"(k-1)-mean significance levels" van de asymptotische versie van de methode voor simultane uitspraken voor het k-steekproevenprobleem gebaseerd op de toets van Kruskal en Wallis

TECHNISCHE HOGESCHOOL EINDHOVEN

Onderafdeling der Wiskunde

SECTIE KANSREKENING, STATISTIEK EN OPERATIONS RESEARCH

Memorandum COSOR 76-27

"(k-1)-mean significance levels" van de asymptotische versie van de methode voor simultane uitspraken voor het k-steekproeven-probleem gebaseerd op de toets van Kruskal en Wallis

door

J.H. Oude Voshaar

Eindhoven, December 1976

Nederland

"(k-1)-mean significance levels" van de asymptotische versie van de methode voor simultane uitspraken voor het k-steekproevenprobleem gebaseerd op de toets van Kruskal en Wallis

door

J.H. Oude Voshaar

1. Inleiding en samenvatting

Laat $\underline{y}_{11},\ldots\ldots,\underline{y}_{1n_1};\ldots\ldots;\underline{y}_{k1},\ldots\ldots,\underline{y}_{kn_k}$ onafhankelijke stochastische grootheden zijn, waarbij $\underline{y}_{ij}$ een kontinue verdelingsfunktie F_i heeft ($k \geq 3$).
Een verdelingsvrije toets voor de nulhypothese H_0: $F_1 = F_2 = \ldots = F_k$ is de toets van Kruskal en Wallis met als toetingsgrootheid:

$$\underline{H} := \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left(\bar{\underline{R}}_{i\cdot} - \bar{\underline{R}}_{\cdot\cdot}\right) = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \bar{\underline{R}}_{i\cdot}^2 - 3N(N+1)$$

waarbij $N := \sum_{i=1}^{k} n_i$ en $\underline{R}_{ij}$ het rangnummer van $\underline{y}_{ij}$ onder alle waarnemingen is. Verder:

$$\bar{\underline{R}}_{i\cdot} := \frac{1}{n_i} \sum_{j=1}^{n_i} \underline{R}_{ij} \quad \text{en} \quad \bar{\underline{R}}_{\cdot\cdot} := \frac{1}{N} \sum_{i,j} \underline{R}_{ij} = \frac{N+1}{2}$$

Bij de onbetrouwbaarheidsdrempel α verwerpen we H_0 als $H > h_\alpha$.
Als H_0 verworpen wordt weten we hiermee echter nog niet, voor welke paren (i,j) F_i en F_j verschillend zijn. Maar gebruik makend van het

lemma:

als $c > 0$, dan geldt:

$$\sum_{i=1}^{k} y_i^2 \leq c^2 \quad \text{d.e.s.d.} \left| \sum_{i=1}^{k} a_i y_i \right| \leq c\left(\sum_{i=1}^{k} a_i^2\right)^{\frac{1}{2}} \text{ voor alle } (a_1,\ldots a_k) \in R^k$$

vinden we dat uit $P_{H_0}\{\underline{H} \leq h_\alpha\} \geq 1-\alpha$ volgt:

$$(1.1) \qquad P_{H_0}\left\{\left|\bar{\underline{R}}_{i\cdot} - \bar{\underline{R}}_{j\cdot}\right| \leq \sqrt{h_\alpha \frac{N(N+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \text{ voor alle i en j}\right\} \geq 1-\alpha$$

We konkluderen dus dat $F_i \neq F_j$ als geldt:

$$|\bar{R}_{i.} - \bar{R}_{j.}| > \sqrt{h_\alpha \frac{N(N+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

en deze methode heeft (onder H_0) een simultane onbetrouwbaarheid kleiner of gelijk aan α.

Opmerking: Deze methode is afkomstig van Nemenyi (1963).

Het nadeel van deze methode is echter, dat de simultane onbetrouwbaarheid, vooral voor grotere waarden van k, veel kleiner is dan α (zie tabel 4.1), omdat in (1.1) niet alle lineaire kombinaties uit het lemma gebruikt worden.
In het geval van steekproeven van gelijke omvang, d.w.z. $n_1 = n_2 = \ldots = n_k = n$, is, voor grotere n, de methode te verbeteren tot:

(1.2) konkludeer $F_i \neq F_j$ als $|\bar{R}_{i.} - \bar{R}_{j.}| > q^{\alpha}_{k,\infty}\sqrt{\frac{k(kn+1)}{12}}$

waarin $q^{\alpha}_{k,\infty}$ de rechterkritieke waarde is van de verdeling van de range van k standaard-normaal verdeelde grootheden (een grootheid met deze verdeling noteren we wel eens als $q_{k,\infty}$).

Deze methode (1.2) heeft voor $n \to \infty$ een simultane onbetrouwbaarheid die exakt gelijk is aan α (zie 4.2).
We zijn echter niet alleen in deze onbetrouwbaarheid onder H_0 geïnteresseerd: stel dat van de verdelingsfunkties $F_1,\ldots,F_k$ er p gelijk zijn en de overige k-p verschillen. Dan gaat onze belangstelling nu vooral uit naar de kans dat uit die p identieke verdelingsfunkties toch sommige verschillend verklaard worden.
Miller (1966) noemt deze kans, gemaximaliseerd over alle mogelijke verdelingsfunkties $F_1,\ldots,F_k$ waarvan er p gelijk zijn, de "p-mean significance level α_p".
Omdat $\bar{R}_{i.}$ diskreet verdeeld is, is het te moeilijk om α_p te bepalen voor de methode gebaseerd op formule (1.1).
Voor methode (1.2) is het wel mogelijk α_p te bepalen als $n \to \infty$. We zullen ons hier beperken tot het geval $p = k-1$, d.w.z. er zijn kontinue verdelingsfunkties F en G zodat geldt: $F_1 = \ldots = F_{k-1} = F$ en $F_k = G$ en we vragen ons af of hierbij (voor $n \to \infty$) geldt:

$$\alpha_{k-1} \leq \alpha$$

waarin

$$\alpha_{k-1} := \sup_{F,G \text{ kontinu}} P\left\{ \max_{1\leq i,j\leq k-1} |\bar{R}_{i\cdot} - \bar{R}_{j\cdot}| > q^{\alpha}_{k,\infty}\sqrt{\frac{k(kn+1)}{12}} \right\}$$

We berekenen hiertoe eerst (in hoofdstuk 2) de verwachting en variantie van $\bar{R}_{i\cdot}$ en $\bar{R}_{j\cdot}$, zodat we, gebruik makend van de in hoofdstuk 3 te bewijzen asymptotische normaliteit van de vektor $(\bar{R}_{1\cdot}, \ldots, \bar{R}_{k-1\cdot})$, de asymptotische verdeling van $\max_{1\leq i,j\leq k-1} |\bar{R}_{i\cdot} - \bar{R}_{j\cdot}|$ kennen (hoofdstuk 4).

Als we, behalve kontinuïteit, geen andere beperkingen aan F en G opleggen, dan zal blijken dat voor de gangbare waarden van α geldt: $\alpha_{k-1} > \alpha$ (zie hoofdstuk 5). Beperken we ons echter tot de klasse van Lehmann-alternatieven ($G = F^t$ voor zekere $t.\epsilon(0,\infty)$, dan geldt wel: $\alpha_{k-1} < \alpha$ (hoofdstuk 6).
Voor verschuivingsalternatieven zullen we in hoofdstuk 7 vinden:
Als F symmetrisch en unimodaal (= ééntoppig) is, dan geldt: $\alpha_{k-1} < \alpha$.

2. Verwachting, variantie en kovariantie van $\bar{R}_{1\cdot}, \ldots, \bar{R}_{k-1\cdot}$

We veronderstellen dus dat $F_1 = F_2 = \ldots = F_{k-1} = F$ en $F_k = G$ (F en G kontinu) en verder $n_1 = \ldots = n_k = n$.
Dan zijn $\bar{R}_{1\cdot}, \ldots, \bar{R}_{k-1\cdot}$ identiek verdeeld.
We definiëren de funktie u door:

$$u(z) := \begin{cases} 0 \text{ als } z < 0 \\ 1 \text{ als } z \geq 0 \end{cases}$$

dan

$$R_{1\cdot} := \sum_{i=1}^{n} R_{1i} = \sum_{i=1}^{n} \sum_{j=2}^{k} \sum_{\ell=1}^{n} u(y_{1i} - y_{j\ell}) + \frac{1}{2} n(n+1)$$

Nu geldt dat:

$$\mathcal{E}[u(y_{1i} - y_{j\ell})] = \begin{cases} \frac{1}{2} \text{ als } j = 2,\ldots,k-1 \\ p \text{ als } j = k \end{cases}$$

(2.1) waarin $p := P(y_{k1} \leq y_{11}) = \int G(y)dF(y)$

zodat:

$$\mathcal{E}(R_1) = \frac{1}{2}n \sum_{j=2}^{k-1} n + pn^2 + \frac{1}{2}n(n+1)$$

en dus $\mathcal{E}(\bar{R}_{1.}) = \frac{1}{2}(kn+1) + (p-\frac{1}{2})n$

$$(= \tfrac{1}{2}(N+1) + (p-\tfrac{1}{2})n)$$

Voor de berekening van var$(\bar{R}_{1.})$ vatten we de waarnemingen op als 3 steekproeven:

$$Y_{11}, \cdots, Y_{1n}; Y_{21}, \cdots\cdots, Y_{2,(k-2)n}; Y_{k1}, \cdots, Y_{kn}$$

met verdelingsfunkties F, F en G.
Dan geldt nu:

$$R_{1.} = \sum_{i=1}^{n} \sum_{j=1}^{(k-2)n} u(Y_{1i}-Y_{2j}) + \sum_{i=1}^{n} \sum_{\ell=1}^{n} (Y_{1i}-Y_{k\ell}) + \tfrac{1}{2}n(n+1)$$

Vanwege de onafhankelijkheid hebben we:

als $i \neq i'$ en $j \neq j'$ dan $\mathcal{E}[u(Y_{1i}-Y_{2j})u(Y_{1i'}-Y_{2j'})] = \frac{1}{4}$

als $i \neq i'$ en $\ell \neq \ell'$ dan $\mathcal{E}[u(Y_{1i}-Y_{k\ell})u(Y_{1i'}-Y_{k\ell'})] = p^2$

en als $i \neq i'$ dan $\mathcal{E}[u(Y_{1i}-Y_{2j})u(Y_{1i'}-Y_{k\ell})] = \frac{1}{2}p$

Verder geldt (omdat $u^2(z) = u(z)$ voor alle z):

$$\mathcal{E}[u^2(Y_{1i}-Y_{2j})] = \frac{1}{2}$$

en $\mathcal{E}[u^2(Y_{1i}-Y_{k\ell})] = P$

Als $j \neq j'$ dan

$$\mathcal{E}[u(Y_{1i}-Y_{2j})u(Y_{1i}-Y_{2j'})] = P(Y_{2j} \leq Y_{1i} \text{ en } Y_{2j'} \leq Y_{1i}) =$$

$$= \int F^2(y)dF(y) = \frac{1}{3} \quad \text{(omdat F kontinu is)}$$

Als $i \neq i'$ dan

$$\mathcal{E}[u(Y_{1i}-Y_{2j})u(Y_{1i'}-Y_{2j})] = \int (1-F(y))^2 \, dF(y) = \frac{1}{3}$$

als $\ell \neq \ell$ dan

$$\mathcal{E}\left[u(\underline{y}_{1i}-\underline{y}_{k\ell})u(\underline{y}_{1i}-\underline{y}_{k\ell'})\right] = \int G^2(y)\, dF(y) = q$$

als $i \neq i'$ dan

$$\mathcal{E}\left[u(\underline{y}_{1i}-\underline{y}_{k\ell})u(\underline{y}_{1i'}-\underline{y}_{k\ell})\right] = \int (1-F(y))^2 dG(y) =$$

$$= \left[(1-F(y))^2 G(y)\right]_{y=\infty}^{y=\infty} + 2\int (1-F(y))G(y)dF(y) =$$

$$= 2\int G(y)dF(y) - 2\int F(y)G(y)dF(y) = 2(p-r)$$

en tot slot geldt dat:

$$\mathcal{E}\left[u(\underline{y}_{1i}-\underline{y}_{2j})u(\underline{y}_{1i}-\underline{y}_{k\ell})\right] = \int F(y)G(y)dF(y) = r$$

Hierbij zijn q en r gedefinieerd door:

(2.2)
$$q := \int G^2(y)dF(y)$$
$$r := \int F(y)G(y)dF(y)$$

We vinden dan:

$$\mathcal{E}(\underline{R}_{1\cdot}^2) = \sum_{i=1}^{n}\sum_{j=1}^{(k-2)n}\sum_{i'=1}^{n}\sum_{j'=1}^{(k-2)n} \mathcal{E}\left[u(\underline{y}_{1i}-\underline{y}_{2j})u(\underline{y}_{1i'}-\underline{y}_{2j'})\right] +$$

$$+ \sum_{i=1}^{n}\sum_{\ell=1}^{n}\sum_{i'=1}^{n}\sum_{\ell'=1}^{n} \mathcal{E}\left[u(\underline{y}_{1i}-\underline{y}_{k\ell})u(\underline{y}_{1i'}-\underline{y}_{k\ell'})\right] +$$

$$+ 2\sum_{i=1}^{n}\sum_{j=1}^{(k-2)n}\sum_{i'=1}^{n}\sum_{\ell=1}^{n} \mathcal{E}\left[u(\underline{y}_{1i}-\underline{y}_{2j})u(\underline{y}_{1i'}-\underline{y}_{k\ell})\right] +$$

$$+ 2\cdot\tfrac{1}{2}n(n+1)\,\mathcal{E}(\underline{R}_{1\cdot}) - \tfrac{1}{4}n^2(n+1)^2 =$$

$$= \tfrac{1}{4}n(n-1)(k-2)n\{(k-2)n-1\} + \tfrac{1}{3}n(k-2)n\{(k-2)n-1\} +$$

$$+ \tfrac{1}{3}n(n-1)(k-2)n + \tfrac{1}{2}n(k-2)n + p^2n(n-1)n(n-1) +$$

$$+ qn^2(n-1) + 2(p-r)n(n-1)n + pn^2 + 2\cdot\tfrac{1}{2}pn(n-1)(k-2)n^2 +$$

$$+ 2\,rn(k-2)n^2 + n(n+1)\{\tfrac{1}{2}n(kn+1) + (p-\tfrac{1}{2})n^2\} - \tfrac{1}{4}n^2(n+1)^2$$

zodat:

$$(2.3)\quad \mathrm{var}(\bar{\underline{R}}_{1\cdot}) = \frac{1}{n^2}\mathcal{E}(\underline{R}_{1\cdot}^2) - (\mathcal{E}\,\bar{\underline{R}}_{1\cdot})^2 =$$

$$= \frac{1}{12}k^2n + (2r - p - \tfrac{1}{4})kn + (4p - 2p^2 + q - 6r + \tfrac{1}{6})n +$$

$$+ \frac{1}{12}k - p + p^2 - q + 2r - \frac{1}{6}$$

Teneinde de kovariantie van $\bar{\underline{R}}_{1\cdot}$ en $\bar{\underline{R}}_{2\cdot}$ te berekenen vatten we de waarnemingen op als 4 steekproeven:

$$\underline{y}_{11},\cdots,\underline{y}_{1n};\underline{y}_{21},\cdots,\underline{y}_{2n};\underline{y}_{31},\cdots,\underline{y}_{3(k-3)n};\underline{y}_{k1},\cdots,\underline{y}_{kn}$$

met verdelingsfunkties F, F, F en G.

Dan geldt:

$$\underline{R}_{1\cdot} = \sum_{i=1}^{n}\sum_{j=1}^{n} u(\underline{y}_{1i}-\underline{y}_{2j}) + \sum_{i=1}^{n}\sum_{\ell=1}^{(k-3)n} u(\underline{y}_{1i}-\underline{y}_{3\ell})$$

$$+ \sum_{i=1}^{n}\sum_{m=1}^{n} u(\underline{y}_{1i} - \underline{y}_{km}) + \frac{1}{2}n(n+1)$$

en $$\underline{R}_{2\cdot} = \sum_{j=1}^{n}\sum_{i=1}^{n} u(\underline{y}_{2j}-\underline{y}_{1i}) + \sum_{j=1}^{n}\sum_{\ell=1}^{(k-3)n} u(\underline{y}_{2j}-\underline{y}_{3\ell})$$

$$+ \sum_{j=1}^{n}\sum_{m=1}^{n} u(\underline{y}_{2j}-\underline{y}_{km}) + \frac{1}{2}n(n+1)$$

Door $\underline{R}_{1\cdot}.\underline{R}_{2\cdot}$ op dezelfde manier als bij de berekening van $\mathcal{E}(\underline{R}_{1\cdot}^2)$ term voor term uit te vermenigvuldigen en de verwachting te nemen, vinden we (hierbij gebruik makend van de definities van p,q en r):

$$(2.4)\quad \mathrm{cov}(\bar{\underline{R}}_{1\cdot},\bar{\underline{R}}_{2\cdot}) = \frac{1}{n^2}\mathcal{E}(\underline{R}_{1\cdot}.\underline{R}_{2\cdot}) - (\mathcal{E}\,\bar{\underline{R}}_{1\cdot})(\mathcal{E}\,\bar{\underline{R}}_{2\cdot}) =$$

$$= -\frac{1}{12}kn + (3p - p^2 - 4r + \tfrac{1}{12})n - \frac{1}{12}$$

Opmerking 1:

Onder H_0 (d.w.z. F = G) geldt:

$$p = \frac{1}{2} \text{ en } q = r = \frac{1}{3}$$

zodat dan

$$\text{var}(\underline{\bar{R}}_{1\cdot}) = \frac{1}{12}k(k-1)n + \frac{1}{12}(k-1)$$

$$\text{en } \text{cov}(\underline{\bar{R}}_{1\cdot},\underline{\bar{R}}_{2\cdot}) = -\frac{1}{12}(kn+1)$$

hetgeen overeenkomt met de bekende uitdrukkingen hiervoor (zie bijvoorbeeld Miller, pag. 171).

Opmerking 2:

Voor $n_1,\ldots,n_k$ niet alle gelijk vindt men op dezelfde wijze:

$$\text{var}\,\underline{\bar{R}}_{1\cdot} = \frac{1}{n_1}\{\frac{1}{12}M^2 + (q-p^2)n_k^2 + (2r-p)Mn_k + \frac{1}{12}M + (2r-p-q+p^2)n_k\}$$

$$+ \frac{1}{12}M + (2p-p^2-m)n_k$$

waarin $M := \sum_{i=2}^{k-1} n_i = N - n_1 - n_k$

$$\text{en } \text{cov}(R_{1\cdot},\underline{\bar{R}}_{2\cdot}) = -\frac{1}{12}(N+1) + (3p - p^2 - 4r + \frac{1}{12})n_k$$

3. Asymptotische normaliteit van $(\underline{\bar{R}}_{1\cdot}, \underline{\bar{R}}_{2\cdot},\ldots,\underline{\bar{R}}_{k-1,\cdot})$

In hoofdstuk 4 zullen we nodig hebben, dat $(\underline{\bar{R}}_{1\cdot},\ldots,\underline{\bar{R}}_{k-1,\cdot})$ asymptotisch multi-normaal verdeeld is voor $n \to \infty$ Dit is d.e.s.d. het geval als elke lineaire kombinatie $\sum_{i=1}^{k-1} c_i\underline{\bar{R}}_{i\cdot}$ asymptotisch normaal verdeeld is

en dit laatste bewijzen we met stelling 2.1 uit Hajek (1968)..
De voorwaarden van deze stelling zijn hier vervuld d.e.s.d. als
$\text{var}\left(\sum_{i=1}^{k-1} c_i\underline{\bar{R}}_{i\cdot}\right)$ naar oneindig gaat voor $n \to \infty$.
Voordat we deze voorwaarde verifiëren formuleren we echter eerst de volgende lemma's:

lemma 3.1:

$$\text{cov}(\underline{\bar{R}}_{1}\cdot,\underline{\bar{R}}_{2}\cdot) \geq -\frac{1}{k-2}\,\text{var}(\underline{\bar{R}}_{1}\cdot)$$

en het gelijkteken geldt d.e.s.d. als $\text{var}\sum_{i=1}^{k-1} \underline{\bar{R}}_{1\cdot} = 0$

Bewijs: Het gestelde volgt uit:

$$\mathrm{var}\left(\sum_{i=1}^{k-1} \bar{R}_{i.} \right) = (k-1) \{ \mathrm{var}(\bar{R}_{i.}) + (k-2)\, \mathrm{cov}(\bar{R}_{i.}, \bar{R}_{2.}) \} \geq 0$$

Q.E.D.

lemma 3.2:

$$\mathrm{cov}(\bar{R}_{1.}, \bar{R}_{2.}) < 0$$

Bewijs: Onder de voorwaarde $\bar{R}_{k} = r_k$ (d.w.z. $\bar{R}_{k1} = r_{k1}, \ldots, \bar{R}_{kn} = r_{kn}$) geldt dat $\sum_{i=1}^{k-1} \bar{R}_{i.}$ = konstant, dus voor $\mathrm{var}(\sum_{i=1}^{k-1} \bar{R}_{i.} \mid \bar{R}_{k}=r_k) = 0$

lemma 3.1 geeft:

$$\mathrm{cov}(\bar{R}_{1.}, \bar{R}_{2.} \mid R_k=r_k) = -\frac{1}{k-2}\ \mathrm{var}(\bar{R}_{1.} \mid \bar{R}_{k}=r_k) < 0$$

waaruit volgt: $\mathrm{cov}(\bar{R}_{1.}, \bar{R}_{2.}) = \sum_{r_k} \mathrm{cov}(\bar{R}_{1.}, \bar{R}_{2.} \mid \bar{R}_{k}=r_k) \cdot P(\bar{R}_{k}=r_k) < 0$

Q.E.D.

lemma 3.3

$$\sum_{i} \sum_{\neq j} c_i c_j \leq (k-2) \sum_{i=1}^{k-1} c_i^{\,2} \text{ en het gelijkteken}$$

geldt d.e.s.d. als $c_1 = c_2 = \ldots\ldots = c_{k-1}$

Bewijs: De **bewering** volgt uit de ongelijkheid van **Cauchy-Schwarz:**

$$(k-1) \sum_{i=1}^{k-1} c_i^{\,2} \geq \left(\sum_{i=1}^{k-1} c_i \right)^2 = \sum_{i} \sum_{\neq j} c_i c_j + \sum_{i=1}^{k-1} c_i^{\,2}$$

Q.E.D.

Als we definiëren:

(3.1) $$a_1 := \frac{1}{12}k^2 + (2r - p - \frac{1}{4})k + 4p - 2p^2 + q - 6r + \frac{1}{6}$$

$$a_2 := \frac{1}{12}k - p + p^2 - q + 2r - \frac{1}{6}$$

$$a_3 := -\frac{1}{12}k + 3p - p^2 - 4r + \frac{1}{12}$$

zodat: $\mathrm{var}\ \bar{R}_{1.} = a_1 n + a_2$

en $\mathrm{cov}\ (\bar{R}_{1.}, \bar{R}_{2.}) = a_3 n - \frac{1}{12}$

dan geldt:

$$\operatorname{var} \sum_{i=1}^{k-1} (c_i \bar{R}_{i\cdot}) = \sum_{i=1}^{k-1} c_i^2 (a_1 n + a_2) + \sum_{i \neq j} \sum c_i c_j \left(a_3 n - \frac{1}{12}\right)$$

$$= (\Sigma c_i^2)(a_1 n + a_2)\left\{1 + \frac{\sum_{i \neq j} \sum c_i c_j}{\Sigma c_i^2} \cdot \frac{a_3 n - \frac{1}{12}}{a_1 n + a_2}\right\}$$

Uit de lemma's 3.1 en 3.2 volgt: $-a_1 \leq (k-2)a_3 \leq 0$ zodat

$\operatorname{var}\left(\sum_{i=1}^{k-1} c_i \bar{R}_{i\cdot}\right)$ naar oneindig gaat voor $n \to \infty$ als de volgende twee voorwaarden zijn vervuld:

(3.3) i $\quad a_1 > 0$

ii $\quad \sum_{i} \sum_{\neq j} c_i c_j < (k-2) \sum_{i=1}^{k-1} c_i^2$ of $-a_1 < (k-2)\, a_3$

Dus als $-a_1 < (k-2)a_3$ (dan ook $a_1 > 0$), dan geldt dat de vektor $(\bar{R}_{1\cdot}, \ldots, \bar{R}_{k-1,\cdot})$ asymptotisch normaal verdeeld is.

Nu geldt: $-a_1 < (k-2)a_3$ d.e.s.d. als:

$$(2p - 2r - p^2)(k-1) + q - p^2 > 0$$

waarin:

$$2p - 2r - p^2 = \int (1-F)^2 dG - \left(\int G dF\right)^2 =$$

$$= \int (1-F)^2 dG - \left(\int (1-F) dG\right)^2 = \int \{F - \int F dG\}^2 dG \geq 0$$

en $\quad q - p^2 = \int G^2 dF - \left(\int G dF\right)^2 = \int \{G - \int G dF\}^2 dF \geq 0$

$2p - 2r - p^2$ en $q - p^2$ zijn beide gelijk aan nul d.e.s.d. als

(3.4) G gelijk is aan 0 of 1 op de drager van F.

We hebben dus dat, behalve wanneer (3.4) geldt, de vektor $(\bar{R}_{1\cdot}, \ldots, \bar{R}_{k-1\cdot})$ asymptotisch normaal verdeeld is.

Als (3.4) wèl geldt, dan is a_1 gelijk aan $\frac{1}{12}k^2 - \frac{1}{4}k + \frac{1}{6}$ en dus positief (want $k \geq 3$), zodat volgens (3.3) en lemma 3.3. $\sum_{i=1}^{k-1} c_i \bar{R}_{i\cdot}$ asymptotisch normaal verdeeld is, als niet alle c_i's gelijk zijn. Verder is dan $\sum_{i=1}^{k-1} \bar{R}_{i\cdot}$ konstant, zodat, ook als (3.4) geldt, $(\bar{R}_{1\cdot}, \ldots, \bar{R}_{k-1\cdot})$ asymptotisch normaal verdeeld is, maar met één dimensie lager, nl. dimensie k-2.

4. De asymptotische verdeling van $\max_{1\le i,j\le k-1} |\bar{R}_{i\cdot} - \bar{R}_{j\cdot}|$

Teneinde de asymptotische verdeling van $\max_{1\le i,j\le 1} |\bar{R}_{i\cdot} - \bar{R}_{j\cdot}|$ te berekenen voor $n \to \infty$, beschouwen we de volgende $(k-1)$-dimensionale stochastische vektor $\underline{v}_n$ gedefinieerd door:

$$\underline{v}_n = (v_{n1},\ldots,v_{n,k-1}) := \left(\frac{\bar{R}_{1\cdot} - \mathcal{E}\bar{R}_{1\cdot}}{\sqrt{n}},\ldots,\frac{\bar{R}_{k-1\cdot} - \mathcal{E}(\bar{R}_{k-1\cdot})}{\sqrt{n}}\right)$$

Volgens hoofdstuk 3 is $\underline{v}_n$ voor $n \to \infty$ asymptotisch multi-normaal verdeeld met kovariantiematrix:

$$\begin{pmatrix} a_1 & a_3 & \cdots & a_3 \\ & a_1 & & \vdots \\ & & a_1 & a_3 \\ & & & a_1 \end{pmatrix}$$

Definiëren we $w_{ni} := v_{ni} - \gamma\bar{v}_n$

waarin $\gamma := 1 \pm \sqrt{\frac{a_1 - a_3}{a_1 + (k-2)a_3}}$ en $\bar{v}_n := \frac{1}{k-1}\sum_{j=1}^{k-1} v_{nj}$

dan is $(w_{n1},..,w_{n,k-1})$ voor $n \to \infty$ asymptotisch multi-normaal verdeeld met als kovariantiematrix $(a_1 - a_3)I_{k-1}$ (waarin met I_{k-1} de identiteitsmatrix van afmeting $k-1$ bedoeld wordt), zodat de range van

$$\frac{w_{n1}}{\sqrt{a_1 - a_3}},\ \ldots\ldots,\ \frac{w_{n,k-1}}{\sqrt{a_1 - a_3}}$$

voor $n \to \infty$ $q_{k-1,\infty}$ verdeeld is.

Hieruit volgt dat ook de range van $\frac{\bar{R}_{1\cdot}}{\sqrt{n(a_1 - a_3)}},\ldots,\frac{\bar{R}_{1}}{\sqrt{n(a_1 - a_3)}}$ voor $n \to \infty$

ook $q_{k-1,\infty}$ verdeeld is zodat we vinden voor $n \to \infty$:

(4.1) $P\{|\bar{R}_{i\cdot} - \bar{R}_{j\cdot}| < q^{\alpha}_{k-1,\infty}\sqrt{n(a_1 - a_3)}$ voor alle i,j met $1 \le i,j \le k-1\} = 1-\alpha$

Op dezelfde wijze bewijst men dat onder $H_0(F=G)$ voor $n \to \infty$ geldt:

$$(4.2) \qquad P\{ |\bar{R}_{i\cdot} - \bar{R}_{j\cdot}| < q^{\alpha}_{k,\infty} \tfrac{1}{12}k^2 n \text{ voor alle } i,j \} = 1 - \alpha$$

(want onder H_0 is $\frac{\bar{R}_{1\cdot}}{\sqrt{n}}, \ldots\ldots \frac{\bar{R}_{k\cdot}}{\sqrt{n}}$ asymptotisch normaal verdeeld

met varianties $\frac{1}{12}k(k-1)$ en kovarianties $-\frac{1}{12}k$)

Onder het door ons beschouwde alternatief $F_1 = \ldots\ldots = F_{k-1} = F$ en $F_k = G$ is de kans om uit uit $\{F_1, \ldots, F_{k-1}\}$ sommige verschillend van elkaar te verklaren gelijk aan:

$$P\{ \max_{1\le i,j\le k-1} |\bar{R}_{i\cdot} - \bar{R}_{j\cdot}| > q^{\alpha}_{k,\infty} \sqrt{\tfrac{1}{12}k^2 n} \} =$$

$$= P\{ q_{k-1,\infty} > q^{\alpha}_{k,\infty} \sqrt{\frac{\frac{1}{12}k^2}{a_1 - a_3}} \}$$

zodat:

$$(4.3) \qquad \alpha_{k-1} = \sup_{\substack{F,G \\ F \text{ en } G \text{ kontinu}}} P\{ q_{k-1,\infty} > q^{\alpha}_{k,\infty} \sqrt{\frac{\frac{1}{12}k^2}{a_1 - a_3}} \}$$

Hierin hangen a_1 en a_3 van F en G af (zie(2.1),(2.2) en (3.1)).

Opmerking:

We kunnen nu ook de echte simultane onbetrouwbaarheid voor $n \to \infty$ berekenen van de methode gebaseerd op formule (1.1). Vanwege de asymptotische χ^2_{k-1} - verdeling van de toetsingsgrootheid van Kruskal-Wallis is die echte onbetrouwbaarheid (die afhangt van α en k) gelijk aan:

$$\lim_{n\to\infty} P\{ \max_{1\le i,j\le k} |\bar{R}_{i\cdot} - \bar{R}_{j\cdot}| > \sqrt{\chi^2_{k-1}(\alpha) \frac{k^2 n + k}{6}} \}$$

$$= \lim_{n\to\infty} P\left\{ \frac{\max_{1\le i,j\le k} |\bar{R}_{i\cdot} - \bar{R}_{j\cdot}|}{\sqrt{\frac{1}{12}k^2 n}} > \sqrt{\chi^2_{k-1}(\alpha) \frac{2(k^2 n + k)}{k^2 n}} \right\}$$

$$= P\{ q_{k,\infty} > \sqrt{2\chi^2_{k-1}(\alpha)} \}$$

Tabel 4.1: Simultane onbetrouwbaarheid van de methode gebaseerd op formule (1.1) als $n \to \infty$

k=	3	4	5	6	7	8	9	10	12	15	20
α=0.10	.081	.060	.042	.029	.019	.012	.0079	.0050	.0019	.00044	.00003
0.05	.038	.027	.018	.011	.0071	.0044	.0027	.0016	.00057	.00011	.000008
0.01	.0068	.0042	.0025	.0014	.00082	.00046	.00025	.00014	.000042	.000007	.0000003

5. Berekening van α_{k-1} zonder restrikties op F en G

Als we, behalve kontinuïteit, geen andere beperkingen opleggen aan de F en G uit formule (4.3) waarover het supremen genomen wordt, dan zal blijken dat voor de gangbare waarden van α geldt $\alpha_{k-1} > \alpha$. (zie tabel 5.1).
Dit bewijzen we als volgt:
de kans in formule (4.3) is maximaal voor die F en G waarvoor $a_1 - a_3$ maximaal is.
Nu geldt:

$$(5.1) \qquad a_1 - a_3 = \frac{1}{12}k^2 + (2r - p - \frac{1}{6})(k-1) + q - p^2 - \frac{1}{12}$$

waarin:

$$2r-p = \int (2F-1)GdF = \int_{\{x|F(x)<\frac{1}{2}\}} (2F-1)GdF + \int_{\{x|F(x)>\frac{1}{2}\}} (2F-1)GdF$$

zodat 2r-p maximaal is als er getallen x_0 en x_1 bestaan ($x_0<x_1$), waarvoor geldt dat:

$$F(x_0) = F(x)_1 = \tfrac{1}{2}$$

$$(5.2) \qquad G(x) = 0 \text{ voor } x < x_0$$

$$\text{en } G(x) = 1 \text{ voor } x > x_1$$

Nu is $q-p^2$ juist ook maximaal in deze situatie, want altijd geldt:

$$(5.3) \qquad q-p^2 = \int G^2 dF - p^2 \leq \int GdF - p^2 = p(1-p) \leq \tfrac{1}{4}$$

en als (5.2) geldt, dan $q - p^2 = \frac{1}{4}$
Voor een F en G als in (5.2) geldt dus:

$$a_1 - a_3 = \frac{1}{12}(k^2 + k + 1)$$

zodat we vinden:

$$(5.4) \qquad \alpha_{k-1} = P\{q_{k-1,\infty} > q^{\alpha}_{k,\infty}\sqrt{\frac{k^2}{k^2+k+1}}\}$$

M.b.v. een tabel van de verdelingsfunktie van de range van **normaal** verdeelde grootheden, bijvoorbeeld Harter (1969), vinden we:

tabel 5.1: α_{k-1} zonder restricties op F en G voor $\alpha = 0.01$, $\alpha = 0.025$, $\alpha = 0.05$ en $\alpha = 0.10$.

k =	3	4	5	6	7	8	9	10	12	15	20
α=0.01	.0153	.0181	.0182	.0178	.0172	.0167	.0162	.0158	.0151	.0143	.0134
0.025	.0303	.0361	.0386	.0385	.0379	.0372	.0365	.0358	.0347	.0334	.0318
0.05	.0512	.0643	.0682	.0690	.0688	.0682	.0674	.0667	.0652	.0633	.0612
0.10	.0877	.1123	.1208	.1240	.1250	.1250	.1245	.1238	.1224	.1202	.1172

6. α_{k-1} voor Lehmann-alternatieven

Nemen we in formule (4.3) het supremum alleen over die paren F,G waarvoor geldt: $G = F^t$ voor **zekere** $t \in (0,\infty)$, dan zal blijken dat wel geldt: $\alpha_{k-1} < \alpha$. (tabel 6.1)

Als: $G = F^t$ dan:

$$p = \int F^t dF = \frac{1}{t+1}$$

$$q = \int F^{2t} dF = \frac{1}{2t+1}$$

$$r = \int F^{t+1} dF = \frac{1}{t+2}$$

zodat invullen in (5.1) geeft:

$$a_1 - a_3 = \frac{1}{12}k^2 + (k-1)\left(\frac{2}{t+2} - \frac{1}{t+1} - \frac{1}{6}\right) \uparrow \frac{1}{2t+1} - \frac{1}{(t+1)^2} - \frac{1}{12}$$

In de volgende tabel zijn voor verschillende waarden van k de maximale waarden van $a_1 - a_3 - \frac{k^2}{12}$ als funktie van t gegeven:

k=	3	4	5	6	7	8	9	10	12	15	20
. $a_1 - a_3 - \frac{k^2}{12}$	.01625	.02112	.02600	.03090	.03579	.04069	.04559	.05049	.06030	.07501	.09954

zodat we voor α_{k-1} vinden:

tabel 6.1: α_{k-1} voor Lehmann-alternatieven:

k=	3	4	5	6	7	8	9	10	12	15	20
α=0.01	.0040	.0057	.0067	.0073	.0078	.0081	.0083	.0085	.0088	.0090	.0093
0.025	.0100	.0144	.0168	.0184	.0194	.0202	.0208	.0213	.0220	.0226	.0233
0.05	.0204	.0291	.0339	.0370	.0391	.0406	.0418	.0427	.0440	.0454	.0466
0.10	.0408	.0597	.0690	.0750	.0790	.0820	.0842	.0860	.0886	.0911	.0935

7. α_{k-1} voor verschuivingsalternatieven

We beschouwen nu alternatieven waarvoor een $a \in \mathbf{R}$ bestaat, zodat $G(x)=F(x-u)$ voor alle $x \in \mathbf{R}$.

Verder nemen we eerst even aan dat $a > 0$ (verschuiving naar rechts).

Dan geldt: $G(x) \leq F(x)$ voor alle x

zodat:

$$2r - p = \int G\,(2F-1)\,dF =$$

$$= \int_{\{x|F(x)\leq\frac{1}{2}\}} G(2F-1)\,dF + \int_{\{x|F(x)>\frac{1}{2}\}} G(2F-1)\,dF \leq$$

$$(7.1) \qquad \leq 0 + \int_{\{x|F(x)>\frac{1}{2}\}} (2F^2-F)\,dF = \left[\frac{2}{3}F^3 - \frac{1}{2}F^2\right]_{F=\frac{1}{2}}^{F=1} = \frac{5}{24}$$

verder geldt nog steeds (zie 5.3):

$$q - p^2 \leq \tfrac{1}{4}$$

We kunnen echter nog wel een scherpere bovengrens voor $q-p^2$ vinden (waarvan bovendien de geldigheid niet alleen beperkt is tot verschuivingsalternatieven):

Stelling 7: 1:

$$(7.2) \qquad q - p^2 \leq 2r - p$$

Bewijs: $$\int G^2 dF + \int F^2 dG = P(\underline{y}_{11} > \underline{y}_{k1} \cap \underline{y}_{11} > \underline{y}_{k2}) + P(\underline{y}_{11} < \underline{y}_{k2} \cap \underline{y}_{12} < \underline{y}_{k2}) =$$

$$= P\{(\underline{y}_{11} > \underline{y}_{k1} \cap \underline{y}_{11} > \underline{y}_{k2}) \cup (\underline{y}_{11} < \underline{y}_{k2} \cap \underline{y}_{12} < \underline{y}_{k2})\} \leq$$

$$\leq P\{\underline{y}_{11} > \underline{y}_{k1} \cup \underline{y}_{12} < \underline{y}_{k2}\} =$$

$$= 1 - P\{\underline{y}_{11} < \underline{y}_{k1} \cap \underline{y}_{12} > \underline{y}_{k2}\} =$$

$$= 1 - p\cdot(1-p) \tag{7.3}$$

Het laatste gelijkteken geldt vanwege de onafhankelijkheid van $\underline{y}_{11}$, $\underline{y}_{12}$, $\underline{y}_{k1}$ en $\underline{y}_{k2}$
We vinden m.b.v. (7.3):

$$q - p^2 = \int G^2 dF - p^2 \leq -\int F^2 dG + 1 - p(1-p) - p^2 = 2r - p$$

waarbij het laatste gelijkteken volgt uit:

$$\int F^2 dG = 1 - 2\int FG dF \quad \text{(partiële integratie)}$$

Q.E.D.

Vullen we (7.1) en (7.2) in in (5.1) dan volgt:

$$a_1 - a_3 \leq \frac{1}{12}k^2 + \frac{1}{24}(k-1) + \frac{1}{8} = \frac{1}{12}(k^2 + \frac{1}{2}k + 1) \tag{7.4}$$

voor $G(x) = F(x-a)$ met $a > 0$.

Als $a < 0$ (verschuiving naar links) daar kijken we naar:

$$-\underline{y}_{11},\ldots, -\underline{y}_{1n}; \ldots\ldots; -\underline{y}_{k1},\ldots; -\underline{y}_{kn}$$

Als F' de verdelingsfunktie is van $-\underline{y}_{11}$ en G' die van $-\underline{y}_{k1}$ dan geldt:

$$F'(x) = 1-F(-x)$$
$$\text{en } G'(x) = F'(x+a)$$

G' is dus een verschuiving van F' naar rechts (omdat $a < 0$). Verder geldt: als $\underline{R}'_{ij}$ het rangnummer is van $-\underline{y}_{ij}$ onder alle $-\underline{y}_{11},\ldots, -\underline{y}_{kn}$ dan:

$$\underline{R}'_{ij} = N - \underline{R}_{ij} + 1$$

zodat $a_1-a_3 = \lim_{n\to\infty} \frac{1}{2n} \operatorname{var}(\bar{\underline{R}}_{1\cdot} - \bar{\underline{R}}_{2\cdot}) = \lim_{n\to\infty} \frac{1}{2n} \operatorname{var}(\bar{\underline{R}}'_{1\cdot} - \bar{\underline{R}}'_{2\cdot})$

$$= a'_1 - a'_3$$

Nu geldt dat $a'_1-a'_3 \leq \frac{1}{12}(k^2 + \frac{1}{2}k+1)$, omdat G' een verschuiving naar rechts is van F', zodat (7.4) ook geldt als $a < 0$.
(7.4) geeft een bovengrens voor a_1-a_3 bij verschuivingsalternatieven. Dat deze bovengrens niet erg veel verlaagd kan worden, blijkt uit het volgende voorbeeld:

voorbeeld 7.1:

Neem F gedefinieerd door

$$F(x) = \begin{cases} x + \frac{1}{2} & \text{als} \quad -\frac{1}{2} \le x \le 0 \\ \frac{x}{c} + \frac{1}{2} & \text{als} \quad 0 \le x \le \frac{1}{2}c \end{cases}$$

en G gedefinieerd door $G(x) = F(x-\frac{1}{2})$
dan geldt (onderstel $c > \frac{1}{2}$):

$$2r - p = \int G(2F-1)dF \to \int_0^{\frac{1}{2}c} (F(x) - \tfrac{1}{2}c)(2F(x) - 1)\, dF(x) =$$

$$= \left[\frac{2}{3}F^3 - \frac{1}{2}(1+c)F^2 + \frac{1}{2}c\,F\right]_{F=\frac{1}{2}}^{F=\frac{1}{2}c} = \frac{5}{24} - \frac{c}{4}$$

en $q - p^2 = \frac{29}{192} + \mathcal{O}(\frac{1}{c})$

zonder geldt:

$$a_1 - a_3 = \frac{1}{12}k^2 + \frac{1}{24}k + \frac{5}{192} + \mathcal{O}(\frac{1}{c})$$

Bij deze F en G geldt dus voor $c \to \infty$:

$$a_1 - a_3 \to \frac{1}{12}(k^2 + \tfrac{1}{2}k + \frac{5}{16})$$

Als we deze waarde van $a_1 - a_3$ en die uit (7.4) invullen in (4.3), dan vinden we de volgende onder- en bovengrenzen voor α_p:

Tabel 7.1

Onder-en bovengrenzen voor α_{k-1} bij verschuivingsalternatieven.

	α= 0.01	α=0.025	α=0.05	α=0.10
k=3	.0079-.0100	.0175-.0213	.0325-.0381	.0612-.0695
4	.0101-.0119	.0230-.0263	.0431-.0483	.0816-.0894
5	.0109-.0123	.0253-.0279	.0478-.0519	.0909-.0972
6	.0113-.0124	.0263-.0284	.0501-.0535	.0958-.1009
7	.0114-.0123	.0268-.0285	.0514-.0541	.0987-.1030
8	.0114-.0121	.0271-.0285	.0521-.0544	.1005-.1040
9	.0114-.0120	.0273-.0284	.0526-.0544	.1019-.1046
10	.0114-.0119	.0273-.0283	.0529-.0544	.1025-.1049
12	.0114-.0117	.0273-.0281	.0531-.0543	.1034-.1053
15	.0112-.0115	.0272-.0277	.0532-.0540	.1039-.1052
20	.0111-.0112	.0270-.0273	.0530-.0534	.1041-.1049

We zien in tabel 7.1 dat voor verschuivingsalternatieven in het algemeen niet geldt dat $\alpha_{k-1} \leq \alpha$(hoewel het niet veel scheelt). Als we echter twee extra eisen aan F opleggen (die in de praktijk wel eens vervuld zijn), dan blijkt dat altijd geldt dat $\alpha_{k-1} \leq \alpha$.

Stelling 7.2: Als G een verschuiving van F is, waarbij F symmetrisch en unimodaal is (en kontinu), dan geldt:

$$(7.5) \qquad 2r - p \leq \frac{1}{6}$$

Bewijs:
Daar het probleem translatie-invariant is, is het geen beperking om F symmetrisch in x = 0 te nemen.

$$2r - p = \int_{-\infty}^{0} F(x-a)(2F(x)-1)dF(x) + \int_{0}^{\infty} F(x-u)(2F(x)-1)dF(x)$$

Nu geldt: (omdat F(-x) = 1 - F(x)):

$$\int_{-\infty}^{0} F(x-a)(2F(x)-1) = \int_{+\infty}^{0} F(-y-a)\ \{2F(-y)-1\ \}dF(-y) =$$

$$= \int_{0}^{\infty} F(-y-a)\ \{1-2F(y)dF(y) = \int_{0}^{\infty}\{F(y+a) - 1\}\{\ 2F(y)-1\ \}dF(y)$$

zodat

$$(7.6) \qquad 2r - p = \int_{0}^{\infty}\{\ F(x+a) + F(x-a) - 1\}\ \{2F(x)-1\}\ dF(x)$$

(7.7) Voor $x \geq 0$ geldt: $F(x+a) + F(x-a) \leq 2F(x)$

Als $x \geq |a|$ geldt dit vanwege de unimodaliteit van F, hetgeen betekent dat F konkaaf is voor $x \geq 0$.
Voor $0 \leq x < |a|$ heeft men de symmetrie en unimodaliteit beide nodig om in te zien dat (7.7) geldt.
Door (7.7) in te vullen in (7.6) vinden we:

$$2r - p \leq \int_{0}^{\infty} (2F(x)-1)^2 dF(x) = \left[\frac{1}{6}(2F(x) - 1)^3\right]_{F(x)=\frac{1}{2}}^{F(x)=1} = \frac{1}{6}$$

Q.E.D.

Met behulp van stelling 7.1 vinden we dat onder de voorwaarden van stelling 7.2 ook geldt:

$$(7.8) \qquad q - p^2 \leq \frac{1}{6}$$

Invullen van (7.5) en (7.8) in (5.1) geeft:

$$a_1 - a_3 \leq \frac{1}{12}k^2 + \frac{1}{12}$$

zodat geldt:

$$\alpha_{k-1} \leq P\{q_{k-1,\infty} > q^{\alpha}_{k,\infty}\sqrt{\frac{k^2}{k^2+1}}\}$$

tabel 7.2:

Bovengrens voor α_{k-1} voor verschuivingen van symmetrische, unimodale verdelingen.

k=	3	4	5	6	7	8	9	10	12	15	20
α=0.01	.0057	.0071	.0077	.0081	.0083	.0085	.0086	.0087	.0089	.0091	.0093
0.025	.0135	.0173	.0190	.0200	.0206	.0211	.0214	.0217	.0222	.0227	.0232
0.05	.0262	.0349	.0375	.0396	.0410	.0421	.0429	.0435	.0445	.0455	.0465
0.10	.0498	.0673	.0749	.0793	.0823	.0845	.0860	.0872	.0893	.0913	.0933

Opmerking: Als we de symmetrie-eis laten vallen, dan zitten we weer in de situatie van tabel 7.1, omdat de F uit voorbeeld 7.1 unimodaal is.

Literatuur:

Hajek J., Asymptotic mormality of simple linear rank statistics under alternatives, The Annals of Mathematical Statistics, vol.39 (1968), pag. 325-346.

Harter H.L, Order statistics and their use in testing and estimation, vol I, Aerospace Research Laboratories, Government Printing Office, Washington (1969)

Miller R.G, Simultaneous statistical inference, McGraw-Hill, New York (1966)

Nemenyi P., Distribution-free multiple comparisons. Unpublished doctoral thesis, Princeton University (1963)