Arbeidsnotater

STATISTISK SENTRALBYRÅ

Dronningensgt. 16, Oslo-Dep., Oslo 1. Tlf. 41 38 20

IO 76/27

16. september 1976

FORELESNINGER OM AVHENGIGHETSMÅL I KONTINGENSTABELLER AV UNIVERSITETSLEKTOR HARALD GOLDSTEIN

Referert av

Tor Haldorsen[x)]

INNHOLD

x) Arne S. Andersen, Harald Goldstein og Hans Viggo Sæbø har kommentert manuskriptet.

FORORD

Hvis enhetene i en bestand er klassifisert m.h.t. ett eller flere kjennetegn, så viser en kontingenstabell hvor mange av enhetene som har hver av de ulike verdiene på kjennetegnet evt. hver av de ulike kombinasjoner av verdier på kjennetegnene. Danskene kaller gjerne denne type tabeller for antallstabeller. En tabell over menn bosatt i Norge etter alder og ekteskapelig status er en typisk kontingenstabell. De fleste tabeller som produseres i Byrået er kontingenstabeller eller lett bearbeidde versjoner av sådane.

Ofte bruker en kontingenstabeller som utgangspunkt for å studere avhengigheten mellom variable som er brukt i klassifiseringen av enhetene i tabellen. Noen ganger ønsker en å se på denne avhengigheten i én tabell mens andre ganger vil en sammenligne denne avhengigheten i ulike materialer. Generelt sett er avhengighet et komplisert fenomen og det kan ofte være nødvendig å forenkle arbeidet ved å se på avledede tabeller som avbilder enkelte avhengighetstrekk eller bruke avhengighetsmål som måler spesielle former for avhengighet.

I dette notat som er et referat av en forelesningsrekke Goldstein holdt i Byrået høsten 1975, drøfter en ulike sider ved avhengighet. I forelesningene ble behandlet noen av de faktorer en bør ta hensyn til når en forenkler problemene en har ved måling av avhengighet. En foretrakk å redegjøre for tankegangen bak noen få avhengighetsmål framfor å behandle flest mulig av det utall av mål som finnes. Det ble understreket at en som mål bør velge populasjonsstørrelser som er lette å tolke. Estimeringsmetoder og samplingproblemer ble stort sett ikke behandlet. Forelesningene ble ført fram til en kort oversikt over log-lineære modeller som i den senere tid har fått en sentral plass i analyse av multivariable diskrete data.

Av litteratur på området ble anbefalt Bishop et al. (1975) som både handler som assosiasjonsmål og log-lineære modeller. Samplingproblemer er også behandlet i denne boka. Goodman og Kruskal (1954 og 1959) er "klassikere" om assosiasjonsmål.

Situasjonene i kapitlene 3-5 i dette notatet er også behandlet i Bjørnstad (1973).

På forelesningene ble det brukt en del eksempler som i notatet er merket "konstruerte tall", disse må på ingen måte oppfattes som realistiske for de forhold som beskrives.

1. INNLEDNING

Eksempel 1.1 (Konstruerte tall)

I et materiale over 141 svulster i hjernen er svulstene delt inn m.h.t. to kjennetegn. X angir om svulsten er av type A=Godartet, B=Ondartet eller C=Annet (ubestemt). Y angir svulstens lokalisering, mulighetene er: I=Tinning, II=Panne og III=Annet sted. Resultatet av klassifiseringen kan settes opp i en tabell.

X \ Y	I	II	III	
A	23	21	34	78
B	9	4	24	37
C	6	3	17	26
	38	28	75	141

Tabellen viser at av de 141 svulstene var 21 godartete og lokalisert til pannen, av alle svulstene var ialt 78 godartete.

Vi er interessert i å studere den statistiske avhengighet mellom X og Y, men må ut fra interesser og formål med undersøkelsen, presisere hva vi mener med "statistisk avhengighet". Hva vi videre gjør med tallene i tabellen, avhenger altså av hvilke forhold vi vil belyse. Hvis vi har interesse av å se hvordan hjernesvulster med gitt lokalisering fordeler seg på kategoriene godartet, ondartet og ubestemt, vil vi prosentuere vertikalt og får tabellen:

X \ Y	I	II	III
A	61	75	45
B	24	14	32
C	16	11	23
	100	100	100

(Prosenttall og andre relative tall vil vi beregne og avrunde hver for seg)

Vi får de betingete fordelinger gitt ulike verdier av Y. Tabellen viser at av svulstene i tinningen var 61 prosent godartete, 24 prosent ondartete og 16 prosent av annen type. Vi skal ikke her ta stilling til om materialet er så stort at vi f.eks. kan påstå at i alminnelighet har svulster i pannen større sannsynlighet for å være godartet enn andre svulster. Et av formålene med å organisere materialet på denne måten kan være at en skal predikere om andre svulster som en bare vet lokaliseringen til, er godartete, ondartete eller annet.

Hvis vi skulle studere hvordan hver av typene godartete, ondartete og andre svulster fordelte seg i hjernen, ville en foreta en horisontal prosentuering. Vi ville se på de betingete fordelinger gitt ulike verdier av X.

X \ Y	I	II	III	
A	29	27	44	100
B	24	11	65	100
C	23	12	65	100

Vi ser at av de ondartete svulstene var 24 prosent i tinningen, 11 prosent i pannen og 65 prosent andre steder.

Hvis svulstens alvorlighet og lokalisering er uavhengige, så vil de tre betingete fordelingene innen hver av tabellene bli like, sett bort fra muligheten for tilfeldige variasjoner.

Beregningen av de betingete fordelingene innebærer en relativt liten bearbeiding av grunnmaterialet. For noen formål kan det være tilstrekkelig med mer summariske uttrykk for sammenhengen. Da er det aktuelt å bruke et assosiasjonsmål som med et enkelt tall gir uttrykk for sammenhengen mellom lokalisering og alvorlighet i materialet.

Vi vil nå innføre en del betegnelser. De vil gi oss mulighet til å vise generelt hvilke avhengighetstrekk som avbildes av spesielle tabeller vi beregner på grunnlag av den opprinnelige kontingenstabell og vi kan med disse betegnelsene vise hva som måles med de enkelte avhengighetsmål.

Eksempel 1.1 og lignende situasjoner vil vi betrakte ved hjelp av en multinomisk modell. Vi har et gitt antall enheter der hver enhet på grunnlag av sin verdi på et kjennetegn faller i en og bare en av t mulige kategorier. Sannsynligheten for at en enhet faller i i'te kategori er p_i i = 1, 2,..., t og $\sum_{i=1}^{t} p_i = 1$. Som i eksempel 1.1 framkommer ofte kategoriene ved at enhetene krossklassifiseres m.h.t. to kjennetegn X og Y. Har kjennetegnene h.h.v. r og k mulige verdier, blir det t = r·k mulige kategorier i krossklassifiseringen. Mulige verdier av kjennetegn X vil vi betegne $A_1, A_2, \ldots, A_r$ eller 1, 2,..., r og mulige verdier av kjennetegn Y betegnes $B_1, B_2, \ldots, B_k$ eller 1, 2,..., k. Hver enhet antar altså en og bare en av verdiene til X samtidig som den antar en og bare en av verdiene til Y. $P(A_i \cap B_j)$ eller $P(X=i \cap Y=j)$ ("$\cap$"="og", begge deler må oppfylles), sannsynligheten for at en enhet har verdiene A_i og B_j, betegnes p_{ij} for i = 1, 2,..., r og j = 1, 2,..., k. Vi har $\sum_{j=1}^{k} \sum_{i=1}^{r} p_{ij} = 1$. Modell og notasjon kan illustreres ved følgende tabell.

X \ Y	B_1 1	B_2 2 ...	B_k k	
A_1 1	p_{11}	p_{12}...	p_{1k}	$p_{1.}$
A_2 2	p_{21}	p_{22}...	p_{2k}	$p_{2.}$
⋮	⋮	⋮	⋮	⋮
A_r r	p_{r1}	p_{r2}...	p_{rk}	$p_{r.}$
	$p_{.1}$	$p_{.2}$...	$p_{.k}$	1

Når en indeks er erstattet med · , har en summert over den indeksen. For i = 1, 2,..., r er $p_{i.}$ ($= \sum_{j=1}^{k} p_{ij}$) den marginale sannsynlighet for at en enhet skal ha verdien A_i mens for j = 1, 2,..., k er $p_{.j}$ ($= \sum_{i=1}^{r} p_{ij}$) den marginale sannsynlighet for at en enhet skal ha verdien B_j.

På tilsvarende måte kan et observasjonsmateriale organiseres.

	B_1	B_2 ...	B_k	
A_1	X_{11}	X_{12}...	X_{1k}	$X_{1.}$
A_2	X_{21}	X_{22}...	X_{2k}	$X_{2.}$
⋮	⋮	⋮	⋮	⋮
A_r	X_{r1}	X_{r2}...	X_{rk}	$X_{r.}$
	$X_{.1}$	$X_{.2}$...	$X_{.k}$	$X_{..}=n$

X_{ij} er tallet på enheter med verdier A_i og B_j . $X_{i.}$ er tallet på enheter i alt med verdi A_i mens $X_{.j}$ enheter i alt har verdi B_j. I et observasjonsmateriale vil X_{ij}/n, $X_{.j}/n$ og $X_{i.}/n$ være estimater for h.h.v. p_{ij}, $p_{.j}$ og $p_{i.}$.

I den generelle modell er det i alt r·k parametre. På disse hviler begrensningen $\sum_{j=1}^{k} \sum_{i=1}^{r} p_{ij} = 1$. Hvis $p_{ij} = p_{i.} \cdot p_{.j}$ for i = 1, 2, ..., r og j = 1, 2, ..., k så sies X og Y å være uavhengige. Hvis likheten ikke er oppfylt for alle i og j så er X og Y avhengige. Hvis en tenker på alle de ulike sett

p_{ij}-er som ikke oppfyller likheten, ser en lett at avhengighet er et komplisert fenomen. En kan tenke seg mange grader og ulike former for avhengighet, og det er viktig å forstå at det i den generelle situasjon ikke finnes noe avhengighetsmål som med et tall kan måle alle former for avhengighet. På den annen side er det lite oversiktlig å bruke de r·k p_{ij}-ene til å uttrykke avhengigheten. Det gjelder å spesifisere avhengigheten ved færre parametre. Et lite steg i den retning kan være å beregne ett av de to settene med betingete fordelinger som vi gjorde i eksempel 1.1. Når vi betinger m.h.t. verdiene av Y, beregner vi

$$P(A_i|B_j) = \frac{P_{ij}}{P_{\cdot j}} = q_{ij} \text{ for } i = 1, 2, \ldots, r \; j = 1, 2, \ldots, k.$$

I et observasjonsmateriale vil $X_{ij}/X_{\cdot j}$ være estimat for q_{ij}.

Avhengighetsstrukturen mellom X og Y studeres ved hjelp av q_{ij}-ene. X og Y er uavhengige hvis

$$q_{i1} = q_{i2} = \ldots = q_{ik} \quad \text{for } i = 1, 2, \ldots, r.$$

Vi har fremdeles r·k parametre men på disse hviler i alt k restriksjoner idet $\sum_{i=1}^{r} q_{ij} = 1$ for $j = 1, 2, \ldots, k$. Den "totale" avhengighet uttrykkes nå ved k·(r-1) frie parametre. Det vi taper i informasjon når vi går fra p_{ij}-ene til q_{ij}-ene er kunnskap om den marginale fordelingen til Y.

En reparametrisering som blir nærmere behandlet i kapittel 8, oppnår vi ved å sette $\log p_{ij} = \mu + \alpha_i + \beta_j + \alpha_{ij}$ $\quad i = 1, 2, \ldots, r$ og $j = 1, 2, \ldots, k$ med bibetingelsene

$$\sum_i \alpha_i = 0 \;, \; \sum_j \beta_j = 0 \;, \; \alpha_{\cdot j} = \alpha_{i \cdot} = 0 \text{ for alle } (i, j).$$

I denne parametriseringen kommer avhengigheten til uttrykk gjennom α_{ij}-ene, så i alt $(r-1) \cdot (k-1)$ frie parametre beskriver avhengigheten. Uavhengighet har vi hvis $\alpha_{ij} = 0$ for alle (i, j). De marginale fordelinger for X og Y kommer til uttrykk gjennom μ, α_i-ene og β_j-ene.

2. FAKTORER AV BETYDNING FOR VALG AV BETRAKTNINGSMÅTE

Vi vil peke på tre av faktorene som har betydning for valg av assosiasjonsmål/betraktningsmåte. Det er

1) Faktor-respons variable
2) Målenivå på variablene
3) Klasseinndeling av variablene

Vi forklarer nærmere hva som ligger bak stikkordene på lista. Eksempler blir gitt i slutten av kapitlet.

a. Faktor-respons

I noen situasjoner kan det falle naturlig å betrakte en variabel (responsvariablen) med bakgrunn i en annen (faktorvariablen), situasjonen er assymmetrisk. Dette skillet kan oppstå fordi faktorvariablen ligger foran responsvariablen i tid eller fordi faktorvariablen sees på som en mulig forklaring for variasjonen i responsvariablen. Det ligger ikke i dette at det nødvendigvis er noe årsak-virkning forhold mellom faktor og respons, men hvis en har en modell som gir uttrykk for et årsak-virkningsforhold, ville vi kalle årsaken en faktorvariabel og virkningen en responsvariabel. Parallellt, men ikke identisk, med forholdet mellom faktor- og responsvariabel, er begrepene eksogen og endogen variabel i økonomien, og regressand og regressor i regresjonsanalysen. Uavhengig/avhengig variabel er også et lignende begrepspar. I noen tilfelle vil det ikke være naturlig å skille mellom de variable på denne måten, situasjonen er symmetrisk. Vi vil da si at begge variable er responsvariable. Inndelingen i respons- og faktorvariabel avhenger av formålet og kan variere med bruken av et sett data. F.eks. vil det i eksempel 1.1 for noen formål være naturlig å si at begge variable er responsvariable, men skal en predikere alvorlighet på grunnlag av lokalisering, vil lokalisering være faktorvariabel og alvorlighet responsvariabel.

b. Målenivå

Det er vanlig å regne med fire ulike målenivåer. Det er

1) Nominalnivå
2) Ordinalnivå
3) Intervallnivå
4) Forholdstallnivå

Hvis klasseinndelingen av en variabel ikke gir mulighet for noen form for ordning, har vi en variabel på nominalnivå. Eksempel på dette er når kategoriene er kjønn, land eller andre regionale inndelinger nevnt ved navn. For en variabel på ordinalnivå eksisterer det en transitiv ordning mellom kategoriene. Det kan avgjøres om elementene i en kategori er "større" eller "mindre enn" elementene i en annen kategori. Men det kan ikke avgjøres om forskjellen mellom to kategorier er større eller mindre enn forskjellen mellom to andre. Hvis vi i en undersøkelse spør folk hvor enige de er i et spesielt utsagn og svaralternativene er "helt enig", "nesten enig" og "uenig" så måler vi graden av enighet på ordinalnivå. På intervallnivå har det mening å snakke om differanser mellom kategoriene. Måler vi temperatur med Celsiusskalaen kan vi si at forskjellen mellom 5^{o} C og 10^{o} C er like stor som forskjellen mellom 10^{o} C og 15^{o} C. Derimot har det ikke mening å si at 10^{o} C er dobbelt så varmt som 5^{o} C. For å komme med slike utsagn må vi ha forholdstallnivå der divisjon og multiplikasjon har mening. Eksempler på variable målt på forholdstallnivå er alder målt i hele år og inntekt målt i kroner. Vi skal senere se at det er naturlig å stille ulike krav til assosiasjonsmål alt ettersom hvilket målenivå de variable har.

c. Klasseinndeling

Hvis grupperingen av en variabel er framkommet ved klasseinndeling av en kontinuerlig (eller finere inndelt diskret) variabel, vil valget av delepunkter influere på resultatene vi får med de fleste assosiasjonsmål. Resultatet kan f.eks. variere ettersom vi bruker 1,5 eller 10 årsgrupper i en aldersgruppering. Vi skal senere belyse dette ved eksempler.

d. Eksempler

Eksempel 2.1 (Plackett (1974))

7 477 personer er gruppert etter synsevne på høyre og venstre øye. Synsevnen er for hvert øye vurdert til 1, 2, 3 eller 4 der 1 står for høyeste og 4 for laveste synsevne. Vi vil se på sammenhengen mellom syn på høyre og venstre øye.

X = Synsevne venstre øye
Y = Synsevne høyre øye

X \ Y	1	2	3	4	
1	1 520	234	117	36	1 907
2	266	1 512	362	82	2 222
3	124	432	1 772	179	2 507
4	66	78	205	492	841
	1 976	2 256	2 456	789	7 477

Det er naturlig med en simultan vurdering av de variable. Vi betrakter begge som responsvariable. Det ville være kunstig f.eks. å studere fordeling av synsevne på høyre øye gitt synsevne på venstre øye. De variable er inndelt på ordinalnivå og tallene 1, 2, 3, 4 er dermed vilkårlige i den forstand at monotont transformerte verdier, f. eks. 2, 3, 7, 10, ville gjøre samme nytte. Mål for sammenheng bør derfor ikke være avhengig av valget av verdiene 1, 2, 3, 4, men samtidig bør det avsløre at høye verdier av den ene variable ofte forekommer sammen med høye evt. lave verdier på den andre variable.

Slike overveielser vil en kunne gjøre i de fleste situasjoner. Faglig innsikt og spesielle ønskemål vil kunne føre til at en setter opp flere krav som assosiasjonsmålet bør oppfylle.

Eksempel 2.2 (Konstruerte tall)

I en gruppe på 95 personer har 16 vært utsatt for påvirkning av stoffer som en har mistanke om kan være kreftframkallende. En undersøkte om de 95 hadde svulster av en bestemt type. Resultatet ble

X \ Y	Svulst	Ikke svulst	
Utsatt	4	12	16
Ikke utsatt	5	74	79
	9	86	95

Vi betrakter X som faktorvariabel og Y som responsvariabel. Det er naturlig å foreta en horisontal prosentuering i tabellen.

X \ Y	Svulst	Ikke svulst	
Utsatt	25	75	100
Ikke utsatt	06	94	100

Tabellen indikerer en viss sammenheng, men en kan ikke av tabellen alene slutte at stoffene er årsak til svulstene. En årsakssammenheng må fastslås på medisinsk grunnlag. Eksemplet viser at selv om vi fastslår sammenheng mellom en faktor og responsvariabel, så vil som regel ikke det alene være nok til å påvise en årsak - virkning.

Eksempel 2.3 (Konstruerte tall)

La X være karakter ved avsluttet utdanning i et skoleslag og Y være nåværende inntekt for en gruppe personer. Vi ser på X som faktorvariabel og Y som responsvariabel. De betingete fordelinger av inntekt gitt karakter er

X \ Y	Høy	Lav	
Høy	0,15	0,85	1,00
Middels	0,25	0,75	1,00
Lav	0,10	0,90	1,00

Vi ser at blant de med middels karakter er det en større andel som har høy inntekt enn blant de med høye(gode) karakterer. Det er neppe karakterene som er direkte årsak til dette. En bedre forklaring har en antakelig i at karakterne i en viss grad bestemmer type arbeid som igjen er delvis bestemmende for inntekten.

Også andre effekter kan tilsløre bildet og gjøre at sammenheng ikke bør utlegges som årsak-virkning. Det illustreres i følgende eksempel.

Eksempel 2.4 (Konstruerte tall)

En gruppe tidligere røykere ble sammenlignet med en gruppe røykere m.h.t. forekomst av en spesiell sykdom.

	Har sykdommen	Har ikke sykdommen	
Tidligere røykere	0,38	0,62	1,0
Røykere	0,48	0,52	1,0

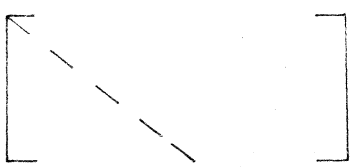
Vi ser at det er en større andel av røykerne som har sykdommen, men kan ikke av tabellen slutte at røyking øker risikoen for å få sykdommen. Forskjellen mellom røykere og tidligere røykere kan skyldes en seleksjon idet de som har sluttet å røyke, kan være mer helsebevisste eller "sterke" enn røykerne, og at det er ett av disse forhold som gjør at de ikke får sykdommen så ofte som røykerne.

I vår inndeling i faktor- og responsvariable er det intet til hinder for at det kan være en årsakssammenheng mellom to responsvariable. I noen situasjoner kan det være naturlig å "snu" et etablert årsak-virkningforhold og si at årsaken er respons og virkningen er faktor. F. eks. i et materiale over ulike sykdommer og symptomer er det sykdommen som er årsak til symptomene, men skal vi måle hvor godt vi kan predikere sykdom på grunnlag av symptomer, kan det være riktig å betrakte symptom som faktor og sykdom som respons.

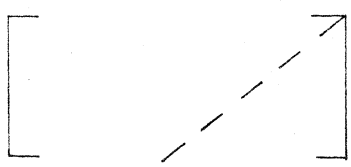
Eksemplene i kapittel 2 illustrerer også at en eventuell påvisning av en statistisk sammenheng ofte bare er ett trinn i en undersøkelse. Når sammenhengen er påvist vil en ofte søke å gå videre og "forklare" sammenhengen ved hjelp av fagkunnskap om etablerte eller hypotetiske årsakssammenhenger. Dette kan igjen avdekke behov for nye statistiske undersøkelser.

3. TO RESPONSVARIABLE PÅ ORDINALNIVÅ

Vi har en symmetrisk situasjon, som eksempel kan vi tenke oss en tabell med personer krysssklassifisert i grupper for utdanning og inntekt (Disse to variable kan måles på høyere målenivå enn ordinalnivå men kan også måles på en slik måte at inndelingen ikke tilfredstiller mer enn kravene til ordinalskalaen). Vi vil ha et mål som gir uttrykk for sammenhengen mellom de to variablene. Vi skal stille spesielle krav til målet i denne situasjonen og vil vise at (Goodman og Kruskals) γ (gamma-koeffisienten), tilfredstiller disse kravene. For det første vil vi kreve at målet er invariant m.h.t. transformasjoner av ordinalskalaen som bevarer ordningen av de ulike responskategoriene. Det betyr f. eks. at målet skal gi samme verdi enten vi kaller kategoriene 1, 2, 3, 4 eller 2, 4, 8, 12. Av målet vil vi også kreve at hvis alle observasjonene ligger på en lengste diagonal f. eks. slik



så skal målet ha verdi +1. Hvis alle observasjonene ligger på en lengste diagonal som er orientert den andre veien f. eks. slik



så skal målet ha verdi -1.

Vi ser på situasjonen med den multinomiske modell for en kryssklassifikasjon m.h.t. to variable. (utdanning og inntekt)

Vi lar (X_1, Y_1) og (X_2, Y_2) være to uavhengige målinger av utdanning og inntekt og definerer

π_s, π_d og π_t.

$$\pi_s = P(X_1 < X_2 \cap Y_1 < Y_2) + P(X_1 > X_2 \cap Y_1 > Y_2),$$
$$= P((X_1-X_2)\cdot(Y_1-Y_2) > 0).$$

$$\pi_d = P(X_1 < X_2 \cap Y_1 > Y_2) + P(X_1 > X_2 \cap Y_1 < Y_2)$$
$$= P((X_1-X_2)\cdot(Y_1-Y_2) < 0).$$

$$\pi_t = P(X_1 = X_2 \cup Y_1 = Y_2)$$
$$= P((X_1-X_2)\cdot(Y_1-Y_2) = 0).$$

"$\cup$" = "det vide eller", enten det ene eller det andre eller begge deler må oppfylles.

π_s er sannsynligheten for at en av personene både har høyere utdannings- og inntektsnivå enn den andre (det er positivt samsvar mellom rangeringen av utdanning og inntekt). π_d er sannsynligheten for at en av personene har høyere utdanningsnivå men lavere inntektsnivå enn den andre (det er negativt samsvar rangeringen av utdanning og inntekt). π_t er sannsynligheten for at utdanningsnivå eller inntektsnivå (eller begge nivåer) er likt for de to personene.

π_s, π_d og π_t kan uttrykkes ved p_{ij} -ene i modellen.

$$\pi_s = 2 \sum_{i=1}^{r-1} \sum_{j=1}^{k-1} p_{ij} \left(\sum_{i'>i} \sum_{j'>j} p_{i'j'} \right).$$

$$\pi_d = 2 \sum_{i=1}^{r-1} \sum_{j=2}^{k} p_{ij} \left(\sum_{i'>i} \sum_{j'<j} p_{i'j'} \right).$$

$$\pi_t = \sum_{i=1}^{r} p_{i\cdot}^2 + \sum_{j=1}^{k} p_{\cdot j}^2 - \sum_{i=1}^{r} \sum_{j=1}^{k} p_{ij}^2$$

I første omgang ser vi på størrelsen $\Delta(=\pi_s - \pi_d)$ som er differansen mellom sannsynligheten for positivt samsvar og sannsynligheten for negativt samsvar. Δ er invariant m.h.t. monotone transformasjoner av verdiene (tallene) som brukes for å betegne klassene i inndelingene. Hvis det er uavhengighet mellom utdanning og inntekt, dvs. $p_{ij} = p_{i\cdot} \cdot p_{\cdot j}$ for alle (i,j), så er $\Delta = 0$. Δ er en populasjonstørrelse, når vi skal estimere denne på grunnlag av et observasjonsmateriale, setter vi inn de observerte hyppighetene $\frac{X_{ij}}{n}$, $\frac{X_{i\cdot}}{n}$ og $\frac{X_{\cdot j}}{n}$ for h.h.v. p_{ij}, $p_{i\cdot}$ og $p_{\cdot j}$ i formelen for Δ.

Eksempel 3.1 (Goodman og Kruskal (1954))

1 438 gifte par er gruppert m.h.t. hustruens utdanning og effektivitet av familieplanleggingen. 1 står for høyeste grad av familieplanlegging og høyeste utdanningsnivå for hustruen er angitt med 1.

		Familieplanlegging				
		1	2	3	4	
Hustruens utdanning	1	102	35	68	34	239
	2	191	80	215	122	608
	3	110	90	168	223	591
		403	205	451	379	1 438

Estimatene for π_s og π_d blir h.h.v. 0,301 og 0,163.

Vi vil lage et avhengighetsmål basert på π_s og π_d. Om disse størrelsene har vi følgende regel.

Regel 3.2

La $m = \min(r, k)$, da vil $-\frac{m-1}{m} \leq \pi_s - \pi_d \leq \frac{m-1}{m}$.

Av regelen ser vi at Δ ikke tilfredstiller vårt krav om at målet skal være +1 eller -1 når alle observasjoner ligger på en lengste diagonal.

Vi setter $\pi_c = (\pi_s - \pi_d)/\frac{m-1}{m}$. π_c antar verdier f.o.m. -1 t.o.m. +1 men har den uheldige egenskap at ytterverdiene oppnås bare hvis $p_{ij} = 1/m$ for alle celler på en lengste diagonal. Vi anser det for naturlig å kreve at avhengighetsmålet gir +1 eller -1 hvis alle celler med positiv sannsynlighet ligger på en lengste diagonal selvom ikke alle cellene på diagonalen har den samme sannsynlighet. Vi anbefaler heller målet γ definert ved:

$$\gamma = (\pi_s - \pi_d)/(1 - \pi_t) =$$

$$P((X_1 - X_2)(Y_1 - Y_2)) > 0 \mid (X_1 - Y_2)(Y_1 - Y_2) \neq 0)$$

$$- P((X_1 - Y_2)(Y_1 - Y_2) < 0 \mid (X_1 - X_2)(Y_1 - Y_2) \neq 0).$$

γ er ikke definert hvis alle celler med positive sannsynligheter ligger på samme rad eller i samme kolonne for da er $\pi_t = 1$. γ tilfredstiller de krav vi satte hvis alle celler med positiv sannsynlighet ligger på en lengste diagonal. Hvis de to variable er uavhengige, så blir $\gamma = 0$.

Siden $1 - \pi_t = \pi_s + \pi_d$ så er $\gamma = (\pi_s - \pi_d)/(\pi_s + \pi_d)$. Av denne skrivemåten ser en at hvis γ er definert, så vil $\gamma = 1$ for $\pi_d = 0$ og $\gamma = -1$ for $\pi_s = 0$. Dette viser at γ kanskje er et noe "grovt" mål f. eks. i tabellen

$$\begin{bmatrix} p_{11} & p_{12} & 0 \\ 0 & p_{22} & p_{23} \\ 0 & 0 & p_{33} \end{bmatrix} \quad \text{så er } \gamma = 1.$$

og i tabellen

$$\begin{bmatrix} p_{11} & 0 & 0 \\ p_{21} & 0 & 0 \\ p_{31} & p_{32} & p_{33} \end{bmatrix} \quad \text{så er } \gamma = 1.$$

Når γ beregnes av én tabell så kan vi tolke tallverdien ved hjelp av definisjonen, men det vil som regel være vanskelig å uttale seg om graden av sammenheng på grunnlag av det ene tallet. Nytte av γ og lignende mål har vi først og fremst når vi skal undersøke sammenhengen mellom to variable i ulike materialer. Da kan γ beregnes separat og en kan foreta en relativ sammenligning. Et moment ved slike sammenligninger er at de marginale fordelinger i tabellene en sammenligner, ikke bør avvike for meget fra hverandre. γ er følsom for forandringer i marginalene i den forstand at hvis en ut fra én tabell lager en ny ved f. eks. å multiplisere radene med ulike positive tall, så vil γ kunne bli forskjellig for de to tabellene. I en del tilfelle er dette en mindre heldig egenskap ved γ og mange andre assosiasjonsmål.

I en 2x2 tabell har vi at

$$\gamma = 2(p_{11}p_{22} - p_{12}p_{21})/2(p_{11}p_{22} + p_{12}p_{21})$$

$$= (p_{11}p_{22} - p_{12}p_{21})/(p_{11}p_{22} + p_{12}p_{21})$$

$$= (\delta - 1)/(\delta + 1) \quad \text{når } \delta = (p_{11}p_{22})/(p_{12}p_{21})$$

δ er en grunnleggende størrelse som i seg selv er et interaksjonsmål.

γ^2 kan tolkes som forklart varians. En del andre avhengighetsmål er utledet ut i fra tankegang om forklart varians.

Vi anbefaler γ som mål i situasjonen med to responsvariable på ordinalnivå. Har vi en assymmetrisk situasjon med to ordinalvariable, dvs. en respons- og en faktorvariabel, kan det være naturlig å bruke andre mål. Somers (1962) har foreslått to slike assymmetriske mål, med våre symboler kan disse skrives:

$$d_{yx} = (P((X_1 - X_2)(Y_1 - Y_2) > 0) - P((X_1 - X_2)(Y_1 - Y_2) < 0))/P(X_1 \neq X_2)$$
$$= (\pi_s - \pi_d)/P(X_1 \neq X_2)$$

når X er faktor - og Y responsvariabel og

$$d_{xy} = (P((X_1 - X_2)(Y_1 - Y_2) > 0) - P((X_1 - X_2)(Y_1 - Y_2) < 0))/P(Y_1 \neq Y_2)$$
$$= (\pi_s - \pi_d)/P(Y_1 \neq Y_2)$$

når Y er faktor- og X responsvariabel. Vi antar vi fremdeles har en multinomisk modell og har da

$$P(X_1 \neq X_2) = 1 - \sum_{i=1}^{r} p_{i\cdot}^2 \quad \text{og}$$

$$P(Y_1 \neq Y_2) = 1 - \sum_{j=1}^{k} p_{\cdot j}^2 .$$

Mens vi med γ så på differensen mellom sannsynlighetene for positivt og negativt samsvar på rangeringene relativt til sannsynligheten for at ingen av rangeringene var like, ser vi med d_{yx} og d_{xy} på den samme differensen relativt til sannsynligheten for at de to tilfeldige enhetene ikke er likt rangert m.h.t. faktorvariablen.

4. TO VARIABLE PÅ NOMINALNIVÅ

Målene som behandles i underkapitlene a, b og c er behandlet i Goodman og Kruskal (1954 og 1959).

a) En faktor- og en responsvariabel.

Situasjonen er assymmetrisk og vi lar X være faktor- og Y responsvariabel. F.eks. kan X angi gymnas i Oslo og Y yrkeskategori (etter fullført utdanning) for en gruppe personer. Vi vil se på sammenhengen mellom X og Y. Målet som vi foreslår framkommer ved at vi for en enhet som trekkes tilfeldig i populasjonen, skal predikere Y (yrkeskategori) i to situasjoner.

1) Uten å kjenne verdien av X for vedkommende (hvilket gymnas vedk. har gått på)

2) Når vi kjenner verdien av X.

For å få vist hvilken populasjonstørrelse vi vil måle sammenhengen med, antar vi at alle p_{ij}-ene i modellen er kjent.

I begge situasjoner vil vi gjette optimalt, vi vil gjette på den Y kategori som har størst sannsynlighet. Det betyr at vi i situasjon 1) gjetter Y = m hvis $p_{\cdot m} = \max_{1 \leq j \leq k} p_{\cdot j}$. I situasjon 2) vil vi når det er oppgitt at X = i, gjette at Y = m hvis $p_{im} = \max_{1 \leq j \leq k} p_{ij}$ for denne gitte i. Om vi i situasjon 2) baserte oss på de betingete sannsynligheter gitt X = i ville vi komme fram til samme Y kategori, da de betingete sannsynligheter framkommer ved å dividere p_{ij}-ene i rad i med samme tall, $p_{i\cdot}$. Hvis vi skulle komme ut for at to eller flere kolonner har den maksimale verdi, kan vi bare gjette på en av dem, f.eks. den med minst kolonnenummer.

I de to situasjoner vil vi ha følgende sannsynligheter for å gjette feil:

$$P_{(1)}\,(\text{feil}) = 1 - p_{\cdot m}$$

$$P_{(2)}\,(\text{feil}) = \sum_{i=1}^{r} P(\text{feil} \cap X = i)$$

$$= \sum_{i=1}^{r} P(\text{feil} \mid X = i) \cdot P(X = i)$$

$$= \sum_{i=1}^{r} (1 - p_{im}/p_{i\cdot})\, p_{i\cdot} = 1 - \sum_{i=1}^{r} p_{im}$$

Som mål på sammenhengen mellom X og Y bruker vi den relative minsking i feilsannsynligheten vi oppnår ved å kjenne X-kategorien. Målet betegner vi λ_b.

$$\lambda_b = \frac{P_{(1)}(\text{feil}) - P_{(2)}(\text{feil})}{P_{(1)}(\text{feil})}$$

$$= (1 - p_{.m} - (1 - \sum_{i=1}^{r} p_{im}))/(1 - p_{.m})$$

$$= (\sum_{i=1}^{r} p_{im} - p_{.m})/(1 - p_{.m})$$

Hvis X er respons- og Y faktorvariabel finner vi på tilsvarende måte fram til målet λ_a.

$$\lambda_a = (\sum_{j=1}^{k} p_{m'j} - p_{m'.})/(1 - p_{m'.}).$$

Vi har da definert $p_{m'j} = \max_{1 \le i \le k} p_{ij}$ for gitt j og $p_{m'.} = \max_{1 \le i \le k} p_{i}$.

Når vi har et sett observasjoner og skal estimere λ_a evt. λ_b, setter vi inn relative hyppigheter for sannsynlighetene i formlene.

Vi skal se på noen av egenskapene til λ_b men vil først peke på følgende relasjon som alltid gjelder:

$$P_{(2)}(\text{feil}) \le P_{(1)}(\text{feil}).$$

Av definisjonen ser en at λ_b bl. a. har følgende egenskaper:

i) Hvis største p_{ij} for hver i er i samme kolonne så blir $\lambda_b = 0$ (X inneholder ingen informasjon om Y).

ii) λ_b ikke definert når $P_{(1)}(\text{feil}) = 0$, da er $p_{.m} = 1$ (Alle celler med positiv sannsynlighet ligger i samme kolonne).

iii) X og Y uavhengige medfører $\lambda_b = 0$.

iv) $\lambda_b = 1$ ekvivalent med $P_{(2)}(\text{feil}) = 0$.

v) $\lambda_b = 0$ ekvivalent med $P_{(1)}(\text{feil}) = P_{(2)}(\text{feil})$.

vi) λ_b er invariant overfor permutasjoner av rader eller kolonner.

λ_a har tilsvarende egenskaper.

Av egenskapene kan en merke seg vi). Fordi det ikke er noen ordning mellom klassene, er dette et krav som et mål i denne situasjon absolutt bør oppfylle.

b) To responsvariable

Eksempel 4.1 (Goodman og Kruskal (1954))

Vi vil se på sammenheng mellom øyenfarge og hårfarge. X angir øyenfarge, mulige kategorier er A_1 (blå), A_2 (grå, grønne) og A_3 (brune). Y angir hårfarge, mulighetene er B_1 (lys), B_2 (brunt), B_3 (sort) og B_4 (rødt). 6 800 personer som er trukket tilfeldig i en populasjon, ble undersøkt, tabellen viser resultatet.

	B_1	B_2	B_3	B_4	
A_1	1 768	807	189	47	2 811
A_2	946	1 387	746	53	3 132
A_3	115	438	288	16	857
	2 829	2 632	1 223	116	6 800

Siden vi ikke har spesifisert annet enn at vi vil undersøke sammenhengen og siden personene trukket tilfeldig i populasjonen, er det naturlig å betrakte både X og Y som responsvariable. Hvis vi derimot hadde hatt et sett med faste marginaler f. eks. ved at vi hadde trukket et bestemt antall personer med hver hårfarge, kunne med fordel hårfargen regnes som faktorvariabel.

Mål på avhengighet i situasjonen med to responsvariable på nominalnivå utvikler vi ved å tenke oss at vi trekker en tilfeldig person fra populasjonen og skal med sannsynlighet 1/2 gjette A-kategori og med sannsynlighet 1/2 gjette B-kategori i situasjonene med:

1) Ingen informasjon om den andre variablen.

2) Kunnskap om verdien på den andre variablen.

Når vi gjetter, følger vi en optimal strategi dvs. vi vil i enhver situasjon gjette på den av de mulige kategoriene som har størst sannsynlighet. Som mål på avhengighet vil vi igjen bruke den relative minsking i feilsannsynligheten vi får i situasjon 2. Målet betegnes λ og vi viser hvordan det kan uttrykkes ved p_{ij} -ene når $p_{m'.}$, $p_{m'j}$, $p_{.m}$ og p_{im} definert som før.

$$P_{(1)}(\text{feil}) = P_{(1)}(\text{feil} \mid \text{skal gjette A-kategori}) \cdot P(\text{skal gjette A-kategori})$$
$$+ P_{(1)}(\text{feil} \mid \text{skal gjette B-kategori}) \cdot P(\text{skal gjette B-kategori})$$
$$= (1 - p_{m'.}) \cdot \tfrac{1}{2} + (1 - p_{.m}) \cdot \tfrac{1}{2}$$
$$= 1 - \tfrac{1}{2} \cdot (p_{m'.} + p_{.m})$$

$$P_{(2)}(\text{feil}) = P_{(2)}(\text{feil} \mid \text{skal gjette A-kategori}) \cdot P(\text{skal gjette A-kategori})$$
$$+ P_{(2)}(\text{feil} \mid \text{skal gjette B-kategori}) \cdot P(\text{skal gjette B-kategori})$$
$$= (1 - \sum_{j=1}^{k} p_{m'j}) \cdot \tfrac{1}{2} + (1 - \sum_{i=1}^{r} p_{im}) \cdot \tfrac{1}{2}$$
$$= 1 - \tfrac{1}{2} \cdot (\sum_{j=1}^{k} p_{m'j} + \sum_{i=1}^{r} p_{im})$$

$$\lambda = \frac{P_{(1)}(\text{feil}) - P_{(2)}(\text{feil})}{P_{(1)}(\text{feil})}$$
$$= (\sum_{j=1}^{k} p_{m'j} + \sum_{i=1}^{r} p_{im} - p_{.m} - p_{m'.})/(2 - (p_{.m} + p_{m'.}))$$

Noen egenskaper ved λ:

i) λ udefinert når $P_{(1)}(\text{feil}) = 0$ dvs. $p_{.m} + p_{m'.} = 2$ dvs. en celle har sannsynlighet 1.

ii) Hvis de to variable er stokastisk uavhengige dvs. $p_{ij} = p_{i.} \cdot p_{.j}$ for alle (i,j) så er $\lambda = 0$. (Det omvendte gjelder ikke).

iii) $\lambda = 1$ hvis og bare hvis hver rad og kolonne har høyst en celle med positiv sannsynlighet dvs. $\sum_{i=1}^{r} p_{im} = \sum_{j=1}^{k} p_{m'j} = 1$.

Beregner vi de relative hyppigheter i eksempel 4.1 og setter inn for p_{ij} -ene, får vi

$$\lambda = 0{,}2076.$$

Tallene i eksemplet kan også brukes til et regneeksempel for λ_a og λ_b, da får en

$$\lambda_a = 0{,}2241 \text{ og}$$
$$\lambda_b = 0{,}1924.$$

c. Andre mål fra samme aktivitetsmodell

Vi kan utvikle tilsvarende mål som λ_a, λ_b og λ ved å ta utgangspunkt i de samme situasjonene, men når vi gjetter, så gjetter vi proposjonalt med sannsynlighetene istedenfor å gjette på den kategori som har størst sannsynlighet. Hvis vi f.eks. har kategoriene A_1, A_2 og A_3 med sannsynligheter 0,1, 0,2 og 0,7, kan vi utføre gjettingene ved å trekke tilfeldig én av 10 lapper hvorav 1 er merket A_1, 2 er merket A_2 og 7 er merket A_3.

I den assymetriske situasjonen med A som faktorvariabel og B som responsvariabel, finner vi:

1) Uten informasjon om A-verdi gjetter vi B_j med sannsynlighet p.j $j = 1, 2, \ldots, k$.

$$P_{(1)}(\text{feil}) = 1 - \sum_{j=1}^{k} p_{.j}^2$$

2) Med informasjon om at A-kategori er A_i, gjetter vi B_j med sannsynlighet $p_{ij}/p_{i.}$.

$j = 1, 2, \ldots, k$. Sannsynlighet for feil når A_i er gitt er $(1 - \sum_{j=1}^{k} (\frac{p_{ij}}{p_{i.}})^2)$

$$P_{(2)}(\text{feil}) = \sum_{i=1}^{r} (1 - \sum_{j=1}^{k} (\frac{p_{ij}}{p_{i.}})^2) \cdot p_{i.})$$

$$= 1 - \sum_{i=1}^{r} \frac{1}{p_{i.}} \sum_{j=1}^{k} p_{ij}^2$$

Målet η_b defineres som den relative minsking i feil

$$\eta_b = \frac{P_{(1)}(\text{feil}) - P_{(2)}(\text{feil})}{P_{(1)}(\text{feil})}$$

$$= (\sum_{i=1}^{r} \frac{1}{p_{i.}} \cdot \sum_{j=1}^{k} p_{i}^{2}{}_{j} - \sum_{j=1}^{k} p_{.j}^2)/(1 - \sum_{j=1}^{k} p_{.j}^2)$$

$$= (\sum_{i=1}^{r} \sum_{j=1}^{k} \frac{p_{ij}^2}{p_{i.}} - \sum_{j=1}^{k} p_{.j}^2)/(1 - \sum_{j} p_{.j}^2)$$

kan også skrives

$$(\sum_{i=1}^{r} \sum_{j=1}^{k} (p_{ij} - p_{i.} \cdot p_{.j})^2/p_{i.})/(1 - \sum_{j=1}^{k} p_{.j}^2)$$

Vi kan utlede tilsvarende formel for η_a. I den symmetriske situasjonen gir denne strategien for gjettingen målet η.

$$\eta = \frac{P_{(1)}(\text{feil}) - P_{(2)}(\text{feil})}{P_{(1)}(\text{feil})}$$

$$= ((1 - \sum_{j=1}^{k} p_{.j}^2 + 1 - \sum_{j=1}^{r} p_{i.}^2) \tfrac{1}{2} - (1 - \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{p_{ij}^2}{p_{i.}} + 1 - \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{p_{ij}^2}{p_{.j}}) \cdot \tfrac{1}{2})/$$

$$(1 - \sum_{j=1}^{k} p_{.j}^2 + 1 - \sum_{i=1}^{r} p_{i.}^2) \tfrac{1}{2}$$

$$= ((\sum_{i=1}^{r} \sum_{j=1}^{k} p_{ij}^2 \cdot (\frac{1}{p_{i.}} + \frac{1}{p_{.j}}) - \sum_{j=1}^{k} p_{.j}^2 - \sum_{i=1}^{r} p_{i.}^2)/(2 - (\sum_{j=1}^{k} p_{.j}^2 + \sum_{i=1}^{r} p_{i.}^2))$$

$$= (\sum_{i=1}^{r} \sum_{j=1}^{k} (p_{ij} - p_{i.}p_{.j})^2 \cdot (\frac{1}{p_{.j}} + \frac{1}{p_{i.}}))/(2 - (\sum_{j=1}^{k} p_{.j}^2 + \sum_{i=1}^{r} p_{i.}^2)).$$

Disse målene har stort sett de samme egenskaper som λ_a, λ_b og λ. Spesielt er at $\eta = 0$ hvis og bare hvis $p_{ij} = p_{i.} \, p_{.j}$ for alle (i, j), dvs. uavhengighet mellom de variable.

d. Sammenligning med andre mål

I situasjoner som i dette kapitlet vil en finne at det ofte brukes avhengighetsmål som er avledet av

$\phi^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} ((p_{ij} - p_{i.}p_{.j})^2/p_{i.} \cdot p_{.j})$. Hvis vi setter inn de observerte hyppighetene i formelen, får vi at dette blir X^2/n der X^2 er kjikvadratobservatoren som vi bruker ved testing av uavhengighet mellom de variable. Av formelen ser vi at ϕ^2 måler avvik fra uavhengighet, men det er vanskelig å finne noen sannsynlighetsmessig eller operasjonell tolkning av målene som er avledet av denne størrelsen. Vi vil derfor i situasjoner der det på en eller annen måte er den "prediktive" sammenheng som er interresant, anbefale λ evt. η-målene. I andre situasjoner bør en undersøke om de interessante avhengighetstrekk kan måles ved avhengighetsmål som har enkle sannsynlighetsmessige eller operasjonelle tolkninger. En fordel slike mål har framfor kjikvadratmålene, er at de som regel er

enkle å modifisere når vi skal studere deltabeller. Hvis vi f. eks. har en krysstabell over yrke til fedre og sønner og vil studere sammenhengen mellom yrkene for de far-sønn par som ikke har samme yrke (dvs. tabellen bortsett fra den ene hoveddiagonalen), så mener vi det er vanskelig å finne en modifikasjon av kjikvadratmålene for denne deltabellen. λ, derimot, kan vi beregne i tabellen over den nye massen. Denne tabellen har 0-er på den ene diagonalen og nye marginaler i forhold til den opprinnelige.

Eksempel 4.2 (Konstruerte tall)

Anta vi har gruppert yrkene til far-sønn par etter en inndeling med tre yrkeskategorier. Den opprinnelige tabell er:

		Sønns yrke 1	2	3	
	1	40	10	10	60
Fars yrke	2	10	30	5	45
	3	5	10	25	40
		55	50	40	145

For denne tabellen er det klart hvordan en skal bruke både "kji-kvadrat-mål" og andre. Men vil vi se på sammenheng mellom yrkene til de par som ikke har samme yrke, ser vi på tabellen

		Sønns yrke 1	2	3	
	1	0	10	10	20
Fars yrke	2	10	0	5	15
	3	5	10	0	15
		15	20	15	50

og det er meningsløst å sette disse tallene rett inn i et "kji-kvadrat-mål". Derimot kan λ og η beregnes direkte fordi det har mening å snakke om sannsynligheten for å predikere sønns(fars) yrke med eller uten informasjon om fars(sønns) yrke gitt at de ikke har samme yrke.

5. SAMSVÅR, EN SPESIELL FORM FOR SAMMENHENG.

Eksempel 5.1

To psykologer, A og B, klassifiserer hver for seg en gruppe personer etter en inndeling i personlighetstyper med r grupper. Resultatet er stilt opp i tabellen.

	B_1	B_2	...	B_r	
A_1	X_{11}	X_{12}	...	X_{1r}	$X_{1.}$
A_2	X_{21}	X_{22}	...	X_{2r}	$X_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_r	X_{r1}	X_{r2}	...	X_{rr}	$X_{r.}$
	$X_{.1}$	$X_{.2}$	...	$X_{.r}$	n

Tabellen viser at av de n personene ble X_{12} av psykolog A sagt å være av type 1 og av psykolog B bedømt til å være av type 2. X_{22} personer ble av begge psykologer klassifisert som type 2. Vi vil måle graden av enighet eller samsvar mellom vurderingene til de to psykologene.

Vi ser på den multinomiske modell for å finne hvilke populasjonsstørrelser som kan tenkes å gi uttrykk for ulike grader av samsvar. p_{ij} er sannsynligheten for at en person klassifiseres av A som type i og av B som type j.

Et naturlig mål er $\rho = \sum_{i=1}^{r} p_{ii}$ som direkte gir sannsynligheten for samsvar i klassiferingene til de to. Dette mål er enkelt og lett forståelig, men i noen situasjoner vil det være en ulempe at det er så følsomt m.h.t. marginalene. F. eks. i tabellen

	B_1	B_2	
A_1	0,6	0,2	0,8
A_2	0,2	0,0	0,2
	0,8	0,2	1,0

er ρ = 0,6 og det er den minste verdi ρ kan ha med disse marginalene, men i tabellen

	B_1	B_2	
A_1	0,2	0,6	0,8
A_2	0,0	0,2	0,2
	0,2	0,8	1,0

er ρ = 0,4 og med marginalene gitt som i denne tabellen, er det den største verdien ρ kan ha. En må vurdere disse forhold før ρ brukes til å sammenligne samsvaret i tabellene.

Hvis den første tabellen viser resultatet av to psykologers klassifisering av en gruppe personer og den andre tabellen viser resultatet av to andre psykologers klassifisering av en annen gruppe og vi vil sammenligne samsvaret mellom klassifiseringen til de to parene med psykologer, mener vi at ρ kan brukes hvis der ikke er noen spesiell spesifisering av problemstillingen. De sterkt ulike marginalene i den andre tabellen er i seg selv indikasjon på stor grad av uenighet og vi vil ikke justere bort dette trekket ved materialet. Men hvis f.eks. den psykolog som sto bak klassifiseringene B_1, B_2 i den andre tabellen, var av en annen "skole" enn de andre tre og dette medfører at marginalene blir så ulike og vår hensikt er å sammenligne samsvaret mellom psykologparene sett bort i fra at den ene tilhører en annen skole, ser vi at ρ er lite egnet.

Hvis marginalene på en eller annen måte er framkommet av årsaker som er uvesentlige for vår problemstilling, er det altså gunstig med samsvarsmål som er uavhengig eller mindre avhengig av marginalene enn ρ. Vi skal se på et slikt mål.

Når to personer skal klassifisere den samme gruppe enheter, vil en forvente at et visst antall enheter av ren tilfeldighet klassifiseres likt. Det neste samsvarsmålet bygger på tankegangen om at det en skal måle, er det relative samsvar ut over det forventede ved stokastisk uavhengighet mellom klassifiseringene. Vi setter

$$\theta_1 = P(\text{samsvar i alt}) = \sum_{i=1}^{r} p_{ii} \quad (= \rho)$$

$$\theta_2 = P(\text{samsvar ved uavhengighet}) = \sum_{i=1}^{r} p_{i\cdot} \cdot p_{\cdot i}$$

$\theta_1 - \theta_2$ måler sannsynligheten for samsvar fratrukket sannsynligheten for samsvar ved ren slump. Vi normerer denne størrelsen ved å dividere med $1-\theta_2$. $1-\theta_2$ angir "mulig samsvar" som ikke skyldes ren tilfeldighet.

Bishop et. al. (1975) påstår på side 395 at $1-\theta_2$ er maksimum av $\theta_1 - \theta_2$ for gitte marginaler. Dette er oftest galt fordi en nødvendig betingelse for at $\theta_1 = 1$, er at de to marginalfordelingene er like. Vårt mål for samsvar blir

$$K = (\theta_1 - \theta_2)/(1 - \theta_2).$$

K vil anta verdier mindre eller lik 1. Minste verdi K kan anta avhenger av marginalfordelingene.

Estimater for K i den multinomiske situasjon får vi ved å sette inn de observerte hyppighetene.

Eksempel 5.2 (Bishop et. al. (1975)).

To psykologer, A og B, klassifiserer hver for seg 72 lærerstudenters "stil" i klasserommet som h.h.v. A(utoritær), D(emokratisk) eller E(ttergivende).

		Psykolog B			
		A	D	E	
Psykolog A	A	17	4	8	29
	D	5	12	0	17
	E	10	3	13	26
		32	19	21	72

Tabellen gir $\hat{K} = 0{,}362$. Ved assymptotisk teori finner vi at (0,156, 0,568) er et 95% konfidensintervall for K.

For å få nærmere innsikt i samsvarsstrukturen kan vi se på sannsynligheten for samsvar gitt en spesiell verdi av den ene marginalen. Da har vi f.eks.

$$P(\text{samsvar} \mid \text{i'te rad}) = p_{ii}/p_{i.}$$

Vi ser på $p_{ii}/p_{i.} - p_{.i}$ som angir økningen i sannsynlighet for at "B vil velge" kategori i når A har gjort det. Siden 1 er maksimum av $\frac{p_{ii}}{p_{i.}}$ velger vi $K_i = (p_{ii}/p_{i.} - p_{.i})/(1 - p_{.i}) = (p_{ii} - p_{i.} \cdot p_{.i})/(p_{i.} - p_{i.} \cdot p_{.i})$ som et betinget samsvarsmål. Med tallene i eksempel 5.2 finner vi $K_1 = 0{,}255$, $K_2 = 0{,}600$ og $K_3 = 0{,}294$.

Fra K_i -ene kommer vi tilbake til K ved å summere over i i nevner og teller for seg.

Vi vil til slutt understreke at samsvar er en spesiell form for sammenheng. I tabellen

	B_1	B_2	B_3	B_4
A_1	0	1	0	0
A_2	0	0	0	1
A_3	1	0	0	0
A_4	0	0	1	0

er sammenhengen maksimal f. eks. i prediktiv forstand men der finnes ikke samsvar.

6. KLASSEINNDELINGEN AV MATERIALET

Sammenheng målt med et eller annet assosiasjonsmål vil som nevnt i innledningen, avhenge av klasseinndelingen av materialet. Hvis vi har en symmetrisk, ordnet situasjon, hvor vi finner det naturlig å bruke γ som mål på sammenheng, kan denne avhengigheten demonstreres ved følgende materiale (Goodman og Kruskal (1954)).

	B_1	B_2	B_3	B_4
A_1	0	0,25	0	0
A_2	0,25	0	0	0
A_3	0	0	0	0,25
A_4	0	0	0,25	0

Her er $\gamma = \dfrac{(0{,}25 \cdot 0{,}5 + 0{,}25 \cdot 0{,}5) - (0{,}25 \cdot 0{,}25 + 0{,}25 \cdot 0{,}25)}{(0{,}25 \cdot 0{,}5 + 0{,}25 \cdot 0{,}5) + (0{,}25 \cdot 0{,}25 + 0{,}25 \cdot 0{,}25)}$

$$= \frac{0{,}25\ (\ 1 - 0{,}5)}{0{,}25 \cdot 1{,}5} = \frac{1/2}{3/2} = \frac{1}{3}$$

Slår vi sammen B_1 og B_2, B_3 og B_4, A_1 og A_2 og A_3 og A_4, blir tabellen

	$B_1\ B_2$	$B_3\ B_4$
$A_1\ A_2$	0,5	0,0
$A_3\ A_4$	0,0	0,5

Da blir $\gamma = 1$.

Slår vi sammen B_2, B_3 og B_4 og A_2, A_3 og A_4 blir tabellen

	B_1	$B_2\ B_3\ B_4$
A_1	0	0,25
$A_2\ A_3\ A_4$	0,25	0,25

Da blir $\gamma = -1$.

Det er neppe trolig at en i praksis kommer opp i en så vrang situasjon som den ovenfor, men den illustrerer at målene må betraktes relativt til klasseinndelingen. Sammenligning av verdier bør kun skje mellom materialer der samme klasseinndeling ligger til grunn.

Vi vil vise at med et gitt antall klasser, har valg av delepunkter innflytelse på verdien av ulike mål. Vi antar at de variable (X, Y) egentlig er kontinuerlige men at vi grupperer målingene i en 2 x 2 - tabell.

Vi definerer, $F(x, y) = P(X \leq x \cap Y \leq y)$, $H(x) = P(X \leq x)$ og $G(y) = P(Y \leq y)$. Med delepunkter (x, y) kan cellesannsynlighetene i en tabell uttrykkes ved F, H og G.

Y / X	$Y \leq y$	$Y > y$	
$X \leq x$	F	H-F	H
$X > x$	G-F	I-G-H+F	I-H
	G	I-G	

Vi har før sett på målet γ. I 2 x 2 situasjonen var $\gamma = (\delta - 1)/(\delta + 1)$ der δ er kryssproduktet dvs. $\delta = p_{11}\, p_{22}/p_{12}\, p_{21}$. Av tabellen får vi, $\delta = F \cdot (I-G-H+F)/(G-F)\cdot(H-F)$.

Av denne formelen kan vi under ulike forutsetninger om simultanfordelingen til (X, Y), beregne hvordan δ (og dermed γ) varierer med ulike valg av delepunkter. I tabellen nedenfor har vi satt opp hvordan δ varierer med valg av delepunkter når (X, Y) er binormale med korrelasjonskoeffisient 0,75.

X \ Y	0,0	0,5	1,0	1,5	2,0
0,0	11,20	13,36	22,86	56,67	200,84
0,5	13,36	12,03	15,38	27,70	69,65
1,0	22,86	15,38	15,04	20,66	38,91
1,5	56,67	27,70	20,66	22,30	33,17
2,0	200,84	69,65	38,91	33,17	40,31

Tabellen er fra Mosteller (1968). Vi vil ikke bruke tabellen som et argument mot δ eller γ som sammenhengsmål, men vil oppfordre til at når de brukes, så må det framgå hvilken klasseinndeling som er brukt.

7. ET SPESIELT SYN PÅ ASSOSIASJON I KONTINGENSTABELLER

Vi skal se på assosiasjon i kontingenstabeller fra en ny synsvinkel.
Denne betraktningsmåten finner en i Mosteller (1968).

Eksempel 7.1 (Konstruerte tall)

Vi skal se på sammenheng mellom kjønn og ytelse (karakterer) ved en eksamen. Resultatet ble

	Kvinner	Menn	
Høy	2	3	5
Lav	1	4	5
	3	7	10

Nå synes Mosteller å mene at sammenhengen ikke forandres om tallene i første kolonne f. eks. multipliseres med 10. En slik multiplikasjon forandrer ikke det relative antall av kvinnene som har h.h.v. høye og lave karakterer. Han påstår altså at assosiasjonen er den samme i tabellen.

	Kvinner	Menn	
Høy	20	3	23
Lav	10	4	14
	30	7	37

Hvis vi i den første tabellen multipliserer tallene i annen rad med 2, framkommer tabellen.

	Kvinner	Menn	
Høy	2	3	5
Lav	2	8	10
	4	11	15

Assosiasjonen er også den samme i denne tabellen, hevder Mosteller. Vi kommer senere tilbake til dette. I første omgang kan en kanskje notere seg at operasjonen ikke forandrer kjønnenes relative andel av de

med h.h.v. høye og lave karakterer.

Tankegangen medfører at tabellene

$$\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \quad \text{og}$$

$$\begin{bmatrix} k_1\, r_1\, p_{11} & k_2\, r_1\, p_{12} \\ k_1\, r_2\, p_{21} & k_2\, r_2\, p_{22} \end{bmatrix}$$

gir uttrykk for samme grad av assosiasjon. Ved vekselvis multiplikasjon av kolonner og rader, kan en fra den første tabellen i eksempel 7.1 komme fram til en tabell der marginalene er like og der assosiasjonen ifølge Mosteller, er den samme som i den første tabellen.

2	3	5
1	4	5
3	7	

↓ (Multipliserer med 0,2 i rad 1 og med 0,2 i rad 2)

0,4	0,6	1,0
0,2	0,8	1,0
0,6	1,4	

↓ (Multipliserer med 5/3 i kolonne 1 og med 5/7 i kolonne 2

2/3	3/7	23/21
1/3	4/7	19/21
1,0	1,0	

↓ (Multipliserer med 21/23 i rad 1 og 21/19 i rad 2)

.
.
.

↓

0,6202	0,3798	1,0000
0,3798	0,6202	1,0000
1,0000	1,0000	

eller

0,3101	0,1899	0,5000
0,1899	0,3101	0,5000
0,5000	0,5000	

hvis vi vil at marginalene skal summere seg til 1,0.

Vi oppnår å få tabellen på en lett lesbar form. Vi kan bruke samme teknikk til å transformere tabellen slik at de får et hvilket som helst sett med oppgitte marginaler.

Vi kan ikke erklære oss helt enige i begrunnelsen for teknikken. I eksempel 7.1 finner vi det naturlig å se på kjønn som faktorvariabel og ytelse som responsvariabel. Det er da naturlig å sammenligne de betingete fordelinger for ytelse gitt kjønn. Disse fordelingene forandres ikke ved multiplikasjon i kolonnene, så det kan vi være med på. Men når radene multipliseres med ulike tall så vil de betingete fordelinger endres. Hvis f. eks. den opprinnelige tabell er

	Kvinner	Menn
Høy	m_{11}	m_{12}
Lav	m_{21}	m_{22}

og vi multipliserer første rad med a blir den nye tabellen

	Kvinner	Menn
Høy	am_{11}	am_{12}
Lav	m_{12}	m_{22}

Vi lar q_{11} stå for den betingete sannsynlighet for høy karakter gitt at kjønnet er kvinne. I den siste tabellen blir $q_{11} = (am_{11})/(m_{21} + am_{11}) = 1 - m_{21}/(am_{11} + m_{21})$. q_{11} vil variere mellom 0 og 1 alt ettersom a er svært stor eller svært liten. Dette betyr at hvis en mener det er de betingete fordelinger som er essensielle, så må det, før en bruker denne teknikken, vurderes hvordan de ulike transformasjoner vil innvirke på de sammenligninger en akter å foreta.

Vi skal se på et eksempel der teknikken brukes på krysstabeller med flere enn to kategorier.

Eksempel 7.2 (Mosteller(1968))

		Sønns yrkesstatus 1	2	3	4	5	
Fars yrkes-status	1	50 18	45 17	8 16	18 4	8 2	129 57
	2	28 24	174 105	84 109	154 59	55 21	495 318
	3	11 23	78 84	110 289	223 217	96 95	518 708
	4	14 8	150 49	185 175	714 348	447 198	1 510 778
	5	3 6	42 8	72 69	320 201	411 246	848 530
		106 79	489 263	459 658	1 429 829	1 017 562	

Tabellen viser to materialer der fedre og sønner er kryssklassifisert m.h.t. yrkesstatus. I hver celle er det øverste tallet fra et britisk materiale og det nederste tallet fra et dansk materiale. Marginalene er forskjellige i de to deltabellene og det gjør det vanskelig å sammenligne strukturen i tallmaterialene.

I de to deltabellene kan en foreta vekselvis kolonne- og radmultiplikasjon til en får marginaler lik 100. Resultatet av disse itereringene er satt opp i tabellen nedenfor.

		Sønns yrkesstatus					
		1	2	3	4	5	
Fars yrkes-status	1	68,5 58,6	20,9 25,0	4,6 12,0	3,7 2,6	2,3 1,8	100
	2	17,8 21,1	37,5 41,6	22,5 21,9	14,7 10,3	7,5 5,1	100
	3	8,0 11,7	19,2 19,3	33,7 33,7	24,3 21,9	14,9 13,5	100
	4	4,1 4,1	14,7 11,4	22,6 20,7	31,1 35,5	27,6 28,4	100
	5	1,6 4,5	7,8 2,7	16,6 11,8	26,2 29,8	47,8 51,2	100
		100	100	100	100	100	

I denne tabellen er det mye lettere å sammenligne strukturene. Materialene har, røft sagt, nokså lik struktur. Denne framgangsmåten innebærer også en form for glatting av den opprinnelige tabell. I den første tabellen ser vi f. eks. i det britiske materialet at tallet 8 i første rad, tredje kolonne gir en "dump" i den horisontale linja. I den glattede tabellen er dette trekk ved materialet forsvunnet.

Teknikken byr på muligheter til prediksjon. Hvis vi f. eks. har observert bare marginalene i ett nytt materiale og vi har et fullstendig eldre materiale der vi kan anta assosiasjonsstrukturen er den samme, kan vi predikere simultanfordelingen i det nye ved å iterere i det gamle materialet til det har marginaler som det nye. I neste kapittel skal vi også se på en metode som gir oss visse muligheter til å glatte empirisk materiale. Metoden vi skal se på har en fordel framfor metoden i dette kapitlet, det er lettere å overskue hva vi gjør idet en konstruerer nye tabeller ved å sette ulike samspills-effekter lik 0.i en spesifisert modell.

8. LOG-LINEÆRE MODELLER

For en utførlig behandling av emnet viser vi til Bishop et. al. (1975).

a) Innledende motivering

Vi skal prøve å motivere innføringen av log-lineære modeller. Først ser vi på hvilken grunnleggende størrelse kryssproduktet er i 2 x 2 tabeller.

I tabellen

	B	$\bar{B}$
A	p_{11}	p_{12}
$\bar{A}$	p_{21}	p_{22}

er kryssproduktet, δ, lik $p_{11}\, p_{22}/p_{21}\, p_{12}$. Som i Lehman (1959, s. 145) kan en lett vise at:

$$P(A|B) = P(A) + (\delta - 1) \cdot \frac{p_{12}\, p_{21}}{P(B)}$$

$$P(A|\bar{B}) = P(A) - (\delta - 1) \cdot \frac{p_{12}\, p_{21}}{P(\bar{B})}$$

$$P(B|A) = P(B) + (\delta - 1) \cdot \frac{p_{12}\, p_{21}}{P(A)}$$

$$P(B|\bar{A}) = P(B) - (\delta - 1) \cdot \frac{p_{12}\, p_{21}}{P(\bar{A})}$$

Ligningene viser bl. a.

1) $\delta = 1$ ekvivalent med A og B uavhengige
2) $\delta > 1$ ekvivalent med $P(A|B) > P(A)$
3) $\delta < 1$ ekvivalent med $P(A|B) < P(A)$
4) Differensen $(P(A|B) - P(A))$ øker med δ.

Vi ser at δ på en naturlig måte belyser avhengighetsforholdet mellom de to variable.

Edwards (1963) viser at hvis vi i en 2 x 2 tabell krever at et avhengighetsmål skal være

i) symmetrisk m.h.t. de to faktorer,
ii) en funksjon av $P(B|A)$ og $P(B|\bar{A})$ og
iii) en funksjon av $P(A|B)$ og $P(A|\bar{B})$,

så må målet være en funksjon av δ.

To av Yules klassiske mål for assosiasjon kan uttrykkes som funksjoner av δ.

Q ("measure of association") = $(\delta - 1)/(\delta + 1)$.

Y ("measure of colligation") = $(\sqrt{\delta}-1)/(\sqrt{\delta}+1)$.

δ er invariant m.h.t. transformasjoner av den type som Mosteller (1968) gjorde bruk av (Se kapittel 7.). I

$$\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$$

har vi $\delta = p_{11}\, p_{22}/p_{12}\, p_{21}$ og i

$$\begin{bmatrix} k_1\, r_1\, p_{11} & k_2\, r_1\, p_{12} \\ k_1\, 2_2\, p_{21} & k_2\, r_2\, p_{22} \end{bmatrix}$$

har vi $\delta = (k_1\, r_1\, p_{11} \cdot k_2\, r_2\, p_{22})/(k_1\, r_2\, p_{21} \cdot k_2\, r_1\, p_{12})$

$= (p_{11} \cdot p_{22})/(p_{21}\, p_{12})$.

δ kan også brukes som mål i 2 x 2 tabeller med en faktorvariabel og en responsvariabel. Vi ser på tabellen

	B_1	B_2
A_1	p_{11}	p_{12}
A_2	p_{21}	p_{22}

Når variabel B er faktorvariabel og A responsvariabel, har vi funnet det naturlig å se på de betingete fordelinger gitt B-verdi, dvs. tabellen

	B_1	B_2
A_1	q_{11}	q_{12}
A_2	q_{21}	q_{22}
	1,0	1,0

der $q_{ij} = p_{ij}/(p_{1j} + p_{2j})$ for i = 1, 2 og j = 1, 2.

En måte å sammenligne q_{ij} -ene på, er å se på $\frac{q_{11}/q_{21}}{q_{12/22}}$ dvs. vi ser på forholdet mellom sannsynligheten for A_1 og A_2 i de to betingete fordelingene og sammenholder forholdene ved å ta kvotienten mellom dem. Men dette er det samme som å se på kryssproduktet i den opprinnelige tabell ettersom

$$\frac{q_{11}/q_{21}}{q_{12}/q_{22}} = \frac{(p_{11}/(p_{11} + p_{21}))/(p_{21}/(p_{11} + p_{21}))}{(p_{12}/(p_{12} + p_{22}))/(p_{22}/(p_{12} + p_{22}))} = \frac{p_{11}/p_{21}}{p_{12}/p_{22}} = \frac{p_{11} \cdot p_{22}}{p_{21} \cdot p_{12}} = \delta$$

Vi kan også strukturere avhengigheten i (r x k) - tabeller ved hjelp av ulike kryssprodukter. Vi har tabellen

$$\begin{matrix} p_{11} & p_{12} \cdots & p_{1k} \\ p_{21} & p_{22} \cdots & p_{2k} \\ \vdots & \vdots & \vdots \\ p_{r1} & p_{r2} \cdots & p_{rk} \end{matrix}$$

En vei å gå er å danne alle subtabeller av typen

$$\begin{matrix} p_{ij} & p_{ik} \\ p_{rj} & p_{rk} \end{matrix}$$

der (i og j) varierer og (r og k) er fast. I subtabellene beregnes kryssproduktene

$$\delta_{ij} = (p_{ij}\, p_{rk})/(p_{rj}\, p_{ik}) \text{ for } i = 1, 2, \ldots, r - 1 \text{ og } j = 1, 2, \ldots, k - 1.$$

δ_{ij} -ene gir nå én beskrivelse av avhengighetsstrukturen. Vi har bl. a.

$p_{ij} = p_{i.} \cdot p_{.j}$ for alle (i, j) er ekvivalent med

$\delta_{ij} = 1$ for $i = 1, 2, \ldots, r - 1$ og $j = 1, 2, \ldots, k - 1$.

En av egenskapene ved denne parametriseringen er at δ_{ij} -ene gir uttrykk for assosiasjon i de betingete subtabellene. Hvis vi lar D være begivenheten at vi er i én av subtabellene, blir de betingete sannsynligheter for cellene i subtabellen,

$$q_{ij} = P(\text{celle}(i, j) | D) = p_{ij}/P(D).$$

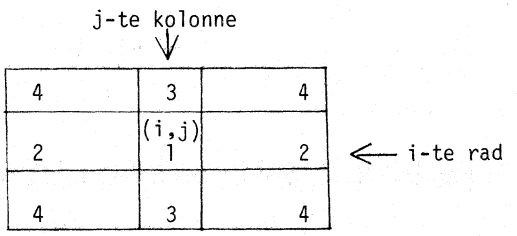
Kryssproduktet i den betingete tabellen blir

$$\frac{\frac{p_{ij}}{P(D)} \cdot \frac{p_{rk}}{P(D)}}{\frac{p_{rj}}{P(D)} \cdot \frac{p_{ik}}{P(D)}} = \frac{p_{ij} \cdot p_{rk}}{p_{rj} \cdot p_{ik}} = \delta_{ij}.$$

En annen parametrisering av avhengigheten oppnår vi ved å beregne kryssproduktene i subtabeller av typen

$$\begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}$$

der cellene a_1, a_2, a_3 og a_4 framkommer ved å slå sammen celler i den opprinnelige tabell etter skjemaet



Vi har ikke sett at noen har prøvd denne formen for parametrisering og vet ikke hvilke egenskaper den har, vi nevner den kun som eksempel for å illustrere de mange muligheter en har til å velge parametrisering etter ens spesielle behov.

b. Momenter om uavhengighet mellom to eller flere variable

Av tabellen

	B_1	B_2	
A_1	4	8	12
A_2	6	12	18
	10	20	30

ville vi tro at det er uavhengighet mellom A- og B- faktoren, men hvis denne tabellen er framkommet ved at vi i tabellen

	B_1	B_2'	B_2''	
A_1	4	1	7	12
A_2	6	6	6	18
	10	7	13	30

har slått sammen kategoriene B_2' og B_2'', så er uavhengigheten tvilsom i det en ikke av den siste tabellen kan slutte at A- og B- faktoren er uavhengige. Regelen er at uavhengighet ved en finere gruppering medfører uavhengighet i en grovere gruppering, men uavhengighet i en grovere gruppering medfører ikke nødvendigvis uavhengighet i en finere gruppering. Når vi har flere enn to variable blir det mer komplisert å vurdere avhengighetsstrukturen. Vi skal illustrere noen sider av problemet ved et konstruert eksempel.

Eksempel 8.1 (Konstruerte tall)

Vi har en mengde dødsfall blant jamnaldrende menn og undersøker disse m.h.t. 3 faktorer. Kategoriene er

A_1, A_2 døde av/døde ikke av lungekreft
B_1, B_2 var/var ikke storrøyker
C_1, C_2 bodde i by/bodde på landet.

Materialet viser en sammenheng mellom faktorene A og B, men sammenhengen går forskjellig vei for de to typene bosted. I deltabeller for de to typene bosted har vi,

C_1	B_1	B_2
A_1	1	5
A_2	4	5

δ = 0,25 dvs. vi har en negativ sammenheng mellom storrøyking og lungekreft i byene

C_2	B_1	B_2
A_1	3	3
A_2	2	7

δ = 3,5 dvs. der en positiv sammenheng mellom storrøyking og lungekreft på landet

Hadde vi studert den marginale tabell over A og B hadde vi "funnet" at A og B var uavhengige i det.

	B_1	B_2
A_1	4	8
A_2	6	12

gir δ = 1.

Eksemplet er konstruert slik at også B og C, A og C er parvis uavhengige.

	B_1	B_2
C_1	5	10
C_2	5	10

gir δ = 1.

	C_1	C_2
A_1	6	6
A_2	9	9

gir δ = 1.

Eksemplet er en advarsel mot å bare studere sammenhenger ved hjelp av ulike to-veistabeller og det viser at når vi skal parametrisere avhengighetsstrukturen mellom tre variable, så må vi ha parametre for annenordens samspill (tre-faktoreffekter) da disse ikke lar seg uttrykke ved samspillene (1. ordens) mellom par av variable.

Et annet moment som kommer til, når vi har tre variable, er at hvis vi har uavhengighet mellom to variable på ethvert nivå av en tredje, så har vi ikke nødvendigvis marginal uavhengighet mellom de to førstnevnte.

I eksempel 8.1 så vi at det var interessant å se på sammenhengen mellom to variable gitt ulike nivåer av en tredje. Vi fikk på den måten avdekket et høyere ordens samspill (tre-faktoreffekt). Et slikt samspill kan foreligge både når vi har marginal uavhengighet og marginal avhengighet mellom de to variable. Vi skal se på parametre som uttrykker dette samspillet.

Vi ser på en (2x2xt) - tabell. De tre variable benevnes A, B og C. Hele tabellen består av t subtabeller av typen

C_ℓ	B_1	B_2
A_1	$p_{11\ell}$	$p_{12\ell}$
A_2	$p_{21\ell}$	$p_{22\ell}$

som viser sannsynlighetene på ℓ-te nivå av variabel C, ℓ = 1, 2, ..., t. For hver av subtabellene vil kryssproduktet,

$$\delta_\ell = (p_{11\ell}\, p_{22\ell})/(p_{21\ell}\, p_{12\ell}),$$

uttrykke avhengigheten mellom A og B for dette nivået av C. Hvis δ_ℓ er konstant for $\ell = 1, 2, \ldots, t$, så sier vi at der ikke er noe annenordens samspill mellom A og B (ingen trefaktoreffekter), dette er ekvivalent med

$$\delta_\ell/\delta_t = 1 \text{ for } \ell = 1, 2, \ldots, t\text{-}1.$$

Hvis vi har annenordens samspill, kan vi bruke forholdet mellom kryssproduktene som mål på dette samspillet. Vi måler samspillet med $\delta^{(2)}_{\ell t}$ der

$$\delta^{(2)}_{\ell t} = \frac{\delta_\ell}{\delta_t} = \frac{p_{11\ell} \cdot p_{22\ell} \cdot p_{12t}\, p_{21t}}{p_{21\ell} \cdot p_{12\ell} \cdot p_{11t}\, p_{22t}} \text{ for } \ell = 1, 2, \ldots, t\text{-}1.$$

I en(2x2xtxv)-tabell vil det være mulighet for tredjeordens samspill (fire-faktoreffekter). Vi kan se på de v deltabellene som er (2x2xt)-tabeller, for hver ℓ sammenligner vi de v $\delta^{(2)}_{\ell t}$-ene i deltabellene, hvis disse ikke like kan vi bruke dem til ny parameterdannelse på tilsvarende måte som vi brukte δ_ℓ -ene til å lage $\delta^{(2)}_{\ell t}$ -ene. De nye parametrene uttrykker tredjeordens samspill.

Samspillsparametrene foran dannes ved multiplikasjon og divisjon av $p_{ij\ell}$ -ene. Tar vi logaritmen av parametrene vil de kunne uttrykkes ved addisjon og subtraksjon av logaritmene til $p_{ij\ell}$ -ene.

c. Log-lineær modell, tre variable.

For i = 1, 2, ..., r, j - 1, 2, ..., k og ℓ = 1, 2, ..., t uttrykker vi den naturlige logaritme til sannsynligheten for at en observasjon faller i celle (i, j, ℓ) ved nye parametre (u-er). Vi setter

$$\log p_{ij\ell} = u + u^1_i + u^2_j + u^3_\ell + u^{12}_{ij} + u^{23}_{j\ell} + u^{13}_{i\ell} + u^{123}_{ij\ell}.$$

På u-ene legges følgende betingelser,

$$u^1_\cdot = u^2_\cdot = u^3_\cdot = 0$$

$$u^{12}_{i\cdot} = u^{12}_{\cdot j} = u^{23}_{\cdot\ell} = u^{23}_{j\cdot} = u^{13}_{i\cdot} = u^{13}_{\cdot\ell} = 0$$

$$u^{123}_{\cdot j\ell} = u^{123}_{i\cdot\ell} = u^{123}_{ij\cdot} = 0.$$

Vi kan tolke u-ene når vi søker å forklare de ulike cellefrekvensene. u-ene med en toppskrift tolkes som hovedeffekter av variable f. eks. u^2_j angir hovedeffekt av at variabel 2 har verdi j. u-ene med to toppskrifter tolkes som to-faktoreffekter, f. eks. $u^{13}_{i\ell}$ angir effekter av at variabel 1 har verdi i samtidig som variabel 3 har verdi ℓ. u-er med tre toppskrifter tolkes som tre-faktoreffekter f. eks. $u^{123}_{ij\ell}$ angir effekt av at variablene 1, 2 og 3 samtidig antar verdiene i, j og ℓ.
Det er ingen tilleggsstruktur i denne modellen i forhold til modellen med $p_{ij\ell}$ -ene. Det er en entydig korrespondanse mellom $p_{ij\ell}$ -ene og u-ene.

Reparametriseringen (det at vi innfører u-ene i stedet for $p_{ij\ell}$ -er) har flere hensikter, momenter er bl. a.

1) Vi får ikke estimater for parametrene som ligger utenfor det område som vi på forhånd vet parametrene må ligge

2) Vi får parametre som gir direkte uttrykk for samspillet mellom to eller flere faktorer. Det er lett å formulere hypoteser ved parametrene.

3) Mange av hypotesene om ulike former for uavhengighet lar seg lett uttrykke ved parametrene (dog ikke alle)

4) Ved å forutsette, evt. etter å ha funnet ved testing, at enkelte samspill er lik 0, kan den observerte tabell glattes. I noen tilfelle kan en på denne måten oppnå estimater for celler som er tomme i den opprinnelige tabell.

Vi skal se nærmere på en del av de hypoteser som kan settes opp i en situasjon med tre variable. $m_{ij\ell}$ står for forventet antall observasjoner i celle (i, j, ℓ) for $i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, k$ og $\ell = 1, 2, \ldots, t$

Hypoteser.

Fullstendig marginal uavhengighet

$$H_a \qquad m_{ij\ell} = \frac{m_{i\cdot\cdot}\, m_{\cdot j\cdot}\, m_{\cdot\cdot\ell}}{m_{\cdot\cdot\cdot}^{\,2}} \qquad P(A_iB_jC_\ell) = P(A_i)\cdot P(B_j)\cdot P(C_\ell)$$

Én uavhengig av de to andre

$$H_{b_1} \qquad m_{ij\ell} = \frac{m_{\cdot j\ell}\, m_{i\cdot\cdot}}{m_{\cdot\cdot\cdot}} \qquad P(A_iB_jC_\ell) = P(A_i)\cdot P(B_jC_\ell)$$

$$H_{b_2} \qquad m_{ij\ell} = \frac{m_{i\cdot\ell}\, m_{\cdot j\cdot}}{m_{\cdot\cdot\cdot}} \qquad P(A_iB_jC_\ell) = P(B_j)\cdot P(A_iC_\ell)$$

$$H_{b_3} \qquad m_{ij\ell} = \frac{m_{ij\cdot}\, m_{\cdot\cdot\ell}}{m_{\cdot\cdot\cdot}} \qquad P(A_iB_jC_\ell) = P(A_iB_j)\cdot P(C_\ell)$$

Marginal uavhengighet

$$H_{c_1} \qquad m_{\cdot j\ell} = \frac{m_{\cdot j\cdot}\, m_{\cdot\cdot\ell}}{m_{\cdot\cdot\cdot}} \qquad P(B_jC_\ell) = P(B_j)P(C_\ell)$$

$$H_{c_2} \qquad m_{i\cdot\ell} = \frac{m_{i\cdot\cdot}\, m_{\cdot\cdot\ell}}{m_{\cdot\cdot\cdot}} \qquad P(A_iC_\ell) = P(A_i)\cdot P(C_\ell)$$

$$H_{c_3} \qquad m_{ij\cdot} = \frac{m_{i\cdot\cdot}\, m_{\cdot j\cdot}}{m_{\cdot\cdot\cdot}} \qquad P(A_iB_j) = P(A_i)\cdot P(B_j)$$

Betinget uavhengighet

$$H_{d_1} \qquad m_{ij\ell} = \frac{m_{ij\cdot}\, m_{i\cdot\ell}}{m_{i\cdot\cdot}} \qquad P(B_jC_\ell|A_i) = P(B_j|A_i)\cdot P(C_\ell|A_i)$$

$$H_{d_2} \qquad m_{ij\ell} = \frac{m_{ij\cdot}\, m_{\cdot j\ell}}{m_{\cdot j\cdot}} \qquad P(A_iC_\ell|B_j) = P(A_i|B_j)\cdot P(C_\ell|B_j)$$

$$H_{d_3} \qquad m_{ij\ell} = \frac{m_{i\cdot\ell}\, m_{\cdot j\ell}}{m_{\cdot\cdot\ell}} \qquad P(A_iB_j|C_\ell) = P(A_i|C_\ell)\cdot P(B_j|C_\ell)$$

$$H_e \qquad \frac{m_{rkt}\, m_{ijt}}{m_{ikt}\, m_{rjt}} = \frac{m_{rk\ell}\, m_{ij\ell}}{m_{ik\ell}\, m_{rj\ell}} \qquad \text{for } i \le r-1,\ j \le k-1 \text{ og } \ell \le t-1.$$

(likhet automatisk oppfylt for $i=r$ eller $j=k$ eller $\ell=t$)

H_e adskiller seg fra de andre hypotesene I 8.A viste vi hvordan avhengighetsstrukturen i en (rxk)-tabell kunne uttrykkes ved kryssproduktene,

$\delta_{ij} = (p_{ij}p_{rk})/(p_{rj}p_{ik})$ for $i = 1, 2, \ldots, r-1$ og $j = 1, 2, \ldots, k-1$. (rxkxt) - tabellen kan en tenke seg som t (rxk)-tabeller lagt over hverandre. I hver av de t lagene kan vi beregne δ_{ij} -er som i 8.A. H_e er en hypotese om at kryssproduktene som tilsvarer hverandre i de t lagene, er like. H_e er ekvivalent med at $U^{123}_{ij\ell} = 0$ for alle (i, j, ℓ).

Av forhold mellom hypotesene nevner vi bl.a. at H_{b_1} er ekvivalent med $H_{d_2} \cap H_{d_3}$. Videre vil H_{b_1} medføre $H_{c_2} \cap H_{c_3}$. Det motsatte gjelder ikke. Men hvis vi i tillegg til at A og C marginalt uavhengige og at A og B marginalt uavhengige, har at H_e gjelder, så vil også H_{b_1} gjelde, dvs. $H_{c_2} \cap H_{c_3} \cap H_e$ medfører H_{b_1}. Dette tar vi som indikasjon på at H_e er en fornuftig hypotese, på en måte utgjør den den logiske differens mellom H_{b_1} og $(H_{c_2} \cap H_{c_3})$.

Vi har ikke sett på om tilsvarende resonnementer kan gjøres i høyere dimensjoner.

For en nærmere drøfting av forholdet mellom ulike hypoteser viser vi til Simpson (1951) og Birch (1963).

Referanser

[1] Birch, M.W. (1963): "Maximum likelihood in three-way contingency tables." J. Roy. Statist. Soc. Ser. B, 25: 220 - 233.

[2] Bishop, Yvonne M.M., Stephen E. Fienberg og Paul W. Holland (1975): "Discrete Multivariate Analysis." The MIT Press, Cambridge, Mass.

[3] Bjørnstad, Jan F. (1973): "En veiledning i valg av avhengighetsmål i kontingenstabeller". Statistisk Sentralbyrå, Arbeidsnotat IO 73/16.

[4] Edwards, A.W.F. (1963): "The measure of association in a 2x2-table." J. Roy. Statist. Soc. Ser. A, 126 : 109 - 114.

[5] Goodman, L.A. og W.H. Kruskal (1954): "Measures of association for cross-classifications. J. Amer. Statist. Assoc., 49: 732 - 764.

[6] Goodman, L.A. og W.H. Kruskal (1959): "Measures of association for cross-classifications, II: further discussion and references. J. Amer Statist. Assoc., 54: 123 - 163.

[7] Lehmann, E.L. (1959): "Testing Statistical Hypotheses" John Wiley, New York.

[8] Mosteller, F (1968): "Association and estimation in contingency tables." J Amer. Statist. Assoc., 63: 1 - 28.

[9] Plackett, R.L. (1974): "The Analysis of Categorial Data." Griffin, London.

[10] Simpson, E.H. (1951): "The interpretation of interaction in contingency tables." J. Roy. Statist. Soc. Ser B, 13: 238 - 241.

[11] Somers, R.H. (1962): "A new asymmetric measure of association for ordinal variables" Amer. Sociol. Rev. 27, 799 - 811.