testene. Kravene som blir stilt til de ulike målemodellene er vist i Fig. 5.2.

	Generell faktor-modell	Konge-nerisk	Essensielt tau-ekvivalent	Parallelle tester
Mengde feilvarians varierer tilfeldig	*	*	*	*
Ukorrelerte feiltermer	*	*	*	*
Ingen korrelasjon mellom feiltermer og latent variabel	*	*	*	*
Bare en latent variabel		*	*	*
Latent variabel virker like sterkt på alle de målte			*	*
Mengde feilvarians like stor for alle ledd				*

Den kongeneriske modellen og den generelle faktormodellen forutsetter selvsagt at hver variabel i det minste i noen grad reflekterer latente variabler.

Fig. 5.2: En oversikt over noen sentrale målemodeller

Det viktigste kriteriet på at variabler kan kombineres til en indeks er den teoretiske eller begrepsmessige begrunnelsen for å foreta en slik kombinasjon. Uten at vi kan begrunne en indeks teoretisk eller begrepsmessig, blir den meningsløs, begrepsmessig uklar eller for sammensatt til å være nyttig. Det begrepsmessige eller teoretiske rasjonalet er derfor alltid det mest avgjørende av de kriteriene vi anvender.

Hvordan vi videre vurderer mulighetene for å kombinerer variabler til en indeks avhenger av om en har å gjøre med en formativ eller en refleksiv situasjon. Dersom vi har å gjøre med en refleksiv situasjon, anvender vi såkalte indre konsistenskriterier. De variablene som skal kombineres til en indeks, skal i slike situasjoner korrelere positivt innbyrdes. Dersom vi har å gjøre med en formativ situasjon, anvender vi i stedet eksterne eller ytre konsistenskriterier (Lazarsfeld, 1959). Dette betyr at de variablene som inngår i en indeks bør korrelere på en konsistent måte med de utenforstående variablene som er relevante i de analysene vi ønsker å gjøre.

Vi skal her gå grundigere inn på analyser som er aktuelle når vi står overfor refleksive indikatorer. Utgangspunktet er interkorrelasjonene mellom samtlige variabler som inngår i en skala. Interkorrelasjonene mellom variabler settes gjerne opp i en korrelasjonsmatrise. Slike korrelasjonsmatriser kan inspiseres visuelt, og forskere som er trenet i å inspisere korrelasjonsmatriser kan ofte se om de har en felles, underliggende dimensjon. En korrelasjonsmatrise der alle korrelasjonene er omtrent jevnstore og alle korrelerer positivt (eller en kan få alle til å korrelere positivt ved å snu på noen av variablene) er et sikkert tegn på en underliggende dimensjon.

Ofte får en imidlertid et mindre konsistent bilde ved å inspisere en korrelasjonsmatrise. Noen korrelasjoner er høye, andre er lave, noen er positive og andre er negative. I slike tilfeller trenger vi hjelp av mer avanserte statistiske teknikker for å finne ut hva vi skal gjøre. Vi trenger en teknikk som kan fortelle hvor mange grupper av variabler (ofte kalt faktorer, dimensjoner eller komponenter) som finnes i variabelsettet, og hvordan de grupperer seg sammen. Det finnes en hel gruppe av slike teknikker. En undergruppe er faktoranalyse. En annen undergruppe er prinsipal komponentanalyse (noen ganger bare kalt komponentanalyse). I denne framstillingen beskrives faktoranalyse og prinsipal komponentanalyse stort sett under ett, selv om en prinsipal komponentanalyse strengt tatt ikke er det samme som en faktoranalyse.

Mens faktoranalysen er basert på en antakelse om at det eksisterer en underliggende eller latent uobserverbar faktor som gir seg utslag i en rekke forhold som kan observeres eller måles, med andre ord en refleksiv relasjon mellom faktor og observerte variabler, bygger ikke den prinsipale komponentanalysen på en slik tankegang. Den prinsipale komponentanalysen kan vi i stedet betrakte som en teknikk som på en nokså ukomplisert måte produserer et sett av sumskårer på grunnlag av et sett variabler.

La oss begynne beskrivelsen av de faktoranalytiske teknikkene med å sitere fra Kerlingers utmerkede bok om metoder i atferdsforskningen:

> Because of its power and elegance, factor analysis can be called the queen of analytic methods. ... factor analysis is an extremely powerful and useful approach to behavioral data, one that can help solve heretofore intractable research problems.
>
> Kerlinger, 1973, s.659

5.2 Eksploratorisk faktoranalyse

5.2.1 Hva er faktoranalyse?

Faktoranalyse er en metode for å bestemme antall underliggende dimensjoner og egenskaper ved disse ut fra et sett variabler. Mer presist kan en si at faktoranalyse er en metode for å bestemme k faktorer fra et sett på n variabler der k er mindre enn n. Det er en metode til å trekke ut felles varians fra et sett målinger. Faktoranalyse er ikke én teknikk, men et sett av statistiske teknikker som alle har det til felles at de skal gi et antall variabler en representasjon i form av et mindre antall hypotetiske variabler (faktorer).

Faktoranalyse kan ha ulike formål:

1) Identifisere underliggende begreper eller faktorer som forklarer interkorrelasjonene mellom et sett variabler

2) Teste hypoteser om strukturene i et variabelsett

3) Redusere et større antall variabler i et mindre antall variabler som er et produkt av det opprinnelige settet

4) Avgjøre hvor mange dimensjoner som skal til for å representere et sett av variabler.

(Norusis, 1985, s.123)

De faktoranalytiske teknikkene kan deles inn i to kategorier. Den ene kategorien er anvendelig når forskeren ikke har noen sikker mening om hvor mange underliggende dimensjoner der finnes i data. Han trenger da et statistisk redskap som kan gi svar på hvor få dimensjoner han kan greie seg med, hvilke variabler som bør inngå i hver dimensjon, og hvordan han best kan konstruere et sett av sumskårer. De typene faktoranalyse som kan anvendes på denne måten kalles gjerne *eksplorerende faktoranalyse*. Eksplorerende faktoranalyse har tradisjonelt vært den vanligste formen for faktoranalyse.

Faktoranalyse kan også brukes til å teste hypoteser en allerede på forhånd måtte ha om antall og sammensetningen av faktorer. Forskeren kan f.eks. forestille seg at det finnes to underliggende dimensjoner og at det er bestemte variabler som tilhører hver av disse. Dersom faktoranalyse brukes til å teste en slik spesifikk hypotese ved å sammenlikne den spesifiserte modellen med data, kalles dette *konfirmatorisk faktoranalyse*. På skikkelig norsk ville vi vel oversette konfirmatorisk faktoranalyse med *bekreftende faktoranalyse*.

5.2.2 Faktoranalysens og den prinsipale komponentanalysens 3 trinn

Både faktoranalyse og prinsipal komponentanalyse kan deles inn i tre trinn:

1) Beregning av en kovarians-matrise (eller kanskje mest vanlig: en korrelasjons-matrise)
2) Ekstraksjon av faktorer (eller prinsipale komponenter)
3) Rotasjon av faktorer (eller komponenter)

Korrelasjonsmatrisen er det råmaterialet som skal bearbeides med tanke på å redusere det opprinnelige antallet variabler til et lavere antall dimensjoner. Vanlige faktoranalytiske teknikker og prinsipal komponentanalyse forutsetter at korrelasjonene er produkt-moment-korrelasjoner eller tilsvarende (for eksempel phi, point-biserial-korrelasjonen eller Spearmans rangkorrelasjonskoeffisient) (Gorsuch, 1988, s.239). Dersom en anvender produkt-moment-korrelasjoner, bør en undersøke om variablene som inngår kan betraktes som intervall-variabler og at relasjonene mellom variablene kan beskrives som lineære sammenhenger.

Ekstraksjonen gir oss holdepunkter for hvor mange faktorer vi bør ende opp med, og er dessuten et nødvendig skritt på veien til en endelig faktorløsning. Ekstraksjonen kan beskrives som en prosess som foregår i et mange-dimensjonalt rom. Hver enhet kan vi forestille oss som punkter i dette rommet. I dette rommet, som har like mange dimensjoner som antall variabler som inngår i analysen, leter datamaskinen etter en retning eller hoveddimensjon. Når denne er funnet, danner denne den første faktoren. Deretter fjernes denne dimensjonen på kunstig vis fra rommet. Deretter går datamaskinen på jakt etter den nest viktigste dimensjonen. Når denne er funnet, fjernes også den fra rommet. Dette medfører at hver ny dimensjon forklarer mindre varians. På ett eller annet tidspunkt forklarer de nye faktorene så lite varians at de ikke lenger er interessante.

Etter at datamaskinen har foretatt en slik ekstraksjon, vil vi oftest oppdage at vi ikke har fått fram noe begrepsmessig interessant resultat. Alle de opprinnelige variablene inngår gjerne i mer enn en faktor, og hver faktor består av et stort antall variabler. For å komme ut av dette kaoset og oppnå et resultat som er enklere og mer oversiktlig, foretar programmet en såkalt rotasjon av de opprinnelige faktorene.

Rotasjonen gir oss en løsning som er enklere å fortolke enn den uroterte. Rotasjonen kan foregå etter mange ulike kriterier eller kombinasjoner av kriterier. Et hovedpoeng er å få til en enklest mulig relasjon mellom opprinnelige variabler og faktorer. Noen rotasjonsmetoder fungerer slik at de leter etter løsninger der hver variabel lader høyt bare på en faktor, og lavt på de øvrige.

5.2.3 Prinsipal komponentanalyse

Når vi skal forklare det som skjer i en faktoranalyse rent teknisk, er det enklest å ta utgangspunkt i prinsipal komponentanalyse. Prinsipal komponentanalyse er matematisk sett en svært enkel teknikk. Idéen til denne analysen ble opprinnelig formulert av Pearson så tidlig som i 1901, men senere utviklet og grundigere beskrevet av Hotelling på begynnelsen av 1930-tallet (Hotelling, 1933; Dunteman, 1989).

I prinsipal komponentanalyse blir hver faktor estimert som en lineær kombinasjon av de opprinnelige variablene.

Den første prinsipale komponenten y_1 er en lineær kombinasjon av variablene x_1, x_2, ..., x_p:

$$y_1 = a_{11}x_1 + a_{12}x_2 + ... + a_{1p}x_p \qquad (5.1)$$

tilsvarende er den andre prinsipale komponenten y_2 en lineær kombinasjon av de opprinnelige variablene:

$$y_2 = a_{21}x_1 + a_{22}x_2 + ... + a_{2p}x_p \qquad (5.2)$$

osv.

a er vekter som avgjør hvor mye hver enkelt variabel skal telle med i sumskåren.

Ved beregning av den første prinsipale komponenten vektes hver variabel slik at variansen til sumskåren y_1 blir så stor som mulig. Den neste prinsipale komponenten dannes ved at variablene vektes slik at mest mulig av den gjenstående variansen tas vare på. Slik fortsetter en å danne nye sumskårer inntil det finnes like mange sumskårer som variabler.

Vi kan også regne den motsatte veien, fra faktorer tilbake til variablene. Hver variabel x_i kan uttrykkes som en lineær kombinasjon av faktorene y_1, y_2, ..., y_k:

$$x_1 = A_{11}y_1 + A_{12}y_2 + ... + A_{1k}F_k \qquad (5.3)$$

osv.

A er vekter som er beregnet slik at vi kan rekonstruere de opprinnelige variablene ved hjelp av de prinsipale komponentene.

Egentlig viser dette hvor enkelt selve prinsippet for prinsipal komponentanalyse er. Hele poenget er å konstruere en og en sumskår. Hver sumskår er et produkt av alle variablene som inngår i analysen, men noen variabler gis høye vekter mens andre får lavere vekter. Faktoranalysen forteller oss størrelsen på vektene med fire-fem desimaler. Hver sumskår vektes slik at den ikke korrelerer med noen av de tidligere sumskårene og slik at den tar vare på maksimalt av den resterende variansen. Vanskeligere er det ikke.

5.2.4 Faktoranalyse

Prinsipal komponentanalyse og faktoranalyse har det til felles at en i begge tilfeller forsøker å redusere det opprinnelige variabelsettet til et mindre antall faktorer. Mens en i prinsipal komponentanalyse tar all varians på de opprinnelige variablene med i beregningen (også den varians som er unik for den enkelte variabel), tar en i faktoranalyse bare hensyn til kovariansen (samvariasjonen). Rent teknisk vises dette ved at en i prinsipal komponentanalyse bruker tallverdien 1,0 på diagonalen i korrelasjonsmatrisen, mens en i faktoranalyse erstatter dette 1,0-tallet med et tall som viser hvor stor andel av variansen i en bestemt variabel den har til felles med de øvrige variablene. Dette tallet er et estimat av det som kalles variabelens kommunalitet og symboliseres gjerne med bokstaven h. En variabels kommunalitet kan også defineres på en litt annen måte. Når en har gjennomført en faktoranalyse eller prinsipal komponentanalyse og bestemt seg for hvor mange faktorer eller komponenter en roterer (hva dette innebærer skal vi komme til nedenfor), kan en variabels kommunalitet defineres som den kvadrerte multiple korrelasjon mellom variabelen (som avhengig) og de faktorene eller komponentene en har bestemt seg for å rotere (som uavhengige variabler). Det er med andre ord den delen av variansen i en variabel som kan forklares av de faktorene en har bestemt seg for å rotere. En variabels endelige kommunalitet kan med andre ord beregnes nøyaktig først når resultatet av faktoranalysen foreligger. I praksis er dette en tallstørrelse som beregnes igjen og igjen underveis i en faktoranalyse.

Thurstone (1931, 1947) var den første som beskrev hvordan en ved å analysere en modifisert korrelasjonsmatrise, der diagonalelementene på 1,0 erstattes med kommunalitetsestimater, ignorerer den varians som er unik for hver variabel, og konsentrerer analysen rundt den varians som er felles for to eller flere variabler.

Formelen for utregning av faktorer blir også noe modifisert. Det samme blir formelen for reproduksjon av de opprinnelige variablene på bakgrunn av faktorene. I tillegg til bidraget fra faktorene, opererer en i faktoranalyse med såkalte unike bidrag. Dette medfører at faktorene ikke lenger er en direkte lineær kombinasjon av variabler. Ligningen for beregning av opprinnelige variabler (x) på grunnlag av faktorene (F) ser slik ut i faktoranalyse:

$$x_1 = b_{11}F_1 + b_{12}F_2 + ... + b_{1k}F_k + d_1u_1 \qquad (5.4)$$

Det nye med denne likningen i forhold til den tilsvarende for prinsipal komponentanalyse er tillegget som handler om unik varians (u). Denne unike variansen er den varians for hver variabel som den ikke har felles med noen annen variabel som inngår. Kvadratet av kommunaliteten på hver varabel (h^2) pluss den unike variansen (d^2) blir alltid lik 1.0.

5.2.5 Signifikanstesting av korrelasjonsmatrisen

Det første trinnet i en faktoranalyse eller en prinsipal komponentanalyser er, som vi tidligere har sagt, å regne ut korrelasjonene mellom alle mulige par av variabler som inngår i analysen. Alle disse korrelasjonene blir stilt opp i det vi kaller en korrelasjonsmatrise. Antall

korrelasjoner er lik antall variabler multiplisert med antall variabler minus en, og deretter må en dele produktet på to. Før en går videre i prosessen og begynner å rotere de faktorene en har ekstrahert, stiller en seg gjerne spørsmålet: Er det i det hele tatt noe å hente gjennom å anvende faktoranalyse? Er det såpass høye interkorrelasjoner mellom variablene at de er signifikant forskjellige fra null? Kan en si dette om hele datamatrisen sett under ett?

Bartlett har konstruert en test som kan vise om korrelasjonsmatrisen inneholder signifikant mye "samvarians". Den kalles "Bartletts test of sphericity" og den tar utgangspunkt i en identitetsmatrise: En matrise der alle verdier på diagonalen er 1,0 og alle andre verdier er 0,0. Nullhypotesen når en anvender Bartletts test tilsvarer en slik identitetsmatrise. Nullhypotesen er med andre ord at det ikke er noen korrelasjon mellom variablene. Jo høyere korrelasjonene er i den matrisen en vil teste, gitt at en holder n konstant, desto mer sannsynlig er det at en vil oppnå signifikans. Eller sagt på en litt annen måte: Jo mer samvarians (kovarians), desto lavere p-verdi, gitt en konstant utvalgsstørrelse.

En annen aktuell test tar utgangspunkt i partielle korrelasjoner. Dersom variablene i stor grad reflekterer underliggende fellesfaktorer, bør de partielle korrelasjonene mellom par av variabler være små dersom effekten av alle de andre variablene er eliminert. De partielle korrelasjonene bør være nær null dersom forutsetningene for faktoranalyse er oppfylt. Kaiser-Meyer-Olkins mål for adekvat sampling er en statistisk størrelse som sammenlikner de observerte korrelasjonskoeffisientene med de partielle. Dette målet kan beregnes for en hel korrelasjonsmatrise så vel som for hver enkelt variabel for seg.

Dersom verdien av Kaiser-Meyer-Olkin's mål for adekvat sampling blir høy, betyr dette at korrelasjonsmatrisen egner seg godt for faktoranalyse. Dersom verdien blir lav, er faktoranalyse en mindre godt egnet teknikk.

Kaiser (1974) angir følgende guide for å vurdere størrelsen på K-M-O:

.90 eller større:	marvellous
.80-.89:	meritorious
.70-.79:	middling
.60-.69:	mediocre
.50-.59:	miserable
under .50:	unacceptable

Et mål på hvor mye felles varians hver variabel har med alle andre variabler i analysen, er, som tidligere nevnt, kommunalitet. Kommunaliteten kan defineres som kvadratet av den multiple R mellom variabelen og alle de andre variablene i analysen. Iflg. Norusis (1985) bør en vurdere å utelate variabler med lav kommunalitet fra faktoranalyser.

5.2.6 Ekstraksjon av faktorer

Det eksisterer flere framgangsmåter når en skal ekstrahere faktorer fra et større antall variabler. Vi skal gjennomgå de vanligste.

Prinsipal komponentanalyse**:** Som allerede forklart tidligere: En danner lineære kombinasjoner av de observerte variablene. Den første prinsipale komponenten er den spesielle kombinasjonen som tar vare på mest av variansen i utvalget. Den andre forklarer mest av den gjenstående variansen, og er ukorrelert med den første. Den tredje forklarer mest av den variansen som gjenstår etter at de to første faktorene er dannet, og er ukorrelert med de to første etc. Prinsipal komponent-analyse tar vare på all varians, også den unike. Dersom en former tilstrekkelig mange faktorer, blir derfor alle variablers kommunalitet lik 1,0. Alt en slik analyse gjør, er å transformere et sett av korrelerte variabler til et sett av ukorrelerte variabler.

5.2.7 Antall faktorer

En viktig statistisk størrelse for å vurdere hvor mange faktorer en bør rotere i en prinsipal komponentanalyse eller en faktoranalyser er de som kalles eigenverdi (eigenvalue). Når vi regner ut en produkt-moment-korrelasjonskoeffisient, foretar vi automatisk en standardisering av alle variablene som inngår i analysen. At vi standardiserer en variabel, vil si at vi transformerer alle enkeltskårene slik at hver variabel får en gjennomsnittsverdi på 0,0 og et standardavvik på 1,0. At en faktor har en eigenverdi på 1,0, betyr at den inneholder like mye varians som en slik standardisert variabel. Et vanlig kriterium for hvor mange faktorer som bør inngå i en faktorløsning er at alle som tas med har en eigenverdi på 1,0 eller mer. Hair, Anderson, Tatham & Black (1992, s.237) påpeker imidlertid at kriteriet først og fremst passer med prinsipal komponentanalyse, og at en i forbindelse med common factoring bør sette grensen ved en eigenverdi som tilsvarer gjennomsnittet av kommunalitetsestimatene for alle variablene som inngår i analysen.

Et annet kriterium er skred-testen (scree-testen) (Cattell, 1965). Eigenverdien for hver faktor plottes på en akse der punktene på x-aksen står for faktor nummer og y-aksen måler størrelsen på eigenverdien. Kurven vil hele veien vise en fallende tendens. Vanligvis er det slik at den først faller nokså bratt, og senere flater ut. En tar med alle faktorene inntil en får en plutselig (og endelig) utflating i forklart varians. Faktorene i det bratte partiet til å begynne med regnes som substansielle, mens faktorene som fordeler seg utover på det flate partiet regnes som "støy".

Kim & Mueller (1978b) mener antall faktorer som bør roteres omfatter alle eigenverdiene i det partiet som viser et forholdsvis bratt fall samt den ene faktoren som markerer stedet der kurven flater ut. Gorsuch (1988, s.246) beskriver skred-testen på en litt annen måte. Han mener at etter den siste reelle faktoren vil kurven falle brattere ned, for deretter å bli mer horisontal. Fra det stedet kurven flater ut regner han faktorene som støy. Han tolker likevel Cattell dithen at en skal ta med den faktoren som danner overgangen mellom og "skredet" og flaten nedenfor. Cattell mener det er bedre å ta med en faktor ekstra for å være på den sikre side. Det har større negative konsekvenser å ta med en faktor for lite enn en for mye. Til dette kan en innvende at også en ekstra faktor kan komme til å påvirke faktorstrukturen på en uheldig måte.

Norusis (1985) mener imidlertid at en ikke skal ta med den faktoren som danner overgangen mellom bratt og flatere parti. Kanskje er det best å se nærmere på hvilken faktorløsning som

gir mest mening teoretisk og begrepsmessig før en velger den ene eller den andre løsningen. Ofte vil en oppdage at det antall faktorer som gir den begrepsmessig beste og klareste løsningen ikke faller sammen med eigenverdi-kriteriet. Oftere vil den være sammenfallende med Norusis fortolkning av skred-testen.

I prinsippet eksisterer det fem måter å avgjøre antall faktorer på:

- Signifikanstester (gjelder for maximum likelihood og least squares-metodene, som blir nærmere forklart nedenfor)
- Eigenverdien (større enn 1,0 når en kjører prinsipal komponentanalyse eller større enn gjennomsnittet av kommunalitetene når en bruker common factoring)
- Den substansielle betydningen av faktorene (prosent "forklart" varians)
- Skred-testen
- Fortolkbarheten og det begrepsmessige innholdet i faktorene.

Faktoranalysens første trinn gir normalt (ved anvendelse av eigenverdi-kriteriet) et antall faktorer som er langt mindre enn antall variabler. Dersom en ønsker, kan en be programmet om å operere med et bestemt antall faktorer. Ved å sammenlikne løsninger der en roterer ulike antall faktorer, kan en se hvilket antall som gir de klareste resultatene, med andre ord hvilken løsning som teoretisk eller begrepsmessig gir mest mening.

De koeffisientene som brukes til å beregne en standardisert variabel fra faktorene kalles faktorladninger. Når faktorene er ukorrelerte (ortogonale), svarer faktorladningene til enkle korrelasjoner mellom faktorer og variabler. Faktorladningene er det samme som de standardiserte regresjonskoeffisientene i en multippel regresjon med de opprinnelige variablene som avhengige (enkeltvis), og faktorene som uavhengige variabler.

Matrisen med korrelasjoner mellom faktorer og variabler kalles en faktorstrukturmatrise. Når faktorene er ukorrelerte, svarer faktorstrukturmatrisen til matrisen av faktorladninger. Faktorkoeffisientene (factor score coefficients) er de vektene som anvendes når en beregner faktorene på grunnlag av variablene. En variabels kommunalitet er, som vi gjorde rede for ovenfor, andelen varians som forklares av faktorene til sammen (alle faktorene eller det antall faktorer en velger å rotere). Den delen av variansen som ikke kan forklares ved hjelp av faktorene, kalles variablenes unikhet (uniqueness). Når faktorene er ortogonale, kan en estimere den opprinnelige korrelasjonsmatrisen ut fra faktorladningene. Hver korrelasjon kan beregnes ved at en multipliserer de to variablenes faktorladninger på en bestemt faktor og deretter adderer produktene over alle faktorene (Pedhazur & Schmelkin, 1991, s.605).

Differansen mellom de virkelige og de estimerte korrelasjonene kalles residualer. Jo færre faktorer en velger å rotere, desto høyere vil residualene bli. Høye residualer betyr at faktorløsningen i liten grad stemmer med data.

Prinsipal komponent-analyse er den enkleste av de faktoranalytiske ekstraksjonsmetodene, og den eneste metoden der faktorene kan beregnes nøyaktig ved bruk av faktorkoeffisienter. Det eksisterer imidlertid mange andre metoder for ekstraksjon av faktorer:

Prinsipal akse faktorisering (Principal axis factoring): er svært lik prinsipal komponent analyse, bortsett fra at diagonalen i korrelasjonsmatrisen inneholder estimerte kommunaliteter. Kommunalitetene estimeres gjentatte ganger underveis i faktoranalysen. Prinsipal akse faktorisering er en teknikk som tilhører kategorien common factor analysis.

Uveiet minste kvadraters metode (unweighted least squares): Minimaliserer summen av kvadrerte differanser mellom observert og reprodusert korrelasjonsmatrise.

Generalisert minste kvadraters metode (generalized least squares method): gjør det samme som unweigthed least squares, men vekter differansene inverst etter hvor unike variablene er.

Maximum-likelihood-metoden gir løsninger som reproduserer den observerte korrelasjonsmatrisen best mulig under forutsetning av at utvalget er trukket fra en multivariat normalfordeling. Også her vektes korrelasjonene med den inverse av variablenes unikhet.

Alpha-metoden antar at variablene i en analyse er et utvalg av variabler fra et univers av mulige variabler. Den maksimaliserer alpha-reliabilitets-koeffisienten for hver faktor (Cronbachs alpha)(for en forklaring på hva alpha er, se senere avsnitt i dette kapittelet).

Under to av metodene, generalized least squares og maximum likelihood, eksisterer det en signifikanstest som kan bruks for å avgjøre hvor mange faktorer som er nødvendig å operere med for å få en god nok tilpasning til data. Signifikanstestene er basert på en statistisk størrelse som er chi-kvadrat-fordelt.

5.2.8 Rotasjon av faktorer

I analysens andre trinn er poenget først og fremst å skape mest mulig orden og oversiktlighet i forholdet mellom variabler og faktorer. For å skape slik orden, er det som regel nødvendig å rotere faktorene. Programmet vekter simpelthen variablene på en ny måte.

Det eksisterer i prinsippet to måter å rotere på:

- Ortogonal rotasjon: Faktorene er ukorrelerte
- Oblik rotasjon: Faktorene tillates å korrelere

De vanligste rotasjonsmetodene er følgende:

a) Ortogonal rotasjon

Varimax: Minimaliserer antall variabler som lader høyt på en faktor
Quartimax: Minimaliserer antall faktorer som trenges for å forklare en variabel (produserer ofte en sterk generell faktor)
Equamax: Er en kombinasjon av Varimax og Quartimax

b) Oblik rotasjon

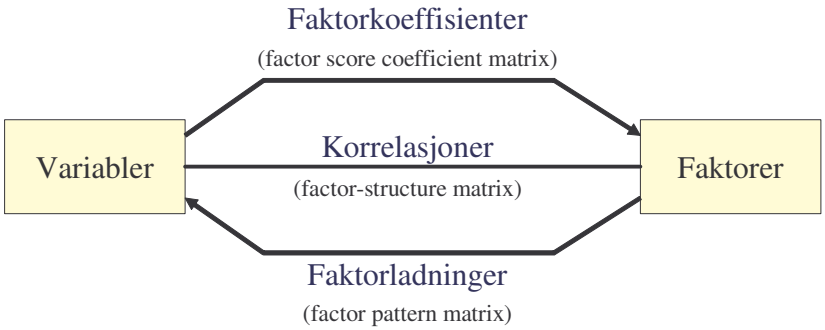
Ved oblik rotasjon, altså en rotasjon der faktorene tillates å korrelere med hverandre, er faktorladninger og faktor-variabel-korrelasjoner ikke lenger identiske. Ved oblik rotasjon eksisterer det tre ulike statistiske størrelser som alle beskriver sammenhengen mellom de opprinnelige variablene og faktorene (se Fig. 5.3):

- Faktorkoeffisienter (faktorkoeffisientmatrise): Vekter som kan brukes for å beregne faktorskårer på bakgrunn av de opprinnelige variablene.

- Faktor-variabel-korrelasjoner (faktorstrukturmatrise): Enkle korrelasjoner mellom de opprinnelige variablene og faktorene.

- Faktorladninger (pattern matrix): Partielle regresjonskoeffisienter som kan benyttes for å beregne de opprinnelige variablene på grunnlag av alle faktorene.

Fig. 5.3: Begreper til beskrivelse av relasjonen mellom variabler og faktorer



Valget mellom ulike rotasjonsmetoder er ikke enkelt for forskere som ikke er godt kjent med disse statistiske teknikkene. Brian Everitt (1995) gir i sin statistikkbok et hint. Han presiserer at oblik rotasjon matematisk sett krever mer kompliserte beregninger. Siden faktorladningene og faktor-variabel-korrelasjonene blir forskjellige ved oblik rotasjon, presiserer han at begge deler bør rapporteres. Ortogonal rotasjon er enklere og kan rapporteres ved bruk av bare ett sett av koeffisienter. Everitt sier også at når faktorene er høyt korrelerte, kan det by på problemer å tolke dem på en entydig og grei måte. Han konkluderer med at en bare bør bruke oblik rotasjon når denne gir en klarere løsning, med andre ord når den deler variablene inn i klarere, ikke-overlappende grupper.

Valget mellom de ulike ortogonale rotasjonsmetodene er heller ikke enkelt. Tradisjonelt er varimax-rotasjon blitt mest anvendt. Varimax-rotasjon ser dessuten ut til oftest å gi de klareste løsningene.

Dersom en variabel bare lader høyt på en faktor, sier vi at den er faktorielt "ren". Dersom en variabel lader høyt på flere faktorer, er den faktorielt komplisert (Kim & Mueller, 1978a).

Faktorer kan være unipolare eller bipolare. At de er unipolare betyr at alle faktorladninger har likt fortegn. At de er bipolare betyr at noen faktorladninger er positive mens andre er negative. Siden den eksploratoriske faktoranalysen produserer faktorer som ikke er gitt på forhånd, er det nødvendig å gi faktorene navn. For å bestemme navnet på en faktor er det nødvendig å studere faktorladningene nøye. Oftest vil en finne at det er variabler som har noe til felles som inngår i samme faktor. Identifikasjon av det som er begrepsmessig eller teoretisk felles bør danne utgangspunktet for å velge navn på faktorene.

5.2.9 Kan en rotere prinsipale komponenter?

Som allerede sagt flere ganger, er prinsipal komponentanalyse og faktoranalyse to forskjellige statistiske teknikker. Mens faktoranalyse baserer seg på en refleksiv situasjon, der enkeltvariablene er mål på et underliggende fenomén eller begrep, er prinsipal komponent-analyse en enkel, lineær transformasjon av de opprinnelige variablene.

Når en kjører faktoranalyse, er rotasjon av faktorene en naturlig del av analysen. Rotasjonen av faktorer gjør at en kan identifisere underliggende faktorer eller dimensjoner. Det er svært vanlig også å rotere prinsipale komponenter. Joliffe (1989) har kritisert den utstrakte bruken av rotasjon i forbindelse med prinsipal komponentanalyse. Brian Everitt (1996) hevder derimot at i mange situasjoner er rotasjon av faktorer i prinsipal komponentanalyse helt akseptabelt.

5.2.10 Hvilke krav må vi stille til data?

For at det skal være meningsfylt å faktoranalysere et sett variabler, må vi være sikre på at datasettet og utvalget av variabler egner seg for slik analyse. Faktoranalysen er vanligvis basert på produkt-momentkorrelasjoner, som forutsetter at variablene er på intervallnivå. I prinsippet kan en bruke andre koeffisienter enn produkt-moment-korrelasjonen i en faktoranalyse. Likevel ser vi ofte at en beregner produkt-moment-korrelasjoner på kategorielle ordinalvariabler der en strengt tatt ikke kan si at kravene til intervallnivå er oppfylt. Det som er særlig viktig å sjekke i slike situasjoner, er om sammenhengene kan beskrives lineært. Dersom det framkommer tydelige ikke-lineære sammenhenger, må en eventuelt transformere skalaene slik at sammenhengen likevel kan beskrives som lineær. Her kan en benytte statistiske teknikker som med utgangspunkt i bestemte modeller regner ut optimale verdier for de ulike kategoriene på kategorielle ordinalvariabler. I SPSS finner en disse teknikkene under det som kalles optmal skalering (*Optimal scaling* under *Data reduction*).

Dikotome variabler er i en viss forstand intervallvariabler. Likevel kan det by på problem å bruke dikotomier når disse er skjevfordelte. Særlig problematisk er det når to dikotomier som korrelerer positivt er skjevfordelte hver sin vei. Når variabler er skjeve i hver sin retning, dannes gjerne faktorer der de som er skjevfordelte i en retning danner en faktor, mens de som

er skjevfordelte i motsatt retning danner en annen faktor. Det finnes statistiske løsninger på dette problemet. Dersom den ene av variablene er dikotom, og en kan anta at den er uttrykk for underliggende normalfordelt variabel (som dessuten må være bivariat normalfordelt i forhold til den andre variabelen), kan en benytte polyserielle korrelasjoner. Dersom begge er dikotome og en kan anta at de er uttrykk for underliggende bivariat normalfordelte variabler, kan en benytte tetrakoriske korrelasjoner. Disse kan beregnes i statistikkprogrammer som LISREL og SAS.

Utvalget av variabler som inngår avgjør i stor grad hvor godt analysen fungerer. Dataene bør alle tilhøre en felles kategori, og de bør være målt på et likt spesifisitetsnivå. Hvis en for eksempel ønsker å faktoranalysere et sett variabler som måler holdninger til det å ta vare på helsa, og en stort sett opererer med bare en påstand per risikofaktor, ville det ødelegge analysen dersom en plutselig opererer med mange påstander rundt en bestemt risikofaktor. Disse spørsmålene ville med høy sannsynlighet danne en egen, triviell faktor.

Et annet viktig spørsmål handler om hvor mange enheter en bør ha per variabel som inngår i faktoranalysen. Hair, Anderson, Tatham & Black (1992) mener at et datasett bør bestå av minst 100 enheter for å kunne faktoranalyseres. Videre mener de at faktoranalyser av mindre enn 50 observasjoner ikke bør forekomme i det hele tatt. Antall enheter som kreves har sammenheng med hvor klar faktorstrukturen er. Med bare moderat sterke korrelasjoner og med en ikke helt krystallklar faktorstruktur, anbefales det minst 10 subjekter eller enheter per variabel. Det skader imidlertid ikke om forholdstallet mellom antall enheter og antall variabler er betydelig større.

Kerlinger uttrykker det slik i sin 1986-utgave:

> Two desiderata, even necessities, of factor analysis are large samples and replication. A general rule is: Use as large samples as possible. Like any statistical procedure, factor analysis is subject to measurement and sampling-error, and the reliable identification of factors and factor loadings requires large N's to wash out error variance. ..... A loose but not bad rule-of-thumb might be: ten subjects for each variable ...." (s.593)

De som har en del erfaring med faktoranalyse har kanskje oppdaget hvor vanskelig det er å få samme faktorstruktur når en gjennomfører analysen på det samme sett av variabler, men på to ulike grupper av individer. Dette illustrerer hvor ustabile resultatene av en faktoranalyse kan være. Det er derfor en god regel å forsøke om en faktoranalytisk løsning lar seg bekrefte på nye utvalg. Når en gjentatte ganger greier å reprodusere den samme faktorløsning på ulike materialer, kan en føle seg trygg på at en har funnet noe som er stabilt og dermed reelt og interessant. Når en tester ut hypoteser om faktorstrukturer på nye utvalg, kan det for øvrig være en god i idé å bruke konfirmatorisk faktoranalyse i stedet for eksplorerende.

5.3 Konstruksjon av sumskårer

Ovenfor har vi gjort rede for hvorfor det ofte er nyttig å konstruere sumskårer og vi har sagt at en langt på vei kan forstå faktoranalytiske teknikker ved å tenke på faktorene som sumskårer av variablene. Vi har imidlertid ikke gitt praktiske eksempler på hvordan vi konstruerer en sumskår. Vi skal derfor gi et slikt eksempel. La oss tenke oss at vi har gjennomført en undersøkelse blant skole-elever og stilt spørsmål om rusmidler og avhengighetsskapende stoffer. Vi har særlig fattet interesse for tre spørsmål. De dreier seg om ukentlig forbruk av tre substanser: Tobakksrøyking, bruk av alkohol og kaffe. Ti av elevenes svar er gjengitt nedenfor (Fig. 5.3).

Vi tenker oss at vi på bakgrunn av elevenes svar på disse tre spørsmålene konstruerer en sumskår som senere skal analyseres mot en del andre forhold, deriblant skoleprestasjoner og skoletrivsel. Det enkleste vi kan foreta oss er å telle opp hvor mange substanser hver enkelt elev har et ukentlig forbruk av. Dette gjør vi ved uttrykket

$$SUM1 = V1 + V2 + V3$$

Fig. 5.4: Enkel datamatrise

Variabel:	V1 Smaker alkohol ukentlig?	V2 Drikker kaffe ukentlig?	V3 Røyker ukentlig?	SUM1 Sumskår (I)	SUM2 Sumskår (II)
(0=nei, 1=ja)					
Elev nr:					
1:	1	0	1	2	3
2:	0	0	0	0	0
3:	0	1	1	2	2
4:	1	1	1	3	4
5:	0	1	0	1	1
6:	0	0	1	1	1
7:	1	0	0	1	2
8:	0	0	0	0	0
9:	1	1	1	3	4
10:	1	1	0	2	3

Resultatet ser vi i kolonnen under SUM1. For hver elev kan vi telle opp at tallene stemmer. Verdien varierer mellom 0 (null) som betyr at en elev med denne skåren ikke har et ukentlig forbruk av noen av disse substansene. En skåre på 3 betyr at eleven har et ukentlig forbruk av alle tre.

Dersom vi av en eller annen grunn gjerne ville vekte opp alkoholbruken slik at den teller sterkere i sumskåren, gjør vi det ganske enkelt ved følgende uttrykk:

*SUM2 = (2*V1) + V2 + V3*

Resultatet ser vi under kolonnen SUM2.
I prinsippet foregår all konstruksjon av sumskårer på samme måte. Det som kan gjøre konstruksjon av sumskårer noe mer komplisert er følgende:

- Ofte har vi ikke å gjøre med dikotome variabler (0/1-variabler), men skalaer på f.eks. fem eller sju punkter.
- Antall variabler som skal inngå er ofte langt større.
- I forbindelse med faktoranalyse og prinsipal komponentanalyse tildeles variablene vekter (faktorskårekoeffisienter) med 4-5 desimalers presisjon.
- På en del av enhetene kan vi risikere å mangle svar, og det finnes gjerne flere alternative måter å løse dette problemet på.
- Når det automatisk konstrueres sumskårer på grunnlag av ekte faktoranalyser (med common factoring), kompliseres bildet noe fordi den variansen som er unik for den enkelte variabel ikke teller med.

La oss nå gå tilbake til den prinsipale komponentanalysen og til faktoranalysen og se hva vi kan gjøre på bakgrunn av de løsningene vi har funnet. Dersom vi har med prinsipal komponentanalyse å gjøre, er det ved bruk av faktorkoeffisientmatrisen enkelt å konstruere sumskårer som matematisk samsvarer perfekt med en bestemt komponentløsning. Dette krever at en lar alle de opprinnelige variablene inngå i hver faktor, og at en vekter med komponent-koeffisienter som stemmer helt ut til siste desimal. Dette er imidlertid en nokså uvanlig måte å konstruere sumskårer på. Intuitivt gir det langt mer mening å dele variablene inn i separate grupper i samsvar med den roterte faktorløsningen og deretter lage en egen sumskår på bakgrunn av hver slik gruppe. Istedenfor nøyaktige fem-sifrede vekter, pleier en å foreta konstruksjonen av sumskåren med "naturlig" vekting. Dersom alle variablene som inngår i sumskåren har samme skala (f.eks. helt enig, ganske enig, verken enig eller uenig, ganske uenig, helt uenig), lar en kategoriene få tallverdier fra 0 og oppover til høyeste tallverdi og snur samtidig skalaene der det er nødvendig for å få samme logiske retning. Deretter lager en sumskåren for hver faktor ved å addere sammen hvert individs tallverdi på hver enkeltvariabel uten noen ytterligere transformasjoner. Alternativt kan en standardisere eller z-transformere hver variabel (trekke fra gjennomsnittsverdien og dele på standardavviket) først og dermed vekte alle variablene matematisk likt.

Konstruksjon av sumskårer kan bidra til å dekke over interessante sammenhenger i datasettet. For å være sikre på at vi ikke går glipp av vesentlig informasjon, kan det være nødvendig å ikke bare analysere sumskårene samlet i forhold til utenforliggende variabler (for eksempel et sett av prediktorer). Vi bør i tillegg se på variablene enkeltvis. Dersom de alle samvarierer på en konsistent måte med det utenforliggende settet av variabler, mister vi ingen vesentlig informasjon ved å benytte sumskårene. Dersom resultatet skulle vise inkonsistens, er det nødvendig å finne ut om det eksisterer spesielle mønstre eller grupper av variabler innen en faktor, eller om det bare er snakk om en mer usystematisk inkonsistens. Slik inkonsistensanalyse i forhold til tredjevariabler er viktig for å avdekke viktige sammenhenger

som kan finnes gjemt i data. Det er dette som tidligere i teksten er kalt for å undersøke ytre konsistens. Når vi inspiserer en korrelasjonsmatrise for å se at det er høye korrelasjoner mellom variablene som skal inngå i en sumskår, undersøker vi indre konsistens. Nedenfor skal vi se at det eksisterer egne statistiske størrelser for å kvanitifisere slik indre konsistens. Når en ser på konsistensen i en skala ved å undersøke hvorda leddene samvarierer med utenforliggende variabler, sier vi at vi undersøker skalaens ytre konsistens.

Noen ganger må vi konstruere sumskårer på grunnlag av variabler som ikke har like skalaer. Noen variabler kan f.eks. være ja/nei-variabler, andre kan være graderte med f.eks. 6-punkts-skalaer, noen kan ha andre antall kategorier og målenivå. I slike situasjoner kan det være fristende å standardisere variablene til z-skårer og deretter addere sammen. På denne måten kan vi hevde å ha gitt alle variablene lik vekt. Dersom en konstruerer slike sumskårer er det imidlertid to viktige problemer en bør være oppmerksom på:

(a) Sterkt skjevfordelte variabler vil medføre at de individene som er plassert i små grupper på ytterpunktene vil få svært store tallverdier på de standardiserte variablene. Dette gjelder også dikotomier der mange havner i den ene kategorien og få i den andre. De få i den minste kategorien vil få svært store tallverdier når disse variablene standardiseres.
(b) En slik prosedyre kan lett føre til at sumskåren får et nærmest uoversiktlig og vanskelig fortolkbart innhold.

Et alternativ kan være å dikotomisere alle variablene som skal inngå i sumskåren og deretter gi ett poeng for å tilhøre den ene kategorien på hver variabel og null poeng for å havne i den andre. På denne måten får en i det minste en sumskår som er lett å gjøre rede for og som har et forståelig innhold. Problemet med en slik framgangsmåte er først og fremst at en kan miste mye varians.

5.4 Reliabilitetsanalyse

5.4.1 Ulike former for reliabilitet

Ordet reliabilitet oversettes som regel med ”pålitelighet”, men kan også oversettes med andre ord, blant annet ”stabilitet”, ”konsistens” og ”nøyaktighet”. Det er, som tidligere beskrevet, flere måter å nærme seg reliabilitetsproblemet på. I det følgende skal vi se på flere av disse tilnærmingene.

For å kunne vurdere et instruments reliabilitet, er det viktig å kjenne til begrepet "varians". Varians har vi tidligere stiftet bekjentskap med når vi presenterte standardavviket. Variansen på en variabel er lik kvadratet til standardavviket på variabelen. Begrepet varians er selvsagt også svært sentralt for å forstå den familien av statistke teknikker som kalles variansanalyse.

Den totale variansen på en variabel kan deles inn i to deler: Sann varians og feilvarians. Dersom feilvariansen på en variabel er null og den sanne variansen dermed utgjør all varians, vil en ha perfekt reliabilitet (1,0). Dersom den sanne variansen er null og feilvariansen

dermed utgjør all varians på en variabel, blir reliabiliteten null. Reliabiliteten (r_v) på en variabel kan defineres som:

$$r_v = 1 - \frac{s_e^{\,2}}{s_t^{\,2}} \qquad (5.5)$$

r_v Reliabiliteten til en variabel v

$s_e^{\,2}$ Feilvarians

$s_t^{\,2}$ Total varians

Et måleinstruments reliabilitet kan med andre ord operasjonelt defineres som den andelen av observasjonenes totale varians som er sann. Jo høyere tallverdi, desto høyere reliabilitet. Reliabilitetskoeffisienten varierer fra 0,0 til 1,0.

Rent praktisk fins det flere måter å gå fram på når en skal undersøke reliabilitet. Havik (1982) refererer til Nunnally (1967) og til Guilford & Fruchter (1978) og beskriver tre kategorier av reliabilitetsmål:

Konsistenskoeffisienter: Estimat basert på konsistensen mellom ulike ledd i en test

Stabilitetskoeffisienter: Estimat basert på repeterte målinger

Ekvivalenskoeffisienter: Estimat basert på parallelle eller sammenlignbare målinger.

Disse kategoriene skiller seg fra hverandre med hensyn til hva som betraktes som sann varians og feilvarians.

5.4.2 Test-retest

Den kanskje aller enkleste måten å beregne reliabiliteten til en skala på er å administrere den to ganger til samme individer og regne ut korrelasjonen mellom skåre første og andre gang. Det er blitt vanlig å bruke en korrelasjon som kalles intraklassekorrelasjon i stedet for den vanlige produkt-moment-korrelasjonen. Dette fordi intraklassekorrelasjonen ikke bare tar hensyn til den relative plasseringen på skalaen, men er sensitiv også overfor endringer i den absolutte plasseringen. En stor fordel med denne metoden (test-retest) er at en kan beregne reliabiliteten på enkelt-spørsmål og ikke bare på skalaer som består av mange ledd. Et alvorlig problem med denne framgangsmåten er imidlertid at en kan risikere at de som deltar i undersøkelsen husker hva de svarte første gangen når de skal svare for andre gang. Nå kan en selvsagt vente så lenge med å administrere datainnsamling nr. 2 at de fleste ikke lenger er i stand til å huske hva de svarte sist. Ulempen med dette er at tidsavstanden mellom datainnsamlingene kan bli for stor. Forutsetningen for at en test-retest-korrelasjon skal være

et gyldig uttrykk for en skalas reliabilitet, er nemlig at selve det fenomènet en studerer ikke har endret seg vesentlig. Jo lenger tid som har gått, desto større er faren for at det virkelig har funnet sted slike endringer. Test-retest metoden er derfor mest aktuell når en skal måle fenomener som endrer seg langsomt over tid, for eksempel personlighetstrekk. Dersom en har utviklet en skala for å måle sinnsstemning (mood), er det mindre nyttig å gjennomføre en test-retest-studie for å undersøke reliabilitet. Dette fordi sinnstemning antas å variere en hel del over tid. Dermed vil avvik fra perfekt test-retest korrelasjon i stor grad kunne skyldes variasjoner i informantenes sinnsstemning og ikke manglende reliabilitet.

En løsning på problemet med at respondentene husker det de svarte første gang er det som kalles alternativ form-metoden. I stedet for å administrere nøyaktig samme skala på nytt, utarbeides en ny skala som skal måle nøyaktig det samme som den første. Imidlertid er ingen av spørsmålene som stilles nøyaktig like i de to testene. Ved å administrere først den ene skalaen og deretter den andre, og så regne ut korrelasjonen mellom dem, får vi et gyldig uttrykk for testenes reliabilitet. Denne metoden er mye brukt i pedagogisk forskning og anbefales av blant andre Carmines & Zeller (1979).

5.4.3 Indre konsistens

Det vanligste av alle mål for reliabilitet er Cronbachs alpha (Cronbach, 1951). Den er et eksempel på en konsistenskoeffisient og kan brukes når en har et sett med ledd som er ment å være indikatorer på ett og samme underliggende fenomen. Den kanskje aller enkleste formelen for Cronbachs alpha er gjengitt nedenfor (formel 5.6). På samme måte som de faktoranalytiske teknikkene, baseres også alpha på alle interkorrelasjonene (vanligvis Pearsons r) mellom et sett variabler.

Nunnally (1967) har vist at alpha kan fortolkes som den korrelasjon en kan forvente å få dersom en korrelerer testen med en annen (hypotetisk) test som måler det samme, og som har samme lengde og helt tilsvarende innhold. Novick & Lewis (1967) har vist at alpha kan betraktes som en nedre grense for reliabiliteten til en uvektet sumskår basert på de aktuelle variablene. Alpha er med andre ord et konservativt estimat av reliabilitet.

$$Alpha = \frac{N\bar{r}}{1+\bar{r}(N-1)} \qquad (5.6)$$

Alpha Reliabilitetskoeffisienten Cronbachs Alpha

$\bar{r}$ Gjennomsnittet av interkorrelasjonene mellom ledd som inngår i skalaen

N Antall ledd i skalaen

Hvor høy må så alpha være for at en skal kunne bruke en skala med god samvittighet? Nunnally (1967; 1978) har hevdet at dersom en befinner seg på et tidlig stadium i

forskningen på et bestemt område, må en kunne forsvare å bruke skalaer med ganske lave alpha-verdier. Dersom en bruker skalaer til å analysere forskjeller mellom grupper, kan en være mer liberal (tillate lavere reliabilitet) enn når en anvender tester der en skal uttale seg om skårene til enkeltindivider. I dette siste tilfellet må en stille svært høye krav til alpha. Nunnally antydet i sin lærebok fra 1967 at en noen ganger må akseptere å bruke skalaer med så lave alpha-verdier som 0,50-0,60. I 1978-utgaven av den samme læreboken mente Nunnally at skalaer bør ha en alpha på minst 0,70. Pedhazur & Schmelkin (1991) harselerer med disse utsagnene og sier spøkefullt at forskere som har brukt skalaer med en reliabilitet på rundt 0,70 dermed bør vise til Nunnally (1978), mens forskere som har brukt skalaer med en reliabilitet på rundt 0,50 bør vise til Nunnally (1967). De konkluderer mer seriøst med å si at det hjelper lite å vise til autoritative kilder. Det må være opp til hver enkelt forsker å vurdere hvor lav reliabilitet en kan tolerere ut fra den spesielle forskningen vedkommende driver med og ut fra de dataene som skal analyseres. Carmines & Zeller (1979) mener at skalaer som brukes av mange forskere, og dermed har stor utbredelse, bør ha en reliabilitet på 0,80 eller høyere (s.51).

Alpha-koeffisienten bygger på flere forutsetninger. For det første bør sammenhengen mellom de leddene som inngår i testen være lineær. For det andre må det kunne hevdes at hvert ledd som inngår i testen er mål på det samme underliggende fenomén. Testen må med andre ord være endimensjonal eller homogén. Dersom de ulike leddene i testen måler ulike, og kanskje lavt korrelerte egenskaper ved et fenomén, vil ikke Cronbachs alpha lenger være et uttrykk bare for høy eller lav reliabilitet. Et avvik fra 1,0 kan dermed deles inn i to komponenter: (i) en ikke-perfekt reliabilitet og (ii) det faktum at del-leddene måler egenskaper som ikke skal korrelere særlig høyt.

Alpha baserer seg på det som tidligere i dette kapittelet er beskrevet som en essensielt tau-ekvivalent målemodell. Dersom denne målemodellen stemmer med den målesituasjonen en faktisk står overfor, er alpha en forventningsrett estimator for skalaens reliabilitet (Lord & Novick, 1968). Dersom forutsetningen om at alle leddene i skalaen skal være parallelle (tau-ekvivalente) ikke holder, blir alpha et minimumsestimat for skalaens reliabilitet.

En viktig og intuitivt fornuftig egenskap ved Alpha er at den har en tendens til å øke når en øker antall ledd. Dette forutsatt at de leddene som føyes til har like høye korrelasjoner med leddene i testen som interkorrelasjonene mellom de leddene som allerede inngikk. Dette illustrerer at en får et mer pålitelig inntrykk av en egenskap dersom den måles ved hjelp av skalaer med mange ledd

Det er viktig å vite at en høy alpha ikke garanterer at skalaen er homogen (endimensjonal) (DeVellis, 1991). Alpha kan bli svært høy selv om et sett av variabler reflekterer to eller flere underliggende dimensjoner. Alpha kan med andre ord ikke erstatte faktoranalyse. Det er viktig å ha sjekket dimensjonaliteten i en skala før en anvender alpha.

Gitt en test med et visst antall ledd, kan en øke alpha på flere måter. En kan ekskludere ledd som ikke bidrar positivt til alpha. En kan også vekte de ulike variablene på spesielle måter. For å bestemme hvilken vekting som gir det aller beste resultatet, kan en foreta en faktoranalyse etter prinsipal-komponent-metoden. Den første faktoren som ekstraheres, gir oss de vektene som skal til for å gi testen en maksimalt høy reliabilitet. Carmines & Zeller

(1979, s.61) presenterer en formel som kan brukes til å beregne alpha med optimal vekting av leddene som inngår. Denne koeffisienten kalles theta.

$$Theta = \frac{N}{N-1} * \left(1 - \frac{1}{e_1}\right) \qquad (5.7)$$

N Antall ledd som inngår i skalaen eller testen

e_1 Eigenverdien (eigenvalue) på den første (og dermed største) ekstraherte faktoren

Et annet reliabilitetsmål ble introdusert av Heise & Bohrnstedt (1970). Deres koeffisient kalles omega, og baserer seg på common factoring-analyse. Det kan vises matematisk at omega alltid er minst like stor som theta, som på sin side alltid er minst like stor som alpha.

Et annet mål for reliabilitet ble introdusert av Kuder & Richardson så tidlig som i 1937. Deres koeffisient var beregnet på dikotome variabler, for eksempel en kunnskapstest der hvert svar kan klassifiseres som enten riktig eller galt. For hvert riktig svar gis det ett poeng, for hvert galt svar gis det null poeng. Det er dette vi tidligere har kalt dummy-koding av variabler. Kuder-Richardson 20 (eller KR20), som koeffisienen deres heter (tallet 20 betegner simpelthen nummeret på formelen i artikkelen de publiserte i 1937) har vist seg å gi nøyaktig samme tall som når en bruker Cronbachs alpha på dummy-kodete variabler. KR20 er med andre ord et spesialtifelle av Cronbahcs alpha. Eller en kan si det på en litt annen måte: Cronbachs alpha er en generalisert KR20.

Cronbachs alpha er ikke den eneste koeffisienten som på grunnlag av en serie variabler som inngår i en skala administrert på ett tidspunkt kan beregne skalaens reliabilitet. En annen tilnærming kalles splitt i halver-metoden (split-halves method). Denne metoden går ut på å dele de leddene som inngår i en skala i to vilkårlige grupper. Etter å ha administrert datainnsamlingen, beregner en så hvert individs skåre på hver av disse to halve testene. Deretter regner en ut korrelasjonen mellom de to skårene. Denne korrelasjonen er imidlertid ikke et direkte uttrykk for hele testens reliabilitet, men heller et uttrykk for hvor reliabel halve testen er. For å få et tall som sier noe om hele testens reliabilitet, må en bruke en formel som kalles Spearman-Browns profeti-formel (Spearman-Brown prophecy formula). Formelen ble utviklet av to statistikere uavhengig av hverandre og publisert i samme nummer av British Journal of Psychology i 1910 (Spearman, 1910; Brown, 1910). Formelen ser slik ut:

$$SB_{xx} = \frac{2r_{xx}}{1 + r_{xx}} \qquad (5.8)$$

SB_{xx} Spearman-Browns reliabilitetskoeffisient

R_{xx} Korrelasjonen mellom de to halve testene

Spearman Brown - koeffisienten varierer mellom 0,0 og 1,0 akkurat slik som Cronbachs alpha.

Denne måten å beregne en skalas reliabilitet på har imidlertid en innebygd svakhet. En kan selvsagt dele inn de leddene som inngår i en skala i to like store grupper på flere forskjellige måter. En skala som består av 10 ledd kan for eksempel deles inn i to grupper på fem ledd i hver på 125 forskjellige måter. Hvor stor korrelasjonen mellom de to halvene blir, vil hele tiden variere. Dette er en usikkerhet som vi unngår ved å bruke Cronbachs alpha. De fleste forskere foretrekker derfor Cronbachs alpha framfor Spearman-Browns koeffisient.

De to statistiske størrelsene er for øvrig i nær slekt med hverandre. Alpha-verdien for en test som har et bestemt antall ledd er lik gjennomsnittet av alle mulige "split half"-reliabilitetskoeffisienter for den samme testen (Cronbach, 1951). En kan derfor betrakte en reliabilitetskoeffisient basert på splitt i halver-metoden som en måte å estimere alpha på (Nunnally, 1978, s.233). Da er selvsagt alpha å foretrekke framfor et estimat som vil være beheftet med en betydelig usikkerhet.

Carmines & Zeller (1979) avslutter sin bok om validitet og reliabilitet med en kort drøfting av hva slags måter de mener det er best å undersøke et instruments reliabilitet på. De stiller seg nokså negative til test-retest korrelasjoner. Dette fordi de mener det er stor fare for at de som deltar lærer noe første gang de svarer på spørsmålene i en skala og at dette påvirker deres skåre ved retest. Det kan ganske enkelt tenkes at de husker noe av det de svarte første gang. Dette vil i så fall bidra til et for høyt estimat av reliabilitet. Carmines & Zeller er også nokså negative til metoder der en deler inn spørsmålene i to grupper (split half) og deretter beregner reliabiliteten på grunnlag av korrelasjonen mellom dem. De viser til at en kan dele inn i to grupper på så mange måter, og at resultatene i praksis vil variere. De viser dessuten til at Spearman-Brown-formelen kan betraktes som en måte å estimere alpha på. Da er det imidlertid bedre å estimere alpha på grunnlag av alle leddene i testen i stedet for å dele den vilkårlig i to grupper av items.

Når en har målt noe med bare ett enkelt spørsmål eller ledd, og ikke har gjennomført noen test-retest-studie, kan det by på problemer å beregne reliabiliteten til dette ene spørsmålet. Måling ved hjelp av enkelt-spørsmål gir dessuten lavere reliabilitet enn når en har flere ledd. Det å måle noe ved hjelp av enkeltspørsmål har derfor ofte blitt kritisert. En risikerer at artikler som presenterer analyser basert på slike enkeltspørsmål blir avvist når de vurderes med tanke på publisering. Wanous og medarbeidere (1997) har imidlertid presentert en elegant metode for å beregne reliabiliteten til enkeltspørsmål ut fra rene tverrsnittsstudier. Dette krever at en både har målt det en ønsker å måle ved hjelp av ett enkelt spørsmål og ved hjelp av en skala (med flere ledd). Utgangspunktet er at korrelasjonen mellom de to målene (enkeltspørsmålet og skalaen) bør være 1,00 når en har korrigert for attenuasjon (korrigert for at reliabiliteten ikke er perfekt). Siden en kan beregne reliabiliteten til skalaen som består av flere ledd ved bruk av Cronbachs alpha, kan en sette opp en likning som bare har en ukjent, nemlig reliabiliteten til enkeltspørsmålet. Wanous og medarbeidere har benyttet denne framgangsmåten for å beregne det de kaller minimums-reliabiliteten til enkelspørsmål som er ment å måle global jobbtilfredshet. De gjorde dette på grunnlag av 17 tidligere studier (totalt n = 7682). De fant at reliabilitetsestimatene varierte fra 0,45 til 0,69. Selv om dette er nokså beskjedne tall, mener de at det i mange tilfeller er forsvarlig å analysere variabler med en så

lav reliabilitet. Og de mener at en ikke uten videre bør avvise artikler som er basert på analyser av globale enkeltspørsmål.

5.4.4 Skalaer og datareduksjon – noen refleksjoner

Etter denne gjennomgangen av statistiske hjelpemidler til å bestemme hvordan en kan foreta datareduksjon ved hjelp av sumskårer, bør det advares mot å basere seg for ensidig på slike teknikker. Selv om en har foretatt faktoranalyse, gruppert variablene på basis av denne, gjennomført reliabilitetsanalyser og deretter ekskludert ledd som ikke bidrar positivt, og selv om antall observasjoner en baserer utregningene på er svært høyt, kan en langt fra være sikker på å ha konstruert sumskårer med et meningsfylt innhold og fornuftige egenskaper. I siste instans må sluttresultatet av en slik prosess gi mening begrepsmessig og teoretisk. Noen ganger vil det være helt malplassert å benytte slike framgangsmåter når en skal konstruere en indeks. I hvilke sammenhenger det er fornuftig å bruke indre konsistens som et kriterium på høy reliabilitet og gode måletekniske egenskaper, avgjør først og fremst de teoriene, modellene og begrepene som blir anvendt.

Når en skal teste ut en skala, ser en gjerne på hva som skjer med alpha når en tar ut et ledd fra skalaen. Dersom leddet positivt til skalaen, forventer en at alpha blir lavere når leddet tas ut. Dersom et ledd bidrar til at alpha går tydelig ned (for eksempel fra 0,80 til 0,75), anbefales det gjerne at leddet fjernes fra skalaen og at en konstruerer en sumskår der leddet ikke inngår. Det er etter hvert blitt ganske vanlig at en under planlegging av en undersøkelse henter ut skalaer fra forskningslitteraturen og bruker disse. Det finnes en rekke bøker som presenterer skalaer som er mye brukt og som har vist seg å ha gode egenskaper, det vil si at de har et kjent antall dimensjoner, at alpha innen hver dimensjon er høy, og at alle enkeltledd som inngår i en dimensjon bidrar positivt til alpha for denne dimensjonen. Noen ganger viser det seg at slike skalaer slett ikke oppfører seg slik som forventet. En klarer kanskje ikke å reprodusere de samme dimensjonene når en faktoranalyserer skalaen, en får kanskje ikke så høy alpha som de tidligere undersøkelsene viser, eller en oppdager at enkelte ledd trekker alpha-verdien ned i stedet for opp. I slike situasjoner kan det være vanskelig å vite hva en skal gjøre. Skal en konstruere sumskårer på sin egen måte, og dermed ødelegge sammenliknbarheten med tidligere undersøkelser, eller skal en følge oppskriftene fra tidligere studier og risikere at en samler inn data som ikke i tilstrekkelig grad er tilpassen den befolkningen en undersøker og det som er formålet med undersøkelsen?

I slike situasjoner er det viktig å huske på at det kan være minst fire årsaker til at egne data ikke stemmer så godt med det en har sett i tidligere undersøkelser. For det første kan forskjellene skyldes at en har gjort en for dårlig jobb med å oversette og pilotteste instrumentene, slik at spørsmålene ikke betyr det samme som i den originale skalaen, eller at de ikke er godt nok formulert. For det andre kan de tidligere studiene være av tvilsom kvalitet. For det tredje kan det tenkes at de fenomenene en studerer har endret seg over tid, eller at det er forskjeller mellom kulturer, kontekster eller befolkningsgrupper som er utslagsgivende. Endelig kan det tenkes at det er tilfeldigheter som fører til at resultatene er forskjellige. Særlig stort er dette problemet dersom en har et lite utvalg. Når utvalgene er små, kan rene tilfeldigheter føre til store forskjeller i resultater. Når n er lavt er det derfor som regel mest fornuftig å konstruere sumskårer i samsvar med det som er anbefalt i

forskningslitteraturen. Dersom utvalget en analyserer på er stort og analysene på en overbevisende måte viser at sumskårene bør konstrueres på en annen måte enn det som er gjort i tidligere forskning, står en overfor et vanskelig valg. Det er umulig å gi generelle råd om hva som er riktig å gjøre i en slik situasjon. En løsning kan ganske enkelt være å skrive en artikkel der en gjør rede for egne resultater og foreslår nye måter å gruppere variablene og bruke skalaen på.

5.4.5 Korrigering for attenuasjon

Når en korrelerer to variabler, og en kjenner variablenes reliabilitet, er det mulig å regne ut hvor sterk sammenhengen ville ha vært dersom begge variablene hadde vært målt med perfekt presisjon (reliabilitet). Formelen for korrigering for attenuasjon for en produkt-moment korrelasjon er vist nedenfor.

$$r_{xt\ yt} = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}} \qquad (5.9)$$

$r_{xt\ yt}$ Korrelasjonen etter korrigering for attenuasjon

r_{xy} Korrelasjonen mellom de to variablene

r_{xx} Reliabiliteten til variabelen *x*

r_{yy} Reliabiliteten til variabelen *y*

Dersom korrelasjonen mellom to variabler er regnet ut til å være 0,20 og begge variablene har en reliabilitet på 0,50, kan vi sette tallene inn i formelen ovenfor slik:

$$r_{xt\ yt} = \frac{0,20}{\sqrt{0,5 * 0,5}} = 0,40$$

Vi ser altså at dersom vi hadde vært i stand til å måle begge variablene med perfekt presisjon, med andre ord uten det minste feil, ville korrelasjonen blitt dobbelt så stor. Resultatet av å analysere variabler som er målt med lav reliabilitet er med andre ord at de sammenhengene vi får er lavere enn de egentlig skulle være. Ikke alle forskere er oppmerksomme på dette, og vil derfor lett undervurdere betydningen av egne funn når de arbeider med skalaer som ikke har høy reliabilitet. Når en analyserer data ved bruk av en type teknikker som kalles modellering av strukturelle likninger (structural equation modeling) som finnes i programpakker og programmer som AMOS, LISREL og MPLUS, kan en få gjort analysene på en slik måte at en korrigerer for attenuasjon. En antar at slike analyser gir et mer realistisk bilde av styrken på sammenhengene enn det en ellers ville få fram. Ut fra de observerte skårene på de leddene som inngår i en skala regner en ut sammenhengene mellom de underliggende latente variablene. Slike sammenhenger er korrigerte for at målingene ikke er gjort med perfekt

reliabilitet. En må imidlertid være påpasselig med bare å benytte slik korrigering for attenuasjon når en har refleksive målemodeller.

5.5 Eksploratorisk faktoranalyse – et eksempel

5.5.1 Helseatferd blant skoleelever

La oss så se nærmere på hvordan faktoranalyse kan brukes i praksis. Eksempelet er hentet fra HEMIL-senterets prosjekt om Helsevaner blant skolebarn i Europa (HBSC - Health Behaviour in School-Aged Children). Prosjektet omfatter nå mer enn 30 land, men vi har her bare benyttet data fra den norske datainnsamlingen som ble gjennomført i 1993-94. Dataene er representative for skoleelever på de aktuelle klassetrinnene i hele landet. Vi bruker her kun data fra elever som gikk i det som nå tilsvarer 10. klasse. Det betyr at elevene var 15-16 år gamle på det tidspunktet dataene ble samlet inn.

Eksempelet er valgt fordi det illustrerer flere av de problemene en ofte står overfor, blant annet vedrørende målenivå, skjevfordelte variabler og usikkerhet om hvor mange faktorer eller dimensjoner som egentlig er mest riktig å operere med.

I undersøkelsen ble det stilt spørsmål om en del vanlige plager, nærmere bestemt:

- Hodepine
- Vondt i magen
- Vondt i ryggen
- Svimmelhet
- Nedstemthet
- Irritabilitet
- Engstelse/nervøsitet
- Søvnproblemer

Svarkategoriene som ble benyttet var følgende:

1 - Daglig
2 - Mer enn en gang i uken
3 - Ukentlig
4 - Månedlig
5 - Sjelden eller aldri

De fleste variablene var skjevfordelte. Enveis-fordelingene på hver enkelt variabel er gjengitt i Tabell 5.1. Vi ser at nesten alle variablene er ganske venstreskjeve. Det vises også av tallene for skjevhet (skewness) som viser verdier varierende fra -1,73 til -0,50. Det må vurderes nærmere om dette får konsekvenser for valg av statistisk teknikk og om det er nødvendig med en eller annen form for variabeltransformasjon slik at variablene blir mindre skjevfordelte.

Slik verdiene er kodet på analysefilen, er "Daglig" gitt tallverdien 1 og "Sjelden eller aldri" er gitt tallverdien 5. Dette medfører at høye tallverdier står for lav forekomst av plager. Før vi

analyserer videre på disse tallverdiene, snur vi derfor skalaen og lar den starte på 0 (null). Etter denne omkodingen har sjelden eller aldri tallverdien 0 og Daglig har tallverdien 4. Dersom vi lager sumskårer baserte på de omkodete variablene, vil høy tallverdi stå for høy forekomst av plager, noe som virker logisk. Videre vil en sumskår få et teoretisk laveste punkt på null. Dette er ryddigere enn sumskårer som starter på tall som avhenger av antall ledd som inngår i skalaen og vektingen av disse leddene. Når alle variablene snus på denne måten, forandres ikke noen av interkorrelasjonene. Men i stedet for å være venstreskjeve, er alle variablene nå høyreskjeve.

Det å nummerere kategoriene fra 0 til 4 slik vi har gjort her, er ikke nødvendigvis den riktigste måten å skalere på. Vi kunne for eksempel ha omkodet alle kategoriene slik at de sier noe om antall plager per uke:

Sjelden eller aldri = 0 (null)
Månedlig = 0,25
Ukentlig = 1
Mer enn en gang i uken = 3 (tallet litt tilfeldig valgt)
Daglig = 7

Dette gir en skala som med større rett kan betraktes som en intervallskala. Imidlertid vil en slik transformasjon føre til at disse variablene, som allerede er ganske skjevfordelte, blir enda skjevere. Det blir større numerisk avstand mellom de kategoriene som har de laveste frekvensene. Vi har derfor valgt å beholde verdiene fra 0 til 4[1].

For å finne ut hvordan vi kan forenkle disse åtte variablene til sumskårer, kjører vi så en faktoranalyse eller en prinsipal komponentanalyse.

Som gjort rede for tidligere, bruker vi faktoranalyse når vi kan anta at variablene reflekterer noe underliggende, noe de har felles. Her er det mulig å tenke seg at variablene alle gjenspeiler individenes subjektive helsetilstand. Når hverdagsbelastningene blir for store, kan vi tenke oss at det må komme til uttrykk på en eller annen måte, men at det kanskje er noe tilfeldig hvordan de kommer til uttrykk. Hos noen kommer tilstanden til uttrykk gjennom en eller flere av de psykiske plagene, hos andre gjennom noen av de kroppslige plagene. Og hos atter andrde kan vi tenke oss at belastningene en har vært utsatt for kommer til uttrykk både psykisk og kroppslig. Dersom dette er vår forståelse av fenoménet, har vi med en refleksiv situasjon å gjøre.

[1] En separat analyse der vi har benyttet optimal skalering (optimal scaling) viser at en ikke får vesentlige endringer i resultatene av den etterfølgende faktoranalysen selv om en benytter de kategoriverdiene som beregnes der.

Tabell 5.1: Enveis prosentfordelinger av en del vanlige plager

	Daglig	Flere gg per uke	Ukentlig	Månedlig	Sjeldnere eller aldri	Til sammen		
	%	%	%	%	%	%	n	Skjevhet
Hodepine	4,0	7,3	12,2	28,2	48,3	100,0	1633	-1,19
Magesmerterer	1,8	4,1	8,0	39,9	46,2	100,0	1628	-1,45
Ryggsmerter	5,0	4,9	8,2	20,1	61,7	100,0	1629	-1,63
Svimmelhet	3,4	6,4	6,9	16,1	67,1	100,0	1633	-1,73
Nedstemt	2,9	5,6	13,6	31,3	46,6	100,0	1624	-1,20
Irritabel	4,7	14,4	24,0	39,4	17,6	100,0	1628	-0,50
Nervøs	2,5	5,0	11,7	24,5	56,3	100,0	1625	-1,42
Søvnprobl.	6,1	9,1	9,6	14,8	60,3	100,0	1631	-1,27

Datasettet (10. klasse) omfatter totalt 1637 subjekter.
Antall manglende svar varierer mellom 4 og 32 (0,2 og 2,0 prosent).

Det behøver imidlertid ikke være så enkelt. Det kan like gjerne tenkes at de ulike plagene har en noe ulik bakgrunn, og at det blir for enkelt å anta en felles underliggende prosess. Vi kan likevel være interessert i å redusere datamengden en del, men samtidig få med oss så mye som mulig av variasjonen i plager, inklusive varians som er unik for den enkelte variabel. I så fall er det naturlig å bruke prinsipal komponentanalyse.

Uansett hva som måtte være mest forsvarlig, skal vi nå forsøke både vanlig faktoranalyse og prinsipal komponentanalyse på disse variablene.

5.5.2 Faktoranalyse

Første trinn i faktoranalysen er å regne ut interkorrelasjonene mellom samtlige variabler som skal inngå i analysen. Korrelasjonsmatrisen er gjengitt i tabell 5.2. Bartletts test viser klart at korrelasjonsmatrisen er signifikant forskjellig fra en enhetsmatrise ($p < 001$). Kaiser-Meyer-Olkins mål for adekvat sampling er på 0.85, noe som kvalifiserer for betegnelsen "meritorious". Dette tyder på at de aktuelle variablene skulle egne seg bra for en faktoranalyse.

Alle korrelasjoner er positive og statistisk signifikante ($p < .001$) og ligger i området mellom 0,17 og 0,50. Det avtegner seg ved første øyekast ikke noen grupper av variabler som er høyere interkorrelerte enn andre grupper av variabler. Kanskje ser det ut til at en enfaktor-løsning vil være riktigst. Siden mange av variablene var temmelig skjevfordelte, er samtlige blitt logaritmetransformert og korrelasjonsmatrisen beregnet på nytt. Transformasjonen

medførte at samtlige korrelasjoner ble noe mindre. Maksimal nedgang var imidlertid bare på 0,03. En valgte derfor å beholde de opprinnelige variablene med tanke på den videre analysen av data. Det er viktig å legge merke til at alle variablene var snudd slik at de ble høyreskeive før logaritmetransformasjonen. Effekten av å transformere logaritmisk er nemlig å gjøre hvert trinn på skalaen mindre etterhvert som en beveger seg mot høyere verdier på skalaen. Dersom skalaen ikke var blitt snudd først, ville en logaritmetransformasjon bare ha virket mot sin hensikt. Før en logaritmetransformerer må en også passe på at skalaen som benyttes starter på 1,0. Logaritmen til 1,0 er nemlig 0,0 En får dermed en skala som går fra null og oppover, og med stadig synkende avstand mellom punktene (kategoriene).

Tabell 5.2: Korrelasjoner (Pearsons produkt-momentkorrelasjoner) mellom åtte plagevariabler.

	Hodepine	Mage-smerter	Rygg-smerter	Svimmel-het	Nedfor	Irritabel	Nervøs	Søvn-problem
Hodepine	1.00							
Magesmerter	.48	1.00						
Ryggsmerter	.32	.31	1.00					
Svimmelhet	.43	.30	.29	1.00				
Nedfor	.34	.37	.25	.31	1.00			
Irritabel	.31	.35	.22	.26	.50	1.00		
Nervøs	.26	.29	.17	.27	.41	.35	1.00	
Søvnprobleme r	.24	.26	.26	.30	.34	.29	.28	1.00

Samtlige korrelasjoner er signifikante på 1%-nivået (to-halet test).
Kaiser-Meyer-Olkins statistikk: 0,85
Bartletts test: p<0,001

Resultatene av en faktoranalyse med "principal axis factoring" (som er en ekte faktoranalytisk ekstraksjonsmetode og ikke må forveksles med prinsipal komponentanalyse) og varimax rotasjon er gjengitt i Fig. 5.5 og 5.6.

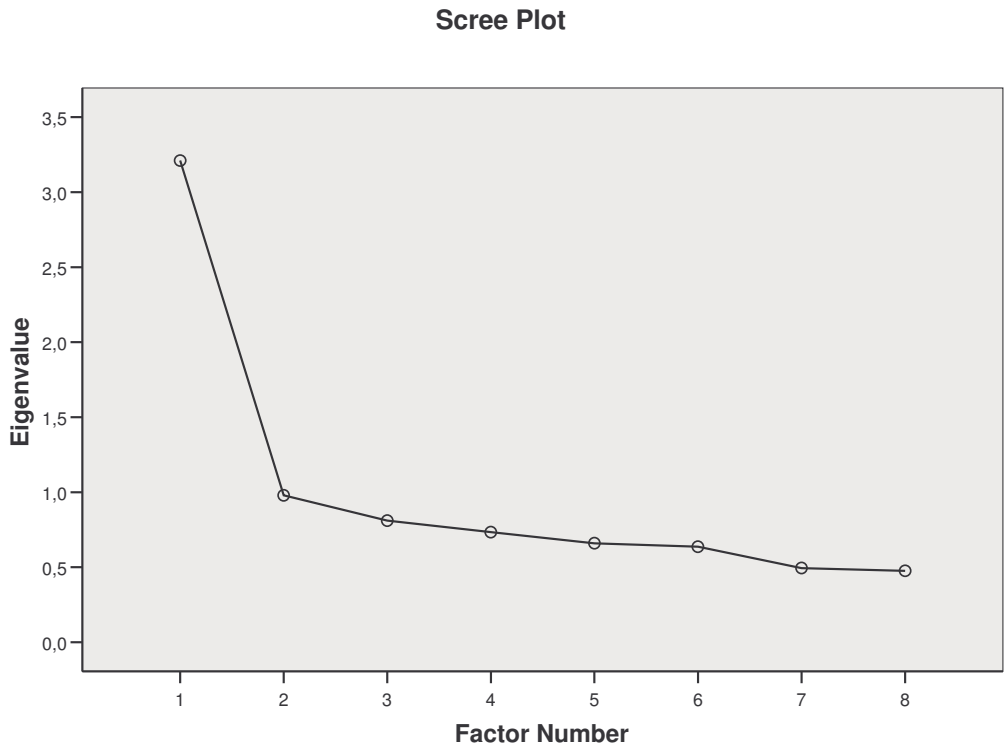
Den første uroterte faktoren forklarer hele 40,1 prosent av variansen i de åtte variablene, noe som her tilsvarer en eigenverdi på 3,21. Den andre faktoren forklarer langt mindre varians, bare 12,2 prosent. Her er eigenverdien mindre enn 1,0, nærmere bestemt 0,979. Etter kriteriet om at eigenverdien bør være større enn 1,0, bør vi her basere oss på en løsning med bare en faktor. Som tidligere nevnt hevder noen at en bør sette en lavere grense, og ta med alle uroterte faktorer som har en eigenverdi som er høyere enn gjennomsnittet av kommunalitetsestimatene. I vårt tilfelle ville dette føre til så mange faktorer at resultatet blir ganske uinteressant. Vi har imidlertid en antakelse om at det her kan vise seg å eksistere to meningsfylte, underliggende faktorer. Vi velger derfor å rotere to faktorer.

For å kjøre en faktoranalyse i SPSS kan en for eksempel gå inn i *Analyze*, deretter velger en *Data reduction*, og så *Factor*. Deretter legger en variablene som skal inngå i analysen inn i vinduet som har overskriften *Variables*. Under *Descriptives* velger en *Initial solution*, *Coefficients*, *Significance levels* og *KMO and Bartlett's test of sphericity*. Videre går en inn in *Extraction* og velger *Principal axis factoring*, *correlation matrix*, *Eigenvalues over 1*, *Unrotated factor solution* og *Scree plot*. Deretter går en inn i *Rotation* og velger *Varimax* og *Rotated solution*. Og til slutt går en inn i *Options* og velger *Exclude cases pairwise*.

Alternativt kan en under *Rotation* velge *Direct oblimin.* Dette vil gi korrelerte eller oblike faktorer til forskjell fra *Varimax* som gir ukorrelerte eller ortogonale faktorer.

Og alternativt kan en under *Extraction* velge *Principal components*. I så fall blir det ikke utført noen faktoranalyse, men i stedet blir det utført en prinsipal komponentanalyse.

Fig. 5.5: Skred-diagram



I Fig. 5.5 er alle eigenverdiene vist i form av et linjediagram. Vi ser at den første faktoren framstår som svært kraftig, mens de etterfølgende ligger langs en nokså langsomt synkende linje. Et slikt mønster kan tyde på at det her bare eksisterer en enkelt faktor, og at de øvrige faktorene bare representerer feilvarians eller støy. Siden vi har en antakelse om at det kanskje eksisterer to underliggende faktorer, holder vi imidlertid fast på beslutningen om å rotere to faktorer.

Fig. 5.6a viser de uroterte faktorladningene. Vi ser at samtlige variabler har høyere ladninger på den første faktoren enn på den andre, og vi ser også at alle har positivt fortegn. Vi kan dermed tolke den første faktoren som en generell plagefaktor.

Fig. 5.6a: Faktoranalyse av plagevariabler. Uroterte faktorladninger.

Factor Matrix[a]

	Factor	
	1	2
Hodepine	,65	,36
Magesmerter	,61	,15
Ryggsmerter	,45	,16
Svimmelhet	,54	,16
Nedfor	,68	**-,30**
Irritabel	,60	**-,25**
Nervøs	,52	**-,19**
Søvnproblemer	,48	**-,08**

Extraction Method: Principal Axis Factoring.

a. 2 factors extracted. 15 iterations required.

Fig. 5.6b: Faktoranalyse av plagevariabler. Roterte faktorladninger etter ortogonal rotasjon.

Rotated Factor Matrix[a]

	Factor	
	1	2
Hodepine	,21	**,72**
Magesmerter	,33	**,54**
Ryggsmerter	,21	**,43**
Svimmelhet	,27	**,50**
Nedfor	**,70**	,26
Irritabel	**,60**	,24
Nervøs	**,50**	,23
Søvnproblemer	**,40**	,28

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Når vi ser på den andre faktoren framkommer det et interessant mønster. Fire av plagene lader med positivt fortegn og de fire andre med negativt. Når vi ser på innholdet i de som lader med positivt fortegn, blir det klart at alle disse handler om kroppslige plager. De øvrige fire handler om psykiske plager. Den andre uroterte faktoren er med andre ord en faktor som viser diskrepans mellom kroppslige og psykiske plager.

Dersom vi nå flytter oppmerksomheten over på den roterte løsningen (Fig. 5.6b), viser det seg at den samme inndelingen i to variabelgrupper kommer til syne på nytt. Alle faktorladninger som er 0.40 eller større er markert med uthevet skrift. Også her har variablene fordelt seg i en gruppe som kan kalles kroppslige plager og en gruppe som kan kalles psykiske plager. Faktorladningene er ikke svært høye, men mønsteret er konsistent. Som vi har sagt tidligere, vil en i eksploratorisk faktoranalyse sette navn på faktorene etter hvilke variabler som lader høyest på en faktor. Vi ser at alle de psykiske plagevariablene lader høyt på første faktor (alle faktorladningene er 0,40 eller høyere), mens de kroppslige plagene lader høyt på andre faktor.

Etter rotasjon er variansen i de opprinnelige variablene som blir forklart av de to faktorene refordelt. Siden vi her har brukt ekte faktoranalyse, fanger faktorene bare opp fellesvarians og neglisjerer den variansen som er spesifikk for den enkelte variabel. Totalt forklart varians blir i dette tilfellet 37,7 %. Første faktor forklarer 19,1 % av variansen mens den andre forklarer 18,6 % av variansen i de opprinnelige variablene.

Fig. 5.7: Faktoranalyse av plagevariabler. Roterte faktorladninger etter oblik rotasjon.

Pattern Matrix[a]

	Factor	
	1	2
Hodepine	-,12	**,83**
Magesmerter	,13	**,52**
Ryggsmerter	,04	**,45**
Svimmelhet	,09	**,50**
Nedfor	**,77**	-,04
Irritabel	**,65**	-,01
Nervøs	**,54**	,02
Søvnproblem	,37	,15

Extraction Method: Principal Axis Factoring.
Rotation Method: Oblimin with Kaiser Normalization.

a. Rotation converged in 12 iterations.

En variabels kommunalitet kan defineres som den andelen av variansen i variabelen som blir forklart av de roterte faktorene. I vårt tilfelle varierer kommunalitetene mellom 0,23 og 0,56. Når en variabel oppnår lav kommunalitet tyder det på at den ikke passer helt inn i noen av de roterte faktorene. Kommunaliteten blir høyere dersom en bruker prinsipal komponentanalyse i stedet for common factoring. Dette fordi den prinsipale komponentanalysen tar vare på variablenes unike varians og ikke bare kovariansen med andre variabler.

Den rotasjonsmetoden vi har benyttet i denne analysen (varimax-rotasjon) fungerer slik at de to faktorene blir ukorrelerte. Ut fra både korrelasjonsmatrisen og den uroterte faktorløsningen kan det imidlertid se ut til at de to faktorene i så fall må være høyt korrelerte. Vi har derfor forsøkt med en oblik rotasjonsmetode. Faktorladningene er gjengitt i Fig. 5.7. Resultatet er nokså likt det vi får ut ved ortogonal rotasjon. Faktorladningene ved oblik rotasjon gir større kontrast mellom høye og lave faktorladninger, og kan virke som en klarere løsning. Imidlertid kommer Søvnproblemer ut med en faktorladning som er lavere enn 0,40. Konklusjonen må vel bli at ingen av de to løsningene er noe bedre enn den andre, og at det dessuten er god overensstemmelse mellom de to.

Når faktorer roteres ortogonalt, er det ikke nødvendig å beregne korrelasjoner mellom faktorene. De er nemlig per definisjon satt til 0,00. Oblik rotasjon innebærer at faktorene får lov til å korrelere. Dermed blir det meningsfylt å beregne korrelasjonene mellom faktorer. I vårt tilfelle er korrelasjonen mellom de to faktorene 0,71. Dette er et tegn på at vi ikke uten videre kan regne med at en tofaktorløsning er det riktigste. Kanskje er det tross alt en enfaktorløsning som stemmer best med data. På den andre side sett er det påfallende hvordan de kroppslige og de psykiske plagene plasserer seg i hver sin gruppe. Dette taler for at løsningen med to faktorer er den beste. For å vite noe mer om dette, bør det gjennomføres konfirmatoriske faktoranalyser (der en tester de to hypotesene – en faktor eller to faktorer) mot hverandre. En kan også bruke eksterne konsistenskriterier. Dersom en sumskår basert på de kroppslige plagevariablene viser et mønster av korrelasjoner med tredjevariabler som er nokså likt det mønsteret en finner for en sumskår basert på de psykiske plagevariablene, taler dette for at enfaktorløsningen er den riktigste.

Dersom en bestemmer seg for å prøve en konfirmatorisk faktoranalyse, må dette gjøres på et nytt datasett. Dersom en bruker de samma dataene på nytt, vil en selvsagt bare få bekreftet det den eksploratoriske faktoranalysen har vist. En annen strategi består i å dele inn dataene i to deler og først gjennomføre en eksploratorisk dataanalyse på den ene halvparten og deretter en konfirmatorisk faktoranalyse på den andre. Dette forutsetter selvsagt at utvalget (antall informanter) er tilstrekkelig stort.

For å finne ut om ekstraksjonsmetoden kan ha noe å bety for resultatet av analysen, ble det gjort tilsvarende analyser med alle de øvrige ekstraksjonsmetodene som er tilgjengelige i SPSS for Windows. Hver gang ble det benyttet varimax-rotasjon av faktorene. Ingen av ekstraksjonsmetodene gav resultater som avvek nevneverdig fra "principal axis factoring". Siden de åtte leddene som inngikk i skalaen var temmelig skjevfordelte, ble det også gjort analyser på logaritmetransformerte variabler. En slik logaritmetransformasjon reduserer skjevheten på høyreskjeve variabler. Disse analysene gav resultater som var temmelig sammenfallende med de resultatene som er beskrevet ovenfor. Skjevfordelingen ser altså ikke ut til å ha noe vesentlig å bety. Dette kommer trolig av at analysene var basert på et svært stort datamateriale (høyt antall observasjoner).

5.5.3 Prinsipal komponentanalyse

Ovenfor er det argumentert for at en ikke uten videre kan anta at det eksisterer et felles underliggende fenomén som de ulike plagevariablene reflekterer. Kanskje har de ulike plagene hver sin egen unike etiologi. De relativt lave korrelasjonene mellom plager kan tyde på det. Kanskje gir det mer mening å tenke på plagevariablene som et sett variabler som vi ønsker å redusere til et mindre antall, men på en slik måte at vi tar vare på maksimalt av variansen i de opprinnelige variablene. I så fall bruker vi prinsipal komponentanalyse, som matematisk sett er enklere enn "common factoring"-teknikkene som vi har sett på ovenfor.

Resultatene av en prinsipal komponentanalyse er gjengitt i tabell 5.8. Utgangspunktet var den samme korrelasjonsmatrisen som er gjengitt i Tabell 5.2.

Fig. 5.8: Prinsipal komponentanalyse av plagevariabler. Roterte faktorladninger etter ortogonal rotasjon.

Rotated Component Matrix[a]

	Component	
	1	2
Hodepine	,22	**,74**
Magesmerter	,34	**,62**
Ryggsmerter	,05	**,71**
Svimmelhet	,24	**,65**
Nedfor	**,75**	,25
Irritabel	**,74**	,18
Nervøs	**,73**	,10
Søvnproblemer	**,50**	,31

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Resultatene er svært like de vi finner når vi kjører faktoranalyse med principal axis factoring og varimax-rotasjon (Fig. 5.7). På nytt identifiserer vi en undergruppe på fire ledd som alle dreier seg om kroppslige plager (plager som kan lokaliseres til en bestemt kroppsdel) og en annen undergruppe som omhandler psykiske plager. Men faktorladningene er høyere når vi bruker prinsipal komponentanalyse. Og kommunalitetene varierer nå mellom 0,35 og 0,63. De er altså høyere enn ved en tilsvarende faktoranalyse. Dette kommer av at en tar vare på varians som er unik for den enkelte variabel og ikke bare varians som er felles for flere variabler.

Her har vi gjengitt resultatene av en analyse der vi har rotert de prinsipale komponentene. Etter Joliffes (1989) kritikk, er det grunn til å advare mot uten videre å anvende slik rotasjon av prinsipale komponenter. Everitt (1996) har imidlertid argumentert for at en bør kunne rotere prinsipale komponenter.

5.5.4 Reliabilitet (indre konsistens)

Forutsatt at vi antar at hver av de to dimensjonene vi mener å ha identifisert gjenspeiler en underliggende, latent faktor, kan vi beregne den indre konsistensen i hver av disse ved å bruke Cronbachs Alpha. Alpha-verdiene er gjengitt i Figurene 5.9 og 5.10.

Vi ser av resultatene at skalaen for somatiske plager oppnår en alpha på 0,68, mens skalaen for psykiske plager oppnår en alpha på 0,69 når en runder av til to desimaler. Dette er ikke en svært høy reliabilitet. En anbefaler at skalaer som anses som så gode at de bør brukes internasjonalt bør ha en alpha-verdi på minst 0,80. Vi kan imidlertid være i tvil om alpha er riktig å anvende i denne sammenhengen. Alpha forutsetter at leddene som inngår i skalaen skal reflektere et underliggende begrep eller en latent størrelse. Avvik fra perfekt overensstemmelse mellom leddene skal ifølge denne logikken skyldes mangel på reliabilitet. Imidlertid kan manglende overensstemmelse mellom skårene på de ulike variablene her skyldes at de handler om fenoméner som er ulike. Vi skal ikke forvente at hodepine og magesmerter skulle korrelere perfekt dersom vi kunne måle perfekt. Tvert imot skulle vi forvente et avvik. Det er derfor mulig at Cronbachs Alpha i dette tilfellet underestimerer reliabiliteten i skalaen.

Fig. 5.9: Cronbachs alpha for kroppslige plager

Reliability Statistics

Cronbach's Alpha	N of Items
,682	4

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Hodepine	12,90	5,172	,549	,557
Magesmerter	12,74	6,225	,486	,610
Ryggsmerter	12,70	5,819	,393	,666
Svimmelhet	12,62	5,727	,447	,628

Tabellene viser også hva som skjer med Alpha dersom enkeltvariabler utelates fra skalaen. Vi ser at på den somatiske siden bidrar alle leddene positivt til alpha, men ryggsmerter bidrar minst. Dette samsvarer godt med det vi tidligere har sett i faktoranalysen og i den prinsipale komponentanalysen. På den psykiske siden bidrar tre av variablene klart positivt. Unntaket er søvnproblemer som praktisk talt ikke gjør noen forskjell. Også dette samsvarer bra med resultatene fra analysene ovenfor.

I dette tilfellet kan vi være i tvil om vi skal redusere hver av sumskårene til å omfatte tre ledd, eller om vi skal la alle inngå. Vi har valgt det siste.

Fig. 5.10: Cronbachs alpha for psykiske plager

Reliability Statistics

Cronbach's Alpha	N of Items
,685	4

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Nedfor	11,91	6,099	,558	,564
Irritabel	12,53	6,174	,499	,599
Nervøs	11,76	6,643	,448	,632
Søvnproblemer	11,90	6,008	,389	,683

5.5.5 Konstruksjon av sumskårer

Når vi skal redusere et større antall variabler til et mindre antall ved å beregne sumskårer, kan vi rent teknisk gjøre dette ved hjelp av faktoranalysene eller de prinsipale komponentanalysene. Faktoranalysen kan produsere faktorskårer og den prinsipale komponentanalysen kan gi oss komponentskårer som kan brukes i den videre analysen av data. Faktorskårer og sumskårer er beregnet med en tilsynelatende meget høy presisjon. Vektene som anvendes når den prinsipale komponentanalysen lager sumskårer har svært mange desimaler. Et annet særtrekk ved slike maskinelt produserte faktorer/komponenter, er at alle variablene inngår i samtlige. Det er størrelsen på vektene som avgjør hvor mye hver enkelt bidrar.

En annen framgangsmåte består i å følge de råd analysen gir om hvordan variablene grupperer seg, og deretter konstruere enkle sumskårer der alle variablene vektes likt eller ikke utstyres med noen bestemt vekting. Denne framgangsmåten har den fordel at sumskårene er enkle å forklare og at de får et mer presist og entydig innhold.

Vi har laget en enkel, additiv sumskår for psykiske plager og en for kroppslige plager. Fordelingene av sumskårene er vist i Fig. 5.11. Begge sumskårene har verdier fra 0-16. De som har verdien null har svart "sjelden eller aldri" på alle de spørsmål som inngår i skalaen. De som har oppnådd verdien 16 har svart daglig på alle fire ledd som inngår i skalaen. Tallverdier på midten av skalaen kan ha framkommet gjennom et stort antall ulike kombinasjoner av avkrysninger. Vi antar imidlertid at jo høyere verdier, desto større forekomst av plager. Dette forutsetter egentlig at de spørsmål som inngår er målt på en

intervallskala, og at vi får fram et riktig bilde ved ganske enkelt å addere sammen svarene på spørsmålene. Dette kan vi ikke alltid regne med, og ideelt bør en undersøke skalaens "additivitet" før en foretar en slik beregning av sumskårer som vi har gjort her.

Vi ser av fordelingene at begge sumskårene er ganske skjevfordelte og at de sannynligvis også er spissere enn en normalfordeling ville vært. Det viser seg da også at skjevheten (skewness) er 1,14 på sumskåren for psykiske plager og 1,41 på sumskåren for somatiske plager. Spissheten (kurtosis) for de to sumskårene er 1,20 for psykiske plager og 1,80 for somatiske plager. Sumskåren for somatiske plager avviker med andre ord sterkere fra normalfordelingen enn det sumskåren for psykiske plager gjør og er ikke særlig godt egnet til bruk i analyser som forutsetter normalfordelte variabler. Også sumskåren for psykiske plager vil få noe bedre egenskaper etter en logaritmetransformasjon.

En mulig løsning på dette problemet er å logaritmetransformere skalaen. Vi adderer da først tallet 1,0 til begge sumskårene. Dette for at skalaen skal starte på 1,0 i stedet for 0,0. Dette er nødvendig når vi skal logaritmetransformere skalaene. Ved å gjøre dette oppnår vi at den logaritmetransformerte skalaen begynner på 0,0. Den naturlige logaritmen til 1,0 er nemlig 0,0. Deretter foretar vi transformasjonen. Resultatet er vist i Fig. 5.12. Vi ser at fordelingene er identiske med de vi finner i Fig. 5.11 med ett unntak. Verdiene på skalaen (som vises i kolonnen lengst til venstre) har endret seg.

Etter denne transformasjonen endrer skjevheten seg til -0,31 på skalaen for psykiske plager og til -0,02 på skalaen for somatiske plager. Spissheten endrer seg til henholdsvis -0,41 og -0,93. Disse verdiene avviker langt mindre fra 0,0 (altså samme skjevhet og spisshet som det vi finner på en normalfordelt variabel) enn verdiene vi regnet ut for de ikke-transformerte variablene.

Det vi imidlertid må huske på når vi foretar slike transformasjoner, er at vi endrer egenskapene til den underliggende skalaen. Sammenliknet med den opprinnelige skalaen har vi nå en skala der intervallstørrelsen reduseres etter hvert som vi beveger oss oppover på skalaen. Det betyr at avstanden mellom lave skårer nå tillegges større vekt enn avstanden mellom høyere skårer (når vi sammenlikner med den opprinnelige skalaen). Det er slett ikke sikkert at dette er ønskelig. Med et så høyt antall observasjoner som vi har i denne studien, er det sannsynlig at resultatene av analyser der vi bruker ikke-transformerte skårer uansett blir pålitelige.

Fig. 5.11: Enveis frekvensfordelinger av psykiske og somatiske plager (sumskårer)

Psykiske plager, sumskår

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	133	8,1	8,2	8,2
	1,00	245	15,0	15,2	23,4
	2,00	263	16,1	16,3	39,7
	3,00	226	13,8	14,0	53,7
	4,00	187	11,4	11,6	65,3
	5,00	146	8,9	9,0	74,3
	6,00	111	6,8	6,9	81,2
	7,00	79	4,8	4,9	86,1
	8,00	64	3,9	4,0	90,0
	9,00	53	3,2	3,3	93,3
	10,00	34	2,1	2,1	95,4
	11,00	27	1,6	1,7	97,1
	12,00	17	1,0	1,1	98,1
	13,00	9	,5	,6	98,7
	14,00	7	,4	,4	99,1
	15,00	7	,4	,4	99,6
	16,00	7	,4	,4	100,0
	Total	1615	98,7	100,0	
Missing	System	22	1,3		
Total		1637	100,0		

Kroppslige plager, sumskår

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	342	20,9	21,1	21,1
	1,00	286	17,5	17,6	38,7
	2,00	272	16,6	16,7	55,4
	3,00	190	11,6	11,7	67,1
	4,00	145	8,9	8,9	76,0
	5,00	104	6,4	6,4	82,5
	6,00	74	4,5	4,6	87,0
	7,00	48	2,9	3,0	90,0
	8,00	58	3,5	3,6	93,5
	9,00	31	1,9	1,9	95,4
	10,00	23	1,4	1,4	96,9
	11,00	17	1,0	1,0	97,9
	12,00	14	,9	,9	98,8
	13,00	7	,4	,4	99,2
	14,00	6	,4	,4	99,6
	15,00	2	,1	,1	99,7
	16,00	5	,3	,3	100,0
	Total	1624	99,2	100,0	
Missing	System	13	,8		
Total		1637	100,0		

Fig. 5.12: Frekvensfordeling av psykiske og somatiske plager etter at skalaene er logaritmetransformert

Psykiske plager - logarimetransformert skala

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	133	8,1	8,2	8,2
	,69	245	15,0	15,2	23,4
	1,10	263	16,1	16,3	39,7
	1,39	226	13,8	14,0	53,7
	1,61	187	11,4	11,6	65,3
	1,79	146	8,9	9,0	74,3
	1,95	111	6,8	6,9	81,2
	2,08	79	4,8	4,9	86,1
	2,20	64	3,9	4,0	90,0
	2,30	53	3,2	3,3	93,3
	2,40	34	2,1	2,1	95,4
	2,48	27	1,6	1,7	97,1
	2,56	17	1,0	1,1	98,1
	2,64	9	,5	,6	98,7
	2,71	7	,4	,4	99,1
	2,77	7	,4	,4	99,6
	2,83	7	,4	,4	100,0
	Total	1615	98,7	100,0	
Missing	System	22	1,3		
Total		1637	100,0		

Somatiske plager - logaritmetransformert skala

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	342	20,9	21,1	21,1
	,69	286	17,5	17,6	38,7
	1,10	272	16,6	16,7	55,4
	1,39	190	11,6	11,7	67,1
	1,61	145	8,9	8,9	76,0
	1,79	104	6,4	6,4	82,5
	1,95	74	4,5	4,6	87,0
	2,08	48	2,9	3,0	90,0
	2,20	58	3,5	3,6	93,5
	2,30	31	1,9	1,9	95,4
	2,40	23	1,4	1,4	96,9
	2,48	17	1,0	1,0	97,9
	2,56	14	,9	,9	98,8
	2,64	7	,4	,4	99,2
	2,71	6	,4	,4	99,6
	2,77	2	,1	,1	99,7
	2,83	5	,3	,3	100,0
	Total	1624	99,2	100,0	
Missing	System	13	,8		
Total		1637	100,0		

5.6 Konfirmatorisk faktoranalyse

Konfirmatorisk faktoranalyse er en teknikk som tradisjonelt har vært langt mindre brukt enn de ulike variantene av eksploratorisk faktoranalyse. Den klassiske versjonen av konfirmatorisk faktoranalyse er Jøreskog's metode LISREL (**Li**near **S**tructural **Rel**ations) (Jöreskog & Sörbom, 1988). Her har vi benyttet et system som er noe enklere å bruke, og som fungerer som et anneks til SPSS, nemlig AMOS (Byrne, 1994). AMOS utfører stort sett de samme analyser og beregner de samme koeffisienter som vi finner i LISREL.

Eksploratorisk faktoranalyse er en teknikk som så og si på egen hånd leter opp en løsning. Slik faktoranalytiske teknikker fungerer i de fleste programpakker, er det programmet selv som finner ut hvor mange faktorer som bør roteres og hvilke variabler som lader høyt på hver faktor. Den er simpelthen eksplorerende eller utforskende. Konfirmatorisk faktoranalyse innebærer at en selv må ha meninger og hypoteser på forhånd. Forskeren spesifiserer selv en modell som deretter blir testet mot data. Analysen gir svar på hvor godt modellen stemmer overens med dataene, og gir holdepunkt for forbedringer i modellen. Bruk av konfirmatorisk faktoranalyse innebærer ofte at en prøver ut flere modeller for å finne en enklest mulig modell som passer tilstrekkelig godt med dataene.

På samme måte som i den eksploratoriske faktoranalysen (altså ikke innen prinsipal komponentanalyse) skilles det mellom observerte variabler og latente variabler. De observerte variablene er de som inngår i dataene som skal analyseres, mens de latente er hypotetiske variabler som vi skal teste mot de observerte.

For å illustrere bruken av konfirmatorisk faktoranalyse, skal vi gå videre med samme problemstilling som ovenfor ved å se på dimensjoner i subjektive helseplager. Dataene vi skal bruke er fra samme prosjektet (Helsevaner blant skolebarn i Europa). Mens vi ovenfor brukte data samlet inn i 1993-94, skal vi nå bruke data som er samlet inn i 1997-98. Siden vi allerede på grunnlag av eksploratorisk faktoranalyse har dannet oss en hypotese om hva slags faktorer vi venter å finne, er det rimelig at vi nå tester ut denne hypotesen med en konfirmatorisk faktoranalyse.

Denne gangen er skalaen utvidet fra 8 til 11 ledd:

- Hodepine
- Vondt i magen
- Vondt i ryggen
- Svimmel
- Vondt i nakken
- Følt deg trist (nedfor)
- Irritabel eller i dårlig humør
- Nervøs
- Vanskelig for å sovne
- Lei og utslitt
- Redd

Svarkategoriene er omtrent de samme som sist:

Omtrent hver dag
Mer enn en gang pr. uke
Omtrent hver uke
Omtrent hver måned
Sjelden eller aldri

I konfirmatorisk faktoranalyse spesifiserer en gjerne modellen en vil prøve ut ved hjelp av en enkel matrise: I Tabell 5.3 vises det hvordan denne ser ut. Ett-tallene symboliserer at en variabel skal inngå i en faktor. Vi ser at under enfaktor-modellen inngår alle variablene i faktoren. Når vi spesifiserer tofaktor-modellen, viser fremdeles ett-tallene hvilke variabler som inngår i en faktor. Nullene viser hvilke variabler som ikke inngår i en bestemt faktor. Variablene 1-5 inngår altså i den første faktoren, mens variablene 6-11 inngår i den andre. Skjematisk framstilt ser tofaktor-modellen ut slik som vist i Fig. 5.13.

Tabell 5.3: Spesifikasjon av modeller i konfirmatorisk faktoranalyse

Opprinnelig variabel nr.:	Enfaktor-modell	Tofaktor-modell	
1. Hodepine	1	1	0
2. Vondt i magen	1	1	0
3. Vondt i ryggen	1	1	0
4. Svimmel	1	1	0
5. Vondt i nakken	1	1	0
6. Følt deg trist (nedfor)	1	0	1
7. Irritabel eller i dårlig humør	1	0	1
8. Nervøs	1	0	1
9. Vanskelig for å sovne	1	0	1
10. Lei og utslitt	1	0	1
11. Redd	1	0	1

I tillegg til å spesifisere hvilke variabler som skal lade på hvilken faktor, kan en også spesifisere om faktorene (dersom en har spesifisert mer enn en) skal tillates å korrelere.

AMOS produserer på grunnlag av kovariansmatrisen og de spesifiserte modellene standardiserte koeffisienter som tilsvarer faktorladninger i eksploratorisk faktoranalyse. Tabell 5.3 viser standardiserte koeffisienter for tre forskjellige modeller. I første kolonne finner vi koeffisientene for enfaktor-modellen. Kolonnene i midten viser en tofaktor-modell der de to faktorene ikke tillates å korrelere med hverandre. Kolonnene lengst til høyre viser en tofaktor-modell der faktorene tillates å korrelere.

Nederst på tabellen gjengis tre statistiske størrelser. χ^2-verdien viser en test av uoverensstemmelse mellom modell og data. Vi ser at alle modellene er signifikant forskjellige fra data. Dersom vi slår opp i en tabell over χ^2-fordelingen finner vi at ved 50 frihetsgrader (de fleste tabeller viser ingen tall for frihetsgrader mellom 40 og 50) er den verdien som tilsvarer en sannsynlighet på $p<.001$ lik 86,66. Våre χ^2-verdier er mye høyere enn dette. Vi kan altså slå fast at diskrepansen mellom modeller og data er signifikant uansett hvilken av de tre modellene vi velger. For å kunne vurdere χ^2-verdien er det viktig å være klar over at den er avhengig av antall observasjoner. Jo større n, desto høyere er χ^2-verdien dersom vi holder alle andre forhold konstante. I vårt tilfelle analyserer vi data fra en undersøkelse som omfatter nesten 1670 skolebarn (elever på 10. klassetrinn). Vi kan derfor regne med at til og med ganske gode modeller vil vise seg å være signifikant forskjellige fra data.

Mer informativt er det å se på ulike koeffisienter som sier noe om modelltilpasning. Kerlinger & Lee, 1999) hevder at "state of the art" i dag er en indeks som kalles Komparativ Tilpasningsindeks (Comparative Fit Index - CFI). Jo høyere den er, desto bedre passer modellen til data. For at en virkelig skal kunne snakke om en god modell, bør CFI være minst 0,95. Dersom den er lavere, betyr det at modellen bør kunne forbedres. Et annet mål som sier noe om hvor god modellen er, er RMSEA (root mean square error of approximation). Jo lavere RMSEA er, desto bedre er modellen. RMSEA bør være 0,05 eller lavere.

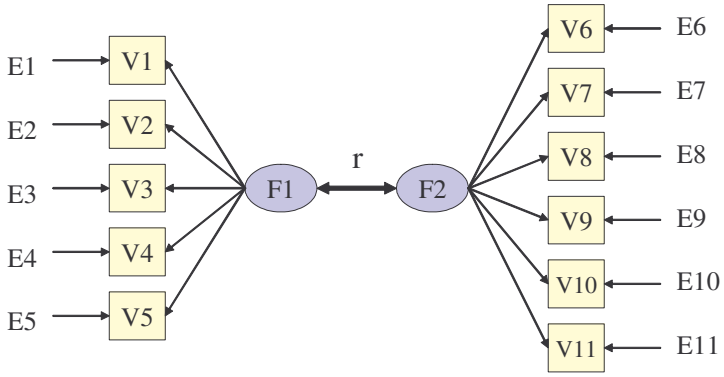
Dersom vi inspiserer tabell 5.4, ser vi at enfaktor-modellen gir en CFI på 0,869 og en RMSEA på 0,085. Modellen ser med andre ord ikke ut til å være tilfredsstillende. Dersom vi i stedet tester ut en tofaktor-modell der vi ikke tillater korrelasjon mellom de to faktorene, blir begge indeksene enda dårligere. CFI går ned til 0,772 og RMSEA øker til 0,113. Vi har derfor prøvd ut en tredje modell der vi opererer med to faktorer, men vi tillater disse faktorene å korrelere. Denne modellen ser bedre ut. CFI øker nå til 0,922 mens RMR synker til 0,067. Fremdeles kan det se ut til at modellen kan forbedres, men det er likevel åpenbart at blant de tre modellene vi har testet ut, er den siste modellen klart å foretrekke. AMOS beregner korrelasjonen mellom de to faktorene. Den er så høy som 0,76.

Det er også mulig å bruke signifikanstesting for å se om en bestemt modell er bedre enn en annen modell. For å kunne teste to modeller mot hverandre, må de være helt identiske, bortsett fra på ett bestemt punkt. Modellene må være "nestede" – den ene modellen må være "innhyllet" i den andre. En bruker i så fall χ^2-tallene for modellene. Diskrepansen i χ^2-verdi og diskrepansen i antall frihetsgrader brukes direkte for å se hvor mye bedre den ene modellen er enn den andre. Idealet er å komme fram til enkle modeller som passer godt med dataene. La oss se på de to siste modellene som er gjengitt i Tabell 5.4. Den ene er en tofaktormodell der de to faktorene ikke tillates å korrelere. Den andre er en tofaktormodell der de tillates å korrelere. Den første av disse er nestet i den andre, og de kan derfor testes mot hverandre. Forskjellen i χ^2-verdi er 618,98 og forskjellen i frihetsgrader er 1. χ^2-verdien er i dette tilfellet skyhøyt over de aktuelle kritiske verdiene χ^2-fordelingen. Vi kan med andre ord trekke den konklusjon at tofaktor-modellen der vi tillater faktorene å korrelere er signifikant bedre enn tofaktor-modellen der vi ikke tillater faktorene å korrelere.

Et nyttig redskap når en skal sammenlikne modeller er Hirotogu Akaikes informasjonskriterium (Akaike's Information Criterion - AIC) (Akaike, 1974). AIC veier en

modells kompleksitet mot hvor godt modellen passer med data. Jo lavere AIC-verdi, desto bedre stemmer modellen med data. Det er ikke særlig vanskelig å få en modell til å stemme godt med data, dersom en bare lager modellen tilstrekkelig kompeks. Idealet er imidlertid å komme fram til modeller som er rimelig enkle, men likevel passer bra med dataene. AIC har ingen høy stjerne blant profesjonelle statistikere[2], men har likevel fått stor utbredelse. Akaikes informasjonskriterium kan anvendes til å sammenlikne modeller også når de er ikke-nestede. I vårt tilfelle var AIC 1047,293 i modellen der vi ikke tillot korrelasjon mellom faktorene og 430,313 i den modellen der vi tillot korrelasjon mellom faktorene. Også AIC bekrefter at den siste modellen er best, og at den også er tydelig bedre enn enfaktor-modellen.

Fig. 5.13: Konfirmatorisk faktoranalyse
grafisk illustrasjon av tofaktor-modell



V - de observerte variablene
E - feiltermer
F - er de latente variablene eller faktorene
r - den estimerte korrelasjonen mellom de to faktorene

Innen AMOS tillates hver variabel å lade på flere enn en faktor. Ved å prøve seg fram, kan en sannsynligvis få til en ytterligere forbedret tilpasning mellom modell og data. En kan også få bedre modelltilpasning ved å tillate korrelasjoner mellom feiltermene (her symbolisert ved E1 – E11). AMOS gir statistikk som indikerer på hvilke punkt en kan forbedre modellen (modification indices).

Resultatene av både den eksploratoriske faktoranalysen som ble gjennomført på et tidligere innsamlet datasett og av den konfirmatoriske analysen som er gjort ovenfor, peker i samme retning. Plagevariablene ser ut til å danne to grupper eller dimensjoner. Den ene gruppen består av psykiske plager og den andre av kroppslige plager. De to faktorene korrelerer imidlertid høyt. Den konklusjonen vi har kommet til her er ikke nødvendigvis den endelige. Dersom det for eksempel viser seg at en får svært like resultater ved å analysere den somatiske sumskåren og den psykiske sumskåren mot andre variabler (eksterne konsistenskriterier), kan det for de fleste formål være akseptabelt å lage en sumskår som omfatter begge grupper av variabler.

[2] http://www.garfield.library.upenn.edu/classics1981/A1981MS54100001.pdf

I det eksempelet på konfirmatorisk faktoranalyse som er presentert her, har vi nøyd oss med å se på en og to faktorer. De viktigste programsystemene for slike analyser (LISREL, EQS, AMOS og MPLUS) kan selvsagt håndtere langt flere variabler og faktorer enn det vi har benyttet her. Dessuten kan disse programmene brukes til uttesting av langt mer kompliserte modeller enn de faktoranalytiske. I disse modellene undersøker en ikke bare hvordan et større sett av observerte variabler kan reduseres til en mindre antall latente variabler, men en kan blant annet også se hvordan de latente variablene henger sammen.

Tabell 5.4: Faktorladninger under ulike modeller. Tofaktor-modell A tillater ikke faktorene å korrelere med hverandre. Tofaktormodell B tillater en slik korrelasjon.

Opprinnelig variabel nr.:	Enfaktor-modell	Tofaktor-modell A		Tofaktor-modell B	
1. Hodepine	.54	.63	0	.60	0
2. Vondt i magen	.53	.58	0	.60	0
3. Vondt i ryggen	.44	.56	0	.52	0
4. Svimmel	.55	.52	0	.56	0
5. Vondt i nakken	.53	.59	0	.60	0
6. Følt deg trist (nedfor)	.66	0	.73	0	.71
7. Irritabel eller i dårlig humør	.59	0	.64	0	.63
8. Nervøs	.56	0	.60	0	.59
9. Vanskelig for å sovne	.45	0	.43	0	.45
10. Lei og utslitt	.61	0	.57	0	.61
11. Redd	.54	0	.57	0	.56
r (korrelasjon mellom faktorer)	-		0,00		0,76
χ^2 for modell (frihetsgrader):	580,16 (44)	981,29 (44)		362,31 (43)	
Comparative fit indeks (CFI):	0,869	0,772		0,922	
RMSEA:	0,085	0,113		0,067	
AIC:	646,164	1047,293		430,313	

Referanser

Akaike, Hirotugu (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.

Bagozzi, R.P., & Fornell, C. (1982). Theoretical concepts, measurement, and meaning. I C.Fornell (red.), *A second generation of multivariate analysis* (Vol.2, s.24-38). New York: Praeger.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296-322.

Byrne, B.M. (1994). *Structural equation modeling using AMOS*. London: Lawrence Erlbaum Associates.

Carmines, E.G. & Zeller, R.A. (1979). *Reliability and validity assessment.* Beverly Hills, California: Sage. (Quantitative Applications in the Social Sciences nr. 17)

Cattell, R.B. (1965). Factor analysis: An introduction to essentials. (I) The purpose and underlying models, (II) the role of factor analysis in research. *Biometrics*, *Vol.21*, 190-215, 405-435.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrica, 16*, 297-334.

DeVellis, R.F. (1991). *Scale development. Theory and applications.* Newbury Park, California: Sage (Applied Social Research Methods Series, Vol.26).

Dunteman, G.H. (1989). *Principal components analysis.* Beverly Hills, California: Sage. (Også gjengitt som eget kapittel i M.S.Lewis-Beck (red.). *Factor analysis and related techniques.* Beverly Hills, California: Sage, 157-245. (International handbooks of quantitative applications in the social sciences, Vol.5)

Everitt, B.S. (1996). *Making sense of statistics in psychology. A second level course.* Oxford: Oxford University Press.

Gorsuch, R.L. (1988). Exploratory factor analysis. I J.R.Nesselroade & R.B.Cattell (red.) *Handbook of multivariate experimental psychology (second edition)*. New York: Plenum Press, 231-258.

Guilford, J.P. & Fruchter, B. (1978). *Fundamental statistics in Psychology and education.* Tokyo: McGraw-Hill Kogahusha.

Hair, J.F.jr., Anderson, R.E., Tatham, R.L. & Black, W.C. (1992). *Multivariate data analysis with readings.* New York: Macmillan Publishing Company.

Havik, O. (1982). Rehabilitering av hjerteinfarktpasienter. Effekten av et informasjonsprogram. Metodenotat IV. ADI-metoder: Utvikling, reliabilitet og validitet. Bergen: Institutt for klinisk psykologi. (Upublisert notat)

Heise, D.R. & Bohrnstedt, G.W. (1970). Validity, invalidity, and reliability. I E.F.Borgatta & G.W.Bohrnstedt (red.). *Sociological methodology*. San Francisco: Jossey-Bass, 104-129.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *Vol.24*, 417-441, 498-520.

Joliffe, I.T. (1989). Rotation of ill-defined principal components. *Applied Statistics*, *38*, 139-148.

Jöreskog, K.G. & Sörbom, D. (1988). *LISREL 7. A guide to the program and application.* Chicago, Illinois: SPSS Inc.

Kaiser, H.F. (1974). An index of factorial simplicity. *Psychometrica*, *Vol.39*, 31-36.

Kerlinger, F.N. (1973). *Foundations of behavioral research.* London: Holt, Rinehart & Winston.

Kerlinger, F.N. (1986). *Foundations of behavioral research (Third edition).* New York: CBS Publishing Japan.

Kerlinger, F.N. & Lee, H.B. (1999). *Foundations of behavioral research (Fourth edition)*. Fort Worth, Texas: Harcourt Collegeg Publishers.

Kim, J.-O. & Mueller, C.W. (1978a). *Introduction to factor analysis: What it is and how we do it.* Beverly Hills, California: Sage. (Også gjengitt som eget kapittel i M.S.Lewis-Beck (red.). *Factor analysis and related techniques.* Beverly Hills, California: Sage, 1-74. (International handbooks of quantitative applications in the social sciences, Vol.5)

Kim, J.-O. & Mueller, C.W. (1978b). *Factor analysis. Statistical methods and practical issues.* Beverly Hills, California: Sage. (Også gjengitt som eget kapittel i M.S.Lewis-Beck (red.). *Factor analysis and related techniques.* Beverly Hills, California: Sage, 75-156. (International handbooks of quantitative applications in the social sciences, Vol.5)

Kuder, G.F., & Richardson, M.W. (1937). The theory of the estimation of test reliability. *Psychometrica, 2*, 151-160.

Lazarsfeld, P.F. (1959). Problems in methodology. I R.K.Merton, L.Broom & L.S.Cottrell (red.). *Sociology today: Problems and prospects. Volume 1.* New York: Harper & Row. (etter Mastekaasa, 1987)

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading Massachussetts: Addison-Wesley.

Mastekaasa, A. (1987). Modellbruk, indekser og konsistenskriterier. *Tidsskrift for samfunnsforskning*, *Vol.28*, 167-188.

Norusis, M.J. (1985). *SPSS-X. Advanced Statistics Guide.* New York: McGraw Hill.

Novick, M. & Lewis, G. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrica*, *Vol.32*, 151-160.

Nunnally, J.C. (1967). *Psychometric theory.* New York: McGraw-Hill.

Nunnally, J.C. (1978). *Psychometric theory* (2. utgave). New York: McGraw-Hill.

Pedhazur, E.J. & Schmelkin, L.P. (1991). *Measurement, design, and analysis. An integrated approach*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Rosenberg, M. (1968). *The Logic of Survey Analysis.* New York: Basic Books.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271-295.

Thurstone, L.L. (1931). Multiple factor analysis. *Psychological Review*, *Vol.38*, 406-427.

Thurstone, L.L. (1947). *Multiple factor analysis.* Chicago: University of Chicago Press.

Wanous, J.P., Reichers, A.E., & Hudy, M.J. (1997). Overall job satisfaction: How good are single-item measures? *Journal of Applied Psychology, 82 (2)*, 247-252.

Kap 6: REGRESJONSANALYSE

Det er utviklet flere regresjonsanalytiske teknikker for kategorielle data. Likevel skal vi begynne dette kapittelet med en kort beskrivelse av vanlig lineær regresjon. Dette gjør vi fordi vanlig lineær regresjon er en velkjent statistisk metode. Ved først å repetere noe som er kjent, vil det for de fleste være lettere deretter å ta fatt på noe som er nytt.

For å kunne forstå hvordan multippel regresjonsanalyse fungerer, er det nødvendig å ha kjennskap til bivariat regresjonsanalyse. Vi skal derfor innledningsvis nokså kort se på noen av de viktigste poengene med bivariat, lineær regresjon.

6.1 Bivariat regresjonsanalyse

6.1.1 Prediksjon av en kriterievariabel

Bivariat regresjonsanalyse er en teknikk til å analysere forholdet mellom to variabler som er målt på en metrisk skala. Nærmere bestemt er det en teknikk som forteller hvordan en på grunnlag av kjennskap til et individs skåre på en variabel kan regne ut hva som er den mest sannsynlige verdien på den andre variabelen. De enkleste formene for regresjon er basert på at sammenhengen mellom to variabler kan beskrives ved hjelp av en rett linje. I to figurer nedenfor viser vi hvordan en lineær og en kurvilineær sammenheng ser ut når vi plotter alle observasjonene i et punktdiagram (Fig. 6.2 og Fig. 6.3).

For å kunne predikere verdier på kriterievariabelen (den avhengige variabelen) ut fra kjennskap til prediktorvariabelen (den uavhengige), har vi bruk for en enkel formel:

$$Y = a + bX \qquad (6.1)$$

Y Kriterievariabel (avhengig variabel)

X Prediktorvariabel (uavhengig variabel)

a Konstant (intercept)

b Stigningskoeffisient (slope)

Dersom det er en perfekt sammenheng mellom variablene, kan vi bestemme verdien av Y helt nøyaktig når vi kjenner X (Fig. 6.1). Et nærliggende eksempel er forholdet mellom en Celsius-skala og en Fahrenheit-skala. Forholdet mellom disse to skalaene til måling av temperatur kan beskrives slik:

$$Y = 32 + \frac{9}{5} * X$$

X er temperaturen angitt på en Celsius-skala mens Y er temperaturen målt på en Fahrenheit-skala. Vi kan da regne ut at når temperaturen er 20 grader celsius blir den 68 grader Fahrenheit (Eksempelet er hentet fra Lewis-Beck, 1980). Vi trenger altså to opplysninger for å konvertere en celcius-verdi til en Fahrenheit-verdi. Det første av disse tallene kalles en intercept, og dette tallet er en konstant (en tallverdi som ikke varierer), mens det andre er en stigningskoeffisient.

Fig 6.1: En lineær sammenheng uten feilvarians

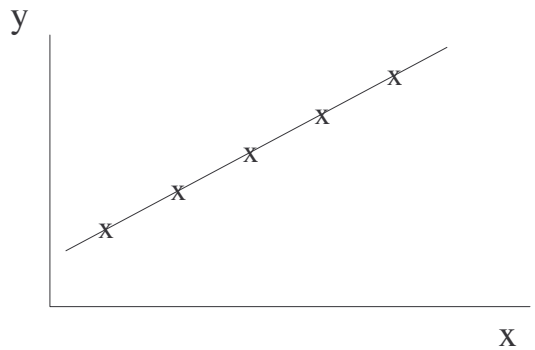


Fig. 6.2: En lineær sammenheng med feilvarians

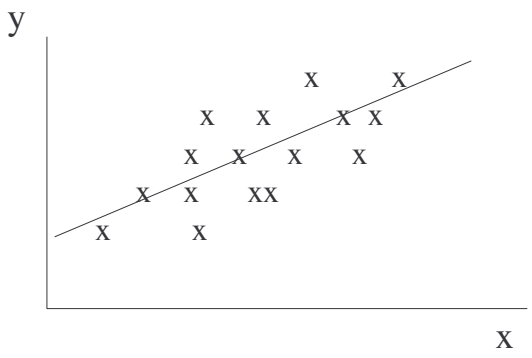
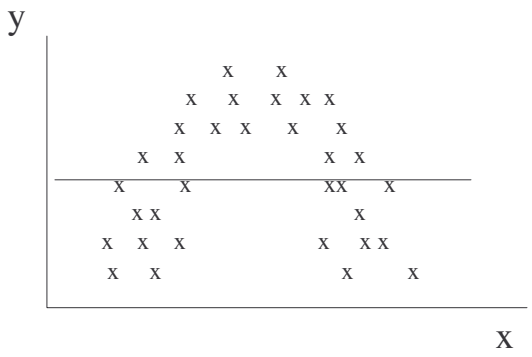


Fig. 6.3: En kurvilineær sammenheng med feilvarians



I det virkelige liv kan forholdet mellom to variabler sjelden beskrives så eksakt som dette. Særlig i studiet av atferd, psykologiske faktorer og sosiale forhold er det langt mer usikkerhet knyttet til å predikere en egenskap ut fra en annen. For å ta vare på denne usikkerheten er det nødvendig å utvide formelen med ett ekstra ledd:

$$Y = a + bX + e \qquad (6.2)$$

Bokstaven *e* står for feil (error) og betegner den justeringen vi må gjøre for å regne ut den nøyaktige verdien av Y når vi har predikert så godt vi kan ved hjelp av konstant og stigningstall. Dersom sammenhengen var perfekt, slapp vi å ha med en feilterm (e). Feiltermen representerer med andre ord observasjonenes avvik fra perfekt sammenheng mellom variablene.

Dersom vi har en situasjon der det finnes feil, men vi neglisjerer feiltermen, får vi en predikert Y-verdi som betegnes *Y'*.

$$Y' = a + bX \qquad (6.3)$$

Y' er til forskjell fra *Y* de predikerte verdiene vi kommer fram til når vi bare har kjennskap til *X*, *a* og *b*. Den linjen vi får ved å tegne inn $Y' = a + bX$ i et punktdiagram, kalles regresjonslinjen. Punktene i et punktdiagram som beskriver sammenhengen mellom to korrelerte variabler, vil for det meste befinne seg i nærheten av denne linjen (forutsatt at sammenhengen er rimelig sterk og lineær eller tilnærmet lineær).

Når vi tar utgangspunkt i to variabler som inngår i en undersøkelse, og lager et punktdiagram for å beskrive forholdet mellom disse to variablene, trenger vi en prosedyre for å plassere regresjonslinjen i diagrammet og for å beregne formelen $Y = a + bX$. Den metoden som vanligvis benyttes heter minste kvadraters metode. La oss tenke oss at vi har tegnet inn en linje som passerer omtrent gjennom tyngdepunktet av punkter i et punktdiagram. Vi kan regne ut hvor godt linjen passer med punktene ved å måle avstanden på y-aksen mellom linjen til hvert punkt i diagrammet. Denne avstanden kvadrerer vi for hvert punkt og adderer sammen. Dette kalles kvadratsummen av feilene (SSE - sum of square of errors) og uttrykkes ved formelen:

$$SSE = \sum_{i=1}^{n} (Y_i - Y_i')^2 \qquad (6.4)$$

SSE Kvadratsummen av feilene (sum of square of errors)

Y_i Subjekt nr. i sin verdi på Y

Y_i' Verdien vi ut fra X forventet at individ nr. i skulle hatt på Y-aksen

$(Y_i - Y_i')$ Avstanden på y-aksen mellom en observasjon og regresjonslinjen

n Antall observasjoner

Regresjonslinjen er den linjen i diagrammet som minimaliserer kvadratsummen av feilene. For å tegne inn regresjonslinjen i et punktdiagram trenger vi å vite to tallstørrelser, nemlig *a* og *b* fra formel 6.3.

Vi beregner regresjonslinjen ved bruk av følgende uttrykk:

$$b = \frac{\sum_{i=1}^{n}((X_i - \overline{X})(Y_i - \overline{Y}))}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \qquad (6.5)$$

$$a = \overline{Y} - b\overline{X} \qquad (6.6)$$

X_i Subjekt nr. i sin verdi på x-aksen

$\overline{X}$ Gjennomsnittet av alle subjekters verdi på x-aksen

Y_i Subjekt nr. i sin verdi på y-aksen

$\overline{Y}$ Gjennomsnittet av alle subjekters verdi på y-aksen

I formlene 6.5 og 6.6 beregner vi det som kalles minste kvadraters estimater av a og b. Stigningskoeffisienten som beregnes på denne måten forteller hvor stor gjennomsnittlig endring i *Y* vi kan regne med for hver enhet på x-aksen. Det er viktig å vite at selv om vi ved hjelp av en slik bivariat regresjonsanalyse kan vise hvordan *X* kan predikere *Y*, så sier ikke dette noe om kausaliteten mellom variablene. Hva som er den riktigste kausalrekkefølge på variablene avgjøres på forhånd og uavhengig av regresjonsanalysen.

En ulempe med stigningskoeffisienten er at den avhenger av skalaen på de variablene som inngår i analysen. Dette betyr at vi ikke uten videre kan sammenlikne to stigningskoeffisienter med hverandre. Dette problemet løses ved at en standardiserer variablene før en regner ut *b* og *a*. Dersom a og b beregnes på grunnlag av standardiserte variabler, kalles de ikke lenger *a* og *b*, men alfa (α) og beta (β). Den standardiserte stigningskoeffisienten kalles med andre ord en beta-vekt eller en beta-koeffisient. Det viser seg at den standardiserte regresjonskoeffisienten i det bivariate tilfellet er identisk med Pearsons r.

En svært anvendbar statistisk størrelse er den såkalte determinasjonskoeffisienten eller R^2. Determinasjonskoeffisienten forteller hvor mye av variansen på variabelen Y som kan forklares ved hjelp av variabelen X. R^2 kan beregnes ved hjelp av kvadratsummer:

$$R^2 = \frac{SSR}{SST} \qquad (6.7)$$

$SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$ Total kvadratsum (sum of squares total)

$SSE = \sum_{i=1}^{n}(Y_i - Y')^2$ Kvadrarsummen av feil (sum of squares error)

$SSR = SST - SSE$ Forklart kvadratsum (sum of squares explained)

Determinasjonskoeffisienten er nært beslektet med produkt-moment-korrelasjonen. Determinasjonskoeffisienten er lik kvadratet av r.

$$R^2 = r^2 \qquad (6.8)$$

Siden vi i det bivariate tilfellet har funnet at den standardiserte regresjonskoeffisienten er lik r, viser det seg også at

$$R = r = \beta \text{ (beta)} \qquad (6.9)$$

6.1.2 Signifikanstesting og konfidensintervall

Den vanligste formen for signifikanstesting i bivariat regresjonsanalyse gjelder testing av om stigningskoeffisienten (b) er større enn 0 (null). For å kunne signifikansteste trenger vi formelen for standardfeilen på b. Den ser slik ut:

$$se_b = \sqrt{\frac{\sum_{i=1}^{n}\frac{(Y_i - Y')^2}{(n-2)}}{\sum_{i=1}^{n}(X_i - \overline{X})^2}} \qquad (6.10)$$

se_b Standardfeilen til stigningskoeffisienten b

Y_i Verdien på Y-variabelen

Y' Verdien på Y estimert ut fra verdien på X

X_i Verdien på X-variabelen

$\overline{X}$ Gjennomsnittet på X-variabelen

n Antall enheter som inngår

s_b er t-fordelt med (n-2) frihetsgrader. Det er derfor enkelt å beregne et konfidensintervall, f.eks. på 95%. Dersom konfidensintervallet til b ikke dekker 0 (null) vet vi at *b* er signifikant forskjellig fra 0 (null) på 5% nivået (to-halet test). Tilsvarende kan vi regne ut om *b* er signifikant for hvilket som helst signifikansnivå ved å benytte t-tabellen. Dette behøves vanligvis ikke gjøres manuelt, siden alle de vanligste statistikkpakkene signifikanstester *b* for oss. Dersom *b* er signifikant forskjellig fra null, er også β også signifikant forskjellig fra null. T-testen for *b* er med andre ord også gyldig for β.

Uteliggere (outliers - enkeltskårer som har verdier som avviker ekstremt fra resten av fordelingen) er ofte et problem i regresjonsanalyse. Slike ekstreme skårer kan ha dramatiske virkninger på resultatene. Ofte kan ekstremt avvikende verdier skyldes kodefeil eller punchefeil. I så fall er problemet løst etter en kontroll av dataene. Dersom uteliggerne skyldes sterkt skjevfordelte variabler eller sjeldent forekommende kombinasjoner av verdier på variablene, kan problemet løses ved transformasjoner av variablene eller ved trunkeringer. Når en transformerer en variabel kan det f.eks. skje ved at en logaritmetransformerer skalaen og gir hver skåre en ny verdi. Effekten av en logaritme-transformasjon vil være å krympe høyre siden av en skala. En høyreskjev fordeling vil dermed få en kortere hale mot høyre. Logaritmetransformering av venstreskjeve fordelinger vil gjøre vondt verre; de blir bare enda skjevere. En venstreskjev skala må derfor først snues, så logaritmetransformeres, og eventuelt snues på nytt. Å trunkere en variabel kan bestå i å omkode alle verdier som er mer ekstreme enn et bestemt avvik fra gjennomsnittsverdien til et mindre ekstremt tall (f.eks. to standardavvik fra gjennomsnittet i positiv retning for uteliggere på positiv side av fordelingen og tilsvarende i negativ retning for negative uteliggere).

6.2 Multippel regresjonsanalyse

6.2.1 Hva kan en multippel regresjonsanalyse gjøre?

Multippel regresjonsanalyse er i likhet med bivariat regresjonsanalyse en statistisk teknikk som baserer seg på at variablene er metriske. I likhet med faktoranalysen tar den utgangspunkt i en matrise av produkt-moment-korrelasjoner mellom alle de variablene som skal analyseres.

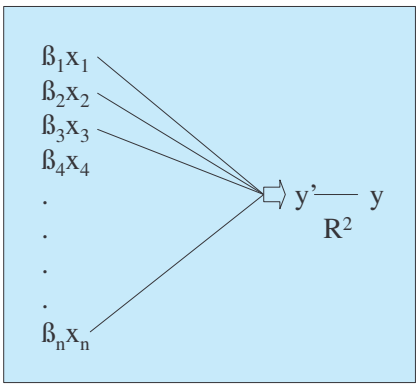
For at multippel regresjonsanalyse skal være til nytte, må en ha en situasjon med en "avhengig" variabel (y) og flere "uavhengige" (x_1-x_n) (Fig. 6.4). Multippel regresjonsanalyse kan vise hvordan flere prediktorer samtidig virker inn på en kriterievariabel. Dersom kriterievariabelen er et mål på psykiske plager blant arbeidstakere, kan prediktorene for eksempel være en serie mål på stress og belastninger i arbeidssituasjonen. Regresjonsanalysen kan i en slik situasjon vise følgende:

- Hvor sterk den samlede effekten av prediktorene er på kriterievariabelen (R^2)
- Hvor sterk er sammenhengen mellom hver enkelt av prediktorene og den avhengige variabelen når en har kontrollert for alle de andre prediktorene (beta).
- Regresjonsanalysen kan plukke ut en mindre gruppe prediktorer som i

kombinasjon med hverandre best predikerer den avhengige variabelen (trinnvis multippel regresjon)

- Å beregne verdier på kriterievariabelen ut fra verdiene på prediktorene.

Fig. 6.4: Multippel regresjonsanalyse: hovedelementer



6.2.2 Ulike koeffisienter

Multippel regresjonsanalyse beregner to sett av vekter: b-vekter og beta-vekter (stigningskoeffisient og standardisert stigningskoeffisient). b-vektene regnes ut på bakgrunn av kovarians-matrisen, mens beta-vektene beregnes på grunnlag av tilsvarende matrise av interkorrelasjoner. Denne siste framgangsmåten medfører at variablene allerede er standardisert og at en kan sammenlikne beta-vektene på tvers av prediktorvariabler. I fortsettelsen skal vi derfor holde oss til beta-vektene og ikke lenger omtale b-verdiene.

$$Y = \alpha_0 + ß_1X_1 + ß_2X_2 + ß_3X_3 + ... + ß_nX_n + e \qquad (6.11)$$

$$Y' = \alpha_0 + ß_1X_1 + ß_2X_2 + ß_3X_3 + ... + ß_nX_n \qquad (6.12)$$

Formelen for multippel lineær regresjon er gitt i 6.11. I stedet for en enkelt beta-verdi får vi nå en beta for hver prediktor. Disse beta-ene har en nyttig egenskap: Hver beta uttrykker en netto-sammenheng, med andre ord sammenhengen mellom en prediktor og den avhengige variabelen justert for effekten av alle de andre prediktorene. De kalles partielle standardiserte regresjonskoeffisienter eller partielle standardiserte stigningskoeffisienter. Ligning 6.12 er identisk med ligning 6.11, bortsett fra at vi har fjernet feiltermen eller residualen. På venstre siden har vi ikke lenger Y, men Y' som er den Y-verdien som predikeres ut fra

konstantleddet og den informasjonen som ligger i prediktorene og prediktorenes sammenheng med Y.

Kvadratet av den enkle korrelasjonen mellom Y' og Y er identisk med R^2 eller det vi tidligere har kalt determinasjonskoeffisienten. Også determinasjonskoeffisienten har fått en ny mening når vi har flere samtidige prediktorer. R^2 uttrykker hvor stor del av variansen på kriterievariabelen (Y) som kan forklares eller predikeres ut fra samtlige prediktorer (X_1-X_n) samtidig. Den kalles nå en multippel determinasjonskoeffisient (multippel R^2).

Vi har tidligere i denne teksten lært hva en interaksjonseffekt er for noe. Det er viktig å være klar over at multippel lineær regresjonsanalyse ikke oppdager eventuelle interaksjonseffekter i data. Multippel regresjonsanalyse forutsetter i utgangspunktet en enkel, additiv modell. Dersom en har hypoteser om interaksjoner mellom variabler, kan en undersøke om de finnes og eventuelt bygge inn slike i de analysene som skal gjøres.

Multippel lineær regresjonsanalyse greier fint å behandle korrelerte prediktorer. Dersom prediktorene ikke er korrelerte i det hele tatt, er vitsen med multippel regresjon borte. Dersom alle interkorrelasjonene mellom prediktorene var nøyaktig null, ville betavektene bli lik de enkle korrelasjonene. Etter hvert som interkorrelasjonene mellom prediktorene blir høyere, blir imidlertid estimatene (beregningene) av beta-verdiene mindre presise (Fox, 1993). Dersom dette går så langt at en prediktorvariabel kan predikeres perfekt ut fra en annen prediktorvariabel eller en bestemt gruppe prediktorvariabler, snakker en om multikollinearitet. Når det oppstår slik multikollinearitet, bryter analysen sammen, og det blir umulig å oppnå en løsning. I slike situasjoner må en velge blant høyt korrelerte prediktorer og utelate en eller flere av disse fra analysen. Alternativt kan en lage sumskårer eller indekser på grunnlag av høyt korrelerte prediktorer, og bruke sumskårene i stedet for de enkelte variablene.

Fox (1993) har sett nærmere på effektene av at prediktorene korrelerer. Han hevder at upresise estimater oftere er et resultat av mye feilvarians og små utvalg enn av kollinearitet. Han har beregnet at kollinearitetet må være temmelig sterk før den medfører sterk reduksjon av presisjonen i estimater. For at presisjonen for en bestemt prediktor skal bli halvert, må den multiple R mellom denne prediktoren og alle eller et subsett av de øvrige prediktorene være nærmere 0,90 (s.254).

Den vanligste nullhypotesen går ut på at regresjonskoeffisientene er lik null. En kan estimere standardfeilen til hver regresjonskoeffisient og denne er t-fordelt eller tilnærmet t-fordelt. En kan også beregne konfidensintervall.

Determinasjonskoeffisienten: R^2 er kvadratet av korrelasjonen mellom variablene når en bare har en prediktor. Dersom en har flere prediktorer, er R^2 kvadratet av korrelasjonen mellom den estimerte Y' og Y. Av alle mulige estimater av Y, beregner multippel regresjonsanalyse den Y' som gir høyest sammenheng med Y. Den justerte R^2 tar hensyn til at en analyserer et utvalg fra en populasjon, og at tilpasningen av regresjonslinjen i utvalget ofte blir kunstig god. For å signifikansteste R^2 bruker en variansanalyse. Den gir en F-verdi og et sett frihetsgrader som kan sammenliknes med kritiske verdier i en F-tabell.

Residualer: Avstanden mellom predikert verdi og observert verdi på Y-skalaen kalles residualen fra regresjonslinjen.

R^2 kan gis en annen fortolkning basert på residualene (eller feilene):

$$R^2 = 1 - \frac{SSE}{SST} \qquad (6.13)$$

R^2 Determinasjonskoeffisienten
SSE Kvadratsummen av feil
SST Total kvadratsum

I forbindelse med regresjonsanalyse møter en ofte på to andre typer korrelasjoner. Semipartiell - korrelasjon defineres som rota av endringen i R^2 når en variabel (i) fjernes fra modellen.

$$R^2_{Change} = R^2 - R^2_{(i)} \qquad (6.14)$$

R^2_{Change} Endring i multippel R kvadrert (semipartiell korrelasjon kvadrert)
R^2 R kvadrert for alle variablene samlet
$R^2_{(i)}$ R kvadrert når variabelen i er fjernet fra modellen

Partiell korrelasjon defineres ved hjelp av følgende formel:

$$R_P{}^2 = \frac{R^2 - R^2_{(i)}}{1 - R^2_{(i)}} \qquad (6.15)$$

$R_P{}^2{}_{Change}$ Partiell korrelasjon kvadrert
R^2 R kvadrert for alle variablene samlet
$R^2_{(i)}$ R kvadrert når variabelen i er fjernet fra modellen

Dette betyr at den partielle korrelasjonen alltid er lik eller større enn den semipartielle korrelasjonen.

Som presisert innledningsvis i dette kapittelet er multippel lineær regresjon en analyse som forutsetter at variablene er målt på metrisk nivå. I første kapittel nevnte vi at dikotomier (0/1-variabler) egentlig kan betraktes som intervallvariabler. I og med at en dikotom variabel bare inneholder ett intervall, tilfredsstiller den kravet om at alle intervall skal ha samme størrelse. Det er derfor ikke overraskende at dikotomier kan benyttes i regresjonsanalyse. Dikotomier benyttet som prediktorer ødelegger ikke regresjonsestimatenes egenskaper i det hele tatt. Dikotome variabler som er laget på grunnlag av kategorielle variabler med flere verdier, kalles i forbindelse med regresjonsanalyse dummyvariabler, og analysen kalles dummy regresjon.

Siden dummyvariabler kan brukes som prediktorer, kan egentlig alle slags variabler benyttes i regresjonsanalyse. Kategorielle variabler på nominalnivå kan nemlig alltid omkodes til dummyvariabler. En kategoriell variabel med k kategorier kan omkodes til k-1 dummyvariabler uten at en mister noe informasjon. Den ene av de k kategoriene må velges som referansekategori, og inngår alltid blant de kategoriene som gis tallverdien 0, mens de andre kategoriene en etter en gis tallverdien 1.

Tilsvarende kan en hvilken som helst ordinalvariabel omkodes til dummyvariabler dersom den er kategoriell med forholdsvis få kategorier. Dersom en ordinalvariabel har svært mange kategorier, kan en dele den inn i grovere kategorier og deretter omkode disse til dummy-variabler. På denne måten kan en risikere å miste en del variasjon, men samtidig blir det mulig å bruke et slikt sett av dummy-variabler i en multippel regresjonsanalyse.

6.2.3 Forutsetninger for bruk av multippel lineær regresjonsanalyse

Multippel lineær regresjonsanalyse bygger på en rekke forutsetninger og krav som må stilles til data. Lewis-Beck (1980) har presentert en oversikt. Den viser hvilke krav som må stilles til en regresjonsanalyse som skal gi riktige estimat av konstant og stigningskoeffisient når en ut fra et utvalg av subjekter skal si noe om tilsvarende parametre i populasjonen:

1. Spesifikasjon av modell:

    a) Relasjonen mellom X_i og Y_i må være lineær
    b) Ingen relevante prediktorer må være utelatt
    c) Ingen irrelevante prediktorer må være med

2. Målefeil:

    a) Variablene X_i og Y_i må være målt nøyaktig

3. Feiltermen:

    a) Den forventede verdien på feiltermen må være 0 (null)
    b) Feilvariansen må være den samme for alle verdier av X_i (homoskedastisitet)
    c) Feiltermene må være ukorrelerte (ingen autokorrelasjon)

d) Prediktoren må være ukorrelert med feiltermen
e) Feiltermen må være normalfordelt.

Når alle disse forutsetningene er oppfylt, er det mulig å beregne såkalte forventningsrette estimater av populasjonsparametre.

Alle som anvender regresjonsanalyse i sin forskning bør bekymre seg over i hvilken grad de bryter forutsetningene for analysen og hvilke konsekvenser dette har for resultatene de kommer fram til. Det eksisterer ingen konsensus på dette området. Noen statistikere mener en kan bryte forutsetningene i stor grad uten at det gjør noe særlig. Representanter for dette synet er f.eks. Kerlinger og Pedhazur (1973). Et motsatt standpunkt blir forfektet av Bibby (1977). Antakelig kommer en nærmest et levelig og fornuftig standpunkt ved å følge anbefalingene hos Lewis-Beck (1980). Han hevder f.eks. at forutsetningen om normalfordeling ikke er viktig dersom antall observasjoner er høyt. Derimot mener han at spesifikasjonen av modell er svært viktig. Dersom det gjøres feil her, vil det få alvorlige konsekvenser for gyldigheten av de resultatene en kommer fram til.

6.2.4 Multippel regresjonsanalyse – et eksempel

En forsker ved høgskolen i Lillehammer gjennomførte en datainnsamling blant 447 personer i to fylker (ett Østlands-fylke og ett Vestlands-fylke) der han blant annet forsøkte å måle følgende forhold:

- Livskvalitet (5 ledd med skala fra 0-6): Påstander av typen ”Jeg er tilfreds med livet mitt”
- Sosial støtte (5 ledd med ulike skalaer, sumskår fra 1-14): Bor sammen med partner, antall gode venner, noen som bryr seg og som en kan snakke med om personlige problemer, samvær med arbeidskolleger.
- Stedstilknytning (9 ledd med skala fra 0-4) – hvor sterk tilknytning føler en til stedet der en bor
- Landskapspreferanser (vurdering av 8 landskapsbilder på skala fra 0-4). Denne variabelen må tolkes som et generelt uttrykk for hvor glad en er i naturen.

Fra tidligere forskning er det et velkjent funn at høy grad av sosial støtte henger sammen med høy livskvalitet. Forskeren var interessert i å finne ut i hvilken grad også variabler som stedstilknytning og landskapspreferanser kan tenkes å bidra til høy livskvalitet. En slik antakelse er nokså lett å finne støtte for i miljøpsykologiske undersøkelser. Men for å vite dette sikkert, er det nødvendig å kontrollere for sosial støtte. Det kunne nemlig tenkes at de som i stor grad skårer høyt på stedstilknytning og landskapspreferanser også skårer høyt på sosial støtte. I så fall kunne det tenkes at sammenhengen mellom landskapspreferanser og stedstilknytning på den ene siden og tilfredshet med livet på den andre kunne forklares av sosial støtte. I en slik situasjon er det nyttig med multippel regresjonsanalyse.

Aller først inspiserer vi korrelasjonene mellom alle de variablene som inngår i analysen. En slik korrelasjonsmatrise er vist i Fig. 6.5. Vi ser der at sammenhengen mellom sosial støtte og tilfredshet med livet er ganske høy, nemlig 0,44 (avrundet til to desimaler). Og pussig nok er

sammenhengen mellom stedstilknytning og tilfredshet med livet nøyaktig like stor som sammenhengen mellom landskapspreferanser og tilfredshet med livet, nemlig 0,23. Det er interessant å legge merke til at de tre prediktorvariablene er positivt korrelert, men at korrelasjonene ikke er svært høye.

Fig. 6.5: Korrelasjoner mellom de fire variablene som skal inngå i den multiple regresjonsanalysen

Correlations

		v1 Tilfredshet med tilværelsen	v2 Sosial støtte	v3 Stedstilknytning	v4 Landskaps preferanser
v1 Tilfredshet med tilværelsen	Pearson Correlation	1	,442**	,233**	,233**
	Sig. (2-tailed)		,000	,000	,000
	N	447	447	446	447
v2 Sosial støtte	Pearson Correlation	,442**	1	,130**	,139**
	Sig. (2-tailed)	,000		,006	,003
	N	447	447	446	447
v3 Stedstilknytning	Pearson Correlation	,233**	,130**	1	,240**
	Sig. (2-tailed)	,000	,006		,000
	N	446	446	446	446
v4 Landskapspreferanser	Pearson Correlation	,233**	,139**	,240**	1
	Sig. (2-tailed)	,000	,003	,000	
	N	447	447	446	447

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

Dersom korrelasjonene mellom prediktorene hadde vært svært høye (kanskje omtrent så store som 0,70-0,90), kunne vi ikke benyttet regresjonsanalyse på disse dataene. Vi kunne få problemer med multikollinearitet, som vi har beskrevet tidligere i dette kapittelet. Multikollinearitet oppstår når en av prediktorene kan predikeres perfekt ut fra en eller flere av de andre. Selv om det bare oppstår tilnærmet multikollinearitet ødelegger det for mulighetene til å bruke multippel regresjon. Men i vårt eksempel er altså ikke dette noe problem.

Regresjonsanalysen kjører vi i SPSS ved å gå inn i *Analyze*, deretter *Regression* og så *Linear*. Deretter legge en inn den avhengige variabelen i boksen som heter *Dependent* og de uavhengige i *Independent(s)*. Velg *Enter* under *Method* og *Estimates* og *Model fit* under *Statistics*. La øvrige verdier være slik de er stilt inn fra før (default).

Når du skal kjøre en trinnvis regresjonsanalyse må du trykke på "Next" og legge inn nye prediktorer under *Independent(s)*. Du må også sette hake i boksen for *R squared change* i *Statistics*.

Det at prediktorene er moderat korrelerte, gjør for så vidt analysen interessant. Dersom de ikke var korrelerte i det hele tatt, ville det ikke være nødvendig å kjøre noen regresjonsanalyse, for alle regresjonskoeffisientene ville da bli identiske med de bivariate korrelasjonene. De ville alle sammen beholde hele sin prediksjonskraft. Slik mønsteret av korrelasjoner mellom prediktorene ser ut nå, er det spennende å se om stedstilknytning og landskapspreferanser beholder noe av sin sammenheng med tilfredshet med livet.

I den første regresjonsanalysen vi presenterer har vi tatt med alle tre prediktorene på en gang. Resultatene er gjengitt i Fig. 6.6. I første deltabell (Model Summary) finner vi statistikk som viser hvor god modell vi har. Det viser seg at den multiple R^2 er lik 0,24. Det vil si at 24 prosent av variansen i "trives med tilværelsen" blir forklart av de tre prediktorene. Det er en nokså sterk sammenheng når en tar i betraktning av vi her har en modell med bare tre prediktorer og dessuten er klar over hvor mange andre forhold som kan tenkes å ha betydning.

Fig. 6.6: Trives med livet etter sosial støtte, stedstilknytning og landskapspreferanser (multippel lineær regresjonsanalyse)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,495[a]	,245	,240	,88277

a. Predictors: (Constant), v4 Landskapspreferanser, v2 Sosial støtte , v3 Stedstilknytning

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	111,932	3	37,311	47,879	,000[a]
	Residual	344,440	442	,779		
	Total	456,372	445			

a. Predictors: (Constant), v4 Landskapspreferanser, v2 Sosial støtte , v3 Stedstilknytning

b. Dependent Variable: v1 Tilfredshet med tilværelsen

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1,360	,237		5,730	,000
	v2 Sosial støtte	,154	,016	,403	9,605	,000
	v3 Stedstilknytning	,201	,058	,147	3,436	,001
	v4 Landskapspreferanser	,233	,070	,142	3,315	,001

a. Dependent Variable: v1 Tilfredshet med tilværelsen

I den andre del-tabellen (ANOVA) er det gjort en signifikanstesting av den totale modellen. Siden det egentlig dreier seg om å signifikansteste hvor mye varians modellen har forklart, benyttes en form for variansanalyse. En F-verdi på 47,879 med 3 og 442 frihetsgrader viser seg å være signifikant på p<0,001-nivået. Det er med andre ord ingen tvil om at de tre prediktorene samlet forklarer så mye varians i kriterievariabelen at vi kan forkaste nullhypotesen om at den multiple R^2 er lik null.

I tredje deltabell i Fig 6.6 vises det som kalles beta-vekter eller standardiserte regresjonskoeffisienter. Disse viser hvor sterkt hver enkelt prediktor predikerer kriterievariabelen når en har justert for de to andre prediktorene. Det viser seg at sosial støtte har mest å bety. Beta-vekten er her på 0,40. Den er signifikant på p<0,001-nivået. Vi ser at testen er utført ved bruk av t-fordelingen. Sosial støtte har mistet noe av sin prediksjonsevne.

Det ser vi når vi sammenligner beta-koeffisienten med korrelasjonen mellom sosial støtte og "tilfredshet med tilværelsen" gjengitt i Fig. 6.5. Korrelasjonen (eller den bivariate regresjonskoeffisienten) var her 0,44, men har altså sunket til 0,40. De to andre prediktorene betyr langt mindre enn sosial støtte, og har oppnådd betavekter på 0,15 (stedstilknytning) og 0,14 (landskapspreferanser). Begge er likevel statistisk signifikante, noe som tyder på at de begge bidrar unikt til å predikere tilfredshet med tilværelsen.

Til venstre for beta-koeffisientene finner vi de ustandardiserte regresjonskoeffisientene. Disse er i denne sammenhengen mindre interessante, for størrelsen på disse avhenger av hvilken skala en har benyttet. I noen sammenhenger gir det likevel mening å bruke de ustandardiserte koeffisientene, og i noen tilfeller er det nødvendig å bruke disse i stedet for de standardiserte.

Ikke alle forskere vil være fornøyde med bare å se på endringer i beta-koeffisientene når en har med flere samtidige prediktorer i modellen. Noen vil si at det i tillegg er viktig å se hvor mye den forklarte variansen øker når vi tar inn de to miljøpsykologiske variablene i tillegg til sosial støtte. Vi har derfor også gjort det som kalles en blokkvis multippel regresjonsanalyse med de samme variablene. Resultatene er gjengitt i Fig. 6.7 nedenfor.

Det som i analysen kalles modell 1 er en analyse der bare sosial støtte inngår som prediktor. Det er med andre ord en bivariat lineær regresjonsanalyse. Under modell 1 i første deltabell ser vi at R er lik 0,442. Dette er identisk med den enkle korrelasjonen mellom sosial støtte og tilfredshet med tilværelsen som vi regnet ut i korrelasjonsmatrisen som er gjengitt i Fig. 6.5. Under modell 1 i den andre deltabellen ser vi at sammenhengen er signifikant ($F = 107,534$; f.g. = 1 og 444; $p < 0,001$). I tredje deltabell ser vi at beta-koeffisienten også er lik den enkle korrelasjonen mellom sosial støtte og trivsel i tilværelsen, nemlig 0,442. I en bivariat regresjonsanalyse er med andre ord korrelasjonskoeffisienten, den standardiserte regresjonskoeffisienten og den multiple R identiske.

Den kanskje mest interessante informasjonen får vi i første deltabell under modell 2. Her ser vi at den multiple R har økt fra 0,442 til 0,495. Den multiple R^2 har økt fra 0,195 til 0,245. Dette betyr at forklart varians har økt med 5% når vi i tillegg til sosial støtte også tar med de to miljøpsykologiske prediktorene. Økningen i forklart varians (eller forbedringen av

modellen) er signifikanstestet (F = 14,727; f.g. = 2 og 442; p < 0,001). Dette betyr at de to miljøpsykologiske prediktorene har bidratt til en signifikant forbedring av modellen. Vi kan med andre ord forkaste nullhypotesen om at de ikke bidrar til modellen i det hele tatt. I deltabellene 2 og 3 finner vi informasjon som vi allerede har sett tidligere.

Hva kan vi så konkludere med? Den aller viktiste konklusjonen fra analysene ovenfor er at de to miljøpsykologiske prediktorene (stedstilknytning og landskapspreferanser) bidrar til å forklare trivsel med tilværelsen ut over det vi fikk til ved å bruke et sammensatt mål på sosial støtte som prediktor. Vi må imidlertid ta en del forbehold. For det første kan det tenkes at vi ikke har med gode nok mål på sosial støtte. Dersom vi kunne måle sosial støtte på en riktigere, mer dekkende eller mer presis måte, kan det tenkes at sammenhengene mellom de miljøpsykologiske prediktorene og trivsel med tilværelsen ville forsvinne helt. Men det kan også tenkes at vi kunne forbedre våre målinger av de to miljøpsykologiske variablene, og dermed enda klarere vise at de har noe å bidra med ut over det som kan forklares ut fra sosial støtte.

Fig. 6.7: Trives med livet etter sosial støtte, stedstilknytning og landskapspreferanser (blokkvis multippel lineær regresjonsanalyse)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,442[a]	,195	,193	,90965	,195	107,534	1	444	,000
2	,495[b]	,245	,240	,88277	,050	14,727	2	442	,000

a. Predictors: (Constant), v2 Sosial støtte

b. Predictors: (Constant), v2 Sosial støtte , v3 Stedstilknytning, v4 Landskapspreferanser

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	88,980	1	88,980	107,534	,000[a]
	Residual	367,393	444	,827		
	Total	456,372	445			
2	Regression	111,932	3	37,311	47,879	,000[b]
	Residual	344,440	442	,779		
	Total	456,372	445			

a. Predictors: (Constant), v2 Sosial støtte

b. Predictors: (Constant), v2 Sosial støtte , v3 Stedstilknytning, v4 Landskapspreferanser

c. Dependent Variable: v1 Tilfredshet med tilværelsen

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2,301	,166		13,867	,000
	v2 Sosial støtte	,169	,016	,442	10,370	,000
2	(Constant)	1,360	,237		5,730	,000
	v2 Sosial støtte	,154	,016	,403	9,605	,000
	v3 Stedstilknytning	,201	,058	,147	3,436	,001
	v4 Landskapspreferanser	,233	,070	,142	3,315	,001

a. Dependent Variable: v1 Tilfredshet med tilværelsen

Noen ganger vil en se at multiple lineære regresjonsanalyser brukes til å bygge noe som kalles sti-modeller. Når regresjonsanalyser brukes slik, kaller en det gjerne for sti-analyse (path analysis). Ved å kjøre flere forskjellige regresjonsmodeller med ulike kombinasjoner av prediktorer, kan en bygge mer kompliserte modeller som også inneholder mediator-variabler. En mer moderne variant er bruken av strukturelle ligningsmodeller, som kan gjøres ved bruk av programpakker som LISREL, EQS, AMOS og MPLUS. Slik programvare forenkler dataanalysene betraktelig, og tillater dessuten uttesting av mer komplekse modeller. De nevnte programpakkene kan også brukes til å analysere relasjoner mellom observerte og latente variabler samt mellom latente variabler innbyrdes. Kompetent bruk av multippel lineær regresjonsanalyse gir et godt utgangspunkt for å sette seg inn i disse mer avanserte statistiske teknikkene.

6.2.5 Analyse av interaksjoner i multippel regresjon

Innen forskningen om faktorer som har betydning for arbeidstakeres helse har en funnet at både de krav som stilles til arbeidstakerne og arbeidstakernes grad av kontroll over egen arbeidssituasjon er viktige prediktorer. Jo høyere krav, og jo lavere kontroll, desto dårligere helse. Karasek og Theorell (1990; Theorell, 2000) har imidlertid lansert en teori om at det også eksisterer en interaksjonseffekt. Dersom det både stilles høye krav og en samtidig har liten grad av kontroll, er den helsemessige effekten enda mer negativ enn det en skulle tro ut fra den effekt hver av variablene separat har på arbeidstakernes helse.

For å teste dette har vi sett nærmere på et datasett som er samlet inn ved en større bedrift. Materialet er stort. Antall personer som har deltatt i undersøkelsen er 3180 personer. Følgende variabler inngår i analysene vi skal presentere:

- Den avhengige variabelen er utmattelse i jobben. Fem ledd inngår, for eksempel følgende: ”Jeg er ofte motløs og tenker på å slutte i jobben”. Svarkategoriene går fra ”Svært uenig” til ”Svært enig” og er kodet med tall fra 1 til 6. Vi konstruerer en såkalt gjennomsnittsindeks (meanscore). Det betyr at vi legger sammen skårene på de

fem leddene og deler på antall ledd. Dermed får også den nye variabelen vi konstruerer verdier fra 1 til 6.

- Krav er målt ved hjelp av ti enkeltledd, for eksempel "Krever arbeidet ditt raske avgjørelser?" og de fem svarkategoriene går fra "Meget sjelden eller aldri" til "Meget ofte eller alltid". Igjen lager vi en meanscore som går fra 1 til 5.
- Kontroll er målt ved hjelp av åtte ledd, og svarkategoriene var de samme som for krav. Eksempel på et ledd som inngikk i skalaen var dette "Kan du selv bestemme ditt arbeidstempo".

Alle de tre skalaene viste tilfredsstillende indre konsistens. Ingen av gjennomsnittsindeksene var ekstremt avvikende fra normalfordeling (ikke ekstrem spisshet eller ekstrem skjevhet).

Vi kjører først en multippel regresjonsanalyse der vi tar inn utmattelse i jobben som avhengig variabel og krav og kontroll som prediktorer. Resultatene er gjengitt i Fig. 6.8.

Fig. 6.8: Utmattelse i jobben etter krav og kontroll.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,354[a]	,125	,125	1,07248

a. Predictors: (Constant), Control, Demands

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	508,834	2	254,417	221,193	,000[a]
	Residual	3558,733	3094	1,150		
	Total	4067,568	3096			

a. Predictors: (Constant), Control, Demands

b. Dependent Variable: Exhaustion

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2,475	,159		15,575	,000
	Demands	,463	,040	,196	11,666	,000
	Control	-,485	,028	-,292	-17,354	,000

a. Dependent Variable: Exhaustion

Det viser seg at de to prediktorene forklarer 12,5% av variansen i Utmattelse (R square = .125). Videre viser variansanalysen at modellen er signifikant på 0,1%-nivået (p<.001). De standardiserte regresjonskoeffisientene er 0,196 for Krav og -0,292 for Kontroll, og begge er signifikante. Vi ser at de ustandardiserte koeffisientene er omtrent like store (bortsett fra fortegnet). Dersom vi hadde regnet ut et 95 prosents konfidensintervall for de ustandardiserte regresjonskoeffisientene, ville vi sett at de overlappet sterkt. Vi kan derfor ikke uten videre trekke den konklusjon at Kontroll betyr mer enn Krav. I dette tilfellet gir det mest mening å sammenlikne de ustandardiserte koeffisientene fordi de er basert på bruk av samme skala med like kategorinavn og samme koding (1-5). En enhets endring på kravskalaen gir omtrent samme effekt på utmattelse som en enhets endring på kontrollskalaen. Beta-koeffisientene påvirkes sterkt av hvor stor spredningen er på variablene, og siden det var langt mindre spredning i svarene på kravskalaen, blir den standardiserte regresjonskoeffisienten for Krav tilsvarende mye mindre enn for Kontroll.

For å analysere interaksjonen mellom Krav og Kontroll er det nødvendig å konstruere et interaksjonsledd. Først må vi imidlertid standardisere begge de uavhengige variablene (trekke fra gjennomsnittet og dele på standardavviket for hver observasjon). Deretter kan vi gange de to standardiserte variablene med hverandre (Cohen et al. 2003). Produktet er interaksjonsleddet. Deretter kjører vi en ny regresjonsanalyse. Denne gangen legger vi inn de to enkle prediktorene som et første trinn (blokk 1). Deretter legger vi inn interaksjonsleddet som et nytt trinn (blokk 2).

Resultatene er gjengitt i Fig. 6.9. Vi ser at den første blokken, selv om vi nå analyserer variabler som er standardiserte, gir nøyaktig like mye forklart varians. Variansanalysen gir nøyaktig samme F-verdi og de standardiserte regresjonskoeffisientene er identiske med de vi fikk i den første analysen. Standardiseringen har altså ingen innvirkning på første trinn. I andre trinn la vi inn interaksjonsleddet. Vi ser at forklart varians (R square adjusted) øker fra 12,5 til 13,5%. Den nye modellen viser seg å være signifikant bedre enn den som bare består av hovedeffektene (altså ingen interaksjonseffekt) (F change = 37,515; d.f. = 1 og 3093; p<0,001). Og i den nederste delen av tabellen ser vi at regresjonskoeffisientene endrer seg lite. Interaksjonsleddet oppnår en beta-koeffisient på -0,103.

Vi har vist at det foreligger en statistisk signifikant interaksjon. Om denne interaksjonen er sterk nok til å være interessant, er en annen sak. En økning i forklart varians på bare 15 virker ikke veldig overbevisende. For å undersøke nærmere hvor sterk interaksjonen egentlig er, kan en for eksempel dele inn de to prediktorene i kategorier og se hvordan gjennomsnittsskåren på den avhengige variabelen ser ut på tvers av kategorier.

Vi har kategorisert både Krav og Kontroll i fire, omtrent lite store kategorier. Ved å krysse disse to variablene, har vi til sammen 16 undergrupper. Vi har regnet ut gjennomstnittlig skår på Utmattelse i hver av de 16 undergruppene, og resultatet av denne anlysen er vist i Fig. 6.10. Vi ser der at sammenhengen mellom Krav og Utmattelse er nokså svak for de to gruppene som har høyest skår på Kontroll. Sammenhengen er noe sterkere for gruppen som skårer nest lavest på Kontroll. Den absolutt sterkeste sammenhengen finner vi i den fjerde gruppen, de som skårer lavest på Kontroll. Kontrasten blant de med høyest skår på Kontroll er 0,19, er kontrasten blant de med lavest skår på Kontroll 1,10, altså nesten 6 ganger så stor.

Fig. 6.9: Utmattelse i jobben etter krav og kontroll samt interaksjonen mellom disse.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,354[a]	,125	,125	1,07248	,125	221,193	2	3094	,000
2	,368[b]	,136	,135	1,06620	,010	37,515	1	3093	,000

a. Predictors: (Constant), ZControl Zscore(Control), ZDemands Zscore(Demands)

b. Predictors: (Constant), ZControl Zscore(Control), ZDemands Zscore(Demands), Dem_Contr

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	508,834	2	254,417	221,193	,000[a]
	Residual	3558,733	3094	1,150		
	Total	4067,568	3096			
2	Regression	551,481	3	183,827	161,707	,000[b]
	Residual	3516,087	3093	1,137		
	Total	4067,568	3096			

a. Predictors: (Constant), ZControl Zscore(Control), ZDemands Zscore(Demands)

b. Predictors: (Constant), ZControl Zscore(Control), ZDemands Zscore(Demands), Dem_Contr

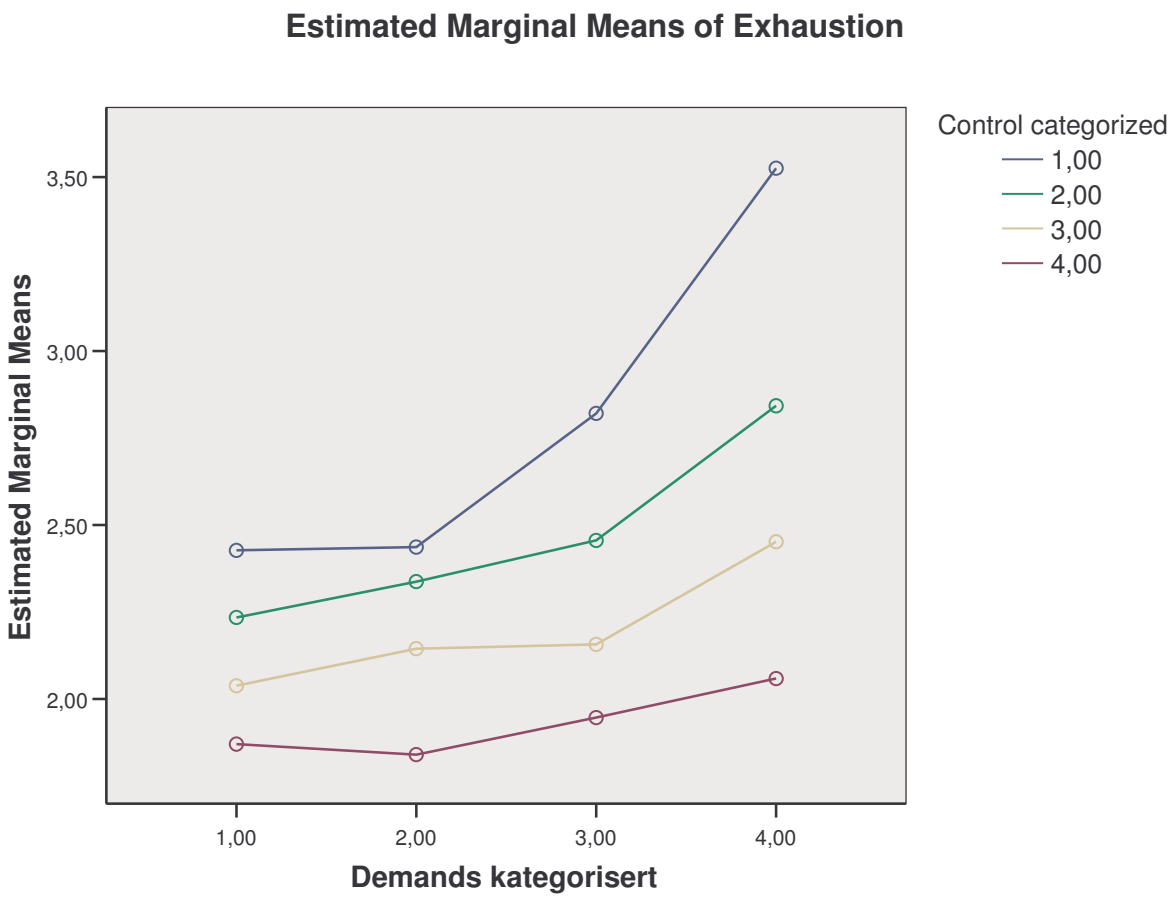
c. Dependent Variable: Exhaustion

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2,351	,019		121,985	,000
	ZDemands Zscore(Demands)	,226	,019	,196	11,666	,000
	ZControl Zscore(Control)	-,335	,019	-,292	-17,354	,000
2	(Constant)	2,350	,019		122,645	,000
	ZDemands Zscore(Demands)	,217	,019	,188	11,215	,000
	ZControl Zscore(Control)	-,334	,019	-,291	-17,428	,000
	Dem_Contr	-,108	,018	-,103	-6,125	,000

a. Dependent Variable: Exhaustion

Fig. 6.10: Utmattelse etter krav og kontoll etter kategorisering av de to prediktorene



Vi kan føye til at resultatene i figuren som er presentert ovenfor kan analyseres ved bruk av toveis variansanalyse. Leseren kan kanskje lure på hvorfor vi ikke allerede i utgangspunktet kunne gjøre analysen på en såpass enkel måte. Svaret er at en slik analyse kan være vanskelig å få til dersom antall observasjoner er lavere, og med en grovere kategorisering kan en miste altfor mye varians. Cohen og medarbeidere (2003) argumenterer sterkt for verdien av å analysere på de opprinnelige variablene i stedet for å kategorisere prediktorene. Et annet poeng er at en sjelden er interessert i analysere med så få prediktorer som her. Når antallet prediktorer blir stort, er regresjonsmodeller mer anvendelige enn variansanalyse.

Analysene har altså bekreftet Karasek og Theorells hypotese om interaksjonen mellom krav og kontroll i arbeideidet. Det er interessant å legge merke til at en økning i forklart varians på bare 1% dekker over en interaksjon som er ganske konsistent på tvers av de kategoriene vi har undersøkt og som dessuten er ganske sterk når vi sammenlikner yttergruppene. Det er lett å neglisjere en interaksjonseffekt som gir en økning i forklart varians på bare 1%. En viktig lærdom vi kan trekke av den analysen vi har gjort her, er at når en oppdager signifikante interaksjoner, bør en gjøre tilleggsanalyser for å forsøke å finne ut hvor sterk interaksjonen

egentlig er. Det er også viktig å huske på at vi har analysert med variabler som slett ikke har en perfekt reliabilitet. Det at reliabiliteten er lavere enn 1,00 på både den avhengige variabelen og på de to prediktorene bidrar sterkt til å redusere beta-koeffisienten til interaksjonsleddet.

Når en bare analyserer hovedeffekter, er det uvesentlig om en standardiserer prediktorene på forhånd. Når en skal legge inn interaksjonsledd, er det derimot viktig å standardisere de prediktorene som skal brukes i interaksjonsleddene. Uten slik standardisering vil en se at interaksjonsleddene korrelerer høyt med de opprinnelige variablene. Resultatet kan bli problemer med multikollinearitet. Et annet problem er at regresjonsvektene til hovedeffektene endrer seg kraftig og ofte blir vanskeligere å fortolke. Cohen og medarbeidere (2003) tar opp denne problematikken nokså grundig, og beskriver konsekvensene av det å ikke standardisere variablene før en konstruerer interaksjonsledd. De beskriver også situasjoner der en kan gjøre unntak fra denne regelen. I regresjonsanalyser med interaksjonsledd er det ikke nødvendig å standardisere den avhengige variabelen.

En kan også analysere interaksjoner der den ene variabelen er en dikotomi. Et typisk eksempel er kjønn. I slike tilfeller må en kode den ene kategorien som 0 og den andre som 1. En bør også her standardisere øvrige prediktorer før en lager interaksjonsledd.

6.3 Multippel logistisk regresjon

Multippel logistisk regresjon er ikke noen ny statistisk teknikk, men har levd en nokså anonym tilværelse i miljøer av statistikere inntil den ble anvendt av Truett, Cornfield & Kannel (1967) til analyse av data fra det kjente Framingham Heart Study. Gjennom en noe nyere innføringstekst (Hosmer & Lemeshow, 1989) er logistisk regresjon blitt tilgjengelig for et stort forsker-publikum. Beskrivelsen her vil særlig baseres på Hosmer & Lemeshows bok, men også på Scott Menards innføring i anvendt logistisk regresjonsanalyse fra 1995 (Menard, 1995).

6.3.1 Odds, betinget odds og odds ratio

Grunnlaget for å forstå logistisk regresjon er begrepet "odds ratio". En odds er ganske enkelt forholdstallet mellom to tall (Knoke & Burke, 1980). Dersom vi har gjennomført en survey blant studenter ved Universitetet i Bergen og funnet at det i utvalget til sammen er 400 studenter som rapporterer at de drikker alkohol ukentlig og 200 studenter som ikke drikker alkohol ukentlig, kan vi regne ut forholdstallet mellom disse to gruppene: Oddsen er da lik antall som drikker alkohol ukentlig delt med antall som ikke drikker alkohol ukentlig, nemlig 400/200=2,0. Dersom antallet hadde vært like stort i begge grupper (300/300), ville OR blitt 1,0.

Vi kan også regne ut oddsen ved å ta forholdstallet mellom de som ikke drikker alkohol ukentlig og de som drikker alkohol ukentlig. I så fall ville vi fått OR=200/400=0,50. Med andre ord er en odds på 2,0 uttrykk for en like sterk skjevfordeling som en odds på 0,5.

Tilsvarende er en odds på 4 uttrykk for en like sterk skjevfordeling som en odds på 0,25. Legg merke til at den inverse av 2 er ½ = 0,5. Den inverse av 4 er ¼ = 0,25.

Dersom en splitter opp et materiale, f.eks. i menn og kvinner, og beregner en odds for bare menn, kalles dette en betinget odds (conditional odds).

Ved hjelp av odds kan en også regne ut styrken på sammenhenger mellom variabler. La oss tenke oss at det i utvalget var med både menn og kvinner, og at vi har fått et resultat som i tabell 6.5. Fra denne tabellen kan vi regne ut betingede odds (antall som drikker alkohol ukentlig dividert med antall som ikke drikker alkohol så ofte som ukentlig) for menn og kvinner separat:

$\text{Odds}_{\text{menn}} = 230/70 = 3{,}29$

$\text{Odds}_{\text{kvinner}} = 170/130 = 1{,}31$

Ved å dele den ene betingede oddsen på den andre får vi en odds ratio:

$$OR_{kjønn*alkoholbruk} = \frac{Odds_{menn}}{Odds_{kvinner}} = \frac{3,29}{1,31} = 2,51 \qquad (6.16)$$

$OR_{kjønn*alkoholbruk}$	Odds ratio-verdi for sammenhengen mellom kjønn og alkoholbruk
$Odds_{menn}$	Oddsen for at menn drikker alkohol
$Odds_{kvinner}$	Oddsen for at kvinner drikker alkohol

En kan fortsette ved å regne ut odds ratio-verdier på høyere nivå. Dersom vi tenker oss at vi har gjennomført samme datainnsamling blant studenter i Oslo, og at vi der har fått en $\text{OR}_{\text{kjønn x alkoholbruk}}$ på bare 1,75, kan vi regne ut et forholdstall mellom disse forholdstallene:

$$OR_{kjønn*alkoholbruk*by} = \frac{OR_{kjønn*alkoholbruk(Bergen)}}{OR_{kjønn*alkoholbruk(Oslo)}} = \frac{2,51}{1,75} = 1,43 \qquad (6.17)$$

Dette betyr med andre ord at sammenhengen mellom kjønn og alkoholbruk er sterkere i Bergen enn i Oslo. Denne forskjellen kan også signifikanstestes, selv om vi ikke skal bruke plass på å beskrive det her.

En kan fortsette ved å bygge høyere ordens interaksjoner inntil det ikke lenger gir noen mening eller inntil dataene setter begrensninger. Hele veien er det slik at en odds ratio på 1,0 er en slags null-verdi. En odds ratio mellom to frekvenser på en variabel som er lik 1,0 forteller at frekvensene er like store. En odds ratio mellom to variabler forteller når den blir

1,0 at det ikke er noen sammenheng mellom variablene. En odds ratio på 1,0 mellom to bivariate sammenhenger forteller at det ikke foreligger noen interaksjonseffekt. På en måte kan vi si at en odds ratio på 1,0 tilsvarer en korrelasjonskoeffisient på 0,0. Dette kan virke uvant i begynnelsen, men siden odds ratio er en lett fortolkbar statistisk størrelse, går det fort å bli fortrolig med den.

Tabell 6.5: Alkoholbruk blant studenter ved Universitetet i Bergen. Et hypotetisk eksempel

	Bruker alkohol:		
	Ukentlig	Sjeldnere	Til sammen
	n	n	n
Menn	230	70	300
Kvinner	170	130	300
Alle	400	200	600

Vi var i kapittel 2 inne på at ett av de største problemene med assosiasjonsmål for krysstabeller er at de ofte påvirkes av marginalfordelingene på variablene. Dette gjør det vanskelig å sammenlikne slike assosiasjonsmål på tvers av undersøkelser og mellom ulike undergrupper i en og samme undersøkelse. Dersom en forsker har langt flere menn enn kvinner med i en undersøkelse og en annen har omtrent like mange menn og kvinner med, og begge regner ut Phi for sammenhengen mellom kjønn og en annen variabel, kan tallene ikke sammenliknes.

På dette punktet har odds ratio et stort fortrinn. Odds ratio er uavhengig av marginalfordelingene. Dette illustreres gjennom konstruerte eksempler i tabell 6.4. Når vi fra eksempel 1 til eksempel 2 reduserer antall kvinner som inngår til halvparten, forblir odds ratio den samme. Når vi deretter halverer antall ikke-røykere som er med i tabellen, får vi heller ingen endring i odds ratio.

Tabell 6.4: Odds ratio med varierende marginalfordelinger

Eksempel 1

	Røyker daglig?		
	ja	nei	Til sammen
Menn	40	60	100
Kvinner	70	30	100
Totalt	110	90	200

$Odds_{menn}$ = 0,677
$Odds_{kvinner}$ = 2,333
$OR_{kvinner*menn}$ = 3,50

Eksempel 2

	Røyker daglig?		
	ja	nei	Til sammen
Menn	20	30	50
Kvinner	70	30	100
Totalt	90	60	150

$Odds_{menn}$ = 0,677
$Odds_{kvinner}$ = 2,333
$OR_{kvinner*menn}$ = 3,50

Eksempel 3

	Røyker daglig?		
	ja	nei	Til sammen
Menn	20	15	35
Kvinner	70	15	85
Totalt	90	30	120

$Odds_{menn}$ = 1,333
$Odds_{kvinner}$ = 4,667
$OR_{kvinner*menn}$ = 3,50

Statistikeren Yule lanserte allerede i 1900 og 1912 assosiasjonsmål som er nært beslektet med odds ratio og kan regnes ut på grunnlag av odds ratio-verdier (Reynolds, 1977; Liebtrau, 1983). Det mest brukte av disse er i dag det som heter Yule's Q. Formelen for Yule's Q er gitt i formel 6.18.

$$Q = \frac{OR - 1}{OR + 1} \qquad (6.18)$$

Q Yules Q

OR Odds ratio-verdien for en firefelts (2x2) tabell

Dette assosiasjonsmålet varierer mellom 0,0 og 1,0, og det påvirkes ikke av endrede marginalfordelinger.

6.3.2 Partiell odds ratio og multippel logistisk regresjonsanalyse

Et viktig begrep i multippel logistisk regresjon er partiell odds ratio. En partiell odds ratio er en odds ratio som er regnet ut under forutsetning av at en eller flere såkalte "tredjevariabler" holdes konstante. En partiell odds ratio kan sammenliknes med betavektene i en multippel lineær regresjonsanalyse. Betavekten mellom en prediktor og en kriterievariabel er justert for de andre prediktorene. En partiell odds ratio i en multippel logistisk regresjon er på tilsvarende vis regnet ut med justering for de andre prediktorene.

En partiell odds ratio kan signifikanstestes. Til dette brukes gjerne Walds test som er basert på chi-kvadrat-fordelingen. Nullhypotesen er at odds ratio er lik 1,0. En signifikant odds ratio vil si at den er signifikant forskjellig fra 1,0. Den kan være lavere enn 1,0 (mellom 0,0 og 1,0) eller høyere enn 1,0 (mellom 1,0 og uendelig).

En kan også beregne konfidensintervallet til en odds ratio. Det vil som regel være et sammenfall mellom signifikanstestingen og konfidensintervallet. Et signifikansnivå på 5% vil tilsvare et konfidensintervall på 95%. Når konfidensintervallet ikke omfatter verdien 1,0, er odds ratio signifikant forskjellig fra 1,0.

I tabell 6.5 ser vi eksempel på resultater fra en multippel logistisk regresjonsanalyse med to prediktorer. Avhengig variabel er en dikotomi. Ett enkelt spørsmål om en føler seg lykkelig er omkodet slik at utvalget er delt i to grupper. De som svarer at de ikke føler seg lykkelige, og de som svarer at de føler seg ulykkelige er slått sammen til en kategori som har fått koden 1. Alle andre har fått skåren 0 (null). Variabelen er nokså skjevfordelt med såpass lite som 6,9 prosent i den gruppen som i minst grad rapporterer at de føler seg lykkelige.

Tabell 6.5: Opplevelse av å være lykkelig etter skoletrivsel og kommunikasjon med mor (Logistisk regresjonsanalyse) (n = 1574)

Variabel	B	S.E.	Walds test (chi-kv.)	fg.	Sig.	Exp(B) O.R.
Trives på skolen:			55.6354	3	.0000	
Svært godt						1.0000
Godt	.7178	.4196	2.9267	1	.0871	2.0499
Ikke særl.	2.0120	.4214	22.7968	1	.0000	7.4779
Dårlig	2.3773	.4641	26.2375	1	.0000	10.7763
Vanskelig for å snakke med mor:			37.8368	3	.0000	
Svært lett						1.0000
Lett	-.4443	.3017	2.1680	1	.1409	.6413
Vanskelig	.5101	.2908	3.0778	1	.0794	1.6655
Svært v.	1.4071	.3216	19.1432	1	.0000	4.0840

Analysen viser at begge de to prediktorene slår ut signifikant. Det ser vi av den såkalte Walds test. Den er chi kvadrat-fordelt og det tilhørende antall frihetsgrader er oppgitt i tabellen under kolonnen f.g. Videre ser vi at vi har valgt å bruke den første kategorien på hver prediktor som referanse for de andre kategoriene. Blant de som rapporterer at de trives svært godt på skolen, er forholdstallet mellom de som ikke er lykkelige og de som er lykkelige satt lik 1,0. Blant de som rapporterer at de ikke trives svært godt, men bare godt, på skolen, er forholdstallet (odds ratio) mellom antall som rapporterer at de ikke er lykkelige og de som rapporterer at de er lykkelige forskjøvet til 2,05 (avrundet) ganger oddsen i referansegruppen. Med synkende skoletrivsel ser vi at oddsen øker via 7,48 til 10,78.

Kommunikasjon med mor slår ut i litt mindre grad. Dessuten observerer vi at sammenhengen ikke er monoton. Odds ratio synker først til 0,64 for deretter å stige via 1,67 til 4,08. Alle disse odds ratio-verdiene er partielle odds ratios. Hver prediktors sammenheng med kriterievariabelen er kontrollert for den andre prediktoren.

Tallstørrelsene i kolonnen under B er den naturlige logaritmen til odds ratio og kan også fungere som assosiasjonsmål. Den ligner i større grad på en korrelasjonskoeffisient. Når odds ratio er 1,00, er B lik 0,0. Når odds ratio er 2,0 eller 0,5, blir B henholdsvis 0,69 og -0,69. B er med andre ord symmetrisk rundt 0 (null). Uheldigvis kan den bli både større enn 1,00 og mindre enn -1,00, og fungerer derfor ikke helt som en korrelasjonskoeffisient.

Ofte rapporteres konfidensintervall i forbindelse med odds ratio. Konfidensintervallet kan beregnes ved å kalkulere et konfidensintervall til B-verdien og deretter finne konfidensintervallet til odds ratio ved å regne ut antilogaritmen til B's konfidensintervallgrenser.

For å regne ut konfidensintervallet til odds ratio-verdien for den gruppen som ikke trives særlig godt på skolen, går vi fram på følgende måte. Vi tar utgangspunkt i at odds ratio-verdien er 7,48 (avrundet). Først kan vi på kalkulatoren kontrollere at B-verdien tilsvarer logaritmen til odds ratio, og får da tallet 2,01. Deretter regner vi ut konfidensintervallet ved å anta at standardfeilen til B er normalfordelt. Et 95-prosent konfidensintervall estimeres ved å regne ut de odds ratio-verdiene som ligger 1,96 x standardfeilen (standard error - S.E.) fra beregnet B-verdi. I vårt tilfelle finner vi da at konfidensintervallets nedre grense er:

2,01 - (1,96 * 0,42) = 1,19

Konfidensintervallets øvre grense ligger på

2,01 + (1,96 * 0,42) = 2,83

Omregnet til odds ratio (antilogaritmen til B), blir tallverdiene

3,29 og 16,95

Vi vet med andre ord at det er 95% sannsynlighet for at den egentlige odds ratio-verdien (i en hypotetisk uendelig stor populasjon) ligger mellom 3,29 og 16,95.

I logistisk regresjon kan en også bruke intervallvariabler som prediktorer. Analysen forutsetter da at spranget i odds ratio er like stort fra intervall til intervall på skalaen, og beregner en slags gjennomsnittlig odds ratio for hele skalaen. Størrelsen på intervallene avgjør størrelsen på odds ratio. For å få en håndterlig størrelse på odds ratio-verdiene, lønner det seg derfor å velge en skala med forholdsvis store sprang (f.eks. alder målt i år transformert til alder målt i hele ti-år).

I likhet med multippel lineær regresjon kan en foreta trinnvise analyser også i logistisk regresjon. I så fall leter programmet seg fram i en hel liste med prediktorer og inkluderer systematisk en og en av disse i modellen. De variablene som bidrar mest, entrer ligningen først. Programmet leter fram til en modell som er slik at alle de som er inne i modellen bidrar signifikant til å forklare den avhengige variabelen mens ingen av de som er utelatt bidrar signifikant. Det eksisterer en rekke måter å foreta slik trinnvis utvelgelse på. En kan starte med hele listen av variabler, og deretter ekskludere en og en inntil alle som gjenstår bidrar signifikant. En kan også prøve ut ulike kombinasjoner av et visst antall prediktorer.

Den kritikken som Cohen & Cohen framsatte i 1975 mot bruk av trinnvis multippel lineær regresjon, gjelder også bruk av trinnvis multippel logistisk regresjon. Hosmer & Lemeshow viser til noe av kritikken som har vært fremsatt (s.87). De mener imidlertid at det ikke er selve prosedyrene det er noe i veien med, men at problemet er forskere som anvender trinnvise prosedyrer uten helt å forstå hva de gjør.

6.3.3 Testing av modeller og mål for multippel assosiasjon

Lesere som er fortrolige med multippel lineær regresjonsanalyse vil vel på dette stadiet forlengst ha spurt seg selv om ikke det finnes en samlet signifikanstest for alle de variablene en har med i multippel logistisk regresjonsanalyse. I tilfellet multippel lineær regresjon benyttes en test som er basert på F-fordelingen. En slik tilsvarende test finnes i multippel lineær regresjon, men denne testen er basert på χ^2 - fordelingen.

$$D_0 = -2((n_{y=0} * \ln(p_{y=0})) + (n_{y=1} * \ln(p_{y=1})) \qquad (6.19)$$

D_0 Den initiale chi-kvadrat-verdien på kriterievariabelen

$n_{y=0}$ Antall som har verdien 0

$n_{y=0}$ Antall som har verdien 1

$p_{y=0}$ Proporsjonen som har verdien 0

$p_{y=1}$ Proporsjonen som har verdien 1

Utgangspunktet er en tallstørrelse som kalles –2 log likelihood. Den er χ^2-fordelt og tilsvarer kvadratsummer i variansanalyse og i multippel lineær regresjon.

For å regne ut det som tilsvarer total kvadratsum for en dikotom kriterievariabel, bruker vi formel 6.17. Denne –2 log likelihood statistikken som viser den totale χ^2-mengden før vi introduserer noen prediktorer, kalles gjerne intial kji-kvadrat-verdi (se formel 6.19).

Dersom vi tenker oss en kriterievariabel der 150 enheter har fått verdien 0 og 81 enheter har fått verdien 1, vil sannsynligheten for å tilhøre gruppe 0 være 0,6494 og sannsynligheten for å tilhøre gruppe 1 være 0,3506. Dersom vi setter disse tallene inn i formelen ovenfor får vi følgende:

D_0 = -2[(150 * ln 0,6494) + (81 * ln 0,3506)] =

D_0 = -2[(150 * (-0,4317) + (81 * (-1,0481)] = 299,30

D_0 tilsvarer altså den totale kvadratsummen (sum of squares total - SST) i variansanalyse, og kan på samme måte dekomponeres i en del som blir forklart i regresjonsanalysen (sum of squares regression – SSR) og en del som forblir uforklart og dermed kan betraktes som feilvarians (sum of squares error – SSE). Når en har kjørt en logistisk regresjonsanalyse vil en få ut en χ^2-verdi (D_m) som sier noe om hvor mye modellen avviker fra data, og en vil få ut en tallstørrelse som forteller om prediktorene samlet sett slår ut signifikant på kriterievariabelen (G_m). Formel 6.20 viser ganske enkelt at D_0 er lik summen av G_m og D_m.

$$D_0 = G_m + D_m \qquad (6.20)$$

D_0 Initial chi-kvadrat-verdi

G_m Modellens chi-kvadrat-verdi

D_m Chi-kvadrat-verdi for avvik mellom modell og data

Det best tenkelige utfallet vil selvsagt være å få en Gm-verdi som er signifikant, som med andre ord viser at selve modellen er signifikant, samtidig med at Dm ikke oppnår signifikans. Dette siste ville bety at det ikke er signifikant avvik mellom modell og data. I praksis er dette et resultat en sjelden ser. Hvis en har gjennomført en undersøkelse med nokså mange enheter eller subjekter, vil D_m så å si alltid vise signifikans. Menard (1995) mener det derfor at en bør legge mest vekt på G_m[1].

Hva så med en koeffisient som kan si noe om hvor god modellen er? Finnes det i multippel logistisk regresjonsanalyse en parallell til den multiple R^2 som vi kjenner fra multippel lineær regresjon? En direkte analog statistikk er foreslått av Hosmer & Lemeshow (1989) (se formel 6.21).

$$R_L^{\,2} = \frac{G_m}{D_0} = \frac{G_m}{G_m + D_m} \qquad (6.21)$$

$R_L^{\,2}$ "Goodness of fit" statistikk for multippel logistisk regresjon

D_0 Den initiale chi-kvadrat-verdien

G_m Modellens chi-kvadrat-verdi

D_m Chi-kvadrat-verdi som angir avvik mellom modell og data

Statistikkpakken SAS pleide tidligere å inkludere en variant av R_L^2 der en korrigerer for antall prediktorer som er med i modellen. Denne er beskrevet i formel 6.22. R_{LA}^2 er analog til den justerte R^2 som vi kjenner fra multippel lineær regresjon (se formel 6.22).

Disse koeffisientene har ikke oppnådd særlig stor popularitet foreløpig. Noen forskere anbefaler at en i stedet beregner en mer direkte multippel R^2. Dette kan gjøres på følgende måte:

1) Først utvikler en den modellen en ønsker å presentere
2) Deretter kjører en analysen og tar vare på de predikerte verdiene, som vanligvis vil være uttrykt som sannsynligheter (probabiliteter) og dermed representerer en intervallskala
3) Endelig beregner man en eta-koeffisient der probabilitetene er kriterievariabel og den dikotome kriterievariabelen fra den logistiske regresjonsanalysen er prediktor.

[1] De modelltilpasning- eller "goodness of fit" – statistikkene vi har gjort rede for her er ikke de eneste som finnes. En del andre er gjort rede for i Hosmer & Lemeshow (1989) og i Menard (1995).

$$R_{LA} = \frac{G_m - 2k}{D_0} = \frac{G_m - 2k}{G_m + D_m} \qquad (6.22)$$

R_{LA} Justert "goodness of fit"-mål for multippel logistisk regresjon
G_m Modellens chi-kvadrat-verdi
D_0 Den initiale chi-kvadrat-verdien
D_m Modellens chi-kvadrat-verdi
k Antall prediktorer som er med i modellen

Noen vil kanskje synes det er rart at kriterievariabelen fra den logistiske regresjonsanalysen her brukes som prediktor. Men dette er det egentlig tradisjon for i diskriminantanalyse, der en gjør det samme (Menard, 1995, s.23-24).

Fordelen med å bruke denne siste framgangsmåten og rapportere den multiple R^2 er at de fleste kjenner denne tallstørrelsen fra multippel lineær regresjon, og at resultatene i en viss forstand blir sammenlignbare med undersøkelser der multippel lineær regresjonsanalyse er brukt. Men ikke alle statistikere liker at en på denne måten blander sammen statistikk fra to ulike verdener, den chi-kvadratbaserte og den variansbaserte.

Ovenfor har vi bare presentert multippel logistisk regresjonsanalyse der den avhengige variabelen er en dikotomi. Det finnes varianter av multippel logistisk regresjon der den avhengige variabelen er en mangekategoriell variabel. Kategoriene kan være ordnede eller det kan være snakk om en ren nominalvariabel.

6.3.4 Trinnvis multippel logistisk regresjonsanalyse

De fleste kjenner nok til at en i multippel lineær regresjon kan velge å kjøre det som kalles en trinnvis analyse. Det samme kan en gjøre i multippel logistisk regresjonsanalyse. Slike trinnvise analyser kan utføres på forskjellige måter. En kan blant annet velge å starte ut med en modell der en ikke har noen prediktorer med, for deretter å inkludere en og en variabel. Analysen fungerer slik at den bare tar med variabler som bidrar til å forbedre modellen. Her kan en bruke forskjellige kriterier som sier noe om hvor mye modellen forbedres når en legger til en ny prediktor, eller hvor mye den forverres når en trekker ut en prediktor. Den enkelte prediktors bidrag til modellen kan signifikanstestes. En kan også velge å starte med en modell der alle prediktorene er med, for deretter å fjerne en etter en blant de som bidrar minst. Dette gjør en inntil bare prediktorer som bidrar signifikant er med i modellen, og bare prediktorer som ikke bidrar er ute.

For forskere som har et forholdsvis stort antall prediktorer, og som gjerne skulle komme fram til hvilke som er de mest sentrale, er det selvsagt svært fristende å bruke trinnvise prosedyrer. Imidlertid har det vært reist kritikk mot bruken av trinnvis regresjon. Dersom en opererer

med forholdsvis små datasett og tester ut et stort antall prediktorer som er innbyrdes korrelerte, kan en komme til å kapitalisere på feil, som det heter. Det vil si at en kan komme til å sitte tilbake med et sett av prediktorer der noen er kommet med i modellen ut fra tilfeldigheter. Dersom en gjentar undersøkelsen med den samme metodologi og de samme målinger på et nytt utvalg (fra den samme populasjonen), risikerer en å få et annet sett av prediktorer neste gang. Problemet er nok noe mindre når en har et stort antall observasjoner og et mer begrenset antall prediktorer.

I forbindelse med trinnvise regresjonsanalyser bør en være særlig oppmerksom på problemet med høyt korrelerte prediktorer (kollinearitet). Dersom den ene i et slikt par av prediktorer blir med i en modell, er det stor sjanse for at den andre utelukkes, og det kan noen ganger være ganske tilfeldig hvilken av de to variablene som blir med i modellen. Når to eller flere prediktorer korrelerer svært høyt, der det sannsynlig at de et stykke på vei er mål på den samme, underliggende faktoren. I slike tilfeller kan det være bedre å konstruere en indeks eller en sumskår der begge eller alle de aktuelle variablene teller med. Alternativt kan en velge å kutte ut variabler dersom en har flere som er likeverdige mål på det samme.

Noen statistikere forsvarer et stykke på vei bruken av trinnvise prosedyrer (for eksempel Agresti & Finlay, 1986 og Hosmer & Lemeshow, 1989).

Referanser

Agresti, A. & Finlay, B. (1986). Statistical methods for the social sciences (2. utgave). San Francisco: Dellen.

Bibby, J. (1977): The general linear model - a cautionary tale. I C.A.O'Muircheartaigh & C.Payne (red.): *The Analysis of Survey Data (Vol.2): Model fitting*. New York: Wiley.

Cohen, J. & Cohen, P. (1975). *Applied multiple regression analysis for the behavioral sciences.* Hillsdale, New Jersey: Lawrence Erlbaum Associate Publishers.

Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (Tredje utgave). Mahwah; New Jersey: Lawrence Erlbaum.

Fox, J. (1993): Regression diagnostics. I Lewis-Beck, M.S. (red.). *Regression analysis*. Thousand Oaks, California: Sage, 245-334.

Hosmer, D.W. & Lemeshow, S. (1989): *Applied logistic regression.* New York: Wiley & Sons.

Karasek, R. & Theorell, T. (1990). *Healthy work: Stress, productivity and the reconstruction of working life*. New York: Basic Books.

Kerlinger, F.N. & Pedhazur, E.J. (1973): *Multiple Regression in Behavioral Research.* New York: Holt, Rinehart & Winston.

Knoke, D. & Burke, P.J. (1980). *Log-linear models*. Beverly Hills, California: Sage. (Vol.20 i serien "Quantitative Applications in the Social Sciences")

Lewis-Beck, M.S. (1980): *Applied regression. An introduction.* Beverly Hills, California: SAGE. (Vol.22 i serien "Quantitative Applications in the Social Sciences"). Inng□r ogs□ som kapittel i Lewis-Beck, M.S. (1993) (red.). *Regression analysis*. Thousand Oaks, California: Sage, 1-68.

Liebtrau, A.M. (1983). *Measures of association.* Beverly Hills, California: Sage.

Menard, S. (1995). *Applied logistic regression analysis*. Thousand Oakes, California: Sage.

Reynolds, H.T. (1977). *Analysis of nominal data.* Beverly Hills, California: Sage. (Vol.7 i serien "Quantitative Applications in the Social Sciences"). Inng□r ogs□ som kapittel i M.S.Lewis-Beck (1993) (red.). *Basic statistics*. Thousand Oaks, California: Sage, 159-234.

Theorell, T. (2000). Working conditions and health. I Berkman, LF. & Kawachi, I. (red.), *Social epidemiology* (pp. 95-117). Oxford: Oxford University Press.

Truett, J., Cornfield, J. & Kannel, W. (1967): A multivariate analysis of the risk of coronary heart disease in Framingham. *Journal of Chronic Diseases*, *Vol.20*, 511-524.

Yule, G.U. (1900). On the association of attributes in statistics. *Philosophical Transcriptions of the Royal Society* A194, 257-319 (etter Liebtrau, 1983).

Yule, G.U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society, 75*, 579-642 (etter Liebtrau, 1983).

Appendiks A: Beregning av utvalgsstørrelse

Enkle beregninger av antall observasjoner

La oss først seg hvordan vi regner ut antall observasjoner når vi ønsker en bestemt presisjon på et prosentestimat. Ved å ta utgangspunkt i formelen for beregning av standardfeilen på prosentestimater, kan vi snu formelen slik at vi kan beregne n (antall enheter i utvalget) på bakgrunn av opplysninger om hvor stor en prosent i populasjonen antas å være og hvor stor standardfeil vi kan tolerere (jfr. Moser & Kalton, 1971, s.147).

Formelen blir da slik:

$$n = \frac{P(100 - P)}{SE_p{}^2}$$

n Antall enheter som må være med i utvalget

P Prosentandelen med et bestemt kjennetegn i populasjonen

SE_p Standardfeilen en estimert prosent

Dersom vi antar at andel som har bestemt seg for å stemme ja til EF er om lag 30%, og dersom vi ønsker et 95% konfidensintervall som skal dekke et område på 4%, må vi først beregne hvor stor standardfeil et konfidensintervall på 4% tilsvarer. Standardfeilen beregnes ved å dele på z-verdien 1,96. Den ønskede standardfeilen blir dermed på 2,04%. Ved å sette tallene inn i formelen får vi:

$$n = \frac{30(100 - 30)}{2{,}04^2} = 505$$

n Antall enheter som må være med i utvalget

Prosentandelen som sier de vil stemme ja: 30

Standardfeilen til andel som vil stemme ja: 2,04

Det er viktig å legge merke til at det nødvendige antall observasjoner blir størst når P (og dermed også 100-P) nærmer seg 50%. Ved å anta at andel som har bestemt seg for å stemme ja er 50%, kan vi beregne den nødvendige utvalgsstørrelsen gitt samme krav til konfidensintervall til 601. Dersom vi tvert imot antar at det er så få som 10% som har bestemt seg for å stemme ja, blir den nødvendige utvalgsstørrelsen 216.

Svært ofte vet en lite om P (og dermed 100-P) på forhånd. Undersøkelsen blir kanskje gjennomført nettopp for å finne ut hvor stor P er i populasjonen. Det kan derfor virke paradoksalt at vi på forhånd skal ha noen sikker mening om dette. Det sikreste er derfor å anta at P (og dermed også 100-P) ligger på 50%. Særlig når en skal estimere mange ulike prosenttall ved bruk av mange forskjellige variabler, er det fornuftig å hele tiden beregne N under forutsetning av at P er nær 50% (Jfr. Cochran, 1963, s.75; Kalton, 1983, s.82).

Tabell A.1: Utvalgsstørrelse som en funksjon av utvalgsfeil og prosenttall

k	P: / Q:	10 / 90	20 / 80	30 / 70	40 / 60	50 / 50
1%		3458	6147	8068	9220	9604
2%		865	1537	2017	2305	2401
2,5%		553	984	1291	1475	1537
5%		139	246	323	369	384
10%		35	62	81	92	96

P - prosent med en bestemt egenskap i populasjonen
Q - 100-P
k - størrelsen på et 95% konfidensintervall
Tallene i tabellen angir utvalgsstørrelsen som tilsvarer P/Q og k

En tysk lærebok i samfunnsvitenskapelig forskningsmetode (Friedrich & Hennig, 1975) gjengir en enkel liten tabell som forteller om sammenhengen mellom krav til konfidensintervall, P og n (tabell A.1). Det er interessant å se hvordan presisjonen avtar sterkt når utvalgsstørrelsen blir mindre enn 100. Det er grunn til å stille spørsmål om nytten av surveys med et antall observasjoner som er så lavt som under 100.

Sammenhengen mellom utvalgsstørrelse og presisjon er formulert i noe som er blitt kalt de store talls lov (som vi har vært inne på flere ganger tidligere i denne teksten) og som refererer seg tilbake til den kjente matematikeren Jacob Bernoulli. Han kom fram til at jo større utvalg en har, forutsatt at hver trekking skjer uavhengig av alle andre trekninger, desto mer vil den statistiske størrelsen en beregner nærme seg den "sanne" verdien, med andre ord nærme seg den verdien en ville fått ved å beregne den statistiske størrelsen på grunnlag av hele populasjonen. De store talls lov kan så og si leses rett ut av den tabellen vi gjengir ovenfor (Tabell A.1).

Dersom vi ønsker å vite hvor mange vi må ha med i et utvalg, kan vi, gitt en del premisser, finne dette ut ved hjelp av tabellen ovenfor. På forhånd må vi bestemme

(a) Hvordan fordelingen på den viktigste parameteren ser ut i populasjonen (P),
(b) hvor stor feilmargin vi kan tolerere (k), og
(c) hva slags krav vi skal stille til konfidensintervallet (95%, 99%, 99,9% etc.).

Tabellen ovenfor er basert på 95% konfidensintervall. Dersom vi bestemmer oss for at fordelingen på parameteren i populasjonen er 50-50 og vi ønsker at feilmarginen skal ligge innenfor området ± 5%, blir vi nødt til å trekke et rent tilfeldig utvalg på minst 384 personer.

Ovenfor så vi hvordan vi ut fra antagelser om prosentfordeling i populasjonen og krav til presisjon kunne regne ut nødvendig utvalgsstørrelse. På tilsvarende måte kan en beregne utvalgsstørrelsen som må til for å sikre at feilmarginene ved estimering av et aritmetisk gjennomsnitt ikke blir større enn en bestemt verdi (Moser & Kalton, 1971). Formelen ser slik ut:

$$n = \frac{SD_x^{\,2}}{SE_x^{\,2}}$$

n Antall enheter som må være med i utvalget

SD_x Standardavviket til x slik vi antar at det er i populasjonen

SE_x Ønsket standardfeil for gjennomsnittet på variabelen x

De to eksemplene ovenfor har dreiet seg om enkeltvariabler. Framgangsmåten er helt tilsvarende dersom en i stedet tar utgangspunkt i forskjellen mellom proporsjoner, forskjellen mellom aritmetiske gjennomsnitt, korrelasjoner eller mer kompliserte statistiske størrelser hentet fra multivariat statistikk. Så lenge det finnes kjente måter å beregne standardfeil på, og så lenge en vet hva slags sannsynlighetsfordeling som skal benyttes, kan en beregne hvor stort utvalget må være (se f.eks. Kraemer & Thiemann, 1987). Ikke alle statistikkbøker gir retningslinjer og formler for beregning av utvalgsstørrelse, men det finnes en rekke bøker som handler nettopp om dette (se. f.eks. Kraemer & Thielmann, 1987 eller Lwanga & Lemeshow, 1991).

Teststyrke

Et viktig begrep når en skal beregne utvalgsstørrelse er teststyrke, og en snakker ofte om styrkeberegninger. Styrken på en statistisk test defineres som sannsynligheten for at nullhypotesen forkastes når den faktisk skal forkastes. Begrepet er nært beslektet med det som kalles type II-feil. Type I-feil er sannsynligheten for å forkaste en null-hypotese når den er riktig. Type II-feil er sannsynligheten for å la være å forkaste en null-hypotese når den er feil. Teststyrken er lik 1,0 minus sannsynligheten for type II-feil. (Shaughnessy & Zechmeister, 1994, s.285). Teststyrken er med andre ord sannsynligheten for ikke å gjøre type II-feil. Jo høyere teststyrke, desto sikrere er en på å påvise en sammenheng eller en forskjell som faktisk finnes.

Cohen (1962, 1965) påpekte at ett av de aller største metodeproblemene innen psykologisk forskning er de små utvalgene de fleste undersøkelser er basert på. En oppsummering av teststyrken på empiriske studier trykket i 1960-utgaven av Journal of Abnormal and Social Psychology viste at den, forutsatt at en antok at studiene ville dreie seg om å påvise middels sterke effekter, gjennomsnittlig lå på 0,48. Dette betyr at sjansen for å oppdage middels sterke sammenhenger som faktisk var til stede, gjennomsnittlig var under 50%. Senere artikler som omhandler samme tema tyder på at situasjonen ikke er blitt særlig bedre (Sedlmeier & Gigerenzer, 1989).

Cohen har i en artikkel fra 1992 gitt en grei veiledning i hvor stort utvalg en bør trekke. Dersom en bestemmer seg for type effektmål, effektstyrke og signifikansnivå, vil en gitt en teststyrke på 0,80 kunne lese antall observasjoner som kreves (n) direkte ut av en tabell. Det er også publisert enklere bøker som på en mer oversiktlig måte gir retningslinjer for å bestemme utvalgsstørrelse ut fra hvilke krav en stiller til teststyrke og ut fra hvor svake sammenhenger en ønsker å kunne oppdage (Kraemer & Thiemann, 1987; Lwanga & Lemeshow, 1991). Forøvrig finnes det i dag en rekke nettsteder med power-kalkulatorer der en kan taste inn opplysninger om teststyrke, effektstørrelse (styrken på sammenhengen eller størrelsen på forskjellen en vil være i stand til å vise) og signifikansnivå. Kalkulatorene gir på bakgrunn av disse opplysningene svar på hvor store utvalg vi har behov for. Vi kan også taste inn opplysninger om antall observasjoner, effektstørrelse og signifikansnivå og få ut beregnet teststyrke. Se for eksempel følgende nettsted: http://calculators.stat.ucla.edu/ Hvis du velger "Sample size calculator", som er nr. 4 på listen, og deretter "proportion", kan du reprodusere alle tallene i tabell A.1. Forskjellig bruk av avrundingsregler gjør at du kanskje vil oppdage forskjeller på en enhet.

Fig. A.1: Type I og type II-feil
(Etter Knoke & Bohrnstedt, 1994)

I populasjonen som utvalget er trukket fra er null-hypotesen ...		Basert på utvalget har en funnet at nullhypotesen ... IKKE SKAL FORKASTES	SKAL FORKASTES
	RIKTIG	Korrekt beslutning	Type I-feil: Forkaster en riktig nullhypotese
	GAL	Type II-feil: Godtar gal nullhypotese	Korrekt beslutning

Andre forhold

Når en skal beregne utvalgsstørrelse, kan en ikke forutsette at alle de som trekkes ut faktisk deltar i den planlagte undersøkelsen. Dette bør en ta hensyn til ved å trekke utvalg som er såpass store at en etter frafall sitter tilbake med et antall som er tilstrekkelig stort. Dette kan selvfølgelig ikke kompensere for den feilkilden som oppstår ved systematisk frafall, f.eks. at røykere unndrar seg deltakelse i røykevaneundersøkelser. Poenget er bare at en ikke bør risikere at usikkerheten blir enda større ved at antall observasjoner blir for lavt.

Når en skal bestemme størrelsen på et utvalg i forbindelse med en survey, er det som regel ikke så enkelt at en velger ut en bestemt variabel eller en bestemt opplysning som kan fungere som kriterium. De færreste surveys har som siktemål å framskaffe ett bestemt tall som er basert på en enkelt variabel. En mer fullstendig oversikt over forhold som må tas i betraktning kan omfatte følgende:

- Hvor presise estimater bør en oppnå? Jo større presisjon vi har bruk for, desto større utvalg må vi ha.

  Eller ekvivalent med dette:

  Hvor svake sammenhenger (eller gruppeforskjeller) bør en være i stand til å registrere som statistisk sikre? Jo svakere statistiske sammenhenger vi ønsker å kunne påvise, desto større utvalg må vi ha.

- Hvor stort frafall kan en forvente? Dersom en for eksempel ønsker å teste sammenhengen mellom et bestemt antall prediktorer (uavhengige variabler) og en kriterievariabel (avhengig variabel), stiller de regresjonsanalytiske teknikkene bestemte krav til antall enheter (f.eks. respondenter) i forhold til antall variabler som kan inngå i analysen. Med stort frafall kan en komme til å avskjære muligheten for å kjøre slik statistikk som en i utgangspunktet hadde planlagt. Jo større frafall, desto større utvalg må en ha. Frafallet i en undersøkelse kan, som nevnt ovenfor, ikke uten videre kompenseres ved at en trekker større utvalg. Problemet er nemlig at frafallet kan være systematisk. I en undersøkelse av illegal kjøp av smuglet sprit eller hjemmebrent, er det fullt mulig at de som i størst grad kjøper og bruker slike produkter er overrepresentert blant de som faller fra. Da hjelper det lite å trekke et større utvalg. Likevel må frafallet tas hensyn til ved bestemmelse av utvalgsstørrelse.

- Hvor mange variabler er en interessert i å bruke i en enkelt multivariat analyse? Jo flere variabler (og jo mer kompliserte modeller), desto større utvalg må en ha. Testing av interaksjonseffekter stiller ofte høye krav til utvalgsstørrelse.

- Hvor små undergrupper ønsker en å gjøre separate analyser på? Jo mer en ønsker å splitte opp et materiale, med andre ord; jo mindre undergrupper en ønsker å kjøre statistikk på, desto større utvalg er det behov for.

- Hvor reliable målinger kan en regne med å få? Jo mindre reliable mål en har, desto større må utvalget være for å påvise en bestemt sammenheng.

Det er kanskje forståelig at mange forskere, stilt overfor en slik liste med spørsmål, velger å bruke sitt skipperskjønn når de skal bestemme utvalgsstørrelse, og for å være på den sikre side heller gjøre n noe høyere enn det de hadde tenkt. Forståelig er det, men langt fra akseptabelt. For lavt n legger begrensninger på hvor mye og hvor god statistikk en kan lage på bakgrunn av en undersøkelse. For høyt n kan bety en unødig hard belastning økonomisk og tidsmessig i en forskningsverden som er kjennetegnet av ressursknapphet.

Designeffekten

For å vurdere hvor gode ulike typer utvalg egentlig er, bruker en gjerne å sammenlikne med rent tilfeldig trekking. Rent tilfeldige utvalg er med andre ord den standarden som andre typer utvalg kan måles mot. For å vise hvor gode de andre utvelgingsmetodene er, bruker en gjerne en statistisk størrelse som kalles designeffekten. Designeffekten er definert som kvadratet av standardfeilen (altså variansen) til den aktuelle estimatoren til det mer kompliserte designet delt på kvadratet av standardfeilen til estimatoren basert på et rent tilfeldig utvalg av samme størrelse. Designeffekten til estimatoren x er med andre ord:

$$D^2_{(x)} = \frac{SE_{(x)}^{\ 2}}{SE_{(x0)}^{\ 2}}$$

$D^2_{(x)}$ Designeffekten

$SD_{(x0)}$ Standardfeilen til en estimator basert på et rent tilfeldig utvalg (simple random sample)

$SE_{(x)}$ Standardfeilen til en estimator basert på et utvalg som er trukket på en mer kompleks måte

Hvor stor designeffekten er, varierer med hvilken variabel i et datasett en tar for seg. Dersom en f.eks. har brukt klyngeutvelging, vil variabler som er homogéne innen klynger og samtidig varierer sterkt mellom klynger få en stor designeffekt. Et eksempel er variabler som måler elevers opplevelse av skolemiljøet. Dersom en har trukket hele skoleklasser, og antar at elever innen en og samme klasse har en nokså lik opplevelse av miljøet i klassen, mens variasjonen fra klasse til klasse er stor, vil designeffekten bli høy. Jo større heterogenitet det er innen klasser, og jo mindre forskjellen er fra klasse til klasse, desto lavere blir designeffekten. Dersom en ikke planlegger å trekke et rent tilfeldig utvalg fra en populasjon, men i stedet er henvist til å benytte mer komplekse framgangsmåter (for eksempel klyngeutvalg eller stratifiserte utvalg), får det konsekvenser for hvordan vi skal beregne utvalgsstørrelse. Dersom designeffekten er større enn 1,0, må utvalgsstørrelsen økes tilsvarende.

Referanser

Cochran, W.G. (1963). *Sampling techniques.* New York: John Wiley & Sons.

Cohen, J. (1962). The statistical power of abnormal-social psychological research. A review. *Journal of Abnormal and Social Psychology*, *Vol.65*, 145-153.

Cohen, J. (1965). Some statistical issues in psychological research. In B.B.Wolman (red.). *Handbook of clinical psychology.* New York: McGraw-Hill, 95-121.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *Vol.112 (No.1)*, 155-159.

Friedrich, W. & Hennig, W. (1975). *Der sozialwissenschaftliche Forschungsprosess.* Berlin: VEB Deutscher Verlag der Wissenschaften.

Kalton, G. (1983). *Introduction to survey sampling.* Beverly Hills, California: SAGE. (Quantitative Applications in the Social Sciences, nr. 35.)

Knoke, D. & Bohrnstedt, G.W. (1994). *Statistics for social data analysis (Third edition).* Itasca, Illinois: F.E.Peacock Publishers.

Kraemer, H.C. & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research.* Newbury Park, California: Sage.

Lwanga, S.K. & Lemeshow, S. (1991). *Sample size determination in health studies. A practical manual.* Geneva: Worls Health Organization

Moser, C.A. & Kalton, G. (1971). *Survey methods in social investigation*. London: Heinemann Educational Books.

Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *Vol.105*, 309-316.

Shaugnessy, J.J. & Zechmeister, E.B. (1994). *Research methods in psychology (Third edition).* New York: McGraw-Hill.