7019 NL

MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM-O.

rangcorr.

Rapport S 223 (V16)
(ZW 1957-018)

Over Ehrenberg's toets voor het probleem van m rangschikkingen

door Ph. van Elteren;

(voordracht in de serie "Actualiteiten" van het Mathematisch centrum op Zaterdag 26 oktober 1957)

1. Inleiding.

Stel dat m waarnemers onderling onafhankelijk ieder één waarneming verrichten aan elk van n objecten. De waarneming van de α^e waarnemer aan het i^e object zij $x_i^{(\alpha)}$. Wij onderstellen, dat $x_i^{(\alpha)}$ een waarneming is van een stochastische variabele $\underline{x}_i^{(\alpha)}$ met verdelingsfunctie $F_i^{(\alpha)}(x)$. Wij beschouwen hier toetsen voor de hypothese:

$$F_1^{(\alpha)}(x) = F_2^{(\alpha)}(x) = \ldots = F_n^{(\alpha)}(x) = F^{(\alpha)}(x)$$

tegen alternatieve hypothesen:

$$(1) \qquad F_i^{(\alpha)}(x) = F^{(\alpha)}(x+\theta_i) \quad , \quad \left(\sum_i \theta_i = 0\right) ,$$

waarbij tenminste één $\theta_i \neq 0$.

In het bijzondere geval:

$$F_i^{(\alpha)}(x) = F(x+\theta_i+\mu^{(\alpha)})$$

waarbij $F(x)$ de verdelingsfunctie van een normale verdeling is, kan hiervoor een standaardtoets uit de variantieanalyse worden gebruikt. Bij benadering kan deze toets ook worden toegepast in het geval $F(x)$ een verdelingsfunctie van een ander type is, doch essentieel is de voorwaarde dat de verdelingsfunctie niet mag afhangen van α. Om deze moeilijkheden te vermijden heeft FRIEDMAN (1937) een toets opgesteld, welke berust op rangschikking van de grootheden $x_1^{(\alpha)}, x_2^{(\alpha)}, \ldots, x_n^{(\alpha)}$ naar opklimmende grootte voor iedere α. De toetsingsgrootheid is invariant tegen monotone transformaties van de waarnemingen en de toets kan dus ook worden toegepast, indien de grootheden $x_1^{(\alpha)}, \ldots, x_n^{(\alpha)}$ niet gemeten, doch alleen gerangschikt kunnen worden. Vandaar de naam: methode van m rangschikkingen.

KENDALL (1948) heeft aangetoond, dat de toetsingsgrootheid van Friedman een lineaire functie is van de som van de rangcorrelatie coëfficiënten van SPEARMAN (1904), berekend voor alle $\binom{m}{2}$ paren vectoren $x^{(\alpha)} = (x_1^{(\alpha)}, \ldots, x_n^{(\alpha)})$ en $x^{(\beta)} = (x_1^{(\beta)}, \ldots, x_n^{(\beta)})$.
Dit heeft EHRENBERG (1952) er toe gebracht een andere toets te

onderzoeken, welke gebaseerd is op de som van de rangcorrelatie coëfficiënten van KENDALL (1938) (eigenlijk van LIPPS (1905)), van genoemde vectorparen. Onafhankelijk van Ehrenberg, is deze toetsingsgrootheid ook voorgesteld door TERPSTRA (1955-1956), die haar in een meer algemeen geval dan het bovenstaande onderzocht heeft. Hij heeft aangetoond, dat haar verdeling onder de hypothese H_0 voor grote n convergeert naar een normale verdeling. Het bestaan van het artikel van Ehrenberg, was door ons tot voor kort niet opgemerkt. Vandaar het feit dat de toets in de aankondiging van deze voordracht naar Terpstra was genoemd. In deze voordracht zullen wij niet ingaan op de generalisatie van Terpstra; het hoofddoel is hier voor het door Ehrenberg beschouwde geval na te gaan wat er gebeurt met de verdeling van de toetsingsgrootheid onder H_0 als m toeneemt.

2. Definitie van de toetsingsgrootheid van Ehrenberg.

De rangcorrelatiecoëfficiënt van LIPPS-KENDALL voor de vectoren $x^{(\alpha)}$ en $x^{(\beta)}$ is op een factor na gelijk aan:

$$T_{\alpha,\beta} \overset{\text{def}}{=} \sum_{i<j} z_{ij}^{(\alpha)} z_{ij}^{(\beta)} \qquad \begin{pmatrix}\alpha=1,2,\dots,m,\ \beta=1,2,\dots,m;\\ i=1,2,\dots,n,\ j=1,2,\dots,n\end{pmatrix}$$

waarin

$$z_{ij}^{(\alpha)} \overset{\text{def}}{=} \operatorname{sign}\left(x_i^{(\alpha)} - x_j^{(\alpha)}\right).$$

Op dezelfde constante factor na, wordt EHRENBERG's toetsingsgrootheid gedefinieerd door:

$$(2) \quad T \overset{\text{def}}{=} \sum_{\alpha<\beta} T_{\alpha,\beta} = \sum_{i<j}\sum_{\alpha<\beta} z_{ij}^{(\alpha)} z_{ij}^{(\beta)} = \frac{1}{2}\sum_{i<j}\Big(\sum_{\alpha} z_{ij}^{(\alpha)}\Big)^2 - \frac{1}{2}\sum_{i<j}\sum_{\alpha}\left(z_{ij}^{(\alpha)}\right)^2.$$

Wij zullen nu aannemen, dat voor iedere α de waarden van de grootheden $x_i^{(\alpha)}$ verschillend zijn. Dit is behoudens een kans 0 het geval, indien de verdelingen $F_i^{(\alpha)}(x)$ continu zijn. Het hier volgende betoog kan met kleine wijzigingen ook gevoerd worden voor het geval, dat er onder de grootheden $x_i^{(\alpha)}$ groepen gelijken optreden; hiervoor wordt verwezen naar een binnenkort te verschijnen publicatie (VAN ELTEREN (1957)).

Onder de gestelde voorwaarde zijn alle $z_{ij}^{(\alpha)} = \pm 1$.
Dus is:

$$(3) \quad T = \frac{m}{2}\sum_{i<j} z_{ij}^2 - \frac{1}{4} m\, n(n-1),$$

als z_{ij} gedefinieerd wordt door:

$$z_{i,j} = m^{-\frac{1}{2}} \sum_{\alpha} z_{i,j}^{(\alpha)}$$

De grootheid $z_{i,j}$ is behoudens de factor $m^{-\frac{1}{2}}$ de toetsingsgrootheid van de tekentoets toegepast op de verschillen $x_i^{(\alpha)} - x_j^{(\alpha)}$. Uit (3) volgt, dat men de toetsingsgrootheid van Ehrenberg kan berekenen door $\binom{n}{2}$ maal de tekentoets toe te passen in plaats van $\binom{m}{2}$ maal de rangcorrelatietoets van LIPPS-KENDALL.

Onder de in de inleiding genoemde alternatieve hypothese zal T in het algemeen grotere waarden aannemen, dan onder de hypothese H_0. Bij een onbetrouwbaarheidsdrempel ε zullen wij dus een kritiek gebied van het type $T \geq T_\varepsilon$ kiezen, waarbij T_ε de kleinste waarde is, die door T kan worden aangenomen, waarvoor geldt: $P[T \geq T_\varepsilon \mid H_0] \leq \varepsilon$.

Voor de bepaling van T_ε dienen wij de kansverdeling van T onder de hypothese H_0 althans bij benadering te kennen. Aangezien H_0 de hypothese H_0' impliceert, inhoudende dat voor iedere α alle permutaties van de gevonden waarden $x_1^{(\alpha)}, \ldots, x_n^{(\alpha)}$ over de n objecten even waarschijnlijk en deze m permutaties onafhankelijk zijn, kan men volstaan met het onderzoek van de verdeling van T onder H_0'.

Voor dit onderzoek maken we gebruik van de eenvoudige identiteit:

(4) $$\sum_{i<j} z_{i,j}^2 = n^{-1}\Big(\sum_i z_i^2 + \sum_{j<k<l} z_{j,k,l}^2\Big)$$

waarin

$$z_i \stackrel{\text{def}}{=} \sum_j z_{i,j}$$

en

$$z_{j,k,l} \stackrel{\text{def}}{=} z_{j,k} + z_{k,l} + z_{l,j} \, .$$

In de uitdrukking (4) is $\sum_i z_i^2$ op een constante factor na de toetsingsgrootheid van FRIEDMAN (1937).

3. Momenten van de variabelen z_i en $z_{j,k,l}$ onder H_0'.

Men bewijst gemakkelijk:

$$\mathcal{E} z_{i,j} = m^{-\frac{1}{2}} \sum_{\alpha} \mathcal{E} z_{i,j}^{(\alpha)} = 0 \, ,$$

$$\mathcal{E} z_{i,j}^2 = m^{-1} \sum_{\alpha} \mathcal{E} (z_{i,j}^{(\alpha)})^2 = 1 \, ,$$

$$\mathcal{E} z_{i,j} z_{i,l} = m^{-1} \sum_{\alpha} \mathcal{E} z_{i,j}^{(\alpha)} z_{i,l}^{(\alpha)} = \tfrac{1}{3} \, , \quad (i,j,l) \neq \, ,$$

$$\mathcal{E} z_{i,j} z_{k,l} = 0 \qquad (i,j,k,l) \neq \, .$$

Daaruit volgt:

(5) $\mathcal{E}\underline{z}_i = \mathcal{E}\underline{z}_{j,k,l} = 0$

(6) $\mathcal{E}\underline{z}_i\underline{z}_{j,k,l} = 0$ voor iedere i, j, k en l,

$$\mathcal{E}\underline{z}_i^2 = \mathcal{E}(\sum_j \underline{z}_{i,j})^2 = \sum_j \mathcal{E}\underline{z}_{i,j}^2 + 2\sum_{j<l}\mathcal{E}\underline{z}_{i,j}\underline{z}_{i,l} = (n-1) + \tfrac{1}{3}(n-1)(n-2) = \tfrac{1}{3}(n-1)(n+1)$$

en voor $i \neq k$

$$\mathcal{E}\underline{z}_i\underline{z}_k = \mathcal{E}(\sum_j \underline{z}_{i,j})(\sum_l \underline{z}_{k,l}) = -\mathcal{E}\underline{z}_{i,k}^2 + \sum_{j(j\neq i,j\neq k)}\mathcal{E}\underline{z}_{ij}\underline{z}_{k,j} +$$
$$\sum_{l(\neq i,\neq k)}\mathcal{E}\underline{z}_{i,k}\underline{z}_{k,l} + \sum_{j(\neq i,\neq k)}\mathcal{E}\underline{z}_{i,j}\underline{z}_{k,i} = -1 + \tfrac{1}{3}(n-2) - \tfrac{2}{3}(n-2) = -\tfrac{1}{3}(n+1).$$

Dus wordt de covariantiematrix van de variabelen $\underline{z}_i$ gegeven door:

(7) $\mathcal{E}\underline{z}_i\underline{z}_j = \frac{1}{3}(n+1)(n\,\delta_{i,j} - 1)$

met $\delta_{i,j} \overset{\text{def}}{=} \begin{cases} 0 & \text{als } i \neq j \\ 1 & \text{als } i = j \end{cases}$.

Nu is: $\sum_i \underline{z}_i = \sum_{i,j}\underline{z}_{i,j} = 0$

dus is:

$$n^{-1}\sum_{i=1}^{n}\underline{z}_i^2 = n^{-1}\left\{\sum_{i=1}^{n-1}\underline{z}_i^2 + \left(\sum_{i=1}^{n-1}\underline{z}_i\right)^2\right\} =$$
$$= n^{-1}\left\{2\sum_{i=1}^{n-1}\underline{z}_i^2 + \sum_{i=1}^{n-1}\sum_{j=1(\neq i)}^{n-1}\underline{z}_i\underline{z}_j\right\}.$$

Dit is een kwadratische vorm met matrix $\mathcal{A} = \| a_{i,j}\|$ gedefinieerd door:

(8) $a_{i,j} \overset{\text{def}}{=} n^{-1}(1 + \delta_{i,j})$.

Men toont gemakkelijk aan, dat het product van deze matrix en de matrix (7) met $i = 1,2\ldots.n-1$; $j = 1,2,\ldots.n-1$ een diagonaal matrix is, met diagonaal elementen: $\frac{1}{3}(n+1)$. Als $\underline{u}$ dus de vector $(\underline{z}_1, \underline{z}_2, \ldots, \underline{z}_{n-1})$ voorstelt, is:

(9) $\mathcal{A}\,\mathcal{E}\underline{u}\,\underline{u}' = \frac{1}{3}(n+1)\,\mathcal{I}_{n-1}$,

waarin $\mathcal{I}_\nu$ de eenheidsmatrix met ν rijen en kolommen voorstelt.

De covariantiematrix van de variabelen $\underline{z}_{j,k,l}$ wordt gegeven door:

$$\mathcal{E}\underline{z}_{j,k,l}^2 = \mathcal{E}(\underline{z}_{j,k} + \underline{z}_{k,l} + \underline{z}_{l,j})^2 = 3\,\mathcal{E}\underline{z}_{j,k}^2 + 6\,\mathcal{E}\underline{z}_{j,k}\underline{z}_{k,l} = 3 - \tfrac{6}{3} = 1$$
$$\mathcal{E}\underline{z}_{i,j,k}\underline{z}_{i,j,l} = \mathcal{E}(\underline{z}_{i,j} + \underline{z}_{j,k} + \underline{z}_{k,i})(\underline{z}_{i,j} + \underline{z}_{j,l} + \underline{z}_{l,i}) =$$
$$= \mathcal{E}\underline{z}_{i,j}^2 + 4\,\mathcal{E}\underline{z}_{i,j}\underline{z}_{j,l} + 2\,\mathcal{E}\underline{z}_{j,k}\underline{z}_{j,l} = 1 - \tfrac{4}{3} + \tfrac{2}{3} = \tfrac{1}{3}.$$

De covariantie van twee variabelen $\underline{z}_{i,j,k}$ is 0, als zij minder dan twee indices hetzelfde hebben.

Dus is:

(10) $\mathcal{E}\underline{z}_{i,j,n}\underline{z}_{k,l,n} = \frac{1}{3}(\delta_{i,k} - \delta_{i,l} - \delta_{j,k} + \delta_{j,l} + \delta_{i,k}\delta_{j,l} - \delta_{i,l}\delta_{j,k})$

indien geen der indices i,j,k en l gelijk is aan n.

Nu is:

$$z_{j,k,l} = z_{j,k,n} + z_{j,n,l} + z_{n,k,l}$$

waaruit volgt:

$$n^{-1} \sum_{j<k<l} z^2_{j,k,l} = \frac{n-2}{n} \sum_{j=1}^{n-2} \sum_{k=j+1}^{n-1} z^2_{j,k,n} - n^{-1} \sum_{j=1}^{n-3} \sum_{k=j+1}^{n-1} \sum_{l=j+1(\neq k)}^{n-1} z_{j,k,n}\, z_{j,l,n} .$$

Dit is een kwadratische vorm in de variabelen $z_{i,j,n}$ met $1 \leq i<j \leq n-1$, met matrix B= $b_{(i,j)(k,l)}$ $(1\leq i<j\leq n-1;\ 1\leq k<l\leq n-1)$ gedefinieerd door:

$$b_{(i,j)(k,l)} \stackrel{\text{def}}{=} n^{-1}(-\delta_{i,k}+\delta_{i,l}+\delta_{j,k}-\delta_{j,l})+\delta_{i,k}\,\delta_{j,l}-\delta_{i,l}\,\delta_{j,k} .$$

Men toont nu gemakkelijk aan, dat het product van deze matrix en de matrix gedefinieerd door (10) met de beperking: $1\leq i<j \leq n-1$ een diagonaal matrix is met diagonaal elementen $\frac{1}{3}$. Als $\underline{v}$ dus de vector$(z_{1,2,n}, z_{1,3,n}, \ldots\ldots z_{n-2,n-1,n})$ voorstelt is dus:

$$(11) \quad B\,\mathcal{E}\,\underline{v}\,\underline{v}' = \frac{1}{3}\, \mathcal{I}_{\frac{1}{2}n(n-1)} .$$

4. Limietverdeling van $\sum_{i<j} z^2_{i,j}$.

Wij beschouwen nu nader de in de vorige paragraaf gedefinieerde vectoren $\underline{u}$ en $\underline{v}$. Volgens een bekende stelling uit de matrix-rekening bestaan er vierkante matrices C en D met resp. n-1 en $\frac{1}{2}(n-1)(n-2)$ rijen en kolommen, zodanig dat:[1])

$$\mathcal{E}(C\underline{u})(C\underline{u})' = C(\mathcal{E}\underline{u}\,\underline{u}')C' = \mathcal{I}_{n-1} \quad \text{en}$$

$$\mathcal{E}(D\underline{v})(D\underline{v})' = D(\mathcal{E}\underline{v}\,\underline{v}')D' = \mathcal{I}_{\frac{1}{2}(n-1)(n-2)} .$$

Uit (9) en (11) volgt nu (omdat: $\mathcal{E}\underline{u}\,\underline{u}'=(C'C)^{-1}$):

$$C'C = \frac{3}{n+1} A \quad \text{en}$$

$$D'D = 3B$$

Laat nu $\underline{w}$ de gecombineerde vector$(C\underline{u}, D\underline{v})$ voorstellen met $(n-1)+\frac{1}{2}(n-1)(n-2)=\frac{1}{2}n(n-1)$ componenten. Deze vector heeft onder de hypothese H_0 de volgende eigenschappen:

a. Het is de som van m stochastisch onafhankelijke stochastische vectoren,

b. De verwachtingswaarden van al zijn componenten zijn 0,

---

1) zie o.a. C.R. Rao (1952) p. 18.

c. De waarden door deze componenten aangenomen, liggen in een begrensd interval en dus voldoen zij aan de Liapounoff-voorwaarden voor de geldigheid van de Centrale limietstelling.

d. Hun covariantiematrix is voor iedere m de eenheidsmatrix $\mathcal{I}_{\frac{1}{2}n(n-1)}$.

Hier uit volgt, dat voor $\underline{w}$ de centrale limietstelling voor stochastische vectoren geldt, en dus dat de simultane verdeling van de componenten van $\underline{w}$ voor $m \to \infty$ nadert tot de multimonale verdeling met covariantiematrix $\mathcal{I}_{\frac{1}{2}n(n-1)}$. Dit is de simultane verdeling van $\frac{1}{2}n(n-1)$ onafhankelijke normaal verdeelde variabelen met verwachtingswaarde 0 en variantie 1.

Nu is:

$$\sum_{i=1}^{\frac{1}{2}(n-1)(n-2)} \underline{w}_i^2 = \underline{w}'\underline{w} = (C\underline{u})'(C\underline{u}) + (D\underline{v})'(D\underline{v}) =$$

$$= \underline{u}'C'C\underline{u} + \underline{v}'D'D\underline{v} = \frac{3}{n+1}\,\underline{u}'A\underline{u} + 3\underline{v}'B\underline{v}$$

Beide termen zijn sommen van kwadraten van componenten van $\underline{w}$, dus hun verdelingen naderen tot χ^2-verdelingen met aantallen vrijheidsgraden gelijk aan de aantallen kwadraten die in de sommen voorkomen (resp: n-1 en $\frac{1}{2}(n-1)(n-2)$)

Dus nadert de verdeling van:

$$\underline{Z}_1 \overset{\text{def}}{=} \frac{3}{n+1}\,\underline{u}'A\underline{u} = \frac{3}{n(n+1)} \sum_i \underline{z}_{j,k,\ell}^2$$

tot de verdeling van een χ^2-variabele $\underline{X}_2$ met $\frac{1}{2}(n-1)(n-2)$ vrijheidsgraden, waarbij $\underline{X}_1$ en $\underline{X}_2$ onafhankelijk zijn.

De verdeling van

$$(12)\quad \underline{Z} = 3\sum_{i<j} \underline{z}_{i,j}^2 = 3n^{-1}\Big(\sum_i \underline{z}_i^2 + \sum_{j<k<\ell} \underline{z}_{j,k,\ell}^2\Big) = (n+1)\underline{Z}_1 + \underline{Z}_2$$

nadert dan tot de verdeling van

$$(13)\quad \underline{X} \overset{\text{def}}{=} (n+1)\,\underline{X}_1 + \underline{X}_2\,.$$

5. Nader onderzoek van de verdelingsfunctie van $\underline{X}$.

Het is van speciaal belang de verdeling van $\underline{X}$ te onderzoeken voor kleine waarden van n. Voor grote waarden van n, wordt de verdeling van $\underline{X}$ bij benadering normaal, omdat dit het geval is met $\underline{X}_1$ en $\underline{X}_2$. Dit volgt trouwens ook uit de onderzoekingen van TERPSTRA (1955), die de limietverdeling van $\underline{T}$ voor $n \to \infty$ beschouwd heeft.

De grootheid $\underline{X}$ kan beschouwd worden als een kwadratische vorm in normaal verdeelde grootheden. In de literatuur worden verscheidene methoden gegeven om de verdeling van dergelijke vormen in een reeks te ontwikkelen welke echter in het geval van $\underline{X}$ slecht convergeren. Voor enkel kleine waarden van n, is de ontwikkeling van ROBBINS en PITMAN (1949) te gebruiken. Hierbij wordt de verdelingsfunctie $F(X)$ van $\underline{X}$ ontwikkeld in verdelingsfuncties $F_\nu(X)$ van χ^2-verdelingen met ν vrijheidsgraden:

$$F(X) = \sum_{j=0}^{\infty} K_j F_{\frac{1}{2}n(n-1)+2j}(X),$$

waarin

$$K_j \overset{\text{def}}{=} \frac{(n-3+2)(n-3+4)\ldots(n-3+2j)}{2\cdot 4\cdot\ldots\, 2j}\cdot\left(\frac{n}{n+1}\right)^j (n+1)^{-\frac{1}{2}(n-1)}.$$

Voor oneven waarden van n kan een ontwikkeling in een eindig aantal termen worden gegeven. Wij beperken ons hier tot het geval n=3. Dan is volgens (13):

$$\underline{X} = 4\underline{\chi}_1^2 + \underline{\chi}_2^2$$

waarbij $\underline{\chi}_\nu^2$ een χ^2 variabele met ν vrijheidsgraden voorstelt, waarvan wij de verdelingsdichtheid aanduiden met $f_\nu(x)$ en de verdelingsfunctie (als boven) met $F_\nu(x)$. De verdelingsfunctie van $4\underline{\chi}_1^2$ wordt nu gegeven door:

$$P[4\underline{\chi}_1^2 \leq x] = F_1\left(\frac{x}{4}\right) = 2^{-\frac{1}{2}}\left\{\Gamma\left(\tfrac{1}{2}\right)\right\}^{-1}\int_0^{x/4} e^{-t/2}\, t^{-\frac{1}{2}}\, dt,$$

en de verdelingsdichtheid van $4\underline{\chi}_1^2$ wordt:

$$\frac{d}{dx}F_1\left(\frac{x}{4}\right) = \frac{1}{4}f_1\left(\frac{x}{4}\right) = 2^{-2\frac{1}{2}}\left\{\Gamma\left(\tfrac{1}{2}\right)\right\}^{-1} e^{-\frac{x}{4}}\left(\frac{x}{4}\right)^{-\frac{1}{2}}.$$

Nu geldt voor de verdelingsdichtheid van de som van 2 onafhankelijk verdeelde continue variabelen, die alleen positieve waarden aannemen, met dichtheden $g(u)$ en $h(v)$[2])

$$f(x) = \int_0^x g(x-y)\, h(y)\, dy.$$

Dus in ons geval met: $g(u) = f_1(u) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}v}\, v^{-\frac{1}{2}}$

en: $h(v) = \frac{1}{4}f_2\left(\frac{v}{4}\right) = \frac{1}{8}e^{-\frac{v}{8}}$:

$$f(x) = \frac{1}{8\sqrt{2\pi}}\int_0^x e^{\left\{-\frac{x-y}{2}-\frac{y}{8}\right\}}(x-y)^{-\frac{1}{2}}\, dy,$$

of met: $t = \frac{3}{4}(x-y)$; $dy = -\frac{4}{3}\, dt$:

----

2) Zie b.v. H. Cramer (1946), p 191.

$$f(x)=\frac{1}{8\sqrt{2\pi}}\sqrt{\frac{4}{3}}\int_0^{\frac{3}{4}x} e^{-\frac{t}{2}}\, t^{-\frac{1}{2}}\, dt = $$
$$= \frac{1}{4\sqrt{3}}\, e^{-x/8}\, F_1(\tfrac{3}{4}x).$$

Dan is:

$$F(X)=\int_0^X f(x)\,dx = -\frac{8}{4\sqrt{3}}\int_0^X F_1(\tfrac{3}{4}x)\, de^{-\frac{x}{8}} =$$
$$= -\frac{2}{\sqrt{3}}\left[e^{-\frac{x}{8}} F_1(\tfrac{3}{4}x\right]_0^X + \frac{\sqrt{3}}{2}\int_0^X e^{-\frac{x}{8}} f_1(\tfrac{3}{4}x)\,dx =$$
$$= -\frac{2}{\sqrt{3}}\, e^{-\frac{X}{8}} F_1(\tfrac{3}{4}X) + \frac{1}{\sqrt{2\pi}}\int_0^X e^{-\frac{x}{2}}\, x^{-\frac{1}{2}}\,dx.$$

Dus is:

$$F(X)=F_1(X)-\frac{2}{\sqrt{3}}\, e^{-\frac{X}{8}} F_1(\tfrac{3}{4}X).$$

Voor grote X wordt dit:

$$F(X)\approx 1-\frac{2}{\sqrt{3}}\, e^{-\frac{X}{8}}.$$

Wij hebben dit **resultaat** vergeleken met de exacte verdeling van $\underline{Z}$ voor n=3 en m=3,4,5 en 6. Men vindt b.v.:

m=3	X=	3	11	19	27
$P[\underline{Z}\geq X]$		1	0,53	0,19	0,03
$P[\underline{X}\geq X]$		0,77	0,29	0,11	0,04
m=4	X=	12	24	27	36
$P[\underline{Z}\geq X]$		0,30	0,069	0,042	0,005
$P[\underline{X}\geq X]$		0,26	0,058	0,040	0,013
m=5	X=	11,4	25,8	30,6	35,4
$P[\underline{Z}\geq X]$		0,38	0,047	0,024	0,009
$P[\underline{X}\geq X]$		0,28	0,046	0,025	0,014
m=6	X=	12	24	28	36
$P[\underline{Z}\geq X]$		0,31	0,061	0,037	0,008
$P[\underline{X}\geq X]$		0,26	0,058	0,035	0,013

Men, ziet, dat de benadering van de verdeling van $\underline{Z}$ door die van $\underline{X}$ inderdaad beter wordt bij toenemende m. Verder blijkt de benadering het beste te zijn bij overschrijdingskansen in de buurt van 0,05 (X in de buurt van 25).

6. Enkele opmerkingen over het verband tussen de toetsen van Friedman en Ehrenberg.

De grootheid $\underline{Z}$ in (12) is een lineaire transformatie van de toetsingsgrootheid van EHRENBERG; $\underline{Z}_1$ een dergelijke transformatie van de grootheid van FRIEDMAN.

Uit de definitie van $\underline{Z}_2$ volgt dat deze grootheid voor n=2 identiek is aan 0. De toetsen van EHRENBERG en FRIEDMAN zijn dus equivalent voor n=2 en komen dan overeen met de bekende tekentoets. Zij zijn eveneens equivalent, indien de waarnemingen groepen gelijk vertonen en wel zo, dat de grootheden $x_1^{(\alpha)},\dots,x_n^{(\alpha)}$ voor iedere α niet meer dan 2 verschillende waarden aannemen. Bij benadering geldt dit ook voor grote n, immers de bijdrage van $\underline{Z}_2$ tot $\underline{Z}$ gaat voor grote n naar 0 (de variantie van $\underline{Z}_2$ is van de orde: n^2, die van $(n+1)\ \underline{Z}_1$ van de orde n^3).

De asymptotische relatieve doeltreffendheid van de toets van Friedman ten opzichte van de corresponderende variantie-analysetoets is onderzocht door van door VAN ELTEREN en NOETHER (1957). Het gaat hierbij om de limiet voor $m\to\infty$ van de verhouding van de waarde van m nodig voor de variantie analyse toets en de waarde m vereist voor de toets van Friedman, om bij dezelfde onbetrouwbaarheidsdrempel eenzelfde onderscheidingsvermogen te bereiken ten opzichte van alternatieven van de gedaante:

$$\theta_i = \theta_{i,m} = \delta_i/\sqrt{m},$$

waarin de δ_i gegeven constanten zijn met $\sum_i \delta_i = 0$.
Genoemde limiet blijkt te worden:

$$e = \frac{3}{\pi}\cdot\frac{n}{n+1},$$

als $F(x)$ (zie par.1) de verdelingsfunctie van een normale verdeling is.

Bij n=2 geldt $e=\frac{2}{\pi}$, overeenkomend met de aymptotische doeltreffendheid van de tekentoets. Voor grote n wordt $e\approx\frac{3}{\pi}$ en dit geldt dan zowel voor de toets van Friedman als voor de toets van Ehrenberg. Een onderzoek is nog gaande naar de doeltreffendheid van Ehrenberg's toets voor kleine $n\geqq 3$; deze hangt in ieder geval af van de gebruikte onbetrouwbaarheidsdrempel en van de waarden der δ_i.

Een ander belangrijk vergelijkingsobject vormen de bruikbaarheidseigenschappen van beide toetsen. Men stelt zich dan de vraag bij welke alternatieve hypothesen de toetsen met zekerheid tot verwerping van H_0 gaan leiden als m toeneemt. Het blijkt, dat de verzameling van alternatieve hypothesen, waarvoor de toetsen bruikbaar zijn, uitgebreider is, dan de door (1) gegevene. Indien wij ons beperken tot continu verdeelde grootheden en de vectoren $\underline{x}^{(\alpha)}$ statistisch onafhankelijk onderstellen vinden wij, dat het bruikbaarheidsgebied van de toets van EHRENBERG bestaat uit alle alternatieve hypothesen waarvoor

$$\pi_{ij} \overset{\text{def}}{=} \lim_{m\to\infty} \frac{1}{m} \sum_{\alpha} P\left[\underline{x}_i^{(\alpha)} < \underline{x}_j^{(\alpha)}\right] \neq \frac{1}{2}$$

voor tenminste één paar (i,j)$(1 \leq i < j \leq n)$.

Bij de toets van Friedman wordt dit gebied gegeven door alle alternatieven waarvoor:

$$\sum_{\substack{j=1 \\ j\neq i}}^{n} \pi_{ij} \neq \frac{1}{2}(n-1)$$

voor tenminste één i.

Het bruikbaarheidsgebied van de toets van Ehrenberg bevat dat van de toets van Friedman, doch niet omgekeerd. Bij n=3 behoort bijvoorbeeld het geval:

$$\pi_{1,2} = \pi_{2,3} = \pi_{3,1} \neq \frac{1}{2}$$

tot het bruikbaarheidsgebied van de toets van Ehrenberg, doch niet tot dat van de toets van Friedman. Dit geval doet zich bijvoorbeeld voor, indien voor iedere α geldt:

$$P\left[\underline{x}_1^{(\alpha)} < \underline{x}_2^{(\alpha)} < \underline{x}_3^{(\alpha)}\right] = P\left[\underline{x}_2^{(\alpha)} < \underline{x}_3^{(\alpha)} < \underline{x}_1^{(\alpha)}\right] =$$
$$= P\left[\underline{x}_3^{(\alpha)} < \underline{x}_1^{(\alpha)} < \underline{x}_2^{(\alpha)}\right] = \frac{1}{3}.$$

Men kan dit grotere bruikbaarheidsgebied van de toets van Ehrenberg niet zonder meer als een voordeel voor deze toets beschouwen, in het bijzonder niet indien men de toets wenst te gebruiken voor de modellen in de inleiding beschreven.

Concluderend kan men zeggen, dat de toetsen van FRIEDMAN en EHRENBERG elkaar weinig ontlopen, wat betreft hun eigenschappen onder alternatieve hypothesen. Om practische redenen zal de toets van FRIEDMAN, omdat zij een eenvoudiger te berekenen toetsingsgrootheid bezit, waarvan de verdeling onder de hypothese H_0 gemakkelijker te benaderen is, althans voor kleine waarden van n en grote van m, de voorkeur verdienen.

Literatuur.

Cramèr, H. (1946), Mathematical Methods of Statistics, Princeton University Press, Princeton.

Ehrenberg, A.S.C. (1952), On sampling from a population of rankers, Biometrika 39, 82-87.

Elteren, Ph. van (1957), The asymptotic distribution for large m of Terpstra's statistic for the problem of m rankings. (wordt gepubliceerd).

Elteren, Ph. van and G.E. Noether (1957), The asymptotic efficiency of the χ^2_r-test for a balanced incomplete block design. (aangeboden voor publicatie).

Friedman, M. (1937), The use of ranks to avoid the asumption of normality implicit in the analysis of variance, Journ. of the American Statistical Ass., 32, 675-699.

Kendall, M.G. (1938), A new measure of rank correlation, Biometrika 30, 81-93.

Kendall, M.G. (1948), Rank correlation methods, Chr. Griffin, London (2e druk, 1955).

Lipps, G.F. (1905), Die Bestimmung der Abhängigkeit zwischen den Merkmalen eines Gegenstandes, Berichte über die verhandlungen der Königlichen Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Math. Phys. Klasse 57, 1-32.

Rao, C.R. (1952), Advanced Statistical Methods in Biometric Research, N.J. Wiley, New York; Chapman and Hall, London.

Robbins, H. and E.J.G. Pitman (1949), Application of the method of mixtures to quadratic forms in normal variates, Ann. of Math. Stat. 20, 552-560.

Spearman, C. (1904), The proof and measurements of association between two things, Am J. Psych. 15, 72-101.

Terpstra, T.J. (1955,1956), A generalization of Kendall's rank correlation statistic,

I Proc. Kon. Ned. Ak. van Wet. A 58 / Indagationes Math. 17 } 690-696,

II Proc. Kon. Ned. Ak. van Wet. A 59 / Indagationes Math. 18 } 59-66.