Hypoteseprøvning i den multinomiske fordeling fra et geometrisk synspunkt

Af Ernst Lykke Jensen*), Sven L. Caspersen**),
Axel Schultz Nielsen***) *og* Jørgen Kai Olsen***)

Resumé:

C. R. Rao har i [6, kapitel 5 og 6] givet en fremstilling af den parametriske teori for den multinomiske fordeling, der i det væsentlige er baseret på dels et krav om, at estimatoren er asymptotisk efficient (af første orden) [6, p. 285], og dels en sætning, der angiver en tilstrækkelig betingelse for, at en kvadratisk form i Karl Pearsons vektor er fordelt efter χ^2-fordelingen [6, p. 318]. Hensigten med den foreliggende artikel er at forenkle teorien yderligere ved udnyttelse af den kendsgerning, at estimatoren er asymptotisk ækvivalent med en projektion.

1. *Indledning*

Det antages, at klassesandsynlighederne $\pi_1, \ldots \pi_k$ i den multinomiske fordeling er funktioner af en parametervektor $\theta = (\theta_1, \ldots, \theta_q)'$ med q elementer, $q < k$. Idet n er antallet af uafhængige gentagelser i det multinomiske eksperiment, sætter vi $D = \sqrt{n}(\hat{\theta} - \theta)$, hvor $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_q)'$ er estimator for θ. Lad $l(\theta)$ betegne logaritmen til likelihood funktionen, og lad $Z = (Z_1, \ldots, Z_q)'$ være en vektor, hvis r'te element er $Z_r = n^{-\frac{1}{2}} \partial l(\theta) / \delta\theta_r$. Estimatoren siges at være asymptotisk efficient, hvis den for $n \to \infty$ er fuldstændig korreleret med den afledede af likelihood funktionen, d.v.s. $D \overset{a}{=} BZ$, hvor B er en matrix med konstante elementer, der gerne må afhænge af θ. Symbolet $\overset{a}{=}$ læses »asymptotisk ækvivalent med« og har betydningen at differensen mellem venstresiden og højresiden konvergerer i sandsynlighed mod nul. Hvis antal observationer i de k klasser er $n_1, \ldots, n_k$ ($n = n_1 + \ldots + n_k$), er $l(\theta) = n_1 \log \pi_1(\theta) + \ldots + n_k \log \pi_k(\theta)$ og

*) Professor, dr. polit.,
**) Afdelingsleder, cand. polit.,
***) Amanuensis, cand. merc., Institut for teoretisk Statistik, Handelshøjskolen i København.

Hypoteseprøvning i den multinomiske fordeling fra et geometrisk synspunkt

Af Ernst Lykke Jensen*), Sven L. Caspersen**),
Axel Schultz Nielsen***) *og* Jørgen Kai Olsen***)

Resumé:

C. R. Rao har i [6, kapitel 5 og 6] givet en fremstilling af den parametriske teori for den multinomiske fordeling, der i det væsentlige er baseret på dels et krav om, at estimatoren er asymptotisk efficient (af første orden) [6, p. 285], og dels en sætning, der angiver en tilstrækkelig betingelse for, at en kvadratisk form i Karl Pearsons vektor er fordelt efter χ^2-fordelingen [6, p. 318]. Hensigten med den foreliggende artikel er at forenkle teorien yderligere ved udnyttelse af den kendsgerning, at estimatoren er asymptotisk ækvivalent med en projektion.

1. *Indledning*

Det antages, at klassesandsynlighederne $\pi_1, \ldots \pi_k$ i den multinomiske fordeling er funktioner af en parametervektor $\theta = (\theta_1, \ldots, \theta_q)'$ med q elementer, $q < k$. Idet n er antallet af uafhængige gentagelser i det multinomiske eksperiment, sætter vi $D = \sqrt{n}(\hat{\theta} - \theta)$, hvor $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_q)'$ er estimator for θ. Lad $l(\theta)$ betegne logaritmen til likelihood funktionen, og lad $Z = (Z_1, \ldots, Z_q)'$ være en vektor, hvis r'te element er $Z_r = n^{-\frac{1}{2}} \partial l(\theta) / \delta\theta_r$. Estimatoren siges at være asymptotisk efficient, hvis den for $n \to \infty$ er fuldstændig korreleret med den afledede af likelihood funktionen, d.v.s. $D \overset{a}{=} BZ$, hvor B er en matrix med konstante elementer, der gerne må afhænge af θ. Symbolet $\overset{a}{=}$ læses »asymptotisk ækvivalent med« og har betydningen at differensen mellem venstresiden og højresiden konvergerer i sandsynlighed mod nul. Hvis antal observationer i de k klasser er $n_1, \ldots, n_k$ ($n = n_1 + \ldots + n_k$), er $l(\theta) = n_1 \log \pi_1(\theta) + \ldots + n_k \log \pi_k(\theta)$ og

*) Professor, dr. polit.,
**) Afdelingsleder, cand. polit.,
***) Amanuensis, cand. merc., Institut for teoretisk Statistik, Handelshøjskolen i København.

$$Z_r = \frac{1}{\sqrt{n}} \sum_{i=1}^{k} \frac{n_i}{\pi_i} \frac{\delta \pi_i}{\delta \theta_r} = \sum_{i=1}^{k} \frac{1}{\sqrt{\pi_i}} \frac{\delta \pi_i}{\delta \theta_r} \frac{n_i - n\pi_i}{\sqrt{n\pi_i}},$$

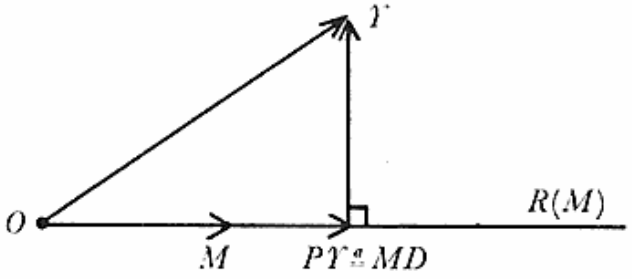
idet $\pi_1 + \ldots + \pi_k = 1$ medfører, at $\delta\pi_1/\delta\theta_r + \ldots + \delta\pi_k/\delta\theta_r = 0$. I matrixformulering kan vi skrive denne relation på formen $Z = M'Y$, hvor M er en kxq matrix, hvis (i,j)'te element er $\pi_i^{-\frac{1}{2}} \delta\pi_i/\delta\theta_j$, og hvor Y er Karl Pearsons vektor

$$Y = \left(\frac{n_1 - n\pi_1}{\sqrt{n\pi_1}}, \ldots, \frac{n_k - n\pi_k}{\sqrt{n\pi_k}} \right)'.$$

Det (r, s)'te element i $M'M$, d.v.s.

$$\sum_{i=1}^{k} \frac{1}{\pi_i} \frac{\delta \pi_i}{\delta \theta_r} \frac{\delta \pi_i}{\delta \theta_s},$$

er det (r, s)'te element i Fishers informationsmatrix $\mathfrak{J}$ for en enkelt multinomisk observation. Vi forudsætter, at søjlerne i M er lineært uafhængige, således at $\mathfrak{J}$ er regulær. Vælger vi nu $B = \mathfrak{J}^{-1}$, er $D \triangleq \mathfrak{J}^{-1}Z = (M'M)^{-1}M'Y$, d.v.s. $MD \triangleq PY$, hvor $P = M(M'M)^{-1}M'$ er en projektionsmatrix. Vi ser altså, at MD er asymptotisk ækvivalent med projektionen af Karl Pearsons vektor på Fishers informationsrum, d.v.s. det vektorrum $R(M)$ med dimension q, som udspændes af søjlerne i M.



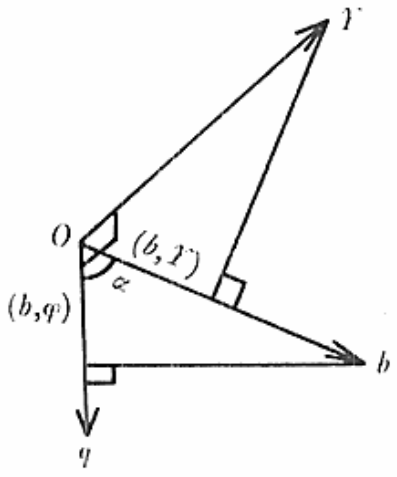
Figur 1

I 1900 beviste Karl Pearson, at en simpel hypotese angående $\pi_1, \ldots, \pi_k$ kan testes ved beregning af Y^2, der under nulhypotesen asymptotisk, d.v.s. for $n \to \infty$, er fordelt efter χ^2-fordelingen med $k-1$ frihedsgrader. Hvis man i Karl Pearsons teststørrelse erstatter π_i med $\pi_i(\hat{\theta})$, hvor $\hat{\theta}$ er maksimum likelihood estimator for θ, fremkommer en stokastisk variabel, der har samme asymptotiske fordeling som $(Y - PY)^2$. Fisher [5] har vist, at fordelingen er en χ^2-fordeling med $k-1-q$ frihedsgrader, når modellen $\pi_i = \pi_i(\theta)$ er rigtig. I Cramér [4, kapitel 30] finder man

relationen $D \triangleq \mathfrak{I}^{-1}\mathcal{Z}$; men bevisførelsen er baseret på den unødvendigt strenge forudsætning, at klassesandsynlighederne har kontinuerte afledede af anden orden. Rao har bevist, [6, p. 296], at eksistensen af de afledede i en omegn af θ og deres kontinuitet i θ (suppleret med en identifikationsbetingelse) er nok til at sikre eksistensen af en løsning af likelihood ligningen $\mathcal{Z} = 0$, der er konsistent og asymptotisk efficient. Holst Andersen [1] tager udgangspunkt i den eksponentielle klasse af fordelinger og viser også, at likelihood ratio testet er asymptotisk ækvivalent med χ^2-testet. I Birch [2] er relationen $D \triangleq \mathfrak{I}^{-1}\mathcal{Z}$ etableret under den svagere forudsætning, at klassesandsynlighederne er differentiable i det sande parameterpunkt θ.

2. *Fordelingslovene for Υ, $P\Upsilon$, $\Upsilon - P\Upsilon$ og D. Karl Pearsons sætning*

Den asymptotiske fordeling for Υ er $N_{k,k-1}(0, I_k - \varphi\varphi')$, d.v.s. en k-dimensional normal fordeling med rang $k-1$, nulvektoren som middelværdivektor og med kovariansmatrix $I_k - \varphi\varphi'$, hvor I_k er enhedsmatricen af orden k, og φ er enhedsvektoren $(\sqrt{\pi_1}, .., \sqrt{\pi_k})'$.



Figur 2

Det er tilstrækkeligt at vise, at fordelingen er en endimensional normal fordeling i en vilkårlig valgt retning. Vi vælger retningen bestemt ved enhedsvektoren b og skal vise, at skalarproduktet

$$(b,\Upsilon) = \sum_{i=1}^{k} b_i \frac{n_i - n\pi_i}{\sqrt{n\pi_i}} = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{k} n_i \frac{b_i}{\sqrt{\pi_i}} - (b,\varphi)\right)$$

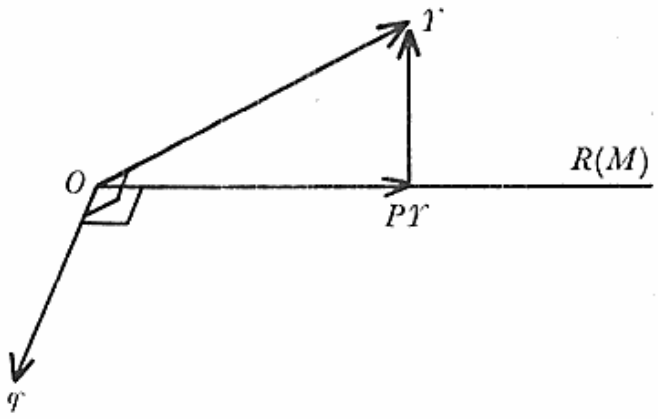
er normalt fordelt for $n \to \infty$. Lad U være en stokastisk variabel med sandsynlighedsfordeling $Pr(U = b_i/\sqrt{\pi_i}) = \pi_i$ $(i = 1, .., k)$. EU, EU^2 og $\mathrm{var}U$ er henholdsvis $EU = (b, \varphi)$ $(= \cos\alpha$, hvor α er vinklen mellem b og $\varphi)$, $EU^2 = b^2$ $(= 1)$ og $\mathrm{var}U = b^2 - (b, \varphi)^2$ $(= 1 - \cos^2\alpha)$. Da nu

$$(b, \Upsilon) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} U_i - (b, \varphi)\right)$$

og $U_1, .., U_n$ er uafhængige og identisk fordelt efter ovennævnte sandsynlighedsfordeling, er $E(b, \Upsilon) = 0$ og $\mathrm{var}(b, \Upsilon) = b^2 - (b, \varphi)^2 = b'(I_k - \varphi\varphi')b$. Heraf følger, at Υ for ethvert n har middelværdi $E\Upsilon = 0$ og kovariansmatrix $\mathrm{cov}\,\Upsilon = I_k - \varphi\varphi'$. Ved hjælp af den centrale grænseværdisætning slutter vi endvidere, at fordelingen af (b, Υ) for $n \to \infty$ er en normal fordeling. Tilbage står at vise, at fordelingen af Υ har rangen $k-1$. Dette er ensbetydende med at vise, at Karl Pearsons vektor i et passende valgt koordinatsystem har én koordinat lig med 0, og at de øvrige koordinater er fordelt efter $N_{k-1}(0, I_{k-1})$. Da $(\varphi, \Upsilon) = n^{-\frac{1}{2}}\sum_{i=1}^{k}(n_i - n\pi_i) = 0$, står φ-vektoren vinkelret på Karl Pearsons vektor. Vi drejer derfor koordinatsystemet omkring begyndelsespunktet 0 over i et nyt koordinatsystem, således at φ bliver basisvektor for en, f.eks. den sidste, af det nye systems koordinatakser. Lad de nye koordinater for Karl Pearsons vektor være $(y_1, .., y_{k-1}, 0)$. Lad A_j være en søjlevektor, hvis elementer er gamle koordinater for den j'te basisvektor i det nye koordinatsystem $(A_k = \varphi)$; da er $y_j = (A_j, \Upsilon)$. Lad A være en $k \times (k-1)$ matrix med søjler $A_1, .., A_{k-1}$. De $k-1$ første nye koordinater for Karl Pearsons vektor er elementerne i vektoren $A'\Upsilon$, der ifølge det ovenfor beviste er fordelt efter $N_{k-1}(0, I_{k-1})$; thi da φ er vinkelret på $A_1, .., A_{k-1}$, og $A'A = I_{k-1}$ $(A_1, .., A_{k-1}$ er enhedsvektorer), fås $\mathrm{cov}(A'\Upsilon) = A'(\mathrm{cov}\,\Upsilon)A = A'(I_k - \varphi\varphi')A = A'A = I_{k-1}$.

Det er herefter en simpel sag at angive fordelingsloven for projektionsvektoren $P\Upsilon$ og residualvektoren $\Upsilon - P\Upsilon = (I_k - P)\Upsilon$.

Det bemærkes, at φ-vektoren står vinkelret på Fishers informationsrum, idet $(\varphi, M_j) = \delta\pi_1/\delta\theta_j + .. + \delta\pi_k/\delta\theta_j = 0$ $(j = 1, .., q)$, hvor M_j er j'te søjle i M. Følgelig er kovariansmatricerne for $P\Upsilon$ og $\Upsilon - P\Upsilon$ henholdsvis $\mathrm{cov}(P\Upsilon) = P(\mathrm{cov}\,\Upsilon)P' = P(I_k - \varphi\varphi')P = P$ og $\mathrm{cov}(\Upsilon - P\Upsilon) = (I_k - P)\,\mathrm{cov}\,\Upsilon(I_k - P)' = (I_k - P)(I_k - \varphi\varphi') = I_k - \varphi\varphi' - P$, idet P og $I_k - P$ er symmetriske og idempotente matricer. Til den ortogonale opspaltning $\Upsilon = (\Upsilon - P\Upsilon) + P\Upsilon$ af Karl Pearsons vektor svarer altså kovariansmatrixopspaltningen $I_k - \varphi\varphi' = (I_k - \varphi\varphi' - P) + P$. Vi lader nu Fishers infor-



Figur 3

mationsrum være udspændt af de q første koordinatakser $A_1, .., A_q$ i det nye koordinatsystem, hvorved projektionsvektorens og residualvektorens nye koordinater er henholdsvis $(y_1, .., y_q, o, .., o)$ og $(o, .., o, y_{q+1}, .., y_{k-1}, o)$. Da nu $y_1, .., y_{k-1}$ asymptotisk er stokastisk uafhængige og fordelt efter den standardiserede normalfordeling, er PY og $Y - PY$ asymptotisk stokastisk uafhængige og fordelt efter henholdsvis $N_{k,q}(0,P)$ og $N_{k,k-1-q}(0, I - \varphi\varphi' - P)$.

Længderne af Karl Pearson vektoren, projektionsvektoren og residualvektoren er invariante over for drejningen af koordinatsystemet. Heraf følger Karl Pearsons sætning, nemlig at

$$\sum_{i=1}^{k} \frac{(n_i - n\pi_i)^2}{n\pi_i} = Y^2 = \sum_{i=1}^{k-1} y_i^2$$

asymptotisk er fordelt efter χ^2-fordelingen med $k-1$ frihedsgrader, samt at

$$Y'PY = (PY)^2 = \sum_{i=1}^{q} y_i^2$$

og

$$Y'(I_k - \varphi\varphi' - P)Y = Y'(I_k - P)Y = (Y - PY)^2 = \sum_{i=q+1}^{k-1} y_i^2$$

asymptotisk er stokastisk uafhængige og fordelt efter χ^2-fordelingen med henholdsvis q og $k-1-q$ frihedsgrader.

Da $Z = M'Y$ og $M'M = \mathfrak{J}$, og da φ står vinkelret på $R(M)$, er $\operatorname{cov} Z = M'(I_k - \varphi\varphi')M = \mathfrak{J}$. Følgelig er Z asymptotisk fordelt efter $N_q(O, \mathfrak{J})$. Heraf afledes umiddelbart fordelingsloven for $D = \sqrt{n}(\hat{\theta} - \theta)$; thi da $D \overset{a}{=} \mathfrak{J}^{-1}Z$, har D samme asymptotiske fordeling som $\mathfrak{J}^{-1}Z$, og følgelig er D asymptotisk fordelt efter $N_q(O, \mathfrak{J}^{-1})$.

3. *χ^2-testet for modelkontrol*

Da π_i er differentiabel i punktet θ, fås ved en Taylorudvikling, at den i'te koordinat i MD er

$$\sum_{r=1}^{q} \frac{1}{\sqrt{\pi_i}} \frac{\delta \pi_i}{\delta \theta_r} \cdot \sqrt{n}(\hat{\theta}_r - \theta_r) \triangleq \frac{n\pi_i(\hat{\theta}) - n\pi_i(\theta)}{\sqrt{n\pi_i(\theta)}},$$

idet restleddet konvergerer i sandsynlighed mod nul. Heraf kan vi slutte, at

$$R = \sum_{i=1}^{k} \frac{(n\pi_i(\hat{\theta}) - n\pi_i(\theta))^2}{n\pi_i(\theta)}$$

har samme asymptotiske fordeling som $(P\Upsilon)^2$, d.v.s. en χ^2-fordeling med q frihedsgrader. Hvis modellen ikke forkastes ved testet for modelkontrol, der omtales nedenfor, har Rao [7, p. 31] foreslået R som teststørrelse ved afprøvning af en simpel hypotese for θ. Testet er asymptotisk uafhængig af χ^2-testet for modelkontrol, der førte til godkendelse af modellen. Da $P\Upsilon \triangleq MD \triangleq M\mathfrak{J}^{-1}\mathcal{Z}$, er R under nulhypotesen asymptotisk ækvivalent med teststørrelserne $(MD)^2 = D'\mathfrak{J}D$, $(M\mathfrak{J}^{-1}\mathcal{Z})^2 = \mathcal{Z}'\mathfrak{J}^{-1}\mathcal{Z}$ og, såfremt θ estimeres ved maksimum likelihood metoden, med likelihood ratio testet.

Den i'te koordinat i residualvektoren $\Upsilon - P\Upsilon$ er asymptotisk ækvivalent med

$$\frac{n_i - n\pi_i(\theta)}{\sqrt{n\pi_i(\theta)}} - \frac{n\pi_i(\hat{\theta}) - n\pi_i(\theta)}{\sqrt{n\pi_i(\theta)}} = \frac{n_i - n\pi_i(\hat{\theta})}{\sqrt{n\pi_i(\theta)}} \triangleq \frac{n_i - n\pi_i(\hat{\theta})}{\sqrt{n\pi_i(\hat{\theta})}},$$

hvor det sidste skridt begrundes med konsistensen af $\hat{\theta}$ og kontinuiteten af π_i. Følgelig er teststørrelsen for modelkontrol

$$\sum_{i=1}^{k} \frac{(n_i - n\pi_i(\hat{\theta}))^2}{n\pi_i(\hat{\theta})}$$

asymptotisk fordelt som $(\Upsilon - P\Upsilon)^2$, d.v.s. som χ^2 med $k-1-q$ frihedsgrader. Den er asymptotisk uafhængig af de ovenfor nævnte teststørrelser for en simpel hypotese vedrørende θ, da Υ og $\Upsilon - P\Upsilon$ er asymptotisk uafhængige.

4. *Test for afvigelse i en enkelt klasse*

Dersom χ^2-testet fører til forkastelse af modellen, kan det have interesse at undersøge, om en bestemt klasse yder et særligt stort bidrag til teststørrelsen. Cochran har i [3] for specielle tilfælde angivet formler

for variansen af $L = n_i - n\pi_i(\hat{\theta})$, der sætter os i stand til at teste afvigelsen ved hjælp af den standardiserede normalfordeling. Det vides ikke om Cochran, som bebudet i artiklen, har publiceret sit bevis. Et simpelt bevis, der kan betragtes som en forenkling af beviset i Rao [6; p. 328], er følgende.

Vektoren med koordinater

$$r_i = \frac{n_i - n\pi_i(\hat{\theta})}{\sqrt{n\pi_i(\hat{\theta})}}, \; i = 1, .., k,$$

har samme asymptotiske kovariansmatrix som $Y - PY$, hvorfor den asymptotiske varians af r_i er det i'te diagonalelement i $I_k - \varphi\varphi' - P$, d.v.s. $1 - \pi_i - P_{ii}$. Hvis π_i er kontinuert differentiabel, kan vi slutte, at Cochrans teststørrelse

$$u = \frac{L}{\sqrt{V(L)}}, \; V(L) = n\pi_i(\hat{\theta})\,(1 - \pi_i(\hat{\theta}) - P_{ii}(\hat{\theta}))$$

asymptotisk følger en standardiseret normalfordeling, hvis afvigelsen i den i'te klasse blot skyldes en tilfældighed. Man bemærker, at $V(L)$ er binomialfordelingens varians korrigeret med det i'te diagonalelement i projektionsmatricen.

Accepteres modellen, men forkastes en simpel hypotese $\theta = \theta_0$ ved Rao's test, nævnt i afsnit 3, kan et signifikant bidrag til R fra den i'te klasse afsløres ved teststørrelsen

$$u = \frac{\pi_i(\hat{\theta}) - \pi_i(\theta_0)}{\sqrt{\pi_i(\theta_0) \cdot P_{ii}(\theta_0)}} \sqrt{n},$$

idet vektoren med koordinater

$$\frac{n\pi_i(\hat{\theta}) - n\pi_i(\theta)}{\sqrt{n\pi_i(\theta)}}$$

har samme asymptotiske kovariansmatrix som PY.

Som eksempel betragter vi et talmateriale bestående af n Poisson-observationer $x_1, x_2, .., x_n$ med parameter θ grupperet i k klasser, således at $\pi_i(\theta) = \theta^i \exp(-\theta)/i!$, $i = 0, 1, .., k-1$. Vi antager, at n er stor nok til at berettige bortkastelse af restsandsynligheden $\sum_{i=k}^{\infty} \pi_i(\theta)$. Da $\delta \log \pi_i(\theta)/\delta\theta = (i-\theta)/\theta$, er $\sum n_i \cdot i/n = \bar{x}$ maksimum likelihood estimatoren for θ. Af formlen $P = M(M'M)^{-1}M'$ finder vi

$$P_{ii} = \left[\sum_{j=0}^{k-1} \frac{1}{\pi_j}\left(\frac{d\pi_j}{d\theta}\right)^2\right]^{-1} \cdot \frac{1}{\pi_i}\left(\frac{d\pi_i}{d\theta}\right)^2 = \pi_i(\theta)\,\frac{(i-\theta)^2}{\theta},$$

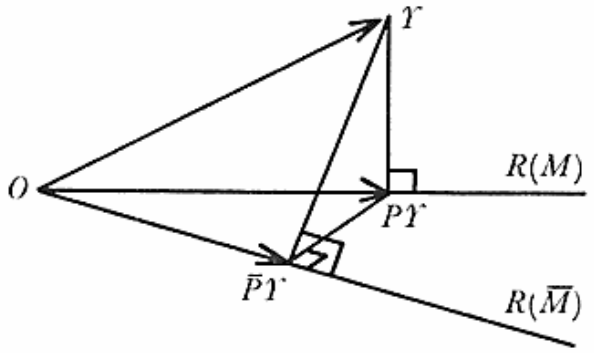
hvilket netop er Cochrans korrektionsled.

5. *Successiv testning*

Vi ønsker på grundlag af specifikationen $\theta_i = g_i(\tau_1, .., \tau_r)$, $i = 1, .., q$, $r < q$, at undersøge om modellen kan udformes med færre parametre. Det antages, at qxr matricen Δ med elementer $\delta g_i/\delta\tau_j$ har rangen r. Lad $R(\overline{M})$ være Fishers informationsrum under τ-modellen. Da

$$\frac{1}{\sqrt{\pi_i}}\frac{\delta\pi_i}{\delta\tau_j} = \sum_{s=1}^{q} \frac{1}{\sqrt{\pi_i}}\frac{\delta\pi_i}{\delta\theta_s}\cdot\frac{\delta g_s}{\delta\tau_j},$$

er $\overline{M} = M\Delta$, hvorfor $R(\overline{M})$ er et underrum af $R(M)$, og Fishers informations matrix $\overline{M}'\overline{M} = \Delta'\mathfrak{J}\Delta$ under τ-modellen er regulær. Lad $\hat{\theta}$ og $\hat{\tau}$ være



Figur 4

efficiente estimatorer under henholdsvis θ- og τ-modellen, og lad PY og $\overline{P}Y$ være projektionerne på henholdsvis $R(M)$ og $R(\overline{M})$. Hvis vi lader $R(\overline{M})$ være udspændt af de r første basisvektorer $A_1, .., A_r$ for det nye koordinatsystem, har vektoren $PY - \overline{P}Y$ de nye koordinater $(0, .., 0, y_{r+1}, .., y_q, 0, .., 0)$. Hvis θ-modellen ikke er forkastet efter modelkontroltestet i afsnit 3, og hvis τ-modellen er rigtig, er $\pi_i(\theta) = \pi_i(\tau)$, og følgelig har den i'te koordinat for vektoren $PY - \overline{P}Y$ den asymptotisk ækvivalente fremstilling

$$\frac{n\pi_i(\hat{\theta})-n\pi_i(\theta)}{\sqrt{n\pi_i(\theta)}}-\frac{n\pi_i(\hat{\tau})-n\pi_i(\tau)}{\sqrt{n\pi_i(\tau)}}=\frac{n\pi_i(\hat{\theta})-n\pi_i(\hat{\tau})}{\sqrt{n\pi_i(\tau)}}\overset{a}{=}\frac{n\pi_i(\hat{\theta})-n\pi_i(\hat{\tau})}{\sqrt{n\pi_i(\hat{\tau})}}.$$

Hvis τ-modellen er rigtig, kan vi heraf slutte, at teststørrelsen

$$\sum_{i=1}^{k}\frac{(n\pi_i(\hat{\theta})-n\pi_i(\hat{\tau}))^2}{n\pi_i(\hat{\tau})}\overset{a}{=}\sum_{i=r+1}^{q}y_i^2$$

asymptotisk er fordelt som $\sum_{i=r+1}^{q}y_i^2$, d.v.s. som χ^2 med $q-r$ frihedsgrader.

En simpel hypotese for τ afprøves ved et χ^2-test med r frihedsgrader på den i afsnit 3 angivne måde. Det bemærkes, at testet på et bestemt trin, på grund af den ortogonale opspaltning af Karl Pearsons vektor, asymptotisk er uafhængig af de tests, der førte til godkendelse af hypoteserne på de foregående trin.

6. *Sammenligning af flere fordelinger*

Der foreligger m multinomiske fordelinger med k_i klasser i den i'te fordeling. Á priori er klassesandsynligheden π_{ij} for den j'te klasse i den i'te fordeling en differentiabel funktion af $\theta_{i1},\ldots,\theta_{iq}$. Vi ønsker at teste homogenitetshypotesen $\theta_{ir}=\theta_r$. Hvis vi anbringer klassesandsynlighederne som elementer i en vektor i rækkefølgen $\pi_{11},\ldots,\pi_{1k_1}$, $\pi_{21},\ldots,\pi_{2k_2},\ldots,\pi_{m1},\ldots,\pi_{mk_m}$ er Fishers informationsrum $R(M)$ et mq-dimensionalt vektorrum, der er udspændt af søjlerne i en $(\sum_{i=1}^{m}k_i)\times(mq)$ blokinddelt matrix af typen

$$M=\begin{pmatrix} M_{11} & 0 & \ldots & 0 \\ 0 & M_{22} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & M_{mm} \end{pmatrix},$$

hvor M_{ii} er en $k_i\times q$ matrix, hvis (r,s)'te element er $\pi_{ir}^{-\frac{1}{2}}\partial\pi_{ir}/\delta\theta_{is}$. Hvis homogenitetshypotesen er rigtig, er Fishers informationsrum $R(\overline{M})$ et q-dimensionalt vektorrum udspændt af $(\sum_{i=1}^{m}k_i)\times q$ matricen

$$\overline{M} = \begin{pmatrix} M_{11} & & & \\ & M_{22} & & \\ & & \ddots & \\ & & & M_{mm} \end{pmatrix}.$$

Lad PY og $\overline{P}Y$ være projektionen af Karl Pearsons $(\sum_{i=1}^{m} k_i)$-vektor på henholdsvis $R(M)$ og $R(\overline{M})$, og lad $\hat{\theta}$ og θ^* være efficiente estimatorer for henholdsvis mq-parametervektoren θ á priori og q-parametervektoren τ under hypotesen. Den (i,j)'te koordinat i PY og $\overline{P}Y$ er asymptotisk ækvivalent med henholdsvis

$$\frac{n_{i.}\pi_{ij}(\hat{\theta}) - n_{i.}\pi_{ij}(\theta)}{\sqrt{n_{i.}\pi_{ij}(\theta)}}$$

og

$$\frac{n_{i.}\pi_{ij}(\theta^*) - n_{i.}\pi_{ij}(\tau)}{\sqrt{n_{i.}\pi_{ij}(\tau)}},$$

hvor $n_{i.} = n_{i1} + \ldots + n_{ik_i}$. Teststørrelsen for modelkontrol

$$\sum_{i=1}^{m}\sum_{j=1}^{k_i} \frac{(n_{ij} - n_{i.}\pi_{ij}(\hat{\theta}))^2}{n_{i.}\pi_{ij}(\hat{\theta})}$$

er under modellen $\pi_{ij} = \pi_{ij}(\theta)$ asymptotisk fordelt som $(Y - PY)^2$, altså som χ^2 med $\sum_{i=1}^{m} k_i - m - mq$ frihedsgrader, medens teststørrelsen

$$\sum_{i=1}^{m}\sum_{j=1}^{k_i} \frac{(n_{i.}\pi_{ij}(\hat{\theta}) - n_{i.}\pi_{ij}(\theta^*))^2}{n_{i.}\pi_{ij}(\theta^*)}$$

under homogenitetshypotesen asymptotisk er fordelt som $(PY - \overline{P}Y)^2$, d.v.s. som χ^2 med $mq - q$ frihedsgrader.

LITTERATUR

1. A. Holst Andersen (1969): »Asymptotiske resultater for exponentielle familier«. Statistiske Interna No. 9, Matematisk Institut, Afdeling for statistik, Aarhus Universitet, og »Asymptotic results for exponential families«. 37th session of the International Statistical Institute, Contributed papers, 259–260, London, 1969.
2. M. W. Birch (1964): »A new proof of the Pearson-Fisher Theorem«. Ann. Math. Statist., 16, 817–824.
3. W. G. Cochran (1954): »Some methods for strengthening the common χ^2 tests«. Biometrics, 10, 417–451.
4. H. Cramér (1945): »Mathematical methods of statistics«, Princeton.
5. R. A. Fisher (1928): »On a property connecting the χ^2 measure of discrepancy with the method of maximum likelihood«. Atti del Congresso Internazionale dei Mathematici, Bologna, 6, 95–100. Genoptrykt i »Contributions to mathematical statistics«, by R. A. Fisher (1950), Wiley, New York.
6. C. R. Rao (1965): »Linear statistical inference and its applications«, Wiley, New York.
7. C. R. Rao (1961): »A study of large sample test criteria through properties of efficient estimates«. Sankhya, A 23, 25–40.