Arbeidsnotater

STATISTISK SENTRALBYRÅ

Dronningensgt. 16, Oslo-Dep., Oslo 1. Tlf. 41 38 20, 41 36 60

IO 73/16

Oslo 11. mai 1973

EN VEILEDNING I VALG AV AVHENGIGHETSMÅL I KONTINGENSTABELLER

Av

Jan F. Bjørnstad

Innhold

I. INNLEDNING

I (i). Noen situasjoner hvor mål for avhengighet anvendes

Problemet med valg av avhengighetsmål opptrer når man ønsker å undersøke avhengigheten mellom to faktorer i en eller flere hyppighetstabeller. Det er særlig to situasjoner hvor mål for avhengighet er av spesiell interesse. Den ene er ved sammenligning av avhengighet i flere tabeller, og den andre ved testing av uavhengighet i en tabell. I Byrået kan det være aktuelt å bruke uavhengighetstester, f.eks. når man avgjør hvilke tabeller som skal publiseres.

Ved testing av uavhengighet mellom to faktorer i en tabell, har det vært vanlig å bruke en kjikvadrat-test på hypotesen om at faktorene er eksakt uavhengige. Nå har det vist seg at når antall observasjoner er stort (f.eks. av størrelsesorden som ved Byråets intervjuundersøkelser) vil den eksakte uavhengighetshypotesen nesten alltid forkastes, selv i situasjoner hvor det synes opplagt at avhengigheten er meget liten.

Det man egentlig ønsker er å akseptere uavhengighet mellom to faktorer også når det foreligger en liten grad av avhengighet, dvs. når graden av avhengighet ikke er fagvitenskapelig betydelig. Vi sier da at faktorene er nesten uavhengige. Istedenfor å teste eksakt uavhengighet ønsker vi altså å teste nesten uavhengighet (når antall observasjoner er stort). Ved bestemmelse av nesten-uavhengighetshypotesen oppstår problemet med å velge et avhengighetsmål.

Dette notatet er ment å være en veiledning for hvilket mål man bør velge. For testing av nesten uavhengighet, og sammenligning av avhengighet i flere tabeller henvises til arbeidsnotatet "Inferensteori i kontingenstabeller" av samme forfatter (IO 73/28).

I (ii). En innledende diskusjon om avhengighetsmål

Begrepet avhengighet mellom to faktorer, A og B, vil ofte være vagt og upresist. Som regel er det, imidlertid, spesielle trekk ved avhengigheten som vi er interessert i å måle i en gitt situasjon. Disse relevante avhengighetstrekk vil ofte kunne spesifiseres som en del av formålet ved en undersøkelse. Et avhengighetsmål bør derfor konstrueres ut fra en relevant modell for den gitte situasjonen, slik at det gir mest mulig informasjon om de interessante avhengighetstrekkene, i den grad disse er tilstede. Dvs. for en gitt situasjon vil vi at avhengig-

avhengighetsmålet er slik at det måler de avhengighetstrekk som er interessante i denne situasjonen. Vi skjerper altså definisjonen av avhengighet ved konstruksjon av relevante, passende mål. Dersom flere mål er konstruert for en gitt situasjon, bør man velge det mål som man mener gir klarest uttrykk for de relevante avhengighetstrekk.

Dessuten bør målene ha en enkel operasjonell (sannsynlighetsteoretisk) tolkning, slik at det bl.a. er meningsfylt å sammenligne verdiene av et mål for flere tabeller.

Ved valg av avhengighetsmål har vi funnet det naturlig å skille mellom følgende fem situasjoner:

1) Ordnet situasjon

Det eksisterer for hver faktor en underliggende ordning mellom kjennetegnene. La f.eks. faktor A være utdanningsnivå og B inntektsnivå.

2) Uordnet symmetrisk situasjon

Det foreligger ingen naturlig eller relevant ordning mellom kjennetegnene. Dessuten opptrer faktorene symmetrisk, de er begge av like stor interesse.

3) Uordnet asymmetrisk situasjon

Denne situasjonen forekommer når en av faktorene, la oss si B, er av primær interesse i forhold til den andre, og når det ikke eksisterer noen ordning mellom kjennetegnene. Det kan f.eks. være at faktoren A "går foran" B kronologisk eller årsaksmessig. Et eksempel kan være: A: yrke og B: stillingstagen til et bestemt problem.

4) Pålitelighets-situasjonen

Denne situasjonen opptrer når v=w, og A og B antar samme kjennetegn, men refererer seg til forskjellige metoder. La f.eks. A og B være to psykologiske tester som klassifiserer mentalt syke individer etter hvilken sykdom de lider av.

5) Blandet situasjon

Den ene faktors kjennetegn innehar en naturlig, relevant ordning, den andre ikke. Et eksempel på denne situasjonen kan være: A: inntektsnivå og B: kommuneinndeling.

Utenom disse fem situasjonene vil vi behandle 2x2-tabellen for seg.

De fleste målene som blir diskutert i forbindelse med situasjonene 1) - 4), er behandlet av Goodman & Kruskal i [2], [3]. Den ordnete situasjonen blir behandlet spesielt grundig, siden den vil opptre ofte. Målene som blir diskutert der vil alle variere i intervallet $[-1,1]$. Som mål for grad av avhengighet i den ordnete situasjon kan man bruke kvadratet av disse målene.

II. UAVHENGIGHETS-SITUASJONEN I EN TOVEIS KONTINGENSTABELL

Følgende situasjon betraktes. To faktorer (eller egenskaper om man vil) A og B kan anta henholdsvis v og w kjennetegn $A_1, \ldots, A_v$ og $B_1, \ldots, B_w$. Ved hvert forsøk vil ett og bare ett av kjennetegnene $A_1, \ldots, A_v$ inntreffe og samtidig ett og bare ett av kjennetegnene $B_1, \ldots, B_w$. Ved hvert forsøk vil altså ett og bare ett av kjennetegnene A_i og B_j, for i=1, ..., v og j=1, ..., w inntreffe. La Y, Z være to tilfeldige variable, definert ved:

$$\begin{aligned} &Y = i \text{ hvis } A_i \text{ inntreffer, for } i=1, \ldots, v \\ &Z = j \text{ hvis } B_j \text{ inntreffer, for } j=1, \ldots, w \end{aligned} \tag{1}$$

Det gjøres n forsøk. Utfallene av de n forsøk er stokastisk uavhengige. I hvert forsøk er sannsynligheten for at A_i og B_j skal opptre lik p_{ij}. Sannsynligheten for A_i er da $p_{i.} = \sum_{j=1}^{w} p_{ij}$ og sannsynligheten for B_j blir $p_{.j} = \sum_{i=1}^{v} p_{ij}$. Dvs. at $p_{ij} = P(Y=1 \cap Z=j)$, $p_{i.} = P(Y=i)$ og $p_{.j} = P(Z=j)$, for i=1, ..., v og j=1, ..., w.

Definisjon 1. Faktorene A og B kalles eksakt uavhengige, hvis Y og Z er stokastisk uavhengige, dvs. hvis $P(Y=i \cap Z=j) = P(Y=i)\,P(Z=j)$ for i=1, ..., v og j=1, ..., w. Hvis Y, Z er stokastisk avhengige, sier vi at A og B er avhengige.

Fra definisjon 1 følger at den eksakte uavhengighetshypotesen for A og B kan formaliseres slik:

$$H: p_{ij} = p_{i.} \cdot p_{.j} \text{ for } i=1, \ldots, v \text{ og } j=1, \ldots, w \tag{2}$$

Cellesannsynlighetene kan stilles opp i en toveis tabell:

A \ B	B_1	B_2	B_w	Sum
A_1	p_{11}	p_{12}.....	p_{1w}	$p_{1.}$
A_2	p_{21}	p_{22}.....	p_{2w}	$p_{2.}$
.	.			
.	.			
.	.			
.	.			
.	.			
.	.			
.	.			
A_v	p_{v1}	p_{v2}.....	p_{vw}	$p_{v.}$
Sum	$p_{.1}$	$p_{.2}$.....	$p_{.w}$	1

La X_{ij} være antall ganger A_i & B_j inntreffer i løpet av de n forsøk,

og la $q_{ij} = X_{ij}/n$. La videre $q_{i.} = \sum_{j=1}^{w} q_{ij}$ og $q_{.j} = \sum_{i=1}^{v} q_{ij}$.

Det statistiske materiale kan stilles opp i en toveis kontingenstabell (hyppighetstabell):

A \ B	B_1	B_2	B_w	Sum
A_1	X_{11}	X_{12}.....	X_{1w}	$X_{1.}$
A_2	X_{21}	X_{22}.....	X_{2w}	$X_{2.}$
.	.			
.	.			
.	.			
.	.			
.	.			
.	.			
.	.			
A_v	X_{v1}	X_{v2}.....	X_{vw}	$X_{v.}$
Sum	$X_{.1}$	$X_{.2}$.....	$X_{.w}$	n

Her er $X_{i.} = \sum_{j=1}^{w} X_{ij}$ og $X_{.j} = \sum_{i=1}^{v} X_{ij}$, slik at $X_{i.}$ er antall ganger A_i inntreffer, og $X_{.j}$ er antall ganger B_j inntreffer.

III. ORDNET SITUASJON

III (i). Tre ordinalinvariante mål

Situasjonen er at det foreligger en relevant ordning mellom kjennetegnene innen begge faktorer. La oss først gi en definisjon av ordinalinvariante mål.

Definisjon 2. Et mål g kalles ordinalinvariant hvis det er uforandret under like typer av monotone transformasjoner av Y og Z, og hvis det forandrer fortegn når transformasjonene er av ulike typer. Dvs. at $g(Y,Z) = g(f(Y), h(Z))$, hvis f og h begge er strengt økende, eller begge er strengt avtagende funksjoner og $g(Y,Z) = -g(f(Y), h(Z))$ hvis den ene funksjonen er strengt avtagende, og den andre er strengt økende.

I denne situasjonen måles Y og Z på en ordinalskala, slik at rekkefølgen av deres verdier, men ikke avstanden mellom verdiene, har mening. Vi vil derfor kreve at et mål for denne situasjonen er ordinalinvariant. Dessuten bør målet g tilfredsstille to krav:

(i) $-1 \leq g \leq 1$

(ii) A, B eksakt uavhengige $\Rightarrow g = 0$

Hvis $g \in [-M,M]$, vil $g/M \in [-1,1]$, slik at hvis variasjonsområdet til g er begrenset, symmetrisk om origo, kan vi alltid få (i) oppfylt ved å normere målet.

Vi vil beskrive tre slike ordinalinvariante mål, som alle er modifikasjoner av en grunnleggende størrelse. De betegnes med:

1) γ, foreslått av Goodman & Kruskal i [2].
2) τ_b, Kendalls rangkorrelasjonskoeffisient modifisert til kontingenstabeller.
3) τ_c, foreslått av Stuart i [7].

Målet γ er også diskutert i [5]. Vi vil senere begrunne hvorfor γ er det mest passende mål av de tre. Alle målene, men spesielt γ, kan gis en enkel sannsynlighetsteoretisk tolkning. La oss først betrakte γ.

III (ii). Konstruksjon av et naturlig mål, γ

La (Y_1, Z_1) og (Y_2, Z_2) være to stokastisk uavhengige variable med samme fordeling som (Y, Z).

Vi sier (Y_1, Z_1) og (Y_2, Z_2) er overensstemmende (eng.: concordant)

hvis Y_1 og Y_2 avviker med samme fortegn som Z_1 og Z_2, dvs. hvis $(Y_1-Y_2)(Z_1-Z_2) > 0$. Tilsvarende sier vi at de variable er uoverensstemmende (eng.: discordant) hvis Y-ene og Z-ene avviker med forskjellig fortegn; dvs. hvis $(Y_1-Y_2)(Z_1-Z_2) < 0$.

γ er definert ved:

$$\gamma = P\{(Y_1-Y_2)(Z_1-Z_2) > 0 \mid Y_1 \neq Y_2 \cap Z_1 \neq Z_2\} - P\{(Y_1-Y_2)(Z_1-Z_2) < 0 \mid Y_1 \neq Y_2 \cap Z_1 \neq Z_2\}.$$

Det ses umiddelbart at γ er ordinalinvariant.

La $\pi_t = P(Y_1=Y_2 \cup Z_1=Z_2)$

$\pi_s = P\{(Y_1-Y_2)(Z_1-Z_2) > 0\}$

$\pi_d = P\{(Y_1-Y_2)(Z_1-Z_2) < 0\}$.

Dvs. at π_s er sannsynligheten for at de variable er overensstemmende, og π_d er sannsynligheten for at de er uoverensstemmende.

I denne situasjonen finner vi det naturlig å utvide definisjonen av eksakt uavhengighet mellom faktorene til:

Definisjon 3

To faktorer A og B sies å være ordningsuavhengige (o.u.) hvis $\pi_s=\pi_d$.

Det ses at γ kan uttrykkes på følgende form:

$$\gamma = \frac{\pi_s-\pi_d}{1-\pi_t} \qquad (3)$$

Dessuten, siden $\pi_t+\pi_s+\pi_d = 1$: $\gamma = (\pi_s-\pi_d)/(\pi_s+\pi_d)$.

Herav sees at $\gamma \in [-1,1]$, slik at (i) er tilfredsstilt.

Det kan vises at:

$$\pi_t = \sum_{i=1}^{v} p_{i\cdot}^{\,2} + \sum_{j=1}^{w} p_{\cdot j}^{\,2} - \sum_{i=1}^{v}\sum_{j=1}^{w} p_{ij}^{\,2}.$$

$$\pi_s = 2\sum_{i=1}^{v-1}\sum_{j=1}^{w-1} p_{ij}\Big\{\sum_{i'>i}\sum_{j'>j} p_{i'j'}\Big\}.$$

$$\pi_d = 2\sum_{i=1}^{v-1}\sum_{j=2}^{w} p_{ij}\Big\{\sum_{i'>i}\sum_{j'<j} p_{i'j'}\Big\}.$$

Videre kan det vises at A, B eksakt uavhengige medfører at $\pi_s=\pi_d$, slik at definisjon 2 faktisk er en utvidelse av definisjon 1.

γ tilfredsstiller dermed kravene (i) og (ii) og har dessuten følgende egenskaper (se [2]).

(iii) γ er veldefinert, såfremt ikke alle positive cellesannsynligheter er konsentrert i en enkel rad eller kolonne.

(iv) A, B eksakt uavhengige $\Rightarrow \gamma=0$, men det omvendte behøver ikke gjelde unntagen i 2x2-tilfellet.

I 2x2-tabellen reduserer målet seg til:

$$\gamma = \frac{P_{11}P_{22}-P_{12}P_{21}}{P_{11}P_{22}+P_{12}P_{21}} = \frac{\Delta-1}{\Delta+1} \tag{4}$$

hvor $\Delta = \frac{P_{11}P_{22}}{P_{12}P_{21}}$ er kryssprodukt-forholdet.

Avhengighetsmål i 2x2 tabellen kommer vi tilbake til i VIII.

III (iii). To alternative mål, τ_b og τ_c

La oss først betrakte følgende situasjon:

La U, V være kontinuerlige tilfeldige variable, Kendalls rang-korrelasjonskoeffisient τ for (U, V) er definert ved:

$$\tau = P\{(U_1-U_2)(V_1-V_2) > 0\} - P\{(U_1-U_2)(V_1-V_2) < 0\} \tag{5}$$

hvor (U_1, V_1) og (U_2, V_2) er to uavhengige variable med samme fordeling som (U, V) ([5], s. 822). τ kan betraktes som korrelasjonskoeffisienten mellom fortegnene til U_1-U_2 og V_1-V_2. La $(u_1, v_1), \ldots, (u_n, v_n)$ være n observasjoner av (U, V). Vi sier at det er ingen sammenfallende verdier hvis $u_i \neq u_j$ og $v_i \neq v_j$ for $i \neq j$, $i=1, \ldots, n$ og $j=1, \ldots, n$. Det engelske uttrykket for sammenfallende verdier er "ties" som vi vil anvende heretter. I en kontingenstabell vil det altså foreligge ties hvis minst to observasjoner faller i samme rad eller kolonne, noe som alltid vil skje dersom $n > \min(v, w)$.

I tilfellet med ingen ties ser man at γ reduseres til τ. M.a.o.: γ er en modifikasjon av τ til situasjonen med ties. Nå vil vi betrakte to andre modifikasjoner av tilfellet med ties.

La situasjonen være som i III (ii).

Kendalls rangkorrelasjonskoeffisient for kontingenstabeller er definert ved (vår definisjon):

$$\tau_b = \frac{\pi_s - \pi_d}{\sqrt{P(Y_1 \neq Y_2) \cdot P(Z_1 \neq Z_2)}} \tag{6}$$

(Legg merke til at $\gamma = (\pi_s - \pi_d)/P(Y_1 \neq Y_2 \cap Z_1 \neq Z_2)$.)

La $\pi_y = P(Y_1 \neq Y_2)$ og $\pi_z = P(Z_1 \neq Z_2)$.

$$\pi_y = 1 - \sum_{i=1}^{v} p_{i.}^2.$$

$$\pi_z = 1 - \sum_{j=1}^{w} p_{.j}^2.$$

I 2x2-tilfellet er

$$\tau_b = \frac{p_{11}p_{22} - p_{12}p_{21}}{\sqrt{p_{1.}p_{2.}p_{.1}p_{.2}}} \tag{7}$$

Siden $\tau_b = 0 \iff \pi_s = \pi_d$, ses at τ_b tilfredsstiller kravene (i) og (ii). Dessuten gjelder følgende:

(iii) τ_b er veldefinert, såfremt ikke alle positive cellesannsynligheter er konsentrert i en enkel rad eller kolonne.

(iv) τ_b er ordinalinvariant.

Angående (i) bør nevnes at grensene ± 1 aldri kan oppnås unntagen i en vxv-tabell hvor $\sum_{i=1}^{v} p_{ii} = 1$.

Det er også verdt å merke seg at τ_b^2 kan betraktes som en generalisering av $\beta = \frac{(p_{11}p_{22} - p_{12}p_{21})^2}{p_{1.}p_{2.}p_{.1}p_{.2}}$ til en vxw-ordnet situasjon, mens det tradisjonelle kjikvadrat-målet

$$\phi^2 = \sum_{i=1}^{v} \sum_{j=1}^{w} \frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}}$$

er en generalisering av β til situasjonen med ingen relevant ordning. For andre tradisjonelle mål i denne situasjonen henviser vil til IV (iii).

Det tredje mål τ_c er definert ved:

$$\tau_c = \frac{\pi_s - \pi_d}{(m-1)/m} \tag{8}$$

hvor $m = \min(v, w)$.

Normeringsfaktoren $\frac{m}{m-1}$ er en følge av setning 1. (For bevis, se appendiks.)

Setn. 1

$-\frac{m-1}{m} \leq \pi_s - \pi_d \leq \frac{m-1}{m}$. Grensene oppnås dersom alle cellesannsynlighetene er lik 0 utenfor en lengste diagonal i tabellen, og lik $\frac{1}{m}$ i diagonalen.

Setn. 1 gir at (i) er oppfylt, hvor nå grensene -1 og +1 også kan oppnås når $v \neq w$. Egenskap (ii) holder også, og τ_c er dessuten ordinalinvariant og alltid veldefinert.

I 2x2-tilfellet er $\tau_c = 4(p_{11}p_{22} - p_{12}p_{21})$. La nå $\hat{\tau}_b$ være estimatoren for τ_b som fås ved å sette inn de relative hyppighetene q_{ij} istedenfor p_{ij} i uttrykket for τ_b. Dvs.

$$\hat{\tau}_b = \frac{P_s - P_d}{\sqrt{P_y \cdot P_z}} \tag{9}$$

hvor

$$P_y = 1 - \sum_{i=1}^{v} q_{i.}^2.$$

$$P_z = 1 - \sum_{j=1}^{w} q_{.j}^2.$$

$$P_s = 2 \sum_{i=1}^{v-1} \sum_{j=1}^{w-1} q_{ij} \{ \sum_{i'>i} \sum_{j'>j} q_{i'j'} \}.$$

$$P_d = 2 \sum_{i=1}^{v-1} \sum_{j=2}^{w} q_{ij} \{ \sum_{i'>i} \sum_{j'<j} q_{i'j'} \}.$$

$\hat{\tau}_b$ kan betraktes som et spesialtilfelle av en generalisert empirisk korrelasjonskoeffisient beskrevet i [4 s. 19]. Vi vil kort gjengi beskrivelsen her. La $(y_1, z_1) \ldots, (y_n, z_n)$ være de n uavhengige observasjonene som tas. Til hvert par $\{(y_i, z_i), (y_j, z_j)\}$ tildeles en Y-score a_{ij} slik at $a_{ij} = -a_{ji}$, og en Z score b_{ij} slik at $b_{ij} = -b_{ji}$. Den generaliserte empiriske korrelasjonskoeffisienten er definert ved:

$$\Gamma = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} b_{ij}}{\sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2 \sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij}^2}} \tag{10}$$

F.eks. fås den vanlige empiriske (produkt) korrelasjonskoeffisienten

$$\frac{\Sigma_i (Y_i-\bar{Y})(Z_i-\bar{Z})}{\sqrt{\Sigma(Y_i-\bar{Y})^2 \cdot \Sigma(Z_i-\bar{Z})^2}},$$ ved å sette $a_{ij}=y_j-y_i$ og $b_{ij}=z_j-z_i$.

Følgende resultat (for bevis, se appendiks) viser hvordan $\hat{\tau}_b$ er et spesialtilfelle av Γ.

Setn. 2

La Y-scorene a_{ij} og Z-scorene b_{ij} være gitt ved:

$$a_{ij} = \begin{cases} +1 \text{ hvis } y_i < y_j \\ 0 \quad " \quad y_i = y_j \\ -1 \quad " \quad y_i > y_j \end{cases}, \quad b_{ij} = \begin{cases} +1 \text{ hvis } z_i < z_j \\ 0 \quad " \quad z_i = z_j \\ -1 \quad " \quad z_i > z_j \end{cases}$$

Da er $\Gamma=\hat{\tau}_b$.

I tilfellet med ingen ties vil vi alltid ha a_{ij}, b_{ij} lik +1 eller -1 for $i \neq j$, og helt tilsvarende vil da $\Gamma=\hat{\tau}$, hvor $\hat{\tau}$ er sampelobservatoren svarende til τ. Dvs. at $\hat{\tau}_b$ er den naturlige modifikasjon av $\hat{\tau}$ basert på Γ.

III (iv). En vurdering av målene γ, τ_b og τ_c

Det første vi noterer oss er at alle tre målene er en modifikasjon av differensen $\pi_s-\pi_d$ til tilfellet med ties. Den naturlige modifikasjon er opplagt γ, hvor man ser på de betingede sannsynligheter gitt ingen ties. Både τ_b og τ_c synes å være noe kunstige modifikasjoner, da kanskje særlig τ_c som bare er en normering av $\pi_s-\pi_d$.

Det andre man bør merke seg (angående τ_b) er at opprinnelig var det den empiriske rangkorrelasjonskoeffisienten $\hat{\tau}$ som ble modifisert til $\hat{\tau}_b$, med utgangspunkt i den generaliserte empiriske korrelasjonskoeffisienten Γ gitt ved (10). ([4], [7]). Definisjonen (6) er vi kommet fram til ved å innsette sannsynligheten p_{ij} istedenfor q_{ij} i $\hat{\tau}_b$. (τ_b er ikke nevnt i noen av de artiklene som vi gir referanser til.) Vi har dermed at, mens γ er den naturlige modifikasjon av τ basert på $\pi_s-\pi_d$, så er $\hat{\tau}_b$ den naturlige modifikasjon av $\hat{\tau}$ basert på Γ. Det man er interessert i er parameteren. Det korrekte må derfor være å modifisere parameteren, og deretter se på problemet med estimering, ikke å gå den omvendte veien som Kendall gjorde med $\hat{\tau}$. Konklusjonen må derfor bli at γ er det mest naturlige og passende mål i den ordnete situasjonen.

Ingen av de foreslåtte mål i denne situasjonen er invariante under permutasjon av rader eller kolonner (av cellesannsynligheter) i tabellen, naturlig nok. I den neste situasjonen vi skal betrakte vil målene være invariante under slike permutasjoner.

IV. UORDNET SYMMETRISK SITUASJON

IV (i). En symmetrisk prediksjonsmodell

To avhengighetsmål λ og η, foreslått av Goodman & Kruskal, [2], [3], vil bli behandlet. Dessuten lister vi opp noen tradisjonelle avhengighetsmål, som imidlertid ikke kan gis noen operasjonell tolkning.

Målene λ og η vil være enkle funksjoner av feilsannsynligheter innen en viss prediksjonsmodell, som vil bli beskrevet. For at prediksjonsmodellen skal ha mening, vil det antas at cellesannsynlighetene p_{ij} er kjente ved konstruksjon av målene λ og η. De to målene er samme funksjon av sannsynligheter for feilprediksjoner, basert på to forskjellige prediksjonsmetoder. Den symmetriske prediksjonsmodell målene er konstruert fra, er følgende ([2], s. 743):

I ett gitt forsøk skal med sannsynlighet 0,5 B's kjennetegn predikeres, og med sannsynlighet 0,5 A's kjennetegn predikeres. (Dvs. enten A eller B's kjennetegn skal forutsies , hver faktor med sannsynlighet lik 0,5 for å bli trukket ut til prediksjon.) Dersom B blir trukket ut, skal prediksjonen foretas på grunnlag av

(1) Ingen informasjon, og

(2) Gitt kjennetegnet til A.

Tilsvarende dersom A skal predikeres.

IV (ii). Målene λ og η basert på henholdsvis optimal og proporsjonal prediksjon

Goodman & Kruskal foreslår to alternative prediksjonsmetoder:

a) Optimal prediksjon

Dersom B blir trukket ut. Predikerer i tilfelle (1) den B_j med $p_{.j} = \max_{j'} p_{.j'}$, og i tilfelle (2) gitt A_i: Predikerer den B_j med $p_{ij} = \max_{j'} p_{ij'}$. Tilsvarende dersom A blir trukket ut.

La nå Q_1 = P(Riktig optimal prediksjon i tilfelle (1))

og Q_2 = P(Riktig optimal prediksjon i tilfelle (2)).

b) Proporsjonal prediksjon

Dersom B blir trukket ut. Predikerer i tilfelle (1) B_j med sannsynlighet $p_{.j}$, for $j=1, \ldots, w$, og i tilfelle (2) gitt A: Predikerer B_j med sannsynlighet $P_{ij}/p_{i.}$ for $j=1, \ldots, w$. Tilsvarende hvis A blir trukket ut til prediksjon.

La P_1 = P(Riktig proporsjonal prediksjon i tilfelle (1))

og P_2 = P(Riktig proporsjonal prediksjon i tilfelle (2)).

Målene λ og η defineres nå slik:

$$\lambda = \frac{(1-Q_1) - (1-Q_2)}{1-Q_1} = \frac{Q_2-Q_1}{1-Q_1} . \tag{11}$$

$$\eta = \frac{(1-P_1) - (1-P_2)}{1-P_1} = \frac{P_2-P_1}{1-P_1} . \tag{12}$$

Man ser at λ og η begge er relativ minsking i sannsynlighet for feilprediksjon fra ukjent til kjent kjennetegn for den faktor som ikke predikeres.

Nå er Q_i = $\frac{1}{2}$ { P(Riktig optimal prediksjon av B's kjennetegn i tilfelle (i))
+ P(Riktig optimal prediksjon av A's kjennetegn i tilfelle (i))}

og tilsvarende for P_i. Vi finner følgende uttrykk for Q_i og P_i, λ og η:

$$Q_1 = \tfrac{1}{2} (p_{.m}+p_{m.})$$

$$Q_2 = \tfrac{1}{2} \left(\sum_{i=1}^{v} p_{im} + \sum_{j=1}^{w} p_{mj} \right)$$

hvor $p_{.m} = \max_j p_{.j}$, $p_{m.} = \max_i p_{i.}$, $p_{im} = \max_{j'} p_{ij'}$ og $p_{mj} = \max_{i'} p_{i'j}$

Herav:

$$\lambda = \frac{\sum_{i=1}^{v} p_{im} + \sum_{j=1}^{w} p_{mj} - p_{.m} - p_{m.}}{2 - p_{.m} - p_{m.}} \tag{13}$$

$$P_1 = \tfrac{1}{2} \left(\sum_{i=1}^{v} p_{i.}^{\,2} + \sum_{j=1}^{w} p_{.j}^{\,2} \right)$$

$$P_2 = \tfrac{1}{2} \left(\sum_{i=1}^{v} \sum_{j=1}^{w} p_{ij}^{\,2} \left(\frac{1}{p_{i.}} + \frac{1}{p_{.j}} \right) \right) .$$

Det sees at η kan uttrykkes på følgende form:

$$\eta = \frac{\sum_{i=1}^{v}\sum_{j=1}^{w}(p_{ij}-p_{i.}p_{.j})^2\left(\frac{1}{p_{i.}}+\frac{1}{p_{.j}}\right)}{2-\sum_{i=1}^{v}p_{i.}^2-\sum_{j=1}^{w}p_{.j}^2} \qquad (14)$$

Noen egenskaper ved λ:

(i) λ er veldefinert, unntagen hvis en $p_{ij} = 1$.

(ii) $0 \leq \lambda \leq 1$.

(iii) A, B eksakt uavhengige $\Rightarrow \lambda = 0$.

(iv) λ er invariant ved permutasjon av rader og kolonner (av cellesannsynlighetene) i kontingenstabellen.

Noen egenskaper ved η:

(i) η er veldefinert, unntagen hvis en $p_{ij} = 1$.

(ii) $0 \leq \eta \leq 1$.

(iii) A, B eksakt uavhengige $\Leftrightarrow \eta = 0$.

(iv) η er invariant ved permutasjon av rader og kolonner.

I 2x2-tilfellet er η lik β. Dvs. at η er lik kjikvadratmålet ϕ^2 i 2x2-tabellen.

Hvilket av målene λ og η som passer best i en gitt situasjon vil bero på hvilken prediksjonsmetode, som er den relevante i denne situasjonen. Vanligvis er det vel mest interessant å forutsi den mest sannsynlige Y- eller Z-verdi, dvs. at optimal prediksjon er den mest relevante som regel. Man bør imidlertid merke seg at λ er et noe "grovere" mål enn η. Med det menes at dersom avhengigheten mellom A og B forandres lite, vil ikke nødvendigvis λ avsløre det.

IV (iii). Tradisjonelle avhengighetsmål

De vanligste tradisjonelle mål for avhengighet er basert på det allerede nevnte kjikvadrat-målet

$$\text{a)}\quad \phi^2 = \sum_{i=1}^{v}\sum_{j=1}^{w}\frac{(p_{ij}-p_{i.}p_{.j})^2}{p_{i.}p_{.j}} = \sum_{i=1}^{v}\sum_{j=1}^{w}\frac{p_{ij}^2}{p_{i.}p_{.j}} - 1 .$$

(også kalt "mean square contingency" i litteraturen).

Tre variasjoner av dette målet er nevnt i [2], s. 739-740.

b) $K = \sqrt{\frac{\phi^2}{1+\phi^2}}$ (foreslått av K. Pearson)

c) $T = \sqrt{\frac{\phi^2}{\sqrt{(v-1)\,(w-1)}}}$ (foreslått av Tschuprow)

d) $C = \frac{\phi^2}{\min\,(v-1,\ w-1)}$ (foreslått av Cramér)

Man ser at K, T, C $\in [0, 1]$, og at: A, B eksakt uavhengige $\Leftrightarrow \phi^2 = K = T = C = 0$. Det er vanskelig å gi en sannsynlighetsteoretisk tolkning av disse målene. Mål basert på ϕ^2 er m.a.o. ikke særlig meningsfylte. Goodman & Kruskal, [2], gir en mer utfyllende diskusjon om slike mål uten tolkning.

Et mål ikke basert på ϕ^2, er foreslått av J.F. Steffensen i 1933. (Se [3], s. 140)

e) $\Psi^2 = \sum_{i=1}^{v} \sum_{j=1}^{w} p_{ij} \frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}(1-p_{i.})p_{.j}(1-p_{.j})}$

Noen egenskaper: (i) $\Psi^2 = 0 \Leftrightarrow$ A, B er eksakt uavhengige og (ii) $0 \leq \Psi^2 \leq 1$. Ψ^2 er et veiet gjennomsnitt (med p_{ij} som vekter) av alle 2x2 "mean square contingencies", dannet fra hver av de vw cellene og deres komplement.

To enkle mål er

f) $H_1 = \max_{i,j} |p_{ij} - p_{i.}p_{.j}|$

og g) $H_2 = \max_{i,j} \left|\frac{p_{ij}}{p_{i.}} - p_{.j}\right|$

H_2 er vel et mer anskuelig-gjørende mål enn H_1, (bl.a. fordi man vanligvis tabellerer $q_{ij}/q_{i.}$ og betrakter differensene $q_{ij}/q_{i.} - q_{.j}$ ved vurdering av avhengighet i tabellen).

For andre avhengighetsmål, se [3].

La oss nå betrakte situasjonen hvor en faktor er av primær interesse.

V. UORDNET ASYMMETRISK SITUASJON

V (i). En asymmetrisk prediksjonsmodell

La oss anta at faktoren B er av primær interesse. To mål, λ_b og η_b, foreslått av Goodman & Kruskal, [2], skal betraktes. Målene λ_b og η_b svarer til λ og η, med den forskjell at de er konstruert ut fra en asymmetrisk prediksjonsmodell. For at prediksjonsmodellen skal ha mening, vil det som i IV (i) antas at p_{ij}-ene er kjente ved konstruksjon av målene λ_b og η_b. Den asymmetriske modellen, gitt i [2], s. 741, er følgende:

I et gitt forsøk skal B's kjennetegn predikeres, gitt

1) Ingen informasjon, og
2) Gitt A's kjennetegn.

Siden B er av primær interesse er de relevante avhengighetstrekk essensielt av typen: "Forskjellen" mellom riktig B-prediksjon gitt A og riktig B-prediksjon gitt ingen informasjon. Altså er den asymmetriske prediksjonsmodell beskrevet ovenfor en relevant modell å konstruere målene ut fra.

V (ii). Målene λ_b og η_b basert på henholdsvis optimal og proporsjonal prediksjon

Optimal og proporsjonal prediksjon for B er helt analogt med definisjonene a) og b) i IV (ii). Dvs.

a) Optimal prediksjon betyr at man predikerer det mest sannsynlige kjennetegn til B i tilfellene (1) ingen informasjon og 2) gitt A_i.

b) Proporsjonal prediksjon betyr at man i tilfellet (1) predikerer B_j med sannsynlighet $p_{.j}$, for j=1, ..., w, og i tilfelle (2), gitt A_i, predikeres B_j med sannsynlighet $p_{ij}/p_{i.}$, for j=1, ..., w.

Definisjonen av λ_b og η_b er helt tilsvarende med definisjonene (11) og (12) av målene λ og η.

La Q_i^b = P(Riktig optimal prediksjon av B i tilfelle (i)), for i=1, 2.

og P_i^b = P(Riktig proporsjonal prediksjon av B i tilfelle (i)), for i=1, 2.

Da er:

$$\lambda_b = \frac{(1-Q_1^b) - (1-Q_2^b)}{1 - Q_1^b} = \frac{Q_2^b - Q_1^b}{1 - Q_1^b} . \tag{15}$$

$$\eta_b = \frac{(1-P_1^b) - (1-P_2^b)}{1 - P_1^b} = \frac{P_2^b - P_1^b}{1 - P_1^b} . \tag{16}$$

λ_b og η_b er relativ nedgang i sannsynligheten for feilprediksjon fra ukjent til kjent A, ved henholdsvis optimal og proporsjonal prediksjon.

Målene kan uttrykkes på følgende form:

$$\lambda_b = \frac{\sum_{i=1}^{v} p_{im} - p_{.m}}{1 - p_{.m}} \tag{17}$$

$$\eta_b = \frac{\sum_{i=1}^{v}\sum_{j=1}^{w} p_{ij}^2/p_{i.} - \sum_{j=1}^{w} p_{.j}^2}{1 - \sum_{j=1}^{w} p_{.j}^2} = \frac{\sum_{i=1}^{v}\sum_{j=1}^{w} \frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}}}{1 - \sum_{j=1}^{w} p_{.j}^2} \tag{18}$$

Noen egenskaper ved λ_b:

(i) λ_b er ubestemt hvis og bare hvis en $p_{.j}=1$.

(ii) $0 \leq \lambda_b \leq 1$.

(iii) A, B eksakt uavhengige $\Rightarrow \lambda_b=0$.

(iv) λ_b er invariant under permutasjon av rader og kolonner.

Egenskapene (i), (ii) og (iv) holder også for η_b, dessuten gjelder:

(iii)': A, B eksakt uavhengige $\Leftrightarrow \eta_b=0$.

Dersom A er den primære faktoren vil vi få helt tilsvarende mål:

$$\lambda_a = \frac{\sum_{j=1}^{w} p_{mj} - p_{m.}}{1 - p_{m.}} \tag{19}$$

$$\eta_a = \frac{\sum_{i=1}^{v}\sum_{j=1}^{w} p_{ij}^2/p_{.j} - \sum_{i=1}^{v} p_{i.}^2}{1 - \sum_{i=1}^{v} p_{i.}^2} \tag{20}$$

Ved vurdering av hvilket mål λ_b eller η_b (eventuelt λ_a eller η_a) som er mest anvendelig i en gitt situasjon vil de samme argumenter gjelde som ved vurderingen av λ og η i IV (ii).

VI. PÅLITELIGHETS-SITUASJONEN

VI (i). Det uordnete symmetriske tilfelle

Situasjonen er beskrevet i I (ii) (se også [2], s. 756). Det spesielle i dette tilfellet er at $A_i=B_i$ for $i=1, \ldots, v$. I denne situasjonen er man ofte interessert i graden av enighet mellom to metoder som A og B vanligvis refererer seg til. For situasjonen hvor kjennetegnene ikke innehar en relevant ordning konstruerer Goodman & Kruskal, [2], et mål ut fra den symmetriske prediksjonsmodell, gitt i IV (i). Prediksjonsmetoden er som følger: I tilfelle (1) predikeres den B_i med $p_{i.}+p_{.i} = p_{M.}+p_{.M} = \max_{i'} (p_{i.'}+p_{.i'})$. Tilsvarende dersom A blir trukket.

I tilfelle (2) gitt A_i, predikeres B_i. Tilsvarende dersom A predikeres.

La Λ_i = P (Riktig prediksjon i tilfelle (i)), for $i = 1, 2$. Målet som foreslås er definert helt analogt med λ, η, λ_b og η_b:

$$\lambda_r = \frac{(1-\Lambda_1) - (1-\Lambda_2)}{1-\Lambda_1} = \frac{\Lambda_2-\Lambda_1}{1-\Lambda_1} \tag{21}$$

Man finner: $\Lambda_1 = \frac{1}{2} (p_{M.} + p_{.M})$

$$\Lambda_2 = \sum_{i=1}^{v} p_{ii} \quad , \text{ slik at}$$

$$\lambda_r = \frac{\sum_{i=1}^{v} p_{ii} - \frac{1}{2} (p_{M.} + p_{.M})}{1-\frac{1}{2} (p_{M.} + p_{.M})} \tag{22}$$

Noen egenskaper:

i) $-1 \leq \lambda_r \leq 1$

ii) λ_r antar ingen spesiell verdi når A og B er eksakt uavhengige, men som Goodman & Kruskal argumenterer, så vil et mål som λ_r bare bli brukt når det er kjent at det er sammenheng mellom metodene A og B, slik at denne uønskede egenskap ikke er noen særlig ulempe.

VI (ii). Det ordnete tilfelle

I denne situasjonen har det vært vanlig å bruke mål av typen

$$\pi_k = \sum_{|i-j| \leq k} p_{ij} \quad \text{for en valgt } k.$$

F.eks., er $\pi_0 = \sum_{i=1}^{v} p_{ii}$ lik sannsynligheten for at metodene "er enige" (dvs. gir samme resultat).

VII. BLANDET SITUASJON

Et tilfelle som ikke er behandlet i noen av de artiklene som det refereres til, er den situasjonen hvor vi har ordinalnivå for den ene av de variable (Y, Z), men ikke for den andre. Her vil vi nå diskutere denne situasjonen, og fremme enkelte forslag på avhengighetsmål. La oss for enkelthets skyld anta at Y har ordinalnivå. Hva slags mål man bør velge vil bero på hvilke trekk ved avhengighet man primært er interessert i. Det synes naturlig å skille mellom følgende tre situasjoner.

a) Asymmetrisk situasjon. B er av primær interesse.
b) Asymmetrisk situasjon. A er av primær interesse.
c) Symmetrisk situasjon.

a) B har primær interesse

Siden det ikke foreligger noen interessant ordning mellom kjennetegnene til B, synes det rimelig at en asymmetrisk prediksjonsmodell som i V (i), er relevant her. Følgelig bør målet være konstruert ut fra denne modellen. λ_b og η_b er derfor passende mål.

b) A har primær interesse

Siden kjennetegnene for den primære faktor innehar en relevant ordning vil det være rimelig å kreve at målet iallfall ikke er invariant under permutasjon av rader i kontingenstabellen. Dermed er alle mål i den uordnete situasjonen ute av betraktning.

Et passende mål synes da å være et som er konstruert for den ordnete situasjonen, dvs. ν, siden vi fant at dette er det naturligste av de tre som ble vurdert i III.

c) Symmetrisk situasjon

Denne situasjonen opptrer som før nevnt, når det er ingen grunn til å gi den ene faktor prioritet framfor den andre. Intuitivt synes det rimelig at et avhengighetsmål i en slik situasjon er en funksjon av to mål D_1, D_2, hvor D_1 er et mål i den ordnete situasjonen ($-1 \leq D_1 \leq 1$), og D_2 er et mål konstruert for den uordnete situasjonen ($0 \leq D_2 \leq 1$).
En slik funksjon $h(D_1, D_2)$ burde da idealistisk sett ha følgende egenskaper:

1) Invariant under permutasjon av kolonner.
2) Ikke invariant under permutasjon av rader.

Dette synes imidlertid å være en altfor ambisiøs forutsetning. En mer upresis betingelse er:

h bør utnytte informasjonen fra D_1 og D_2 i "like stor grad".
Dessuten kan det være ønskelig at

$$h(D_1, D_2) = 0 \iff D_1 = D_2 = 0 \tag{23}$$

Eksempler på slike mål er:

i) $h(D_1, D_2) = a(|D_1| + D_2)$, $a > 0$

ii) $h(D_1, D_2) = b(D_1^2 + D_2)$, $b > 0$.

Målene i) og ii) vil være ikke-negative. Dersom man synes betingelsen (23) er uvesentlig, kan andre mål av formen $c(D_1 + D_2)$ og $d(D_1 \cdot D_2)$ komme på tale.

Til slutt skal vi spesielt betrakte 2x2-tilfellet.

VIII. 2x2-TABELLEN

VIII (i). Utledning av et avhengighetsmål

2x2-kontingenstabellen kan settes opp slik:

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

(24)

A og B er de to egenskapene vi skal måle avhengighet mellom. $\overline{A}$ og $\overline{B}$ er deres negasjoner (komplementer). Cellesannsynlighetene kan uttrykkes på følgende form (vises i appendikset):

Setn. 3

$$p_{11} = p_{1.}p_{.1} + (\Delta-1)\, p_{12}\, p_{21}$$
$$p_{12} = p_{1.}p_{.2} - (\Delta-1)\, p_{12}\, p_{21}$$
$$p_{21} = p_{2.}p_{.1} - (\Delta-1)\, p_{12}\, p_{21}$$
$$p_{22} = p_{2.}p_{.2} + (\Delta-1)\, p_{12}\, p_{21}.$$

hvor $\Delta = \dfrac{p_{11}\, p_{22}}{p_{12}\, p_{21}}$ er kryssprodukt-forholdet.

Den eksakte uavhengighetshypotesen kan formuleres slik:

$$H: \quad p_{11}\, p_{22} = p_{12}\, p_{21} \quad (\Longleftrightarrow \Delta=1) \tag{25}$$

Det er visse rimelige forutsetninger et avhengighetsmål for (A, B) bør tilfredsstille i en 2x2-tabell (se [1] og [6], s. 4). I de fleste tilfeller vil følgende tre krav være rimelige:

1) Målet må være en funksjon av den betingede sannsynlighet for B gitt A, $p_{11}/p_{11}+ p_{12}$ og den betingede sannsynlighet for B gitt $\bar{A}$, $p_{21}/p_{21}+ p_{22}$, eller alternativt av den betingede sannsynlighet for A gitt B, $p_{11}/p_{11}+ p_{21}$, og den betingede sannsynlighet for A gitt $\bar{B}$, $p_{12}/p_{12}+ p_{22}$.

2) De alternative mål i 1) skal være like.

3) Målet bør forandre seg monotont, for et gitt sett av marginaler $p_{1.}$ og $p_{.1}$, når avhengigheten blir sterkere.

Kravene 1), 2) og 3) leder til at avhengighetsmålet må være en en-entydig funksjon H av kryssprodukt-forholdet Δ. (Se [1].) H (Δ) vil være invariant under multiplikasjonen av rader og/eller kolonner, dvs. H (Δ) gir samme verdi til tabellen (24) og tabellen

	B	$\bar{B}$
A	$r_1c_1p_{11}$	$r_1c_2p_{12}$
$\bar{A}$	$r_2c_1p_{21}$	$r_2c_2p_{22}$

(26)

for alle ikke-negative r_1, r_2, c_1, c_2 slik at $r_1c_1p_{11}+ r_1c_2p_{12}+ r_2c_1p_{21}+ r_2c_2p_{22}= 1$. Det er dermed vist at det naturlige valg av avhengighetsmål i 2x2-tabellen essensielt er kryssprodukt-forholdet Δ.

Her vil vi nå nevne fire mål som er en-entydige funksjoner av Δ.

Yules "coefficient of assocation":

$$d_1 = \frac{p_{11}p_{22} - p_{12}p_{21}}{p_{11}p_{22} + p_{12}p_{21}} = \frac{\Delta-1}{\Delta+1} = 1 - \frac{2}{\Delta+1} \tag{27}$$

d_1 er det ordinalinvariante målet γ i 2x2-tilfellet.

Yules "coefficient of colligation":

$$d_2 = \frac{\sqrt{p_{11}p_{22}} - \sqrt{p_{12}p_{21}}}{\sqrt{p_{11}p_{22}} + \sqrt{p_{12}p_{21}}} = \frac{\sqrt{\Delta}-1}{\sqrt{\Delta}+1} = 1 - \frac{2}{\sqrt{\Delta}+1} \tag{28}$$

$$\rho = \ln \Delta \tag{29}$$

og selfølgelig Δ selv.

Yules to mål er strengt økende når Δ øker.
La oss nå definere hva vi mener med positiv og negativ avhengighet mellom A og B i tabellen (24).

Definisjon 3. Dersom $\Delta > 1$ $(p_{11}p_{22} > p_{12}p_{21})$ sier vi det er positiv avhengighet (p.a.) mellom A og B. Dersom $\Delta < 1$ er A og B negativt avhengige (n.a.).

Noen egenskaper ved Yules to mål:

(i) $-1 \leq d_i \leq 1$, $d_i > 0$ hvis p.a., $d_i < 0$ hvis n.a., for i=1, 2.

(ii) $d_i = 0 \iff$ A og B eksakt uavhengige.

(iii) d_i antar verdien -1, dersom $p_{11} = 0$ eller $p_{22} = 0$, for i=1, 2.
d_i antar verdien +1, dersom $p_{12} = 0$ eller $p_{21} = 0$, for i=1, 2.

Dersom vi ikke er interessert i hvilken retning avhengigheten går, men bare i graden av avhengighet vil et av målene d_1^2, d_2^2, eller ρ^2 være passende.

VIII (ii). Et alternativt avhengighetsmål

Det kan selfølgelig forekomme situasjoner hvor andre mål enn de basert på Δ kan være anvendelige. Her vil vi nevne ett:

Kendalls rangkorrelasjonskoeffisient: $$\tau_b = \frac{p_{11}p_{22} - p_{12}p_{21}}{\sqrt{p_{1.}p_{2.}p_{.1}p_{.2}}} \tag{30}$$

eller τ_b^2 hvis vi bare er interessert i graden av avhengighet mellom faktorene. (For andre mål se [1] og [3].)

IX. AVSLUTTENDE KOMMENTARER

Som vi har sett, vil de fleste mål som er konstruert ut fra en gitt modell ha den egenskapen at de er null hvis det foreligger ingen avhengighet relativt til de relevante avhengighetstrekk målet er konstruert for, selv om andre typer av avhengighet muligens er tilstede. Dette må vi vente siden vi skjerper "definisjonen av avhengighet" i de forskjellige situasjonene. Merk at for alle situasjonene, unntagen VI (i), vil eksakt uavhengighet mellom A og B medføre at målet er lik null.

Til slutt vil vi igjen, som i I (ii), presisere at ved valg av avhengighetsmål for en gitt kontingenstabell, bør man velge det mål som gir best informasjon om de avhengighetstrekk som er av interesse.

APPENDIKS

La oss vende tilbake til III (iii). Setningene 1 og 2 vil bli bevist.

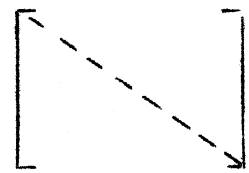
Setn. 1

$-\frac{m-1}{m} \leq \pi_s - \pi_d \leq \frac{m-1}{m}$. Grensene oppnås dersom alle cellesannsynlighetene er lik 0 utenfor en lengste diagonal i tabellen, og lik 1/m i diagonalen (m=min (v, w)).

Bevis

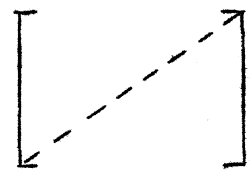
Antall celler i en lengste diagonal er lik m. Anta først v=m. Da vil $\pi_s - \pi_d$ anta sin maksimumsverdi dersom de positive cellesannsynlighetene er konsentrert i en lengste diagonal av typen:

(1)



Videre vil $\pi_s - \pi_d$ anta sin minimumsverdi dersom de positive cellesannsynlighetene er konsentrert i en lengste diagonal av typen:

(2)



Ekstremumsverdiene blir oppnådd hvis cellesannsynlighetene er like i diagonalene (iflg. [7]). Dvs. at $\max(\pi_s - \pi_d)$ inntreffer når $\sum_{i=1}^{m} p_{i,i+k} = 1$ for en k, slik at $0 \leq k \leq w-m$ og $p_{i,i+k} = \frac{1}{m}$ for $i=1, \ldots, m$.

Herav fås

$$\max(\pi_s - \pi_d) = 2\sum_{i=1}^{m-1} p_{i,i+k}\Big(\sum_{i'>i}\sum_{j'>i+k} p_{i'j'}\Big) - 2\sum_{i=1}^{m-1} p_{i,i+k}\Big(\sum_{i'>i}\sum_{j'<i+k} p_{i'j'}\Big)$$

$$= 2\sum_{i=1}^{m-1}\frac{1}{m}\Big(\sum_{i'>i}\frac{1}{m}\Big) - 2\sum_{i=1}^{m-1}\frac{1}{m}\Big(\sum_{i'>i}\sum_{j'<i+k} 0\Big) = \frac{2}{m^2}\sum_{i=1}^{m-1}\sum_{i'>i} 1 = \frac{2}{m^2}\{(m-1) + (m-2)$$

$$+ \ldots + 1\} = \frac{2}{m^2}\frac{(m-1)\,m}{2} = \frac{m-1}{m}.$$

Tilsvarende vil min $(\pi_s - \pi_d)$ inntreffe når

$$\sum_{i=0}^{m-1} p_{m-i,k+i} = 1 \quad \text{for en} \quad k \quad \text{slik at} \quad 1 \leq k \leq w-m+1$$

og $$p_{m-i,k+i} = \frac{1}{m}, \quad \text{for } i=0, 1, \ldots m-1.$$

Dette gir at

$$\min (\pi_s - \pi_d) = 2 \sum_{i=0}^{m-1} p_{m-i,k+i} \Big(\sum_{i'>m-i} \sum_{j'>k+i} p_{i'j'} \Big)$$

$$- 2 \sum_{i=0}^{m-1} p_{m-i,k+i} \Big(\sum_{i'>m-i} \sum_{j'<k+i} p_{i'j'} \Big) .$$

Det ses at $p_{i'j'} = 0$ for $i' > m-i$ og $j'>k+i$ fordi de positive cellesannsynlighetene er $p_{i',k+m-i'}$, og hvis $j'>k+i$ så er $j'>k+m-i'$ siden $i>m-i'$. Dessuten ses at

$$\sum_{i'>m-i} \sum_{j'<k+i} p_{i'j'} - \sum_{i'>m-i} p_{i',k+m-i'} = \sum_{i'>m-i} \frac{1}{m} \quad \text{siden}$$

$j' = k+m-i' \Rightarrow j'<k+i$

Herav: $$\min (\pi_s - \pi_d) = -2 \sum_{i=1}^{m-1} \sum_{i'>m-i} \frac{1}{m^2} = -\frac{2}{m^2}\{1+2+\ldots+(m-1)\} = -\frac{m-1}{m} .$$

Dersom $w = \min (v, w)$ blir beviset helt analogt. Forskjellen er bare den at $\max (\pi_s - \pi_d)$ inntreffer når $\sum_{j=1}^{m} p_{j+k,j} = 1$ hvor $0 \leq k \leq v-m$ og $p_{j+k,j} = \frac{1}{m}$, og $\min (\pi_s - \pi_d)$ inntreffer når $\sum_{j=1}^{m} p_{v-k-j,j} = 1$ for $-1 \leq k \leq v-m-1$ og $p_{v-k-j,j} = \frac{1}{m}$. Q.E.D.

Neste resultat viser at $\hat{\tau}_b$ er et spesialtilfelle av Γ definert ved (10).

Setn. 2

La Y-scorene a_{ij} og Z-scorene b_{ij} være gitt ved:

$$a_{ij} = \begin{cases} +1 & \text{hvis } y_i < y_j \\ 0 & " \quad y_i = y_j \\ -1 & " \quad y_i > y_j \end{cases} \quad \text{og} \quad b_{ij} = \begin{cases} +1 & \text{hvis } z_i < z_j \\ 0 & " \quad z_i = z_j \\ -1 & " \quad z_i > z_j \end{cases}$$

Da er $\Gamma = \hat{\tau}_b = \dfrac{P_s - P_d}{\sqrt{P_y P_z}}$.

Bevis

Fra (10): $\Gamma = \dfrac{\sum\limits_{i\neq j} a_{ij} b_{ij}}{\{\sum\limits_{i\neq j} a_{ij}^2 \cdot \sum\limits_{i\neq j} b_{ij}^2\}^{\frac{1}{2}}}$

Antall ordnete par blant k individer er $k(k-1)$.

Det gir da at antall ordnete par blant de i alt $n(n-1)$ ordnete par hvor $a_{ij} = 0$ er $\sum\limits_{i=1}^{v} X_{i.}(X_{i.} - 1)$, slik at $\sum\limits_{i\neq j} a_{ij}^2 = n(n-1) - \sum\limits_{i=1}^{v} X_{i.}(X_{i.} - 1)$

$= n^2 - \sum\limits_{i=1}^{v} X_{i.}^2$.

Analogt blir: $\sum\limits_{i\neq j} b_{ij}^2 = n(n-1) - \sum\limits_{j=1}^{w} X_{.j}(X_{.j}-1) = n^2 - \sum\limits_{j=1}^{w} X_{.j}^2$.

Dette medfører at nevneren i Γ kan uttrykkes slik:

$$\{\sum_{i\neq j} a_{ij}^2 \cdot \sum_{i\neq j} b_{ij}^2\}^{\frac{1}{2}} = \{(n^2 - \sum_{i=1}^{v} X_{i.}^2)(n^2 - \sum_{j=1}^{w} X_{.j}^2)\}^{\frac{1}{2}}$$

$$= n^2\{(1 - \sum_{i=1}^{v} q_{i.}^2)(1 - \sum_{j=1}^{w} q_{.j}^2)\} = n^2(P_y \cdot P_z)^{\frac{1}{2}} .$$

La U være summen av scorene $a_{ij} b_{ij}$ for alle $n(n-1)$ ordnete par, dvs. $U = \sum\limits_{i\neq j} a_{ij} b_{ij}$.

Nå er det klart at paret $\{(y_i, z_i), (y_j, z_j)\}$ gir samme score som $\{(y_j, z_j), (y_i, z_i)\}$, siden $a_{ij} b_{ij} = a_{ji} b_{ji}$.

Dermed

$U = 2 \sum\limits_{i<j} a_{ij} b_{ij}$ ($\sum\limits_{i<j} a_{ij} b_{ij}$ er hva Kendall, [4], kaller total-scoren S)

Videre ser man at for hver observasjon (y_i, z_i) i celle (rk) så er $a_{ij} b_{ij} = 1$ for alle observasjonene (y_j, z_j) i celle (r', k') hvor $r' > r$ og $k' > k$ eller $r' < r$ og $k' < k$. Tilsvarende er $a_{ij} b_{ij} = -1$ for alle observasjoner (y_j, z_j) i celle (r'', k'') hvor $r'' > r$ og $k'' < k$ eller $r'' < r$ og $k'' > k$.

Ellers vil alle par gi score 0. Herav:

$$\sum_{i<j} a_{ij}b_{ij} = X_{11}\Big(\sum_{i>1}\sum_{j>1} X_{ij}\Big) + X_{12}\Big(\sum_{i>1}\sum_{j>2} X_{ij}\Big) + \ldots + X_{v-1,w-1}X_{vw}$$

$$- \Big(X_{12}X_{21} + \ldots + X_{v-1,w}\Big(\sum_{j<w} X_{vj}\Big)\Big)$$

$$= \sum_{r=1}^{v-1}\sum_{k=1}^{w-1} X_{rk}\Big(\sum_{i>r}\sum_{j>k} X_{ij}\Big) - \sum_{r=1}^{v-1}\sum_{k=2}^{w} X_{rk}\Big(\sum_{i>r}\sum_{j<k} X_{ij}\Big)$$

Dette gir:

$$U = 2n^2\Big(\sum_{r=1}^{v-1}\sum_{k=1}^{w-1} q_{rk}\Big(\sum_{i>r}\sum_{j>k} q_{ij}\Big)\Big) - 2n^2\sum_{r=1}^{v-1}\sum_{k=2}^{w} q_{rk}\Big(\sum_{i>r}\sum_{j<k} q_{ij}\Big)$$

$$= n^2(P_s - P_d)$$

Dermed er

$$\Gamma = \frac{U}{n^2\sqrt{P_y P_z}} = \frac{P_s - P_d}{\sqrt{P_y . P_z}} \qquad \text{Q.E.D.}$$

Det siste resultatet som skal bevises er setn. 3 fra VIII (i) i (2x2-tabellen):

Setn. 3

$$P_{11} = P_{1.}P_{.1} + (\Delta-1)P_{11}P_{21}$$

$$P_{12} = P_{1.}P_{.2} - (\Delta-1)P_{12}P_{21}$$

$$P_{21} = P_{2.}P_{.1} - (\Delta-1)P_{12}P_{21}$$

$$P_{22} = P_{2.}P_{.2} + (\Delta-1)P_{12}P_{21}, \quad \text{hvor } \Delta = \frac{P_{11}P_{22}}{P_{12}P_{21}}$$

Bevis

a) $P_{11} - P_{1.}P_{.1} = P_{11} - (P_{11}+P_{12})(P_{11}+P_{21}) = P_{11} - P_{11}(P_{11}+P_{12}+P_{21}) - P_{12}P_{21}$

$= P_{11} - P_{11}(1-P_{22}) - P_{12}P_{21} = P_{11}P_{22} - P_{12}P_{21} = (\Delta-1)P_{12}P_{21}$.

b) $P_{12} - P_{1.}P_{.2} = P_{12} - P_{12}(1-P_{21}) - P_{11}P_{22} = P_{12}P_{21} - P_{11}P_{22} = -(\Delta-1)P_{12}P_{21}$.

c) $P_{21} - P_{2.}P_{.1} = P_{21} - P_{21}(1-P_{12}) - P_{11}P_{22} = P_{12}P_{21} - P_{11}P_{22} = -(\Delta-1)P_{12}P_{21}$.

d) $P_{22} - P_{2.}P_{.2} = P_{22} - P_{22}(1-P_{11}) - P_{12}P_{21} = P_{11}P_{22} - P_{12}P_{21} = (\Delta-1)P_{12}P_{21}$

Q.E.D.

REFERANSER

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

[2] Goodman, L.A. & Kruskal, W.H. (1954): "Measures of association for cross classifications", J. Am. Statist. Ass., Vol.49, 732-764.

[3] Goodman, L.A. & Kruskal, W.H. (1959): "Measures of association for cross classifications II. Further discussions and references", J. Am. Statist. Ass., Vol.54, 123-163.

[4] Kendall, M.G, (1955): "Rank Correlation Methods", Charles Griffin & Co. Lim.

[5] Kruskal, W.H. (1958): "Ordinal measures of association", J. Am. Statist. Ass., Vol. 53, 814-861.

[6] Mosteller, F. (1968): "Association and estimation in contingency tables", J. Am. Statist. Ass., Vol. 63, 1-28.

[7] Stuart, A. (1953): "The estimation and comparison of strengths of association in contingency tables", Biometrika, Vol. 39, 105-110.