

Jon Sørgaard

INNFØRING I
UTVALGTE GRUNNLEGGENDE
STATISTISKE TEKNIKKER

STS-arbeidsnotat 13/94

ISSN 0802-3573-90

arbeidsnotat
working paper

Innhold

Innledning	2
Hva kan vi oppnå med statistisk analyse ...2	
Om variabler og verdier ...4	
Datamatriksen ...4	
Omkodning ...5	
Indekser ...6	
Frekvenser ...7	
Målenivåer ...8	
Variabler på nominalnivå ...8	
Variabler på ordinalnivå ...9	
Variabler på intervall-/forholdstallsnivå ...10	
Sannsynlighetsberegning	10
Sannsynlighet og proporsjoner ...11	
Sannsynlighet og uavhengighet ...12	
Effekt og samspill	15
Effekt med én uavhengig variabel (bivariat analyse) ...16	
Effekt og samspill i multivariat analyse ...18	
'Hernes-metoden' ved beregning av effekter ...18	
'Hellevik-metoden' ved beregning av effekter ...19	
Beregning av samspill ...21	
Beregning av prediksjonseffekt ...24	
Kji-kvadrat(χ^2) (bivariat hypotesetesting)	26
Om bruken av kjikvadrat ...26	
Framgangsmåte ved kjikvadrat-testing ...27	
Tabellstørrelse og antall frihetsgrader ...27	
Valg av signifikansnivå ...27	
Beregning av kjikvadratet ...29	
Korrelasjonsmål	34
Valg av korrelasjonsmål ...34	
Normering av korrelasjonsmål ...35	
Fi; korrelasjon på nominalnivå ...36	
Utrekning av fi ...37	
Tau og gamma; korrelasjon på ordinalnivå ...39	
Variabler på ordinalnivå; klyngediagram ...40	
Beregning av like og ulike par ...42	
Utrekning av tau ...43	
Utrekning av gamma ...46	
Rho (rangkorrelasjonskoeffisienten); ordinalnivå ...47	

Jon Sørgaard:

INNFØRING I UTVALGTE GRUNNLEGGENDE STATISTISKE TEKNIKKER

1. Innledning

I dette heftet tar jeg sikte på å gi en framstilling av en del av de mest brukte statistiske målene i samfunnsvitenskapene. Jeg legger vekt på å vise når og hvorfor vi bruker de enkelte målene, hvilken type spørsmål de enkelte målene gir oss muligheter til å besvare, og hvordan vi foretar utregningene.

Det er selvfølgelig sjelden at vi faktisk sitter og regner ut dette manuelt. Gjennom å forstå hvordan utregningene foregår, kan vi imidlertid også forstå hvorfor de enkelte målene måler det de gjør. Å trenge inn i formlene og brette de ned til de enkelte regneoperasjonene betyr at vi kan få et mindre 'fremmedgjort' forhold til disse teknikkene. Det er derfor ikke å huske formlene som er det viktigste, men å kjenne deloperasjonene som de uttrykker.

HVA KAN VI OPPNÅ MED STATISTISK ANALYSE

Statistisk analyse er bare en liten del av det metodegrunnlaget som er tilgjengelig i samfunnsvitenskapene. I utgangspunktet mener jeg at vi skal forholde oss til det vi undersøker med et mest mulig åpent sinn, og se i øynene at alt kan bære med seg informasjon som kan være nyttig i forskningsprosessen: Enkeltindividers utsagn og atferd, holdninger og følelser, eksisterende sosiale strukturer og prosesser, materielle føringer som ligger i gjenstandene vi omgir oss med, strukturelle forhold ved geografiske og romlige omgivelser osv. - alt dette er med å forme den sosiale virkeligheten som vi skal prøve å gripe tak i. Vi bør ha et metodespekter både for innsamling og registrering og til analyse av alle disse data- eller informasjonstypene, og vi bør bruke dem hver for seg eller i kombinasjon alt etter hvilke spørsmål vi vil ha svar på.

Bare en begrenset del av all denne informasjonen lar seg (eller bør) samles inn gjennom bruk av kvantitative metoder, og bare en begrenset del lar seg meningsfylt analysere gjennom bruk av statistikk. Når vi imidlertid står overfor slike typer data, og slike metoder er gunstige i forhold til de spørsmålene vi har reist, er de statistiske metodene ofte svært presise verktøy.

Jeg vil trekke fram to alvorlige feil som kan gjøres når vi bruker kvantitative metoder og statistisk analyse.

Den første feilen jeg vil advare mot, er å la valg av metode gå forut for valg av problemstilling. Det er ikke slik at vi skal la for eksempel fagtradisjoner med tanke på metodevalg være en dominerende rettesnor for hvilke metoder vi skal velge. Ved å ta utgangspunkt i hva som er fagets tema, eller undersøkelsens formål, velger vi metode ut fra spørsmålsstillingen. Velger vi metode først - og foreksempel bestemmer oss for at når vi skal forske eller gjennomføre undersøkelser så skal vi bruke kvantitative metoder - da legger vi også begrensninger på oss sjøl om hvilke problemstillinger og spørsmål vi i det hele tatt kan jobbe med.

Den andre feilen jeg vil advare mot er å tro at kvantitative metoder og statistiske analyseteknikker nødvendigvis er mer pålitelige enn kvalitative. Dette er ikke tilfelle. Kvantitative metoder har sin styrke når det gjelder å foreta pålitelige generaliseringer om spesielle sammenhenger, det vil si i situasjoner hvor vi med utgangspunkt i å ha jobbet med et utvalg trekker konklusjoner som skal gjelde også en større gruppe.

I de kvantitative metodene finnes det regler for hvordan slike utvalg skal trekkes, det finnes muligheter til å beregne feilmarginer og det finnes forskrifter for hvordan slike generaliseringer kan gis en meningsfylt form. Forutsatt at det innledende arbeidet er tilstrekkelig godt gjennomført - med tanke på utvalgstrekking, spørsmålsformulering, begrepsklarhet, pålitelighet i registreringsfasen osv. - vil riktig bruk av riktige statistiske mål gi godt grunnlag for generaliseringer. Hvorvidt alle de nødvendige forutsetningene er tatt hensyn til i tilstrekkelig grad er imidlertid et spørsmål som ofte bare kan besvares skjønnsmessig. Faren er derfor til stede for at statistiske analyser blir som et byggverk på sandgrunn (jfr. det skjeve tårn i Pisa) - selve overbygningen kan være raffinert og nærmest fullkommen, men grunnlaget svikter.

Kvalitative metoder gir ikke det samme grunnlaget for den samme typen pålitelige generaliseringer. Når vi velger ut et utvalg til en kvalitativ undersøkelse er det oftest ikke et representativt utvalg vi trekker. Ofte kan vi velge ut de vi tror har mest å si som infomanter, de som kan gi oss spesielle opplysninger, de som har stått i sentrum av begivenheter og videre. Oftest er selve formålet for undersøkelsen forskjellig, i kvalitative undersøkelser er det gjerne forståelsen, bredden og spekteret av og innholdet i et fenomen (for eksempel med tanke på holdninger) vi er ute etter: Vi vil ha svar på hvordan holdningene oppstår, hvorfor de varierer som de gjør, og hva som faktisk ligger i dette spekteret. Disse

resultatene kan vi generalisere forutsatt at forarbeidet og gjennomføringen er god nok. Bruker vi derimot kvantitative metoder på de samme holdningene vil det være utbredelsen og eventuelt styrken vi er ute etter å kartlegge, og det er den typen resultater som vi eventuelt kan generalisere ved hjelp av disse metodene.

Som et eksempel kan vi tenke oss å studere rasisme i Norge. Gjennom en kvalitativ undersøkelse kan vi få vite hva som ligger i de enkelte standpunktene, vi kan skaffe oss innsikt i hvordan enkeltpersoner begrunner og legitimerer sine standpunkter, og vi kan få forståelse over det spekteret av synspunkter og standpunkter som har betydning for dette fenomenet: Hva er rasisme, hvordan motiveres rasisme, hvor ekstreme er de mest ekstreme holdningene, hvor langt er folk villig til å gå, hvordan argumenteres det for og mot rasistiske holdninger. Vi oppnår forståelse av fenomenet og vil kunne dokumentere dets karakter og innhold. Ved å bruke kvantitative metoder er det i første rekke utbredelsen og styrken av enkelte av disse punktene på holdningsskalaen vi vil finne. Vi oppnår kunnskap om spredningen, og vil kunne være i stand til å dokumentere fenomenets omfang.

Skillet mellom de to ulike metodene er derfor ikke et skille mellom pålitelighet eller ikke, men et skille mellom hvordan de er pålitelige: De kan ligge til grunn for pålitelige svar på ulike spørsmål.

OM ENHETER, VARIABLER OG VERDIER

- Datamatriksen

Vi snakker om enheter, variabler og verdier i metodelæren. Forholdet mellom disse er svært viktig å ha på det rene.

Vi kan tenke oss ethvert datasett som en matrise eller som et rutesystem. Enhetene ligger linje for linje i dette rutenettet, variablene ligger kolonne for kolonne, og verdiene er plassert i hvert enkelt skjæringspunkt mellom enhet og variabel:

	Variabel 1	Variabel 2	Variabel 3	Variabel 4	Variabel 5
Enhet nr. 1					
Enhet nr. 2					
Enhet nr. 3					
Enhet nr. 4					
Enhet nr. 5					
Enhet nr. 6					
Enhet nr. 7					

I denne matrisen står enhet for den enheten vi har informasjon fra eller om, altså en enhet på det nivået vi undersøker. Ofte er enhetene hver enkelt person som inngår i undersøkelsen, men det kan også være for eksempel nasjoner, grupper e.l. Variabel er den egenskapen vi måler, den faktoren vi har informasjon om. For hver egenskap eller faktor må vi ha en kolonne, på samme måte som vi for hver enhet må ha en linje. Dette betyr at vi ender opp med et system hvor vi for hver enhet og for hver egenskap eller faktor kan fylle ut i rutenettet med én verdi.

Arbeider vi med kvalitative metoder er det ikke bestandig at det er på denne måten vi setter opp lageret av informasjon som vi har samlet, men matriseformen svarer i slike tilfeller stort sett til hvordan vi forestiller oss informasjonsmengden. Når vi arbeider med kvantitative metoder foregår oppsettet før selve databehandlingen på denne måten, ved at vi legger inn i skjemaet de verdiene vi har klart å samle inn:

	Kjønn	Alder	Inntekt	Yrke	Bosted
Enhet nr. 1	Mann	30	240 000	Fysioterapeut	Mosjøen
Enhet nr. 2	Kvinne	28	270 000	Rådmann	Namsos
Enhet nr. 3	Mann	45	120 000	Sosiolog	Trondheim
Enhet nr. 4	Kvinne	21	265 000	Sykepleier	Fosnes
Enhet nr. 5	Kvinne	26	184 000	Lege	Namsos
Enhet nr. 6	Mann	42	64 000	Sjåfør	Steinkjer
Enhet nr. 7	Kvinne	37	162 000	Lærer	Rørvik

Ved maskinell behandling av datamaterialet må vi rent teknisk gi hver verdi for hver variabel en kode som maskinen er i stand til å forstå. Som regel bruker vi tallkoder, og setter opp en kodebok hvor hver kode får oppgitt sin betydning. For kjønn bruker vi for eksempel 1 = mann, 2 = kvinne og 9 = ubesvart, for alder taster vi kanskje inn tallene direkte osv.

Ut fra dette skulle det også framgå at vi for hver kombinasjon mellom enhet og variabel bare kan oppgi en verdi. Verdien er den informasjonen vi har om en gitt egenskap for en gitt enhet. Dersom vi har flere verdier som kunne ha vært eller burde ha vært brukt for en enhet på for eksempel yrke for å få med oss eventuelle bistillinger og ekstrajobber, må vi opprette flere variabler: Yrke 1, yrke 2, yrke 3 osv.

- Omkoding

Når vi omkoder en variabel, betyr det at vi omgrupperer verdiene på variabelen. Oftest gjør vi dette ved at vi med utgangspunkt i de opprinnelige verdiene grupperer dem i verdiklasser hvor hver nye verdiklasse omfatter flere av de opprinnelige verdiene på variabelen. Hovedhensikten med omkoding er vanligvis

å få mer håndterlige data. Når vi omkoder en variabel, betyr derfor dette ofte at vi mister en del informasjon, vi blir stående igjen med en grovere og mindre eksakt inndeling.

Konkret kan vi tenke oss at vi omkoder variabelen inntekt til tre verdiklasser. Vi setter verdien 1= lav inntekt for alle enheter som har lavere inntekt enn 150 000 kr. (dette vil dreie seg om 2 enheter, nr. 3 og nr. 6. Den nye verdien 2= middels inntekt definerer vi til inntekter f.o.m. 150 000 og opp til 230 000 kr. Enhetene 5 og 7 vil derved få verdien 2. De resterende enhetene vil få verdien 3= høy inntekt.

	Kjønn	Alder	Inntekt	Yrke	Bosted	Innt.gr.
Enhet nr. 1	Mann	30	240 000	Fysioterapeut	Mosjøen	3
Enhet nr. 2	Kvinne	28	270 000	Rådmann	Namsos	3
Enhet nr. 3	Mann	45	120 000	Sosiolog	Trondheim	1
Enhet nr. 4	Kvinne	21	265 000	Sykepleier	Fosnes	3
Enhet nr. 5	Kvinne	26	184 000	Lege	Namsos	2
Enhet nr. 6	Mann	42	64 000	Sjåfør	Steinkjer	1
Enhet nr. 7	Kvinne	37	162 000	Lærer	Rørvik	2

For at ikke den opprinnelige informasjonen skal gå tapt for alltid, oppretter vi gjerne en ny variabel når vi omkoder en eksisterende, slik at vi får to variabler som inneholder informasjon om inntekt: En med den opprinnelige detaljerte informasjonen, og en basert på en grovere inndeling. Vi kan da seinere i løpet av analysen bruke den variabelen som er best egnet til ulike analyseteknikker. (Enkelte dataprogrammer lar oss ikke i det hele tatt omkode direkte eksisterende variabler, men vil automatisk opprette en ny.)

Omkoding av en variabel er noe som gjøres for hele datasettet under ett, det vil si at omkodinger vil gjelde for alle enheter i undersøkelsen. Det samme gjelder hvis vi sletter en variabel: Da vil kolonnen for den aktuelle variabelen bli fjernet for alle enheter i undersøkelsen, og informasjonen går tapt. Sletter vi en enhet, betyr det at all informasjon - det vil si opplysningene for alle variablene - for den aktuelle enheten fjernes fra datasettet.

- Indekser

Å konstruere indekser vil si at vi slår sammen informasjonen fra to eller flere av de opprinnelige variablene i en ny variabel. Vi kan for eksempel tenke oss at vi slår sammen den informasjonen som vi har om inntektsforhold med den informasjonen vi har om yrke, og lager en ny variabel som vi kaller status (som mål på status blir dette kanskje litt knapt...). I den nye variabelen vil vi for eksempel la høy inntekt gi høy status, og vi lar forestillinger vi har om de forskjellige yrkesgruppene prestisje også bestemme noe av hver enkelt enhets

verdi på denne nye variabelen.

I praksis vil konstruksjon av indekser også bety at vi får et mer håndterlig datamateriale, ved at vi i stedet for å være henvist til å bruke flere av de opprinnelige variablene kan få fram hovedpoengene gjennom bruken av den ene indeksen. Imidlertid mister vi også her en del av den eksakte informasjonen. Ser vi på en tenkt verdi, for eksempel 'relativt høy status', på indeksvariabelen kan vi ikke lenger vite hva det er som først og fremst trekker 'opp': Er det høy inntekt, som gir litt ekstra status i forhold til yrket, eller er det yrkesgruppas prestisje som trekker statusen opp til tross for eventuelt lav inntekt?

- Frekvenser

En (univariat) frekvensfordeling vil si at vi får en liste over hyppigheten av forekomsten av hver enkelt verdi på en variabel. En frekvensfordeling for kjønn i vårt eksempel vil derfor gi som resultat 3 menn og 4 kvinner. Dette resultatet kan også prosentueres, slik at vi får en andel av menn på 43 % og en andel kvinner på 57 %. I krysstabeller får vi på samme måte oppgitt antallet i undersøkelsen som har en gitt kombinasjon av verdier:

Inntektsgruppe	Menn	Kvinner	Totalt
1 'Lav inntekt'	2	0	2
2 'Middels inntekt'	0	2	2
3 'Høy inntekt'	1	2	3
	3	4	7

Også i dette tilfellet kan opplysningen i tabellen prosentueres, i utgangspunktet på 3 måter:

a)

Inntekt	M	K	T
Lav	29	0	29
Middels	0	29	29
Høy	14	29	43
	43	58	101

b)

Inntekt	M	K	T
Lav	67	0	29
Middels	0	50	29
Høy	33	50	43
	100	100	101

c)

Inntekt	M	K	T
Lav	100	0	100
Middels	0	100	100
Høy	33	67	100

a) Totalprosent

For det første kan vi oppgi totalprosenten, og beregne hver enkelt gruppes andel ut fra denne. Da vil vi få prosenttall som gir oss grunnlag for å sammenlikne hver enkelt gruppes andel av hele utvalget: Menn m. lav inntekt vs. kvinner m. middels inntekt osv.

b) Vertikal prosenttering; kolonneprosent

Hvis det er den sammenlikningen som tydeligst viser forskjellen mellom kjønnene med tanke på inntektsforskjeller vi er interessert i, bør vi velge loddrett prosenttering. Da vil hver enkelt av kjønnene bli summert til 100 %, slik at vi kan lese andelen av kvinnene i de ulike inntektsgruppene direkte ut av tabellen og sammenlikne disse med andelen av mennene.

c) Horisontal prosenttering; rekkeprosent

I den tredje prosentteringsmåten er det de ulike inntektsgruppene som utgjør prosentteringsgrunnlaget, og som hver for seg skal summeres til eller utgjøre 100 %. Med denne prosentteringsmåten kan vi lese rett ut av tabellen hvordan den kjønnsvis fordelingen er for hver enkelt inntektsgruppe.

Valg av prosentteringsmåte må vi foreta etter hvilket sammenlikningsgrunnlag vi er ute etter.

MÅLENIVÅER

Hvordan vi kan beskrive forholdet mellom verdiene på en gitt variabel er avgjørende for hvilket målenivå variabelen kan sies å befinne seg på. En rekke av de ulike statistiske teknikkene forutsetter at en variabel (minst) er på et gitt målenivå. Vi kan si at variabelens målenivå er avgjørende for hvilke typer statistiske metoder som meningsfylt kan brukes i analysen.

Vi skiller mellom nominal-, ordinal- og intervall-/forholdstallsnivå, hvorav nominalnivå betegnes som det laveste nivået og intervall-/forholdstallsnivå som det høyeste. Vi stiller stadig strengere krav til en variabel for å si at den kan godkjennes som en variabel på et høyere nivå.

- Variabler på nominalnivå

Det laveste målenivået er nominalnivå, dette betyr at alle variabler minst befinner seg på dette nivået. For at en variabel skal kunne sies å befinne seg på dette nivået forutsetter vi bare at det er mulig å skille mellom de ulike verdiene på variabelen, uten at forholdet mellom verdiene er av betydning. Kravet er med andre ord atskillbare kategorier, uten at disse nødvendigvis står i noen spesiell orden i forhold til hverandre, og uten at vi trenger å kjenne til noe størrelsesforhold eller avstanden mellom hver enkelt verdi.

Blant variabler på dette nivået finner vi for eksempel kjønn, nasjonalitet, bosted og en rekke andre. Vi skal være klar over at enkelte variabler som vanligvis er på nominalnivå i visse tilfeller kan opptre som variabler på ordinalnivå. Ett eksempel på dette kan være yrke. Dette er en variabel som vanligvis vil være på ordinalnivå. Imidlertid kan det tenkes at vi gjennomfører en analyse hvor vi legger

inn i yrkesvariabelen ett eller annet mål på de ulike yrkenes prestisje. I slike tilfeller kan yrke opptre som en ordinal variabel. Et liknende eksempel kan bosted (bygd, bygdesenter, tettbebyggelse osv.) være. vanligvis vil det være mest meningsfylt å behandle dette som en variabel på nominalnivå, mens den kan behandles som en variabel på ordinalnivå hvis vi legger inn en forutsetning om urbanitetsgrad.

I analyse av statistiske sammenhenger på nominalnivå vil det ikke være meningsfylt å kommentere noe annet enn selve styrken av sammenhengen. Når vi ikke har noen logisk rangorden mellom verdiene, og de i utgangspunktet kan stå i tilfeldig rekkefølge, kan vi ikke snakke om positive eller negative sammenhenger (sammenhengens retning), eller om sammenhengenes art utover dette (sammenhengens form, uttrykt f.eks. som proporsjonalitet o.l.).

Når det gjelder mål for sentraltendens er modus og modalprosent de mest aktuelle målene for nominale variabler. Mål for sentraltendens som median og aritmetisk gjennomsnitt kan ikke brukes, heller ikke spredningsmålene variasjonsbredde, kvartildifferanse, gjennomsnittsavvik, varians og standardavvik.

- Variabler på ordinalnivå

Det er allerede antydnet at en variabel på ordinalnivå er en variabel hvor vi på en meningsfylt måte kan snakke om en logisk rangordning mellom verdiene på variabelen. Slike rangordninger kan f.eks. være slike som uttrykker grader av enighet, grader av positive og negative holdninger osv.

For at en variabel skal kunne sies å være på ordinalnivå trenger vi ikke å kjenne den eksakte avstanden mellom verdiene, og det er ikke noe krav om at det skal kunne være mulig å si noe om den relative styrken eller graden mellom dem. For variabler på ordinalnivå vil det være mulig å snakke om positive og negative sammenhenger. En positiv sammenheng vil vi oppnå når vi avdekker en situasjon hvor økning av verdiene på den ene variabelen faller sammen med økning av verdiene på den andre variabelen, eller når begge reduseres. En negativ sammenheng vil vi få når økning på en variabel finner sted sammen med svekking eller reduksjon på den andre. Når det gjelder ordinale variabler kan vi ikke si noe om hvor sterk endring på den ene variabelen som henger sammen med hvor sterk endring på den andre.

Vanlige variabler på ordinalnivå er for eksempel utdanningsnivå, uttrykt gjennom verdier som lav, middels og høy. Mange intervall-/forholdstallsvariabler blir ofte ordinalvariabler når vi koder dem om: Inntekt i antall kroner omkodet til inntektsgrupper, alder omkodet til aldersgrupper osv.

Om slike omkodete variabler ender opp på ordinalnivå, eller om de forblir på intervall-/forholdstallsnivå, avhenger av hvilke verdiklasser vi benytter.

Uttrykker vi aldersgruppe som ung, middelaldrende og gammel får vi en ordinal variabel, men hvis vi fremdeles angir de enkelte verdiene slik at vi kjenner avstanden mellom dem, vil variabelen (under visse omstendigheter) kunne brukes som en intervall-/forholdstallsvariabel.

Det er ikke meningsfylt å beregne aritmetisk gjennomsnitt eller standardavvik for en variabel på ordinalnivå. Median er et mye brukt mål for sentraltendens som forutsetter minst ordinalnivå. Modus kan også brukes. Av mål for spredning er modalprosent det som uten videre kan brukes, variasjonsbredde og kvartildifferanse kan også benyttes under visse forutsetninger (f.eks. ved holdningsskalaer selv om vi strengt tatt ikke bestandig kjenner avstanden mellom de forskjellige verdiene).

- Variabler på intervall-/forholdstallsnivå

Når vi snakker om variabler på intervall-/forholdsnivå snakker vi om variabler hvor den eksakte avstanden mellom verdiene på variabelen er kjent. Det vil si at vi kan forholde verdiene til hverandre og omtale en verdi som så og så mye større enn andre verdier. Noen av disse variablene er også kontinuerlige (i motsetning til diskrete) - dvs. at verdiene i utgangspunktet kan plassere seg hvor som helst langs en lang skala (eks. inntekt, fra kr. 0,- til kr.).

Når det gjelder valg av statistiske teknikker ligger spekteret åpent når vi arbeider med variabler på dette nivået. Det vil si at målenivået ikke er den avgjørende begrensende faktoren for valget, noe som åpner for at hvilke spørsmål vi vil ha svar på blir (som regel og i praksis) det eneste vi trenger å forholde oss til.

De samme teknikkene kan brukes for variabler både på intervall- og forholdstallsnivå. Skillet mellom dem gir sjelden praktiske konsekvenser. Skal vi si noe om dette skillet ligger det i at på intervalltallsnivå er avstanden mellom verdiene kjent, mens vi ikke har noe fast nullpunkt å forholde oss til. Vi kan derfor si noe om avstanden mellom verdiene, men ikke om proporsjonene mellom dem. På forholdstallsnivå er dette nullpunktet kjent - det betyr at vi kan uttrykke forholdet mellom verdiene ved hjelp av begreper som 'dobbelt så stor' osv.

2. Sannsynlighetsberegning

Statistikk dreier seg om sannsynligheter, og ikke absolutte sannheter. Gjennom ulike statistiske teknikker beregner vi sannsynligheten for at noe skal inntreffe eller opptre gitt visse forutsetninger: Vi beregner sannsynligheten for at konklusjonene våre er riktige eller gale gitt at resten av undersøkelsen holder mål (spørsmålsformulering, utvalgsmetoder osv.), vi beregner sannsynligheten for at

sammenhenger vi har avdekket i en undersøkelse stemmer med virkeligheten, vi beregner sannsynligheten for at visse personer eller grupper gjør spesielle valg ut fra visse karakteristika vi kjenner til.

Svært ofte dreier statistikk seg om å vurdere de konkrete resultatene i en undersøkelse opp mot det som rent matematisk var mest sannsynlig. Når resultatene avviker fra den beregnede sannsynligheten tyder det på at det kan være en sammenheng på et eller annet nivå. 'Noe' har virket inn slik at matematiske sannsynlighetsutregninger ikke kan brukes til å forutsi resultatene. Med dette i bakhånden går vi så videre og prøver å finne ut hva det er som virker inn, og beregner sannsynligheten for at det er spesielle faktorer som vi identifiserer.

Teorien rundt normalfordelingskurven må ut fra dette ikke først og fremst sees som en oppsummering av hvordan 'noe', f.eks. karakterer, faktisk fordeles eller bør fordeles i en gruppe. Teorien bygger på en antakelse om at hvis gruppen er stor nok, og ingen spesielle faktorer spiller inn, så vil vi få en tendens til at karakterer fordeles i henhold til denne kurven: Flest vil oppnå de midterste karakterene, mens antallet som oppnår de øvrige karakterene - de dårligste og de beste - vil avta gradvis og like mye på hver side, jo lenger vi kommer fra midtpunktet på karakterskalaen. Hvis en konkret undersøkelse i en relativt stor gruppe (den må være relativt stor for at vi skal kunne redusere sannsynligheten for at tilfeldigheter ved utvalget skal spille inn) viser at karakterene ikke fordeler seg etter normalfordelingskurven, vil vi ha grunn til å tro at det finnes spesielle faktorer som spiller inn: Er elevene spesielt dyktige, er det en god lærer, er pensumkravet lavt, er sensor snill?

Vi skal ikke gå mye inn på beregning av sannsynligheter, men må innom noen hovedtrekk. Husk: Begrepet sannsynlighet har ikke nødvendigvis noe med virkeligheten å gjøre; det må i denne sammenhengen oppfattes som et rent matematisk begrep.

SANNSYNLIGHET OG PROPORSJONER

Hovedtanken i sannsynlighetsberegning er at den matematiske sannsynligheten for at noe skal inntreffe, svarer til hvor ofte dette inntreffer i det hele tatt. Sannsynlighet blir derfor et 'speilbilde av proporsjoner' eller andeler i en gitt gruppe.

Hvis vi skal trekke tilfeldig en person fra en gruppe mennesker hvor kjønnsfordelingen er 60 kvinner og 40 menn, er sannsynligheten for at vi trekker en kvinne $60/100$, og sannsynligheten for at vi trekker en mann $40/100$. Regner vi ut brøkene får vi en sannsynlighet på 0,6 for at vi trekker en kvinne og 0,4 for at vi trekker en mann. Sannsynligheten kan også uttrykkes som prosenttall: 60 % sannsynlighet for at vi trekker en kvinne, og 40 % sannsynlighet for at vi trekker en mann.

Hvis gruppen utelukkende består av norske statsborgere, er 'sannsynligheten' for å trekke en norsk statsborger 100 %, eller 1. 'Sannsynligheten' for å trekke en svenske blir 0. En sannsynlighet lik 1 er det samme som sikkerhet.

SANNSYNLIGHET VED UAVHENGIGHET

Når vi i denne sammenhengen snakker om uavhengighet, betyr det at resultatet fra en 'trekking' ikke påvirker sannsynligheten i de neste trekkingene. Hvis vi i eksemplet ovenfor har trukket 10 personer - 8 kvinner og 2 menn - vil sannsynlighetsfordelingen i resten av utvalget forandre seg:

$$P_{(kvinne)} = (60-8)/(100-10) = 52/90 = 0,58$$

$$P_{(mann)} = (40-2)/(100-10) = 38/90 = 0,42.$$

Hvis vi etter å ha trukket de 10 lar de vende tilbake til gruppa før neste trekking, ville sannsynligheten vært den samme.

Hvis vi i tillegg til kjønnsfordelingen også vet noe om alderen til personene i gruppa, men ingenting om sammenhengen mellom alder og kjønn, kan vi beregne en uavhengig sannsynlighet for kombinasjonen mellom disse faktorene. La oss si at 30 personer er under 30 år og 70 personer er over 30 år. Vi står da igjen med en sannsynlighet for å trekke en person under 30 år på 0,3, og sannsynligheten for å trekke en person over 30 år er 0,7. Sannsynligheten er hele tiden

$$\frac{(\text{Antall personer som oppfyller kriteriet})}{(\text{Antall personer totalt})}$$

eller

$$\frac{K}{N}$$

Vi har nå 4 forskjellige sannsynlighetsangivelser:

$$\begin{array}{ll} P_{(kvinne)} & = 0,6 \\ P_{(mann)} & = 0,4 \\ P_{(under 30 \text{ år})} & = 0,3 \\ P_{(over 30 \text{ år})} & = 0,7 \end{array}$$

Hvordan skal vi så finne ut sannsynligheten for å trekke en person med en spesielt angitt kombinasjon av disse faktorene? Forutsatt at det ikke er noen spesiell sammenheng mellom kjønn og alder, finner vi dette som et produkt av de to delsannsynlighetene:

$P_{\text{(kvinne under 30 år)}}$	$= P_{\text{(kvinne)}}$	$\times P_{\text{(under 30 år)}}$	$= 0,6 \times 0,3 = 0,18$
$P_{\text{(kvinne over 30 år)}}$	$= P_{\text{(kvinne)}}$	$\times P_{\text{(over 30 år)}}$	$= 0,6 \times 0,7 = 0,42$
$P_{\text{(mann under 30 år)}}$	$= P_{\text{(mann)}}$	$\times P_{\text{(under 30 år)}}$	$= 0,4 \times 0,3 = 0,12$
$P_{\text{(mann over 30 år)}}$	$= P_{\text{(mann)}}$	$\times P_{\text{(over 30 år)}}$	$= 0,4 \times 0,7 = 0,28$

Dette er den matematiske sannsynligheten for hver av de mulige utfallene basert på uavhengighet mellom de to faktorene. Uansett hvem vi trekker, vil det være en person som faller inn i en av disse gruppene. Derfor: Summen av sannsynligheter for de enkelte gruppene må bli lik 1, fordi 'sannsynligheten for å trekke en person hvis vi trekker en person' er lik 1, og andre personer enn en av disse kan vi ikke trekke!

Skal vi lage en fullstendig formel for å beregne sannsynligheten for en gitt kombinasjon, tar vi utgangspunkt i K/N , og får

$$\frac{K_1}{N} \times \frac{K_2}{N} = P(K_1 \text{ og } K_2)$$

der K_1 og K_2 er antallet personer som oppfyller kriteriet vi er ute etter for de to egenskapene (kjønn og alder) - for eksempel antall kvinner og antall over 30 år. Høres dette kjent ut? Ja - det er omtrent det samme som vi gjør når vi beregner f_u (uavhengig frekvens) i forbindelse med kji-kvadratet. Den eneste forskjellen er at vi her skal finne en sannsynlighet uttrykt i prosent eller som et tall mellom 0 og 1, mens vi i beregning av uavhengig frekvens i kjikvadrat-testen skal ha antallet i hver gruppe. Derfor deler vi her på N to ganger, mens vi deler på N en gang ved beregning av f_u .

Med utgangspunkt i eksempelet ovenfor har vi ulike typer situasjoner som vi skal beregne sannsynligheten for. Disse kan beskrives som

- både/og-situasjoner - både A og B inntreffer
- enten/eller-situasjoner - A eller B inntreffer, men ikke i kombinasjon
- og/eller-situasjoner - A og/eller B inntreffer.

Ved beregning av sannsynligheten i en 'både/og-situasjon' som ovenfor, altså når vi skal se på sannsynligheten for at visse forhold opptrer i kombinasjon, skal de enkelte delsannsynlighetene multipliseres.

Vi må passe på at vi hele tiden har klart for oss skillene mellom variabler og verdier: En person eller en enhet i undersøkelsen kan bare ha en verdi på hver variabel (dette svarer til definisjonen på hva en variabel er). Sannsynligheten for en verdi på en variabel kan multipliseres med sannsynligheten for en verdi på en annen variabel, men vi kan lage logiske umuligheter hvis vi multipliserer sannsynligheten for to ulike verdier på samme variabel.

Hvis vi skal trekke ut en person fra utvalget, kan vi selvfølgelig ikke bruke formelen $P(\text{mann og kvinne}) = 0,4 \times 0,6 = 0,24$. Det er ikke 24 % sannsynlighet for at den personen vi trekker ut er både mann og kvinne. Trekker vi to personer, stiller saken seg annerledes. Vi må imidlertid huske på at det er flere mulige utfall når vi trekker to personer:

Person A	Person B
Kvinne	Kvinne
Mann	Mann
Kvinne	Mann
Mann	Kvinne

Sannsynlighetsfordelingen for disse utfallene blir

$P_{(2 \text{ kvinner})}$	$= 0,6 \times 0,6 = 0,36$
$P_{(2 \text{ menn})}$	$= 0,4 \times 0,4 = 0,16$
$P_{(A \text{ er kvinne, B er mann})}$	$= 0,6 \times 0,4 = 0,24$
$P_{(A \text{ er mann, B er kvinne})}$	$= 0,4 \times 0,6 = 0,24$

Dette er de fire mulige utfallene, og den sammenlagte sannsynligheten blir derfor 1. Total sannsynlighet for å trekke en mann og en kvinne er summen av delsannsynlighetene for utfall som oppfyller det ønskede resultatet: $0,24 + 0,24 = 0,48$. Sannsynligheten for å få et resultat som ikke består av en mann og en kvinne er $0,36 + 0,16 = 0,52$, eller $1 - 0,48 = 0,52$.

Når antallet mulige utfall øker, øker også problemene med å sette opp en komplett oversikt over dem. Det finnes en spesiell formel som hjelper oss å beregne mulighetene i slike tilfeller, men vi skal ikke komme inn på den her.

Dette leder oss inn på neste situasjon: Sannsynligheten i en enten/eller-situasjon. I det ovenstående eksemplet skal vi finne den totale sannsynligheten for å trekke ut en kvinne og en mann, men vi bryr oss ikke om vi trekker kvinnen eller mannen først. Med andre ord er det sannsynligheten for å trekke først en kvinne og så en mann, eller først en mann og så en kvinne vi skal finne. Den samlede sannsynligheten finner vi ved å addere de to delsannsynlighetene. Hvis vi skal finne sannsynligheten for å trekke ut enten en kvinne under 30 år, eller en mann uansett alder, men ikke kvinner over 30 år, må vi gå fram på samme måte: Finne de ulike mulige utfallene, og beregne delsannsynlighetene (og holde tunga rett i munnen):

$P_{(\text{kvinne over 30 år})}$	$= 0,6 \times 0,7 = 0,42$
$P_{(\text{kvinne under 30 år})}$	$= 0,6 \times 0,3 = 0,18$
$P_{(\text{mann over 30 år})}$	$= 0,4 \times 0,7 = 0,28$
$P_{(\text{mann under 30 år})}$	$= 0,4 \times 0,3 = 0,12$

Vi er her ute etter den totale sannsynligheten for å trekke en person som enten er kvinne under 30 år, eller er en mann. Vi adderer de aktuelle delsannsynlighetene $0,18 + 0,28 + 0,12 = 0,58$. I og med at det bare er ett utfall vi ikke er interessert i, nemlig kvinne over 30 år, kan vi også sette opp regnestykket som $1 - 0,42 = 0,58$ (den totale sannsynligheten for alle mulige utfall - summen av delsannsynlighetene for ikke ønskede resultater).

3. Effekt og samspill

Når vi skal beregne effekt ser vi på hvilken innvirkning ulike verdier på en eller flere variabler har på en eller flere andre variabler. Vi måler altså samvariasjonen mellom ulike verdier på ulike variabler, og på en slik måte at dette kan inngå i en forklaringsmodell for å forklare variasjon og spredning mellom de ulike verdiene på enkelte av variablene.

Måling av effekt og samspill kan gjøres på alle målenivåer, men det vil bestandig være en fordel med ikke alt for store tabeller. Ofte vil det derfor være gunstig å omkode til færre verdier, dersom det i utgangspunktet er svært mange verdier på en eller flere av variablene.

De variablene som vi mener virker inn på andre variabler kaller vi uavhengige variabler (årsaks- eller forklaringsvariabler), mens de variablene som blir påvirket kalles for avhengige variabler (effektvariabler). Hva som er avhengige variabler og hva som er uavhengige er kontekstavhengig - det vil si at vi må vurdere dette ut fra sammenhengen. Skal vi for eksempel se på sammenhengen mellom kjønn, utdanning og inntekt er det naturlig å benytte kjønn og utdanning som uavhengige variabler, og forklare hvordan disse virker inn på inntektsfordelingen. I andre sammenhenger er det kanskje utdanningsulikheter som skal forklares, dette vil da bli den avhengige variabelen, mens f.eks. bosted, kjønn og foreldres utdanning kan være uavhengige variabler.

Måling av samspill betyr at vi måler i hvilken grad effekten av den ene uavhengige variabelen betinges av hvilken verdi den har på den andre uavhengige variabelen (for eksempel ved at vi måler forskjellen i betydning som utdanning har for de to kjønnene når vi skal forklare inntektsforskjeller).

Enkelte rettesnorer gjelder for valg av uavhengige variabler. En del 'biologiske' variabler av typen kjønn og alder vil som regel være forklarings- eller uavhengige variabler (da tenker vi ikke på situasjoner der vi skal forklare aldersfordelinger i en gitt sammenheng). Rekkefølge i tid vil ofte være til god hjelp - som regel vil det som inntreffer først være det som påvirker det som inntreffer sist (dersom det i det hele tatt er noen sammenheng). Imidlertid kan dette også kompliseres noe, ved at forutsigelser om konsekvensene av noe en vet skal skje kan påvirke handlingsvalg før dette faktisk inntreffer (dette gjelder vel blant annet deler av norsk næringsliv som i noen år har lagt opp nye strategier for

å være forberedt når EFs indre marked blir iverksatt).

Ved beregning av effekt og samspill vil dette vanligvis inngå i en modell om årsakssammenhenger, enten denne er klart uttrykt eller ikke. Den viktigste rettesnoren er en logisk oppbygging av modellen med tanke på årsaks- og virkningsmekanismer. Her er det viktig å huske på at en slik modell kan ha flere trinn, med mellomliggende variabler som i første omgang påvirkes av noen uavhengige (og derved i første omgang betraktes som avhengige variabler), men som i neste trinn igjen påvirker andre variabler (og derved må betraktes som uavhengige).

EFFEKT MED ÉN UAVHENGIG VARIABEL (BIVARIAT ANALYSE)

Den enkleste analysen av effekt får vi når vi skal analysere bivariate sammenhenger, med en uavhengig og en avhengig variabel.

a) Inntektsfordeling; absolutte tall

INNTEKT	Menn	Kvinner	Totalt
< 150 000	37	95	132
>= 150 000	191	88	279
	228	183	411

Når vi skal beregne effekten som kjønn har på inntekt, må vi regne om de absolutte tallene til prosenttall. Vi må vite noe om andelen av hver gruppe.

b) Inntektsfordeling; prosenttall

INNTEKT	Menn	Kvinner	Totalt
< 150 000	16	52	32
>= 150 000	84	48	68
	100	100	100

Effekten av kjønn for å forklare inntektsfordeling beregner vi så ved å beregne prosentdifferansen mellom menn og kvinner for en av inntektsgruppene. Vi trenger i dette tilfelle - når vi har bare to verdier på den avhengige variabelen - bare å beregne én prosentdifferanse, fordi prosentdifferansen på den ene linja nødvendigvis må være et speilbilde av prosentdifferansen på den andre linja.

Hadde vi hatt flere verdier på den avhengige variabelen, burde vi ha regnet et snitt (dvs. å legge sammen absoluttverdien av prosentdifferansen for hver linje,

og delt på antall linjer), eller vi kunne ha angitt prosentdifferansen mellom kjønnene for å se hvilken effekt kjønn har for å forklare lav inntekt, midlere inntekt og høy inntekt.

I vårt tilfelle ser vi at prosentdifferansen mellom kjønnene blir på

$$16 - 52 = -36$$

Dette forteller oss at det er 36 % mindre sjanse for at en mann i dette utvalget skal ha lav inntekt enn at en kvinne skal ha det. Det er 36 % større sjanse for at en mann skal ha høy inntekt enn at en kvinne skal ha det. Tabellen forteller oss også at det er omtrent dobbelt så mange i utvalget som har relativt høy inntekt enn de som har relativt lav inntekt. I og med at variabelen er på ordinalnivå, og vi ikke vet hvordan hver enkelt plasserer seg innafor gruppene, kan vi ikke uttrykke noe om lønnsforskjellen mellom kvinner og menn utover dette.

Vi merker oss også at for å beregne effekt i slike tabeller prosentuerer vi hver av gruppene på den uavhengige variabelen for seg, slik at vi får andelen av kvinner og andelen av menn med lav og høy inntekt. Det er ingen lov, men oftest plasserer vi også den uavhengige variabelen på toppen av tabellen, og den avhengige variabelen vertikalt. Dette betyr at vi må ha horisontal prosentuering.

At vi har beregnet effekten av kjønn på denne måten, gir oss ikke grunnlag for at det er kjønn som er den direkte årsaken til såvidt store inntektsforskjeller. For det første kan det være at effekten av kjønn tilfeldigvis faller sammen med effekten av en annen, ukjent variabel som vi ikke kjenner (tilfeldig sammenheng). For det andre kan det tenkes at kjønn først og fremst har en indirekte effekt på inntektsforskjeller (f.eks. ved at kjønn påvirker utdanningsvalg som så igjen er det som virker inn på inntektsforskjellene).

I andre tilfeller, hvor det ikke er såvidt klart hva som er den uavhengige og hva som er den avhengige variabelen, kan det også tenkes at vi lager en analyse basert på en påvirkningsretning, mens den virkelige påvirkningen eller effekten faktisk går i motsatt retning. Endelig ville det i andre sammenhenger vært tenkelig at det ikke eksisterer noe årsaksforhold mellom den uavhengige og den avhengige variabelen i analysen, men at begge to i reliteten påvirkes av en tredje, ukjent variabel (spuriøs sammenheng). I dette siste tilfellet vil det si at vårt valg av uavhengig variabel ikke er korrekt; at vi i realiteten har valgt en avhengig variabel og satt den opp som uavhengig.

Samspill vil det være meningsløst å beregne i tabeller med bare en uavhengig variabel, da samspill dreier seg om hvordan to eller flere uavhengige variabler virker sammen.

EFFEKT OG SAMSPILL I MULTIVARIAT ANALYSE

I multivariat analyse av effekt og samspill viser vi hvordan to eller flere variabler har effekt på den eller de avhengige variablene.

I dette tilfelle skal vi se på hvordan kjønn og utdanning, hver for seg og sammen, påvirker inntektsnivået. Vi starter med en prosentuert tabell, hvor vi plasserer variablene slik at vi får en kolonne for hver kombinasjon av verdiene for de to uavhengige variablene:

a) Inntekt; betinget av kjønn og utdanning. Prosenttall

KJØNN UTDANNING	Menn		Kvinner		Totalt
	<= 12 år	>12 år	<=12 år	>12 år	
INNTEKT					
<=150 000	33	2	76	31	32
>150 000	67	98	24	69	68
	100	100	100	100	100
(N=)	(105)	(123)	(85)	(98)	(411)

Av denne tabellen ser vi for det første at det er betydelige forskjeller mellom andelen av de ulike gruppene som har henholdsvis lav og høy inntekt. Sterkest slår det oss at andelen av menn med høy eller lang utdanning som har lav inntekt er meget liten (2 %), og at andelen av kvinner med lav utdanning og som har høy inntekt også er meget lav. Vi får et inntrykk av at både kjønn og utdanning spiller inn. Samtidig ser vi at det er en viss forskjell i størrelsen på de enkelte gruppene.

- 'Hernes-metoden' for beregning av effekter

Hvis vi velger å se bort fra ulikhetene i andel som hver gruppe utgjør av hele utvalget, kan vi velge Hernes-metoden for å beregne effekt og samspill. Valg av Hernes-metoden betyr at vi lar effektene for hver gruppe veie like tungt i beregningen av gjennomsnittseffektene - noe som innebærer at gjennomsnittseffekten strengt tatt ikke blir matematisk korrekt. Dersom størrelsesforskjellen mellom gruppene er store vil dette kunne gi seg betydelige utslag. (Hvis vi i vårt tilfelle velger å se bort fra dette og altså velger Hernes-metoden, vil jeg betegne det som 'på kanten' av hva som vil passere.) Dersom vi velger å ta hensyn til ulikhetene i gruppestørrelse, velger vi 'Hellevik-metoden' for beregning av gjennomsnittseffekt, da denne metoden innebærer at vi veier gjennomsnittet.

Også når vi beregner effekt og samspill i multivariate tabeller må vi velge ut en verdi på den avhengige variabelen som vi skal forklare av gangen. Også her

gjelder at med mer enn to verdier på den avhengige variabelen kan det oppstå systematiske skjevheter som gjør at vi må vurdere om vi må gjennomføre beregningen for flere av verdiene.

Med to verdier slipper vi dette, og i utgangspunktet vil vi få samme resultat uansett hvilken verdi vi velger ut på den avhengige variabelen. Vi kan imidlertid som en hjelp til oss selv bruke følgende huskeregel når vi skal velge verdi som skal forklares og den avhengige variabelen er på ordinalnivå eller høyere: Når vi tenker på effekt tenker vi på noe som fører til at noe øker. Språkbruken vår vil være rettet inn på dette. Det vil derfor bestandig være lettere å forklare med ord effekter på den høyeste av verdiene på den avhengige variabelen. I vårt tilfelle blir dette høy inntekt - når vi tenker på faktorer som påvirker inntekt tenker vi først og fremst på hva som fører til høy inntekt. (Denne hjelperegelen er kun til hjelp, og må ikke oppfattes som en lov.)

En annen huskeregel uansett målenivå: Når vi skal konstruere hjelpetabellen vi trenger for beregning av effekt og samspill, skal vi bestandig velge ut den gruppen hvor effektene er størst og la denne gruppen fylle ut plassen øverst til venstre i tabellen (dette er vesentlig for fortegnet vi vil få ved utregning av samspill). Gruppen hvor effektene er størst finner vi ved å velge ut den gruppen som har den største prosentandelen 'av' det vi skal forklare. Her vil dette si menn med høy utdanning, hvor hele 98 % har høy inntekt.

b) Forklaring av høy inntekt

	KJØNN		diff.
	Menn	Kvinner	
UTDANNING			
> 12 år	98	69	29 (d ₁)
≤12 år	67	24	43 (d ₂)
diff.	31 (d ₁)	45 (d ₂)	

I denne tabellen ser vi at når vi skal forklare høy inntekt, er effekten av kjønn for de med høy utdanning 29, og effekten av kjønn for de med lav utdanning er 43. Vi ser også at effekten av utdanning for menn er 31, og effekten av utdanning for kvinner er 45.

I avsnittet om samspill skal vi komme tilbake til hva dette forteller oss, men i og med at beregning av samspill er lik for de to metodene skal vi gjøre ferdig beregning av effekt først.

Formelen for beregning av gjennomsnittseffektene etter Hernes-metoden (i en 2x2-tabell) er

$$E = 1/2 (d_1 + d_2)$$

det vil si at vi legger sammen effektene for den aktuelle variabelen og dividerer på to. Dersom vi har flere verdier på den uavhengige variabelen må vi dele på antallet prosentdifferanser (ant. verdier) for å finne gjennomsnittet.

Ut fra dette kan vi beregne effekten av kjønn som

$$E_K = 1/2 (29 + 43) = 1/2 (72) = 36$$

og effekten av utdanning på tilsvarende måte:

$$E_U = 1/2 (31 + 45) = 1/2 (76) = 38.$$

Vi leser ut av tabellen at det er menn som 'får' effekt av kjønn med tanke på høy inntekt, og det er de med høy utdanning som 'får' effekt av utdanning. Dette følger av at kvinnene i større grad får lav inntekt, noe som også gjelder for de med lav utdanning. Det er med andre ord verdiene mann og høy utdanning som gir effekter i vårt tilfelle.

Ettersom vi bestandig skal føre opp den største prosentandelen øverst til venstre i hjelpetabellen, vil de verdiene som gir effekt være de verdiene som svarer til denne ruta - altså den første kolonnens verdi og den første linjas verdi.

Menn har med andre ord 36 % større sjanse til å oppnå høy inntekt enn kvinner, og personer med høy utdanning har 38 % større sjanser til å oppnå høy inntekt enn personer med lav utdanning. Dette er høge effekter, og effekten av utdanning er altså noe høyere enn effekten av kjønn.

- 'Hellevik-metoden' for beregning av effekter

'Hellevik-metoden' for beregning av effekter atskiller seg som nevnt fra 'Hernes-metoden' ved at vi her tar hensyn til de enkelte gruppenes størrelse. Effektene regnes derved ut som et veid gjennomsnitt slik at de større gruppene som er involvert trekker gjennomsnittet mere i 'sin' retning enn de mindre gruppene.

Beregning av effekt for hver enkelt gruppe (som når vi beregner effekt av utdanning for menn og effekt av utdanning for kvinner) gjøres på samme måten som ved Hernes-metoden. Det er altså de samme prosentdifferansene (d_1 og d_2) som vi skal sette inn i formelen. Dette skyldes at så lenge vi skal beregne effektene for hovedkategoriene hver for seg trenger vi ikke å ta hensyn til størrelsesforholdet. Dette behovet melder seg først når vi skal beregne gjennomsnittseffekter.

Forutsetningen for å bruke Hellevik-metoden er altså at vi kjenner andelen som hver gruppe utgjør av hele utvalget. I tabell a) i dette kapitlet har vi ikke regnet ut den prosentvise andelen, men vi har n-verdien for hver enkelt gruppe:

Menn m. lav utdanning	105	Menn	105 + 123 = 228
Menn m. h��g utdanning	123	Kvinner	85 + 98 = 183
Kvinner m. lav utdanning	85	Lav utdanning	105 + 85 = 190
Kvinner m. h��g utdanning	98	H��g utdanning	123 + 98 = 221

Med utgangspunkt i dette kan vi s   beregne gjennomsnittseffektene etter formelen

$$E = \frac{n_1}{N} (d_1) + \frac{n_2}{N} (d_2)$$

hvor n_1 og n_2 er antallet personer i de gruppene som effekten gjelder for.

Uttrykker vi dette i ord f  r vi ogs   gjennomsnittseffekten av utdanning ved    ta

$$E_U = \frac{\text{Ant. menn}}{N} (\text{eff. av utd. for menn}) + \frac{\text{Ant. kvinner}}{N} (\text{eff. av utd. for kvinner})$$

N  r vi setter de riktige tallene inn i disse formlene f  r vi ogs   f  lgende beregning av gjennomsnittseffektene (vi tar med oss prosentdifferansene fra tab. b) p   side 22):

$$\begin{aligned} E_{\text{UTDANNING FOR MENN}} &= 31 \\ E_{\text{UTDANNING FOR KVINNER}} &= 45 \\ E_{\text{KJ  NN FOR DE MED LAV UTDANNING}} &= 43 \\ E_{\text{KJ  NN FOR DE MED H  G UTDANNING}} &= 29 \end{aligned}$$

$$E_U = \frac{228}{411} (31) + \frac{183}{411} (45) = 0,555 (31) + 0,445 (45) = 17,2 + 20,0 = \underline{37,2}$$

$$E_K = \frac{190}{411} (43) + \frac{221}{411} (29) = 0,462 (43) + 0,538 (29) = 19,9 + 15,6 = \underline{35,5}$$

Vi ser at dette resultatet bare i liten grad avviker fra resultatet vi fikk ved    bruke Hernes-metoden. Forskjellen i gruppenes st  rrelse ga ogs   ikke s  rlig utslag i dette tilfellet.

- Beregning av samspill

Effekten av kj  nn er ogs   forskjellig for de to utdanningsgruppene, og effekten av utdanning er forskjellig for de to kj  nnene. Av dette kan vi umiddelbart sl   fast at det eksisterer samspill eller statistisk interaksjon i tabellen, fordi samspill defineres som det som oppst  r n  r effekten av den ene uavhengige variabelen (f.eks. kj  nn) er betinget av hvilken verdi enheten har p   den andre uavhengige variabelen (f.eks. utdanning).

En annen m  te    si dette p   er at effekten av en uavhengig variabel varierer etter hvilken verdi vi holder konstant p   den andre uavhengige variabelen.

Vi snakker om positivt samspill når effekten av de to uavhengige variablene virker sammen på en slik måte at de forsterker hverandre. Negativt samspill får vi når effekten av de to uavhengige variablene virker sammen på en slik måte at de svekker hverandre. I positivt samspill får vi altså en tilleggseffekt når effektene virker sammen, mens vi i negativt samspill får en redusert effekt i forhold til summen av de to effektene hver for seg.

I vårt eksempel er det effekten av det å være mann og effekten av det å ha høy utdanning som forklarer høy inntekt (i den grad vi har forklart det; jfr. neste underkapittel). Vi kan altså tenke oss at vi ville hatt positivt samspill mellom de uavhengige variablene hvis forholdet mellom disse hadde vært slik at en mann (som får uttelling for det kjønnet han tilhører) også tilhører den gruppa som får mest uttelling for høy utdanning. Da ville effekten av kjønn gitt en effekt, effekten av utdanning ville gitt en effekt, og i tillegg ville vi fått effekten av at kjønn plasserte ham i den gruppa som gir mest uttelling for utdanning.

Nå vet vi at dette ikke er tilfelle, fordi det ikke er menn, men kvinner som får størst uttelling for utdanning. Det at han er mann har derfor plassert ham i den gruppa som får minst uttelling for høy utdanning, noe som betyr at effekten av kjønn og effekten av høy utdanning blir noe svekket: Vi må 'legge til' effekten av at kjønn har plassert ham i den gruppa som får minst ut av høy utdanning, og ettersom dette er en negativ effekt (og et negativt tall) vil det i praksis si at vi trekker fra absoluttverdien av samspillseffekten. Dette betyr at vi i vårt tilfelle har negativt samspill.

Vi skal se på hvordan vi beregner dette. Utgangspunktet er de samme prosentdifferansene som vi fant da vi skulle beregne effekter. Framgangsmåten ved beregning av samspill er den samme enten vi velger å bruke Hernes- eller Hellevik-metoden. Vi skal bruke de opprinnelige prosentdifferansene (effektene for hver enkelt gruppe), og trenger derfor ikke å tenke på om gjennomsnittseffekten er veiet eller ikke.

Vi tar utgangspunkt i tab. b) fra side 22:

	KJØNN		diff.
	Menn	Kvinner	
UTDANNING			
> 12 år	98	69	29 (d ₁)
<=12 år	67	24	43 (d ₂)
diff.	31 (d ₁)	45 (d ₂)	

Når vi beregner samspill gjør vi dette ut fra ett av settene med prosentdifferanser. Det er det samme hvilket sett vi velger - de skal uansett gi samme resultat. Det som

er av vesentlig betydning er, som tidligere nevnt, at vi har ført opp den gruppa som 'nyter godt av' den største totaleffekten (dvs. det høyeste prosenttallet) øverst til venstre i tabellen. Dette er viktig fordi det er forskjell i tolkningen av negativt og positivt samspill, og ettersom fortegnet vil avhenge av rekkefølgen av verdiene på de to uavhengige variablene i tabellen så må vi ha en fast regel for hvilken rekkefølge vi skal velge.

Formelen for samspill er

$$S = 1/2 (d_1 - d_2)$$

I vårt tilfelle kan vi altså beregne samspillet til å bli

$$S = 1/2 (29 - 43) = 1/2 (-14) = -7$$

eller

$$S = 1/2 (31 - 45) = 1/2 (-14) = -7$$

Med andre ord et negativt samspill mellom de uavhengige variablene, men et relativt svakt negativt samspill. Dette stemmer også med det vi forutså ut fra vurderinga av hvordan effektene fordelte seg.

Når vi skal tolke dette samspillet må vi sammenholde dette med det vi vet om effekten av de uavhengige variablene hver for seg. Gjennomsnittseffekten av kjønn er ut fra Hellevik-metoden, som gir det matematisk mest korrekte resultatet) 35,5, og gjennomsnittseffekten av utdanning er 37,2. Med Hernes-metoden er de tilsvarende tallene 36 og 38. Dersom det ikke hadde vært hverken positivt eller negativt samspill mellom disse variablene kunne vi ha beregnet den samlede effekten av dem til å være

$$35,5 + 37,2 = 72,7$$

eller

$$36 + 38 = 74$$

Imidlertid vet vi at vi har en samspillseffekt, og at denne er negativ. Den samlede effekten av de to variablene (og den samlede effekten opptrer når variablene 'spiller sammen') blir derfor

$$35,5 + 37,2 + (-7) = 72,7 - 7 = 65,7$$

eller

$$36 + 38 + (-7) = 74 - 7 = 67$$

Vi ser med andre ord at den samlede effekten av de to variablene når de opptrer sammen blir noe redusert enn summen av de to variablenes effekter hver for seg,

ved at vi legger til en negativ samspillseffekt. Hadde samspillet vært positivt ville samspillseffekten kommet i tillegg.

- Beregning av prediksjonseffekt

Vi har nå vært gjennom beregningen av den effekten to uavhengige variabler har på den avhengige variabelen. Denne beregningen forutsetter at vi på forhånd har valgt ut de to forklaringsvariablene. Den effekten vi finner av de to variablene kan forklare en del av den variasjonen vi har funnet på den avhengige variabelen. Spørsmålet er imidlertid: Kan disse variablene sies å være en tilstrekkelig forklaring på inntektsforskjeller, eller er det andre faktorer som spiller inn. Og hvor sterkt spiller disse faktorene inn?

I første omgang kan vi beregne det som kalles grunneffekt. Beregningen forutsetter at vi har benyttet Hernesmetoden ved beregning av gjennomsnittseffekten av de to variablene. Grunneffekten kan forstås som summen av effekter av andre, ukjente variabler - variabler som ikke er trukket inn i analysen men som allikevel påvirker enhetene i undersøkelsen sin sjanse til å oppnå høy inntekt. Dette er altså snakk om faktorer som mer til stede i virkeligheten, men som vi ennå ikke har identifisert. Hvor stor betydning har så disse faktorene?

Vi finner grunneffekten ved å ta summen av de effektene som vi har identifisert og som gjelder for hvilken som helst av gruppene i undersøkelsen, og trekke denne summen fra den totale prosentandelen i den aktuelle gruppa som har høy inntekt. For menn med høy utdanning (98 % av disse har høy inntekt) vi regnestykket se slik ut

$$G = 98 - (E_U + E_K + S_{UK}) = 98 - (38 + 36 + (-7,0)) = 98 - 67 = 31$$

Den enkleste gruppa å beregne grunneffekten ut fra, er den gruppa hvor ingen av de uavhengige variablene gir effekt. Når vi vet at verdien mann og verdien høy utdanning gir effekt, vet vi også at det er for kvinner med lav utdanning at de to uavhengige variablene ikke gir effekt. Vi finner grunn-effekten ved å ta prosentandelen i gruppa som har høy inntekt, og trekke fra den eneste effekten som spiller inn, nemlig samspillseffekten:

$$G = 24 - (-7) = 31$$

Årsaken til at samspillseffekten spiller inn her, er at de uavhengige variablene her opptrer sammen - riktignok ved at de aktuelle verdiene som gir effekt i denne gruppa er fraværende sammen.

Med utgangspunkt i effektene av de enkelte variablene, samspillseffekten og grunneffekten kan vi så beregne prediksjonseffekten. Prediksjonseffekten er

altså summen av de effektene vi vet virker inn på hver enkelt av de forskjellige gruppene. Summen av disse effektene vil tilsvare den andelen i hver gruppe som har høy inntekt, noe som også svarer til sannsynligheten for at en person, gitt en spesiell kombinasjon av verdier på de uavhengige variablene, faktisk vil ha høy inntekt. Vi kan derfor benytte prediksjonseffekten til å predikere eller angi sannsynligheten for at en person som vi vet kjønn og utdanningsnivået til, vil befinne seg i høginntektsgruppa. Vi kan sette opp beregningen for alle grupper i undersøkelsen.

	KJØNN	
	Menn	Kvinner
UTDANNING > 12 år	$G + E_K + E_U + S_{KU}$ $= 31 + 36 + 38 + (-7)$ $= 105 - 7 = 98$	$G + E_U$ $= 31 + 38 = 69$
≤12 år	$G + E_K$ $= 31 + 36 = 67$	$G + S_{KU}$ $= 31 + (-7) = 24$

Vi ser altså at grunneffekten virker inn på alle gruppene i undersøkelsen, nettopp fordi at dette er en effekt som ligger utenfor eller bakenfor de variablene vi har trukket inn i analysen. Vi ser også at samspillseffekten bare virker når de uavhengige variablene virker sammen, enten ved at de begge gir effekt eller når de begge ikke gir effekt. Når én av dem virker og ikke den andre trenger vi ikke å ta hensyn til samspillseffekten. Ellers ser vi at kvinnene ikke får effekt av kjønn (fordi det jo er mennene som nyter godt av kjønnsulikhetene i denne sammenhengen) og de med lav utdanning får ikke effekt av utdanning (fordi det er høy utdanning som gir høy inntekt).

Vi kan gå videre i tolkninga av denne tabellen og se på hvordan et medlem av en av gruppene kan forbedre sine sjanser til å få høy inntekt. En mann med lav utdanning har for eksempel (i alle fall i teorien som en individuell løsning) muligheten til å skaffe seg høyere utdanning, og derved øke sine sjanser til høy inntekt med en effekt tilsvarende effekten av utdanning og samspillseffekten (som er negativ), fordi de to uavhengige variablene da vil spille sammen. En kvinne med lav utdanning kan øke sine muligheter til å få høyere inntekt på samme måte, her legger vi til effekten av høy utdanning men uten å ta hensyn til samspillseffekten. En kan også gå direkte inn på holdninger i yrkeslivet knyttet til kjønn, og søke å påvirke disse slik at kjønnsforskjellene mellom kjønnene blir redusert.

Utover dette kan vi gå inn i hvilke faktorer som ligger i grunneffekten, og analysere hvorvidt en gjennom å påvirke disse faktorene kan øke muligheten til å få høyere inntekt. I denne grunneffekten vil i vårt tilfelle den generelle

inntektsfordelingen i samfunnet ligge, på samme måte som det i effekten av kjønn ligger samfunnsmessige holdninger og mekanismer.

Hvis vår analyse tar sikte på nettopp å analysere sammenhengen mellom utvalgte uavhengige variabler i forhold til en utvalgt avhengig variabel, vil det ikke svekke resultatet selv om grunneffekten er relativt høy. Vi vil allikevel kunne si at vi har funnet svar på de spørsmålene vi faktisk har stilt. Hvis problemstillinga vår derimot ikke presiserer hvilke uavhengige variabler vi skal analysere, men derimot er en problemstilling hvor vi tar sikte på å forklare mest mulig av de inntektsforskjellene vi finner i samfunnet, vil en høy grunneffekt svekke resultatet. I vårt tilfelle er det faktorer med effekter som tilsvarende 31 som påvirker inntektsforskjellene, og som vi ikke har identifisert eller kan si at vi vet noe som helst om.

4. Kji-kvadrat (χ^2)

OM BRUKEN AV KJIKVADRAT

Kjikvadrat-testen brukes i bivariat analyse for å gi svar på om en statistisk sammenheng slik den framkommer i en tabell basert på et utvalg, er uttrykk for at det finnes en slik sammenheng også i virkeligheten. Spørsmålet som bruken av denne testen vil gi svar på er altså: Kan vi med en viss grad av sannsynlighet se bort fra muligheten for at utslagene i tabellen skyldes tilfeldigheter ved utvalget? Hvis vi i tabellen har med hele populasjonen som er berørt er det ikke aktuelt å bruke kjikvadrat-testen: Denne testen vurderer sannsynligheten for at det er berettiget å generalisere ut fra utvalgsresultater. Har vi hele populasjonen med i tabellen er det årsaksforhold og ikke signifikansnivå vi eventuelt skal konsentrere oss om, når vi arbeider med utvalg må vi imidlertid først beregne signifikansnivået før en eventuell årsaksanalyse.

Kjikvadrat-testing sier oss ikke noe om hvor sterk en sammenheng er i virkeligheten (det er altså ikke et korrelasjonsmål). Heller ikke får vi svar på hva slags sammenheng det er: Bruken av kjikvadrat vil ikke i seg selv si noe om hva som påvirker hva i tabellen, eller hvorvidt utslagene på variablene i tabellen kanskje skyldes en tredje utenforliggende variabel. Vi kan med andre ord ikke trekke en konklusjon om årsakssammenheng basert utelukkende på en vellykket kjikvadrat-test. Heller ikke sier kjikvadratet noe om retningen på en eventuell sammenheng, dette betyr at den kan brukes på alle målenivåer.

En vellykket kjikvadrat-test - det vil her si en test som viser at en statistisk sammenheng er signifikant - vil i første rekke fortelle oss at det er liten sannsynlighet for at den statistiske sammenhengen skyldes tilfeldigheter. Det igjen

betyr at det kan være snakk om en sammenheng som bør undersøkes nærmere. Det vil også gi oss en viss hjelp når vi skal formulere våre kommentarer: En signifikant sammenheng vil kunne rettferdiggjøre at vi generaliserer fra utvalget i tabellen til hele populasjonen (dersom håndverket ellers er bra nok), og det vil hjelpe oss i å stramme opp formuleringene våre.

I utregningen av kjikvadrat er det avviket i tabellen fra det vi kunne forvente ved statistisk uavhengighet som måles. Tankegangen er at hvis dette avviket er relativt stort vil det være liten sjanse for at dette avviket skyldes tilfeldigheter. Slik sett er kjikvadrat-testing en form for hypotesetesting; vi formulerer en alternativ hypotese om at det ikke foreligger noen sammenheng mellom de to variablene, og forkaster denne hypotesen dersom kjikvadratet blir tilstrekkelig stort.

FRAMGANGSMÅTE VED KJIKVADRAT-TESTING

- Tabellstørrelse og antall frihetsgrader

Når vi skal beregne kjikvadratet og signifikansnivået i en gitt tabell, er det enkelte forutsetninger som vi må ta hensyn til.

Utregningen baserer seg på å summere avviket i tabellen rute for rute. Dette betyr at jo flere ruter det er i tabellen, jo større kan kjikvadratet bli. Tolkningen av selve kjikvadratet - det tallmessige uttrykket for det avviket vi finner - må derfor korrigeres i forhold til tabellstørrelsen eller antall ruter i tabellen. Jo større tabellen er, jo flere frihetsgrader er det i den. Antall frihetsgrader forkortes ofte til df ('degrees of freedom').

Vi bestemmer antall frihetsgrader i en tabell ut fra formelen

$$(\text{Antall kolonner} - 1) \times (\text{Antall linjer} - 1) = \text{Antall frihetsgrader}$$

I en firefeltstabell er det derfor 1 frihetsgrad (dvs. $(2-1) \times (2-1) = 1$), i en seksfeltstabell 2 frihetsgrader (dvs. $(2-1) \times (3-1) = 2$), og i en nifeltstabell 4 frihetsgrader (dvs. $(3-1) \times (3-1) = 4$).

Jo flere frihetsgrader det er i en tabell, jo større må kjikvadratet være for at de sammenhengene vi finner skal være signifikant på et gitt nivå.

- Valg av signifikansnivå

Vi velger selv signifikansnivå. Signifikansnivået angir hvor stor sannsynligheten er for at et gitt avvik kan skyldes tilfeldigheter. Vanligvis brukes et signifikansnivå på 0.01 i samfunnsvitenskapelig forskning. Et slikt nivå innebærer at det ikke er mer enn 1 % sannsynlighet for at et gitt resultat skyldes tilfeldigheter ($p \leq 0.01$).

Ut fra statistisk teori kan dette forklares på en annen måte: Hvis vi har valgt et signifikansnivå på 0.01, og vi ut fra dette finner en signifikant sammenheng, vil det innebære at det er mindre enn 1 % sjanse for at vi tar feil hvis vi sier at det faktisk eksisterer en sammenheng mellom disse variablene også i virkeligheten. Dette er et relativt strengt krav; vi vil statistisk sett ha rett i 99 av 100 tilfeller.

På den andre siden vil en forsker som konsekvent opererer med et signifikansnivå på 0.01, og som bruker kjikvadrat-testing jevnt og trutt, kanskje i snitt avdekke og kommentere 10 signifikante sammenhenger pr. dag. Da vil faktisk vedkommende forsker rent statistisk kunne ta feil en gang hver 10. dag. Som regel vil det kanskje ikke føre til så store konsekvenser, men kanskje av og til...

En tredje måte å anskueliggjøre dette på er ved å ta utgangspunkt i alle de forskjellige utvalgene som kan trekkes av en gitt populasjon. Dette er i praksis et nærmest uendelig utvalg. Har vi i en tabell basert på ett utvalg funnet en sammenheng som er signifikant på 0.01-nivået, vil vi rent statistisk kunne regne med å finne en liknende sammenheng i tilsvarende tabeller basert på 99 av hver 100. utvalg. (Men sjansen er jo til stede for at det utvalget vi jobber med er det 100., det som gir skjevheter og som lurar oss til å tro at en sammenheng er signifikant selv om den ikke er det. Da vil vi jo i 99 av 100 tilfeller oppleve at vi ikke finner liknende sammenhenger.)

Velger vi et signifikansnivå på 0.05 - som også av og til kan være akseptabelt i forbindelse med samfunnsvitenskapelig forskning - øker sjansen vår til å trekke feil konklusjon. Den statistiske sjansen for at vi får 'godkjent' et tabellresultat på dette nivået uten at det er en sammenheng i virkeligheten er nå økt til fem prosent - dvs. at det i snitt vil forekomme hver tjuende gang. Dette betyr at det fremdeles er 95 % sjanse for at vi har rett hvis vi med utgangspunkt i en signifikant tabell forkaster 'null-hypotesen' om ingen sammenheng mellom variablene.

Ettersom valg av signifikansnivå påvirker muligheten for feilslutninger og derved feilaktige konklusjoner, må dette valget sees i lys av hvor viktig det er å ikke trekke slike feilaktige konklusjoner. Konsekvensene som oppstår ved feilaktige konklusjoner må vurderes. Her vil ulike disipliner kunne ha hver sine 'stiger' med akseptable nivåer.

Det vi bruker kjikvadrat-testen til er å sikre oss mot at vi for ofte skal tro at det er en sammenheng uten at det i virkeligheten er det. I andre situasjoner - der hvor vi skal sikre oss mot å se bort fra muligheten for at det kan være en sammenheng, bør vi velge andre metoder for hypotesetesting.

Ett eksempel på dette: Vi kjenner fra media debatten om det er slik at økt antall volds- og seksualforbrytelser henger sammen med forbruket av videoer med sex-/voldinnslag. I dette tilfellet ser vi bort fra at dette kan dreie seg om relativt kompliserte årsakssammenhenger som sannsynligvis også omfatter en del

andre variabler som virker inn. Vi tar et utvalg med norske menn mellom 16 og 40 år (med andre ord den aldersgruppa som kanskje ser mest på video uansett type). For dette utvalget lager vi en tabell som omfatter a) lavt eller høyt forbruk av sex-/voldsvideoer, og b) befatning med seksual- og voldsforbrytelser. (Hvordan vi skal skaffe oss pålitelige data om dette er en annen sak.)

Dersom vi er redde for å påvise en sammenheng mellom disse faktorene (dersom vi for eksempel er videoutleiery av den mer lugubre typen), velger vi et strengt signifikansnivå: Vi er mest redd for å få et resultat hvor vi trekker en konklusjon om at det er en sammenheng mellom disse faktorene uten at det er det i virkeligheten, derfor lager vi en test hvor det skal ekstreme utslag til for at vi skal forkaste en nullhypotese om ingen sammenheng.

Hvis vi derimot tilhører en foreldregruppe som er bekymret for konsekvensene av økt forbruk av videoer av denne typen, vil vi være tilbøyelige til å velge et mindre strengt signifikansnivå: Vi tror i utgangspunktet at det er en sammenheng, og derfor vil vi gjerne forkaste en hypotese om at det ikke er det. Ut fra hva som kan skje med barna våre, vil vi ikke ta risikoen på at det er en sammenheng som vi blir nødt til å se bort fra pga. strenge forskningsmetodiske krav.

Dette dilemmaet kan avspeile noen av de frustrasjonene en kan møte som forsker: Dersom en har mistanke om en sammenheng som det bør settes i gang tiltak for å unngå de negative konsekvensene av, er det ofte svært vanskelig med strenge krav til forskningsmetodikk å påvise disse sammenhengene. Dette gjelder for eksempel i ett tilfelle som beskrevet ovenfor, det gjelder i miljøvern (hva skyldes drivhuseffekten og hvor farlig er det), det kan dreie seg om næringsmiddel- og legemiddelkontroll ('vi kan ikke stoppe dette medikamentet for det er ikke påvist signifikante bivirkninger') osv.

Løsningen på slike dilemmaer ligger ikke i bruken av kjiqvadrat-testen. Enkelte andre kvantitative teknikker kan være mer velegnet, og ulike kvalitative metoder kan avdekke og forklare sammenhenger som høyst sannsynlig er til stede uten at det er mulig å bekrefte dem gjennom en slik test.

- Beregning av kjiqvadratet

Utgangspunktet for beregning av kjiqvadratet er en tabell som viser den faktiske fordelingen av utvalget i forhold til verdiene på to variabler. Denne tabellen kan være prosentuert, eller den kan være uttrykt i absolutte tall. Hvis tabellen er prosentuert må vi ha nok opplysninger om utvalget til at vi kan regne den om til absolutte tall - det vil si at vi i alle fall må vite utvalgets størrelse (populasjonens størrelse trenger vi imidlertid ikke å kjenne).

Trinn 1: Eventuell omregning fra prosenttall til absolutte tall

Vi starter med en prosentuert tabell, hvor det i hver rute er angitt gruppens andel av det totale utvalget:

TILBUD OM KURS SISTE HALVÅR	YRKESGRUPPE		
	Sykepleiere	Leger	
Ja	8.0	22.0	30.0
Nei	62.0	8.0	70.0
(N=300)	70.0	30.0	100.0

Hvor mange personer inngår så i hver gruppe? Vi beregner det ut fra formelen

$$\text{Gruppens andel} \times \text{Antall personer i utvalget} = \text{Antall personer i gruppen}$$

Vi får da følgende tabell, som viser den aktuelle tabellen i absolutte tall (faktisk frekvens, eller f)

TILBUD OM KURS SISTE HALVÅR	YRKESGRUPPE		
	Sykepleiere	Leger	
Ja	24	66	90
Nei	186	24	210
(N=300)	210	90	300

Trinn 2: Beregning av forventet frekvens ved statistisk uavhengighet

Vi har da funnet den faktiske fordelingen i vårt utvalg. Samtidig kjenner vi nå den relative fordelingen mellom sykepleiere og leger, og den relative fordelingen mellom de som har fått tilbud om kurs og de som ikke har fått tilbud. Dette siste hjelper oss til å beregne den forventede frekvensen (f_v) hvis vi forutsetter statistisk uavhengighet.

Variant A) Fra kapitlet om sannsynlighetsberegning vet vi at sannsynligheten for at et gitt utfall skal inntreffe er produktet av delsannsynlighetene. Samtidig husker vi at sannsynlighet som matematisk begrep kan sees som et

speilbilde av proporsjonene. Dette betyr at vi kan multiplisere sannsynligheten for å tilhøre en av gruppene på den ene variabelen med sannsynligheten for å tilhøre en av gruppene på den andre variabelen. Hvis vi så multipliserer denne sannsynligheten med antallet i utvalget finner vi det forventede antall personer i hver gruppe.

P(sykepleier)	= 70 % = 0,7
P(lege)	= 30 % = 0,3
P(kurs)	= 30 % = 0,3
P(ikke kurs)	= 70 % = 0,7

Dette gir oss følgende delsannsynligheter:

P(sykepleier med tilbud om kurs)	= 0,7 x 0,3 = 0,21
P(lege med tilbud om kurs)	= 0,3 x 0,3 = 0,09
P(sykepleier uten tilbud om kurs)	= 0,7 x 0,7 = 0,49
P(lege uten tilbud om kurs)	= 0,3 x 0,7 = 0,21

Dette igjen gir oss følgende forventede frekvenser (f_u):

TILBUD OM KURS SISTEHALVÅR	YRKESGRUPPE		
	Sykepleiere	Leger	
Ja	(0,21 x 300) 63	(0,09 x 300) 27	90
Nei	(0,49 x 300) 147	(0,21 x 300) 63	210
(N=300)	210	90	300

Variant B) En annen måte å finne den forventede frekvensen på er å unngå veien om beregning av delsannsynligheter. Det vi gjør er egentlig det samme, men vi setter opp utregningen på en annen måte. Framgangsmåten er

Antall personer med den aktuelle verdien på den ene variabelen	x	Antall personer med den aktuelle verdien på den andre variabelen
---	---	---

Antall personer i utvalget

I praksis betyr dette at vi får 4 utregninger:

$$\begin{aligned}(\text{Antall sykepleiere} \times \text{Antall med kurstilbud})/\text{Utvalget} &= (210 \times 90)/300 = 63 \\(\text{Antall leger} \times \text{Antall med kurstilbud})/\text{Utvalget} &= (90 \times 90)/300 = 27 \\(\text{Antall sykepleiere} \times \text{Antall uten kurstilbud})/\text{Utvalget} &= (210 \times 210)/300 = 147 \\(\text{Antall leger} \times \text{Antall uten kurstilbud})/\text{Utvalget} &= (90 \times 210)/300 = 63\end{aligned}$$

Trinn 3: Kalkulering av kjikvadratet

Kjikvadratet kan defineres som summen av de kvadrerte avvikene i tabellen delt på de forventede frekvensene. Formelen for dette blir

$$\chi^2 = \sum \frac{(f - f_u)^2}{f_u}$$

Kalkulering av kjikvadratet foregår i følgende rekkefølge: Først finner vi avviket for hver rute og kvadrerer dem, så deler vi det kvadrerte avviket for hver rute med den forventede frekvensen for ruta, og så summerer vi resultatet.

Faktisk frekvens f	Forventet frekvens f _u	f - f _u	(f - f _u) ²	$\frac{(f - f_u)^2}{f_u}$
24	63	-39	1521	24.1
66	27	39	1521	56.3
186	147	39	1521	10.3
24	63	-39	1521	24.1

$$\chi^2 = \frac{1521}{63} + \frac{1521}{27} + \frac{1521}{147} + \frac{1521}{63} = 24.1 + 56.3 + 10.3 + 24.1 = \underline{114.8}$$

Trinn 4: Vurdering

Etter å ha kalkulert kjikvadratet står vi igjen med ett tall, som er et tallmessig uttrykk for avviket i den aktuelle tabellen fra hva vi kunne ha forventet hvis det ikke var noen sammenheng mellom variablene - i dette tilfelle at sykepleiere og leger får den samme mengden med kurstilbud.

Jo større kjikvadratet er, jo større er avviket i tabellen, og jo mindre sannsynlig er det at vi rent tilfeldig ville ha fått et slikt resultat. Vi kan sette opp

dette som ved hypotesetesting, og få

H_0 : Det er ingen sammenheng mellom yrkesgruppe og tilbud om kurs, og

H_{ALT} : Det er en sammenheng mellom yrkesgruppe og tilbud om kurs.

Vi er ute etter å finne ut om det er sannsynlig at det er en slik sammenheng, med andre ord skal vi finne ut om det er grunnlag for å forkaste null-hypotesen.

Vi vil ikke konkludere med at det er en sammenheng mellom variablene uten at vi er rimelig sikre på at vi har rett. Denne sikkerheten uttrykkes gjennom valg av signifikansnivå.

Når vi har valgt signifikansnivå og bestemt antall frihetsgrader i tabellen, må kjikvadratet vurderes ut fra en tabell over kritiske verdier. I disse tabellene er antall frihetsgrader fra 1..... listet opp for hver linje, og kolonnene angir ulike signifikansnivåer. Vi må finne riktig plass i tabellen (under det signifikansnivået vi har valgt, og rett til høyre for det riktige antall frihetsgrader) og finne hvilken kritisk verdi for kjikvadratet som står der.

Hvis 'vårt' kjikvadrat er større enn eller lik den kritiske verdien, er tabellen signifikant på det angitte nivået, og vi kan forkaste null-hypotesen. Hvis kjikvadratet vårt er mindre enn den kritiske verdien, har vi ikke grunnlag for å forkaste null-hypotesen.

Nedenfor er anført et utsnitt av en slik tabell over kritiske verdier.

Antall frihetsgrader	Signifikansnivå		
	0.1	0.05	0.01
1	2.71	3.84	6.64
2	4.60	5.99	9.21
3	6.25	7.82	11.34
4	7.78	9.49	13.28

I vårt tilfelle ser vi at vårt kjikvadrat er betydelig større enn de kritiske verdien i tabellen, uansett hvilket signifikansnivå vi hadde valgt av disse. Valg av signifikansnivå er forøvrig noe en gjør før testen gjennomføres, slik at konklusjonene blir signifikant eller ikke-signifikant på det nivået vi har valgt. Vi uttrykker altså ikke konklusjoner av typen 'mere signifikant' eller 'mindre signifikant'.

Vi ser forøvrig av tabellen at jo større tabellen blir, og jo strengere krav vi stiller til signifikans, jo større blir den kritiske verdien for χ^2 .

4. Korrelasjonsmål

VALG AV KORRELASJONSMÅL

Når vi måler eller beregner korrelasjon beregner vi styrken av en sammenheng mellom to eller flere variabler. Det finnes en rekke ulike korrelasjonsmål, som kan gi svar på ulike spørsmål. Valg av hvilket korrelasjonsmål som er riktig å bruke til enhver tid avhenger for det første av hvilket svar vi er ute etter (eller hvilket spørsmål vi har stilt oss), og for det andre av hvilket målenivå variablene som inngår i analysen er på.

Spørsmålene vi kan stille oss er

- hvor sterk er sammenhengen?
- hvilken retning har sammenhengen, er den positiv eller negativ?
- i hvilken grad følges en endring på den ene variabelen av tilsvarende endringer på den andre (proporsjonalitet)?

Vi må vite hvilke korrelasjonsmål som kan gi svar på disse ulike spørsmålene, og hvilke som ikke kan gjøre det. I tillegg må vi kjenne variablene våre på en slik måte at vi forstår hvilke av disse spørsmålene som det er aktuelt å stille. Ut fra hvilket målenivå variablene har, vil noen av disse spørsmålene bli meningsløse.

Vi kan for eksempel spørre om hvor sterk sammenheng det er mellom kjønn og politisk interesse, men vi kan ikke spørre om denne sammenhengen er positiv eller negativ (da måtte vi hatt et begrep om en rangordning mellom kjønnene). Heller ikke kan vi spørre om det er slik at politisk interesse øker med kjønn - vi kan alltså måle hvordan politisk interesse øker, men hvordan kan vi måle hvordan kjønn øker?

De 'svakeste' korrelasjonsmålene er de som bare kan gi svar på det første spørsmålet - hvor sterk er sammenhengen. Av disse korrelasjonsmålene er fi det mest brukte. Fi kan brukes på variabler på alle målenivåer, det vil si i forhold til variabler hvor det kan være meningsfullt også å stille de andre typene spørsmål, men vi vil ikke få svar på disse spørsmålene gjennom bruken av fi. Bruken av fi er derfor først og fremst knyttet til variabler på nominalnivå (det vil si kategoriseringer hvor vi ikke kan snakke om rangordning eller hvor vi kan måle de forskjellige verdiene på variabelen i forhold til hverandre).

På neste trinn i hierarkiet har vi tau og gamma. Disse målene forutsetter variabler på ordinalnivå, og det betyr at begge variablene minst må være på ordinalnivå. Hvis en av variablene er på nominalnivå kan ikke disse målene brukes.

Både tau og gamma gir svar både på spørsmålet om hvor sterk en eventuell sammenheng er, og på spørsmålet om denne sammenhengen er positiv eller negativ. Dette siste spørsmålet vil kunne være av typen 'Er det slik at høy verdi på den ene variabelen tenderer til å gi høy verdi også på den andre variabelen?'

(f.eks. høy utdanning - høy inntekt). Vi trenger ikke å vite forholdet mellom verdiene på hver enkelt variabel for å kunne bruke tau og gamma meningsfylt ut over rangordninga mellom dem - vi trenger for eksempel ikke å vite hva som er høy utdanning i forhold til hva som er lav utdanning, og vi trenger ikke å være i stand til å måle forskjellen på dette.

En annen sak er at vi selvfølgelig bør kunne knytte noen kommentarer til dette i forbindelse med tolkningen, slik at vi kan gi rimelig fornuftige kommentarer.

Rho er et mål vi kan bruke når vi har variabler med relativt stort antall verdier. Vi trenger ikke å kjenne det eksakte forholdet mellom verdiene, men vi må kjenne rangordenen mellom dem.

Det at vi ikke trenger å vite forholdet mellom de ulike verdiene betyr også at hverken tau, gamma eller rho kan gi svar på spørsmål som forutsetter at dette er kjent. Vi kan bruke tau, gamma og rho også i slike sammenhenger, men vi vil bare kunne måle om eventuelle sammenhenger er positive eller negative.

Er vi derfor ute etter å måle sammenhenger mellom variabler på intervall- og forholdstallsnivå der dette forholdet er kjent (noe som ligger i definisjonen av disse målenivåene), bør vi velge andre korrelasjonsmål. Det viktigste av disse er det som kalles Pearson's r eller produktmomentkorrelasjons-koeffisienten.

Dette korrelasjonsmålet vil kunne gi svar på hva som vil skje med verdiene på den ene variabelen hvis vi øker med så og så mye på den andre. Ett klassisk eksempel: Sammenhengen mellom hastighet og bremsestrekning. Er vi ute etter å måle denne sammenhengen kan vi bruke r hvis vi vil vite hvor sterk sammenhengen er, tau eller gamma hvis vi vil vite om det er slik at når farten øker øker også bremselengden, og Pearson's r hvis vi vil vite noe om hvor mye bremselengden øker når vi øker hastigheten med en viss størrelse.

NORMERING AV KORRELASJONSMÅL

Når vi snakker om at et korrelasjonsmål er normert, betyr det at verdiene på korrelasjonsmålet bestandig vil bli ett eller annet sted fra og med -1 til og med $+1$. At verdiene blir innenfor dette spranget, betyr ikke annet enn at en har utviklet formlene for å prøve å oppnå akkurat dette.

Skal vi tolke en gitt korrelasjonsverdi må vi derfor tolke den angitte verdien i forhold til denne skalaen. Vi må også vite om det aktuelle korrelasjonsmålet er normert, og hva som ligger i de ulike verdiene.

I et fullstendig normert korrelasjonsmål vil $+1$ bety en perfekt positiv sammenheng, 0 vil bety ingen sammenheng eller korrelasjon, og -1 vil bety en perfekt negativ sammenheng. Dette vil i praksis bety at vi finner for eksempel 1) at alle med høy utdanning også har høy inntekt, eller 2) at det ikke er noen sammenheng mellom utdanning og inntekt, eller 3) at alle med høy utdanning har lav inntekt.

Fordelen med normering av korrelasjonsmålene er at vi da kan sammenlikne korrelasjonsverdiene fra tabell til tabell direkte med hverandre.

Hvilke korrelasjonsmål er det da som er normert, og hva er det som gjør at de som ikke er det ikke er normert?

Fi er ikke normert. Den laveste verdien som fi kan oppnå er 0, vi kan ikke få en negativ verdi. Dette er greit nok, fordi vi kan ikke bruke fi til å beregne eventuelle negative sammenhenger allikevel (og delvis nettopp derfor). Årsaken til at fi ikke kan oppnå negative verdier ligger i formelen for fi: Vi skal dele kjikvadratet på antall personer i utvalget, og så trekke kvadratroten av dette. Ettersom kjikvadratet ikke kan bli negativt, kan heller ikke fi bli det. Heller ikke er fi normert andre verdien - det er ikke slik at den maksimale verdien for fi bestandig er +1, uavhengig av tabellstørrelse og hvordan resultatet i tabellen er. Ut fra en gitt tabell kan den teoretiske maksimalverdien for fi kunne være lavere enn 1.

Tau er ikke normert. I motsetning til fi kan tau imidlertid gi oss negative verdier. Problemet er at gitte tabellresultater kan gi lavere teoretisk mulige ekstremverdier. Dette skyldes at vi i formelen for tau skal dele på alle i utvalget, mens vi over streken kan risikere å få med oss en langt lavere andel av det totale utvalget (det vil inntreffe dersom to eller flere enheter har samme verdi på en av eller begge variablene). Dersom dette skjer (og det vil gjerne skje) vil tau i praksis for en gitt tabell variere over en mindre skala.

Gamma korrigerer for svakheten ved tau, ved at det bare er de enhetene som blir med i telleren i brøken som også blir med i nevneren. Gamma vil derfor i de fleste sammenhenger fungere som om det var normert. Vi kan imidlertid risikere å få en korrelasjon på 0 ved bruken av gamma, selv om det er en sammenheng. Dette kan skje når vi får likt antall like og ulike par (i gjennomgangen av gamma skal vi se hva dette betyr).

Rho er et fullstendig normert korrelasjonsmål, det samme gjelder for Pearson's r .

Fi; KORRELASJON PÅ NOMINALNIVÅ

Vi har sett at kjikvadratet setter oss i stand til med en viss angitt sannsynlighet å vurdere om det eksisterer en sammenheng mellom to variabler eller ikke. I tolkningen av χ^2 får vi imidlertid ikke svar på hvor sterk en eventuell sammenheng er, og heller ikke - hvis det er meningsfullt å vurdere det - om det er snakk om en positiv eller negativ sammenheng.

Det er altså ikke slik at et høgt kjikvadrat nødvendigvis tilsier at det er en sterkere sammenheng mellom variablene enn med et lavere kjikvadrat. Fi kan derimot gi oss svar på hvor sterk en eventuell sammenheng er.

Fi er en direkte videreføring av kjikvadratet. Det vi gjør når vi regner ut fi er at vi korregerer kjikvadratet for utvalgsstørrelsen. Vi skal se nærmere på hva det betyr.

- Utregning av fi

Vi må ta utgangspunktet i kjikvadratet når vi skal beregne fi. Formelen for fi er

$$F_i = \sqrt{\frac{\chi^2}{N}}$$

Hvis vi nå i første omgang ser på denne formelen, ser vi at jo større utvalget blir, jo lavere blir resultatet av brøken, og jo lavere vil fi bli - gitt en fast verdi for kjikvadratet. Hvorfor er det slik?

La oss vende tilbake til kapitlet om kjikvadratet. I eksemplet der hadde vi et utvalg som besto av 300 personer, 210 sykepleiere og 90 leger. Vi så på sammenhengen mellom yrkesgruppe og tilbud om kurs.

I den første tabellen anga vi den relative - eller den prosentmessige - fordelingen på disse to variablene:

TILBUD OM KURS SISTEHALVÅR	YRKESGRUPPE		
	Sykepleiere	Leger	
Ja	8.0	22.0	30.0
Nei	62.0	8.0	70.0
(N=300)	70.0	30.0	100.0

Vi beregnet så den faktiske frekvensen for tabellen, med et utvalg på 300 personer.

TILBUD OM KURS SISTEHALVÅR	YRKESGRUPPE		
	Sykepleiere	Leger	
Ja	24	66	90
Nei	186	24	210
(N=300)	210	90	300

Den forventede frekvensen ved statistisk uavhengighet ble beregnet til

TILBUDOMKURS SISTEHALVÅR	YRKESGRUPPE		
	Sykepleiere	Leger	
Ja	63	27	90
Nei	147	63	210
(N=300)	210	90	300

og kjikvadratet ble regnet ut til å være 114,8.

Hvis vi nå setter dette inn i formelen for fi, vil vi få

$$F_i = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{114,8}{300}} = \sqrt{0,383} = \underline{0,62}$$

Med et utvalg på 50 personer og den samme relative fordelingen, ville den faktiske fordelingen blitt

TILBUDOMKURS SISTEHALVÅR	YRKESGRUPPE		
	Sykepleiere	Leger	
Ja	4	11	15
Nei	31	4	35
(N=300)	35	15	50

Den forventede frekvensen ved statistisk uavhengighet ville blitt

TILBUDOMKURS SISTEHALVÅR	YRKESGRUPPE		
	Sykepleiere	Leger	
Ja	10,5	4,5	15
Nei	24,5	10,5	35
(N=300)	35	15	50

Dette ville i sin tur ha gitt oss et kjikvadrat = 19,2.

Hvis vi nå setter dette inn i formelen for fi, vil vi få følgende resultat:

$$F_i = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{19,2}{50}} = \sqrt{0,384} = \underline{0,62}$$

Med andre ord: Den samme relative fordelingen gir et kjikvadrat som øker når utvalgsstørrelsen øker. Samtidig er det klart at når den relative fordelingen holdes konstant vil også sammenhengen mellom variablene være tilsvarende i begge tilfeller. Bruken av χ^2 korrigerer derfor for effekten av at utvalgsstørrelsen øker.

Betyr så ikke dette at kjikvadratet kan være misvisende. Nei, fordi det ikke er styrken av sammenhengen som måles. Med kjikvadratet får vi et mål på hvorvidt tilfeldigheter ved utvalget har slått ut i tabellen slik at vi tilsynelatende har en sammenheng uten at dette er tilfelle i virkeligheten. Jo større utvalget er, jo mindre er sjansen for at det oppstår tilfeldigheter i den grad at det vil slå ut i tabellen. Derfor er kjikvadratet et godt mål på denne sannsynligheten, men forutsatt at vi ikke legger mere i tolkningen av det enn vi har grunnlag for.

TAU OG GAMMA; KORRELASJON PÅ ORDINALNIVÅ

Når begge variablene våre er på ordinalnivå eller høyere kan vi bruke tau eller gamma som korrelasjonsmål. I hovedsak kan vi si at vi velger gamma framfor tau: Gamma er, i motsetning til tau, normert. Årsaken til at tau ikke er normert er at tau i visse tilfeller, uansett hvor sterk en eventuell sammenheng er, ikke vil kunne oppnå ekstremverdiene 1 og -1. Vi kan si at med visse sammenstillinger av enhetene vil skalaen som tau kan bevege seg innenfor bli snevret inn. Konsekvensen av det er at vi aldri vil kunne vite hvor store verdier vi kunne ha fått ved en perfekt sammenheng, og vi har derfor ikke noen målestokk som vi kan vurdere den aktuelle tau-verdien i forhold til. Dette er ikke noe som vil skje bare i ekstreme tilfeller, men faktisk noe som oftest vil inntreffe, og jo færre verdier vi har på de to variablene jo oftere vil dette skje og jo større utslag vil det gi. Vi skal komme tilbake til hvorfor det er slik.

Hvorfor bruker vi da tau i det hele tatt? Svaret ligger at vi bruker tau i liten grad i forskning, men forståelsen av hvordan tau beregnes, svakhetene ved tau og den korrigeringen av tau som ligger i utviklingen av gamma kan gi god innsikt i det tallbehandlingsmessige grunnlaget som ligger til grunn for statistikken. Slik sett kan vi si at tau lever videre av pedagogiske grunner, og ikke fordi bruken av denne koeffisienten er et viktig bidrag til forskningen.

Tau og gamma forutsetter variabler minst på ordinalnivå, og de kan brukes på slike variabler uansett hvor mange verdier hver av variablene har. Nå når vi har beveget oss opp til det nivået hvor det vil være meningsfylt å snakke om retningen på sammenhengen, er det nettopp slike spørsmål bruken av disse målene gir svar på, i tillegg til det grunnleggende spørsmålet om hvor sterk sammenhengen er.

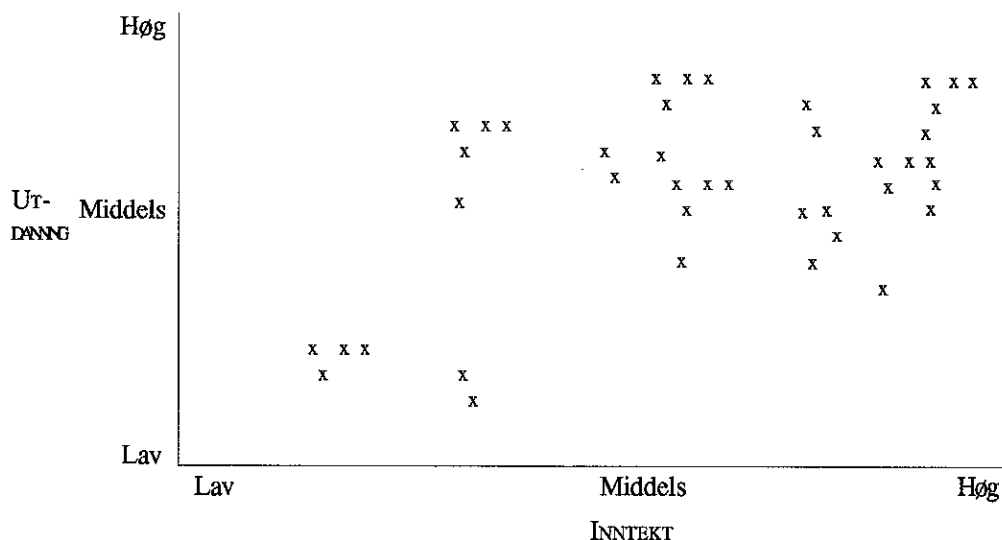
Spørsmål som har med retningen på sammenhengen å gjøre, er av typen 'Hvis verdien øker på den ene variabelen, øker den da (stort sett) også på den andre?'. Svarer vi ja på dette har vi en positiv sammenheng, mens en negativ sammenheng oppstår når det å øke verdien på den ene variabelen betyr at vi

reduserer verdien på den andre. Hvis vi vender tilbake til sammenhengen som vi tidligere har sett på i kapitlet om effekt og samspill betyr dette at vi kan spørre om det er slik at høyere inntekt har en sammenheng med høyere utdanning. Vi kan imidlertid ikke spørre om inntektene øker med kjønn, for kjønn er ikke en variabel på ordinalnivå.

Et annet spørsmål som vi ikke kan stille når variablene er på ordinalnivå er av typen 'Hvordan er sammenhengen mellom hvor stor økning i utdanning vi har i forhold til hvor stor inntektsøkning vi har?'. Grunnen til at vi ikke kan stille dette spørsmålet er at datamaterialet ikke gir oss tilstrekkelig med informasjon: Vi vet ikke den eksakte utdanningslengden og den eksakte inntekten til hver enkelt enhet i materialet, vi kan derfor ikke vite for eksempel hvor mye lenger utdanning en enhet i høgutdanningsgruppa har i forhold til en i lavutdanningsgruppa - det eneste vi vet er at utdanninga er lengre.

- Variabler på ordinalnivå; klyngediagram

Som et utgangspunkt for tankegangen når vi skal finne retningen av sammenhengen mellom variabler på ordinalnivå kan vi tenke oss et klyngediagram der aksesystemet utgjøres av de to variablene og verdiene på dem, og hvor hver enkelt enhet kan plottes inn i forhold til disse verdiene.

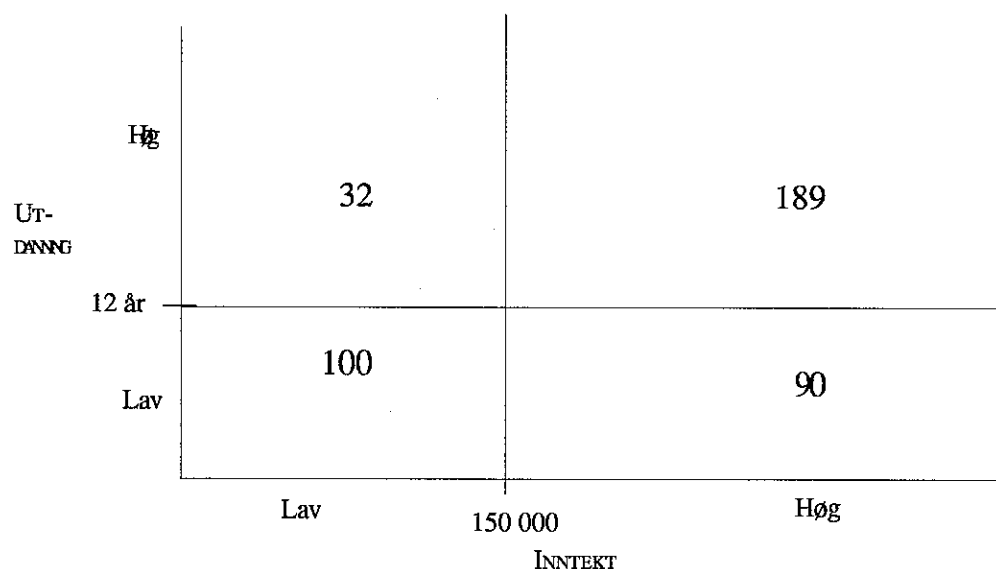


Hvert punkt i dette diagrammet svarer til en bestemt kombinasjon av verdier på de to variablene. Disse verdiene finner vi ved å trekke en horisontal linje fra punktet og ut til utdanningsaksen, og en vertikal linje fra punktet og ned til inntektsaksen. Skjæringspunktene på de to aksene bestemmer enhetens verdi på de to variablene.

I det ovenstående klyngediagrammet kan vi tenke oss at de to aksene er inndelt i punkter som angir alle verdier på de to variablene: Inntekt for eksempel

i 1000 kr. og oppover, utdanning i ett punkt for hvert år. Hvis dette er tilfelle er vi i stand til å beregne hver enhets nøyaktige verdi på de to variablene - og vi står i tilfelle overfor to variabler på forholdstallsnivå. Svært ofte vil dette være tilfelle i utgangspunktet: I datamatriksen har vi ofte såpass detaljert informasjon om verdiene på denne typen variabler.

Når vi skal gjennomføre analysen er det imidlertid ofte hensiktsmessig å forenkle informasjonen. I kapitlet om effekt og samspill var dette forenklet på en slik måte at vi hadde bare to verdier på hver av de to variablene. Hvis vi går tilbake til klyngediagrammet kan vi si at vi delte opp hver av de to aksene i det antall verdier vi var interessert i, vi valgte grensene mellom verdiklassene, og vi beregnet antallet enheter som hører hjemme i hver rute i diagrammet:



Det vi har gjort i dette diagrammet, er at vi har omkodet de to variablene: Vi har gitt variabelen utdanning to verdier; høg og lav, og satt grensen mellom dem til 12 års utdanning. Samtidig har vi gitt variabelen inntekt to verdier; høg og lav, med 150 000 kr. som grense.

Vi kjenner etter dette ikke lenger til hver enkelt enhets nøyaktige verdi på de to opprinnelige variablene, men må nøye oss med å beregne antallet enheter som har hver av verdiene på de omkodete variablene. Variablene er nå på ordinalnivå.

Vi kan også tenke oss at vi hadde foretatt en annen omkoding, og gitt hver av variablene 3 verdier:

UT- DANNING	Høg	0	21	102
	15 år			
	Middels	32	33	33
	12 år			
	Lav	100	66	24
		Lav	150 000	220 000
		Høg		
		INNTEKT		

I stedet for en todeling har vi her splittet kategoriene høg inntekt og høg utdanning fra forrige eksempel opp i to, slik at vi har fått inndelingen lav, middels og høg på begge variablene.

Diagrammet kan overføres direkte til en tabell, slik:

		INNTEKT			
		Lav	Middels	Høg	
UTDANNING	Lav	100	66	24	190
	Middels	32	33	33	98
	Høg	0	21	102	123
		132	120	159	411

Vi har her snudd litt på rekkefølgen av verdiene på den ene variabelen, slik at de enhetene som har lavest verdi på begge variablene befinner seg oppe til venstre, og verdiene øker når vi beveger oss til høyre og nedover i tabellen.

- Beregning av like og ulike par

Utrekning av tau og gamma vil bli gjennomgått for begge de ovenstående eksemplene. I utregningen av begge de to målene tar vi utgangspunkt i å kartlegge andelen like par i forhold til andelen ulike par. Bak disse begrepene skjuler følgende betraktninger seg:

Et likt par har vi i datamaterialet når vi kan sammenlikne to tilfeldige enheter og finner at den ene enheten har høyere verdi på begge variablene enn den andre. Dette vil si at når verdien øker på den ene variabelen har den også økt på den andre.

Et ulikt par finner vi derimot når denne 'symmetrien' bortfaller: En enhet har høyere verdi på den ene variabelen og lavere verdi på den andre variabelen.

I diagrammet kan vi tenke oss at vi trekker linjer fra en hvilken som helst enhet i undersøkelsen og til alle de andre enhetene. Alle linjene som innebærer at vi beveger oss i samme retning på begge variablene vil utgjøre antallet like par som akkurat denne enheten inngår i. I praksis vil dette i vårt diagram si alle linjer som går mot høyre og oppover (/) i diagrammet. Dette kan gjentas for alle enhetene i diagrammet, slik at vi finner summen av alle like par i undersøkelsen.

Likeledes kan vi for hver enkelt enhet finne antallet ulike par. Disse parene vil bli representert av linjer som går nedover mot høyre (\): Verdien øker på en variabel og synker på den andre. Summen av alle slike kombinasjoner eller linjer i diagrammet for alle enhetene vil være summen av ulike par i undersøkelsen.

Dersom to enheter har samme verdi på en av eller begge variablene vil vi ikke få linjer som går på skrå, hverken den ene eller andre veien. Slike kombinasjoner eller forhold mellom enheter vil derfor ikke være med hverken blant de like eller de ulike parene. Jo færre verdier vi har på de to variablene, jo mere vil enhetene måtte klumpe seg sammen på de få verdiene vi har, og jo flere kombinasjoner av enheter vil det være som hverken kommer med blant de like eller de ulike parene.

Når det er mange enheter, vil det i praksis være en uoverkommelig måte å beregne antallet forbindelser gjennom å trekke linjer mellom enhetene. Vi har da stor nytte av å sette tallene inn i en tabell. Ved hjelp av denne kan vi beregne antallet like par og antallet ulike par gjennom direkte utregning. Det er den samme tankegangen som ligger bak: For å finne antallet like par må vi for hver enhet i tabellen finne antallet forbindelser til andre enheter som karakteriseres ved at vi øker verdien på begge variablene - og summere sammen disse, og for å finne antallet ulike par må vi for hver enhet i tabellen finne antallet forbindelser til andre enheter karakterisert ved at verdien øker på den ene variabelen og reduseres på den andre. I tabellen vet vi at verdiene øker når vi beveger oss til høyre og nedover.

La oss se på tabellen igjen:

	INNTEKT			
	Lav	Middels	Høg	
UTDANNING				
Lav	100	66	24	190
Middels	32	33	33	98
Høg	0	21	102	123
	132	120	159	411

Vi har seks ruter i tabellen, i forhold til enhetene som befinner seg i øverste venstre hjørne (100 stykker) vet vi at alle som befinner seg nedenfor og til høyre for disse

- altså i de fire rutene nederst og til høyre - har høyere verdi på begge variablene ($33 + 33 + 21 + 102 = 189$). For å finne antallet like par som enhetene i ruta øverst til venstre inngår i, må vi multiplisere antallet enheter i den ruta med summen av antallet i de fire rutene nederst til høyre.

Neste skritt blir å beregne antallet like par som enhetene i neste rute (øverst i midten) inngår i. I denne ruta finner vi 66 personer eller enheter, og det er to ruter som inneholder enheter med høyere verdi på begge variablene ($33 + 102 = 135$).

Prosessen med å beregne antallet like par må gjøres for hver enkelt av de rutene hvor det går an å finne andre enheter som har høyere verdi på begge variablene, m.a.o. for hver av rutene som har en eller flere andre ruter til høyre og nedenfor seg. Dette betyr at vi må gjenta prosessen for hver av de fire rutene øverst til venstre i tabellen.

Antallet like par i tabellen kalles for P, og beregningen av P blir derfor slik:

$$\begin{array}{rclcl}
 100 \times (33 + 33 + 21 + 102) & = & 100 \times 189 & = & 18\,900 \\
 66 \times (33 + 102) & = & 66 \times 135 & = & 8\,910 \\
 32 \times (21 + 102) & = & 32 \times 123 & = & 3\,936 \\
 33 \times 102 & & & = & 3\,366 \\
 & & & & P = 35\,112
 \end{array}$$

Når antallet like par i tabellen er beregnet på denne måten, går vi videre og beregner antallet ulike par, som kalles Q. I beregningen av antallet ulike par bruker vi samme framgangsmåte, men vi skal finne antallet forbindelser som karakteriseres ved at en enhet har høyere verdi på den ene variabelen og lavere på den andre. Vi starter da med kombinasjonen høy inntekt/lav utdanning, og tar utgangspunkt i ruta øverst til høyre i tabellen. For hver rute som har en eller flere ruter til venstre og nedenfor seg må vi nå multiplisere antallet enheter i den aktuelle ruta med summen av de enhetene som befinner seg nedenfor og til venstre (lavere verdi på den ene variabelen til venstre i tabellen, høyere verdi på den andre variabelen nedover i tabellen).

$$\begin{array}{rclcl}
 24 \times (32 + 33 + 0 + 21) & = & 24 \times 76 & = & 1\,824 \\
 66 \times (32 + 0) & = & 66 \times 32 & = & 2\,112 \\
 33 \times (0 + 21) & = & 33 \times 21 & = & 693 \\
 33 \times 0 & & & = & 0 \\
 & & & & Q = 4\,629
 \end{array}$$

Går vi tilbake til den første omkodingen vi gjorde i forbindelse med disse målene (på s. 41) kan vi beregne P og Q også for denne tabellen:

	INNTEKT		
	Lav	Høg	
UTDANNING			
Lav	100	90	190
Høg	32	189	221
	132	279	411

Her blir beregningen av P og Q enklere, fordi vi har færre verdier på begge variablene. Vi får også flere kombinasjoner av enheter som hverken telles med i P eller Q, fordi en større andel av enhetene har samme verdi på en av eller begge variablene (de som tidligere var skilt ut som 'middels' og 'høg' på hver av variablene er nå samlet i en kategori 'høg').

$$P = 100 \times 189 = 18\,900$$

$$Q = 90 \times 32 = 2\,880$$

Det vi nå skal beregne er følgende: Jo flere like par vi har i forhold til antallet ulike par jo sterkere positiv sammenheng eller korrelasjon vil vi finne. Og omvendt: Hvis vi har flere ulike enn like par, så vil sammenhengen bli sterkere negativ jo flere ulike par vi har i forhold til antallet ulike par.

Vi skal se hvordan dette fortoner seg i forbindelse med utregning av tau og gamma for våre eksempler.

- Utregning av tau

Når vi skal beregne tau, setter vi verdiene vi har funnet for P og Q inn i følgende formel:

$$\text{Tau} = \frac{P - Q}{1/2 N (N - 1)}$$

I telleren i brøken inngår (antallet like par) - (antallet ulike par), mens alle de forbindelsene mellom enhetene som hverken er like eller ulike par altså forsvinner. I nevneren, derimot, inngår alle mulige par gitt en spesiell størrelse på utvalget. Nevneren er altså basert på en teoretisk matematisk beregning av det maksimalt antall mulige par som kan tenkes.

I eksemplet med to verdier på hver av de to variablene fant vi at $P = 18\,900$ og $Q = 2\,880$. Vi setter disse verdiene inn i formelen:

$$\text{Tau} = \frac{18\,900 - 2\,880}{1/2 N (N - 1)} = \frac{16\,020}{84\,255} = \underline{0.190}$$

I eksemplet med tre verdier på hver av de to variablene fant vi at $P = 35\,112$ og $Q = 4\,629$. Nevneren i brøken vil bli den samme, mens det er færre av forbindelsene mellom enhetene som har falt ut av telleren (i og med at det er flere verdier enhetene kan fordele seg på, vil det som regel være færre enheter med samme verdi på en av eller begge variablene):

$$\text{Tau} = \frac{35\,112 - 4\,629}{1/2 N (N - 1)} = \frac{30\,483}{84\,255} = \underline{0.362}$$

Vi ser at egenskapene ved tau gir en langt høyere verdi med dette materialet når vi får flere verdier på de to variablene. I tillegg til en eventuell effekt ved at materialet kanskje reelt sett er sterkere korrelert, får vi altså en høyere korrelasjonsverdi pga. at antall verdier øker. Dette utgjør et problem, sammen med det faktum at tau ikke er normert og at vi derfor ikke kjenner de teoretiske yttergrensene for tau i forbindelse med disse tabellene. Det er derfor vanskelig å vurdere hvor sterk korrelasjonen er ut fra de verdiene vi har funnet. Vi kjenner ikke de aktuelle tau-verdiens størrelse i forhold til det som er yttergrensene for disse tabellene.

- Utregning av gamma

Dette problemet unngår vi når vi benytter gamma som korrelasjonsmål. Ved at nevneren i brøken i formelen for tau (det teoretisk mulige antall like og ulike par) erstattes med summen av P og Q (det faktiske antallet like og ulike par i det aktuelle materialet) oppnår vi et normert korrelasjonsmål. Gamma vil alltid bevege seg mellom $+1$ og -1 , og gamma vil uansett i alle tabeller kunne oppnå ekstremverdiene $+1$ og -1 . Når vi derfor får oppgitt en gammaverdi eller en gammakoeffisient, vil vi bestandig kunne relatere denne til $+$ eller -1 , og får dermed en målestokk til hjelp i vurderingen.

Formelen for gamma er

$$\text{Gamma} = \frac{P - Q}{P + Q}$$

For våre to eksempler vil gammatesten gi følgende resultater:

a) 2 verdier på variablene

$$\text{Gamma} = \frac{18\,900 - 2\,880}{18\,900 + 2\,880} = \frac{16\,020}{21\,780} = \underline{0.736}$$

b) 3 verdier på variablene

$$\text{Gamma} = \frac{35\,112 - 4\,629}{35\,112 + 4\,629} = \frac{30\,483}{39\,741} = \underline{0.767}$$

Vi ser altså at vi får en langt høyere verdi for gamma enn for tau. Det er da også slik problemet med tau arter seg (å gi oss for lave korrelasjonsverdier). Vi ser også at det er en liten forskjell mellom de to gammaverdiene, men det skyldes ikke annet enn at vi faktisk har to ulike tabeller. Størrelsesordenen for gamma er imidlertid temmelig lik for de to tabellene, slik at vi kan si at vi finner akkurat de samme tendensene i begge. Dette var ikke tilfelle i forbindelse med tau, tabellen med tre verdier på de to variablene ga en langt høyere 'korrelasjon' enn tabellen med to verdier på variablene.

Forøvrig kan vi si at gammaverdier av denne størrelsesordenen viser en sterk positiv korrelasjon. En slik konklusjon kunne vi vanskelig ha trukket ut fra tau-verdiene. Hovedregelen er derfor: Bruk gamma framfor tau!

RHO (RANGKORRELASJONSKOEFFISIENTEN); ORDINALNIVÅ

Beregning av rho eller rangkorrelasjonskoeffisienten forutsetter at variablene er på ordinal- eller intervall-/forholdstallsnivå. Når vi bruker rho får vi svar på hvorvidt det er samsvar mellom de ulike enhetenes rang på de to variablene. Vi tenker oss med andre ord at vi setter opp en liste hvor vi rangerer enhetene etter hvilken verdi de har på den ene variabelen, og så beregner vi samsvaret mellom denne lista og en tilsvarende liste over enhetenes rang etter hvilken verdi de har på den andre variabelen.

Dette betyr egentlig at vi lager eller konstruerer to nye variabler, som vi kan kalle Rang1 og Rang2. Informasjonen som ligger i de variablene som vi tar utgangspunkt i utgår av analysen, den eneste informasjonen vi benytter er selve rangordenen på de to variablene.

Person nr.	Inntekt i 1000 kr.	Skatteprosent	Rang (inntekt)	Rang (skatte%)	Differanse d	d ²
1	220	22	4	9	-5	25
2	170	29	11	2,5	8,5	72,25
3	140	28	14	4	10	100
4	130	20	15	10,5	4,5	20,25
5	240	18	2	12,5	-10,5	110,25
6	200	27	7	6	1	1
7	195	27	8	6	2	4
8	180	26	9,5	8	1,5	2,25
9	180	27	9,5	6	3,5	12,25
10	230	14	3	14	-11	121
11	270	0	1	15	-14	196
12	160	18	12	12,5	-0,5	0,25
13	145	20	13	10,5	2,5	6,25
14	210	29	6	2,5	3,5	12,25
15	215	30	5	1	4	16
					0	$\Sigma d^2 = 699$

I tabellen ovenfor er de nødvendige elementene i utregningen av rho satt opp. De tre første kolonnene er 'hentet fra det opprinnelige datasettet'; her finner vi enhetene (personene) i undersøkelsen, vi finner data om inntekten deres oppgitt i 1000 kr. og vi finner den gjennomsnittlige skatteprosenten deres.

I de to neste kolonnene har vi listet opp den rangen hver enkelt person har på de to variablene. Personen med høyest inntekt har fått rang 1, personen med nest høyest inntekt rang 2 osv. Likedan er det gjort med skatteprosenten, den høyeste skatteprosenten gir rang 1, nest høyeste gir rang 2 osv. Der to eller flere personer har samme verdi (og dermed samme rang) gir vi hver av dem gjennomsnittet av de plasseringene de opptar. For inntekt ser vi at to personer har 180 000 kr. inntekt, disse etterfølger personen som har rang 8, og får som verdi i denne kolonnen 9,5, dvs. snittet av 9 og 10. For skatteprosent ser vi at tre personer har fått rangorden 6, dette er gjennomsnittsverdien for de tre som etterfølger nr. 4 - dvs. gjennomsnittet av 5, 6 og 7.

I den nest siste kolonnen har vi beregnet verdien d, det vil si differansen mellom de to rangverdiene for hver person. Enkelte av disse verdiene blir positive, enkelte blir negative. Som en kontroll kan vi summere dem sammen, hvis vi har gjort det riktig skal dette svaret bli 0.

0 i en formel er bestandig vanskelig - og ofte meningsløst - å arbeide med, vi har derfor i siste kolonne kvadrert d-verdien for hver enkelt person. Dermed har

vi blitt kvitt problemet med at fortegnene opphever hverandre. Når vi summerer disse kvadrerte d-verdiene får vi verdien som kan betegnes som Σd^2 . Denne verdien (summen av de kvadrerte rangdifferansene) går inn i formelen for rho.

Formelen for rho skrives slik:

$$\text{Rho} = 1 - \frac{6 \Sigma d^2}{N^3 - N}$$

Når vi setter inn tallene fra tabellen, vil utregningen av rho gi følgende resultat

$$\text{Rho} = 1 - \frac{6 \Sigma d^2}{N^3 - N} = 1 - \frac{6 (699)}{15^3 - 15} = 1 - \frac{4194}{3375} = 1 - 1,243 = \underline{-0,243}$$

Vi får altså en verdi for rho som er negativ, - 0,243. Dette betyr for det første at det er en viss korrelasjon mellom hver enhets rang på de to variablene i dette utvalget, denne korrelasjonen eller sammenhengen er klar, men relativt svak. For det andre er korrelasjonen negativ, noe som innebærer at i den grad det er en sammenheng så er hovedtendensen at de som tjener mest har den laveste skatteprosenten.

Det eksakte forholdet mellom inntekt og skatteprosent kan vi ikke si noe om etter bruken av rho.