

Technische Hogeschool Eindhoven
Onderafdeling der Wiskunde
COSOR-Memorandum 79-14
november 1979

MEDIAANScores IN SIMULTANE VERDELINGSVRIJE METHODEN

Kees van der Hoeven

SAMENVATTING:

In deze studie wordt de bruikbaarheid van mediaanscores in simultane verdelingsvrije methoden onderzocht. Beschouwd wordt de situatie van k steekproeven met in elke steekproef n waarnemingen. Er worden drie klassen van toetsingsgrootheden bekeken, nl. Q gebaseerd op kwadraatsommen, R gebaseerd op de range van steekproefscores en P gebaseerd op paarsgewijze twee steekproefgrootheden. Voor het toetsen van gelijkheid van populaties bleek de Q toets het geschiktst, voor het simultaan toetsen van alle populatieparen op gelijkheid bleek zowel een stapsgewijze Q toets als de P toets goed te zijn. R toetsen, door diverse auteurs voorgesteld bij simultane vergelijking van populatieparen bleken in geen situatie de voorkeur te genieten. Bij de genoemde klassen van toetsen werden mediaan scores en Wilcoxon scores onderzocht. Grootheden met mediaan scores kunnen veel minder verschillende waarden aannemen dan grootheden met Wilcoxon scores. Hierdoor is de asymptotische benadering bij de laatstgenoemden veel beter. Bij de Q grootheid gebaseerd op mediaan scores is de exacte verdeling echter vrij gemakkelijk te berekenen. Voor de P grootheid voldoet de afschatting van Bonferroni redelijk.

Er werd gevonden dat het onderscheidingsvermogen bij mediaan scores in ongeveer dezelfde mate verschilt van Wilcoxon scores als het geval is bij twee steekproeven. Hieruit kan worden geconcludeerd dat een onderzoeker slechts aan een simultane toets met mediaanscores begint als hij er van overtuigd is dat zijn waarnemingen uit een verdeling met een langere staart dan de dubbelexponentiele afkomstig zijn.

1. INLEIDING

Laten $F_1, F_2, \dots, F_k$ continue verdelingsfuncties behorende bij k populaties zijn. We beschouwen uit elke populatie een steekproef van n herhalingen.

De realisatie van de steekproeven is een matrix van waarnemingen

$$X = \{x_{ij}\}, \quad i = 1, 2, \dots, k, \quad j = 1, 2, \dots, n.$$

De nulhypothese is

$$H_0 : F_1 = F_2 = \dots = F_k$$

en het alternatief

$$H_{alt} : \text{er is een } i, l : F_i \neq F_l.$$

We zullen twee situaties nader beschouwen, nl. het toetsen van H_0 en het paarsgewijs vergelijken van alle populaties.

Toetsen van H_0 : Er wordt alleen getoetst of alle k populaties gelijk zijn.

Met de grootte T wordt met onbetrouwbaarheid α getoetst als H_0 verworpen wordt bij een uitkomst $T(X) > T(\alpha)$, waarbij

$$(1.1) \quad P(T(X) > T(\alpha) | H_0) = \alpha.$$

Het onderscheidingsvermogen bij een alternatief H is

$$f(H; T(\alpha)) := P(T(X) > T(\alpha) | H).$$

Paarsgewijze populatie toetsen: Elk populatiepaar wordt op gelijkheid getoetst. Het paar i, l wordt ongelijk verklaard bij een uitkomst

$T_{il}(X) > T_p(\alpha_p)^{1)}$ van toetsingsgrootte T_{il} , waarbij α_p de onbetrouwbaarheid van de toets is, gedefinieerd als

$$\begin{aligned} \alpha_p &:= \text{de kans dat tenminste één populatiepaar ten onrechte verschillend} \\ &\quad \text{wordt verklaard bij welke nulhypothese dan ook,} \\ &= \max_H P(\exists_{i,l} T_{il}(X) > T_p(\alpha_p), F_i = F_l | H). \end{aligned}$$

1) dit kunnen vektoren zijn, zie 2.2.1.

We kunnen de volgende maten van onderscheidingsvermogen van de toets hanteren (zie Ramsey):

- Any Pair Power: de kans dat tenminste een van de in werkelijkheid ongelijke populatieparen verschillend wordt verklaard.
- Per Pair Power: de kans dat een nader gespecificeerd populatiepaar verschillend wordt verklaard,

$$f_{pp}(H; i, \ell; T_p(\alpha_p)) := P(T_{i\ell}(X) > T_p(\alpha_p) | H)$$

- Mean Pair Power: $f_{mp}(H; T_p(\alpha_p))$, het percentage populatieparen dat terecht verschillend van elkaar wordt verklaard bij alternatief H en gebruik van paarsgewijze T test met onbetrouwbaarheid α_p .
- All Pair Power: de kans dat alle werkelijke ongelijke populatieparen verschillend worden verklaard.

We zullen voor deze beide gevallen een aantal verdelingsvrije simultane toetsen bekijken en de onderscheidingsvermogens met elkaar vergelijken. Hierbij is onze aandacht speciaal gericht op bruikbaarheid van toetsen met mediaanscores.

Definieer voor de steekproeven de overallrangnummers

$$r_{ij} := \# \{x_{pq} | x_{pq} \leq x_{ij}, p = 1, 2, \dots, k, q = 1, 2, \dots, n\}$$

$$i = 1, 2, \dots, k, j = 1, 2, \dots, n,$$

en de paarsgewijze rangnummers van de samengevoerde i^e en ℓ^e steekproef

$$r_{ij, \ell} := \# \{x_{pq} | x_{pq} \leq x_{ij}, p = i \text{ of } \ell, q = 1, 2, \dots, n\}$$

$$i \neq \ell, i, \ell = 1, 2, \dots, k, j = 1, 2, \dots, n.$$

Onder H_0 zijn de rangordevektoren aselekte permutaties. Voor een rangnummer r uit een permutatie van $(1, 2, \dots, N)$ wordt door een functie $a(r)$ een score gedefinieerd.

Zo is

$$a_W(r) = r \quad \text{de Wilcoxon score}$$

$$a_M(r) = \begin{cases} 1 \\ 0 \end{cases} \quad \begin{matrix} \text{als } r \leq N/2 \\ \text{elders} \end{matrix} \quad \text{de Mediaanscore}$$

Deze scores worden voortaan met de letters W en M aangegeven.

2. EEN AANTAL KLASSEN VAN SIMULTAAN VERDELINGSVRIJE TOETSEN

We behandelen nu drie klassen van toetsen, aan te duiden als P, Q en R toetsen. In 2.1 en 2.2 geven we aan hoe daarmee respectievelijk H_0 en paarsgewijze verschillen getoetst kunnen worden.

2.1. Het toetsen van H_0

2.1.1. Q-toetsen (Hajek blz. 96 e.v.)

Deze zijn gebaseerd op de som van kwadraten van populatiescores:

$$Q(X) = \sum_{i=1}^k \left(\sum_{j=1}^n a(r_{ij}) \right)^2.$$

De grootheid Q is onder H_0 voor grote n bij benadering in te schalen als een chikwadraat verdeelde grootheid. Met de Wilcoxon scores is dit de Kruskal-Wallis test.

2.1.2. R-toetsen (Range)

Deze gaan uit van

$$R(X) = \text{range}_{i=1, \dots, k} \left(\sum_{j=1}^n a(r_{ij}) \right) = \max_i \sum_{j=1}^n a(r_{ij}) - \min_i \sum_{j=1}^n a(r_{ij}).$$

Als asymptotische benadering is deze grootheid onder H_0 als een studentised range in te schalen. Voor de R toets met Wilcoxon scores zie Dunn en Tobach/Smith/Rose & Richter en met mediaanscores zie Miller pag. 182-185.

2.1.3. P toetsen (Paarsgewijs)

Deze gaan uit van

$$P(X) = \max_{i \neq l} \sum_{j=1}^n a(r_{ij,l})$$

zie bijvoorbeeld Steel (met Wilcoxon scores).

Bij een speciale structuur van de correlaties tussen de getallen

$$\sum_{j=1}^n a(r_{ij,l}) \quad , \quad i, l = 1, 2, \dots, k$$

is deze in te schalen en asymptotisch te benaderen door een studentised range grootheid.

We geven nu een toets aan door twee letters waarbij de eerste slaat op de gebruikte score en de tweede op de klasse van toets.

De bijbehorende toetsingsgrootheid geven we ook zo aan, bijvoorbeeld

$MR(X) = a$, betekent : de toetsingsgrootheid behorende bij een Range toets met mediaan scores heeft waarde a .

2.2. Paarsgewijze populatietoetsen

De P en R toetsen zijn met de door (1.1) bepaalde kritieke waarden $P(\alpha)$ en $R(\alpha)$ ook eenvoudig te gebruiken als paarsgewijze populatie toetsen.

Voor P toetsen geldt dan dat gebruik van de kritieke waarde $P_p(\alpha_p) = P(\alpha)$ voor paarsgewijze vergelijkingen een onbetrouwbaarheid $\alpha_p = \alpha$ levert, maar bij R toetsen levert gebruik van $R_p(\alpha_p) = R(\alpha)$ een α_p die groter is dan α . Voor de Q grootheid is een stapsgewijze procedure vereist om tot uitspraken over verschillen in populatieparen te komen. We zullen hier nu nader op ingaan.

2.2.1. (Stapsgewijze) Q toetsen

Noem Q_s de Q -grootheid van s gelijke populaties en $Q_s(\alpha)$ de bijbehorende kritieke waarde:

$$P(Q_s(X) > Q_s(\alpha) \mid H_0) = \alpha.$$

Geven we met B_s een deelverzameling van s populaties uit $\{1, 2, \dots, k\}$ aan dan definiëren we

$$Q_s(B_s) = \sum_{i \in B_s} \left(\sum_{j=1}^n a(r_{ij}^*) \right)^2$$

waarbij

$$r_{ij}^* := \# \{x_{pq} \mid x_{pq} \leq x_{ij}, p \in B_s, j = 1, 2, \dots, n\}$$

het rangnummer van x_{ij} onder de steekproeven van B_s is.

Noteer voor populaties $i, \ell \in \{1, 2, \dots, k\}$ met $O_s(i, \ell)$ de verzameling van alle deelverzamelingen van s populaties die i en ℓ omvatten:

$$O_s(i, \ell) = \{B \subset \{1, 2, \dots, k\} \mid i \in B, \ell \in B, \#B = s\}.$$

We verklaren een populatiepaar i, ℓ slechts dan verschillend als

$$\forall s = 2, \dots, k \quad \forall B_s \in O_s(i, \ell) \quad Q_s(B_s) > Q_s(\alpha(s))$$

waarbij de getallen $\alpha(2), \alpha(3), \dots, \alpha(k)$ de obetrouwbaarheid α_p bepalen.

Het aantal verzamelingen dat door de bovenstaande quantoren wordt gedefinieerd is

$$\binom{k-2}{0} + \binom{k-2}{1} + \dots + \binom{k-2}{k-2} = 2^{k-2}.$$

We kunnen dus als Q toetsingsgrootheid voor een paarsgewijze populatietoets een vektor met 2^{k-2} componenten beschouwen. De ongelijkheid $Q_{i\ell} > Q_p(\alpha_p)$ moet dan componentsgewijs gelden.

Bij de keuze

$$\begin{aligned}\alpha(s) &:= 1 - (1-\alpha)^{s/k} & s = 2, 3, \dots, k-2 \\ \alpha(k) &:= \alpha(k-1) := \alpha\end{aligned}$$

is de onbetrouwbaarheid van de paarsgewijze toets $\alpha_p \leq \alpha$.

Bij het uitvoeren van de stapsgewijze Q toets moeten nu achtereenvolgens alle deelverzamelingen van $k, k-1, \dots, 2$ populaties worden afgelopen. Als er een stel populaties B_s niet verschillend wordt verklaard dan hoeven deelverzamelingen van B_s niet meer op ongelijkheid getest te worden.

Dit kan worden uitgevoerd met behulp van een $k \times k$ matrix A met

$$a_{ij} = 1 \text{ als er tot dan toe nog geen deelverzameling } B \text{ is geweest}$$

waarvoor $i \in B$ en $j \in B$ en B is niet significant verschillend verklaard.

Dan is A aan het begin van de toets dus een matrix met allemaal enen, en na afloop bevat A alleen enen op plaatsen i, j waarbij de populaties i en j volgens de stapsgewijze Q test significant verschillend worden verklaard. Een deelverzameling B hoeft slechts onderzocht te worden als er een i en $j \in B$ zijn met $a_{ij} = 1$.

2.2.2. De P toets

We verklaren populaties i en l verschillend als

$$P_{il}(X) = \max \left(\sum_{j=1}^n a(r_{ij,l}), \sum_{j=1}^n a(r_{lj,i}) \right) > P_p(\alpha_p) := P(\alpha).$$

Hiermee is de onbetrouwbaarheid van de toets α_p gelijk aan α .

Met behulp van de P test is het mogelijk simultane betrouwbaarheidsintervallen voor de populatie verschillen te geven (zie b.v. Lehmann voor Wilcoxon scores).

2.2.3. De paarsgewijze R toets

Populaties i en l worden verschillend van elkaar verklaard als

$$R_{il}(X) := \left| \sum_{j=1}^n a(r_{ij}) - \sum_{j=1}^n a(r_{lj}) \right| > R_p(\alpha_p) := R(\alpha) .$$

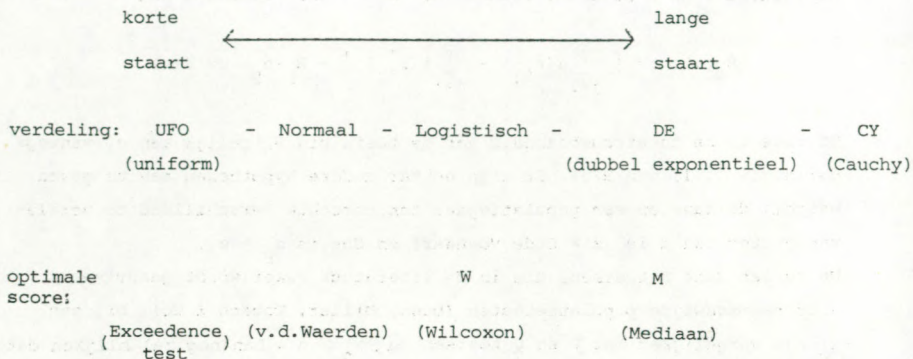
Hiermee is de onbetrouwbaarheid van de toets bij H_0 gelijk aan α , vanwege definitie (1.1) van $R(\alpha)$. Er zijn echter andere hypothesen aan te geven waarbij de kans om een populatiepaar ten onrechte verschillend te verklaren groter dan α is (zie Oude Voshaar) en dus is $\alpha_p > \alpha$.

We zullen deze R toetsen, die in de literatuur vaker wordt geadviseerd voor paarsgewijze populatietesten (Dunn, Miller, Tobach & Co), bij een $\alpha_p > \alpha$ vergelijken met P en Q toetsen bij $\alpha_p \leq \alpha$. Dan nog zal blijken dat de R toetsen inferieur zijn aan de anderen, zodat we kunnen concluderen dat deze geen nut hebben voor paren vergelijkingen.

2.3. Enkele opmerkingen over de P,Q, en R toetsen

Voor een Q of R toets moet slechts eenmaal de overall-rangnummering worden aangebracht, voor een P toets moet $\binom{k}{2}$ maal een steekproevenpaar van rangnummers worden voorzien, en voor een stapsgewijze Q toets kan maximaal $2^k - k - 1$ verzamelingen van steekproeven geordend moeten worden. Enige trucjes kunnen het een en ander wel iets bekorten, maar het verschil in rekenwerk bij P, Q of R blijft aanzienlijk. De vraag is of dit nog meetelt bij de huidige stand van automatisering. Een P toets is altijd zeer eenvoudig te programmeren, voor een stapsgewijze Q toets is dit wat lastiger en vereist veel meer rekentijd naarmate k groter wordt.

In geval van twee steekproeven ($k=2$) zijn de methodes P,Q en R identiek. Bij verschuivingsalternatieven is bij kennis van de soort verdeling een score te kiezen die een optimaal onderscheidingsvermogen heeft voor het geval $\Delta \rightarrow 0$, waarbij Δ de verschuivingsparameter is (zie Hajek). Het komt erop neer dat de scorekeuze afhankelijk is van de lengte van de staart van de verdeling en wel:



Op het gebied van paarsgewijs vergelijkende toetsen publiceerden Lin/Haseman een vergelijking tussen WP, WR, en een verdelingsvrij analogon van de LSD test. Ze kwamen tot de conclusie dat de laatstgenoemde duidelijk superieur is en vergelijken daarbij het resultaat van Carmen/Swanson voor het normaal verdeelde geval. Ze hebben dan ook dezelfde 'fout' als Carmen en Swanson gemaakt, welke in Einot/Gabriel duidelijk aan het licht wordt gebracht, nl. dat de toetsen worden vergeleken bij een verschillende onbetrouwbaarheid α_p .

Bij paren vergelijkingen kunnen inconsistente uitspraken van de vorm: "populatie i is beter dan j, j beter dan l" opleveren zonder dat "i beter dan l" kan worden verklaard. Voor een voorbeeld van inconsistentie van WP zie Lehmann blz. 245. Bij R toetsen zijn zulke inconsistenties niet mogelijk, omdat als $R_{ij} > R(\alpha)$ en $R_{jl} > R(\alpha)$ dan ook $R_{il} = R_{ij} + R_{jl} > R(\alpha)$. Ook bij MP toetsen zijn inconsistenties onmogelijk.

3. KRITIEKE WAARDEN

In hoofdstuk 4 zal op verschillen in onderscheidingsvermogen van de genoemde testen worden ingegaan. Er worden daar voornamelijk bij $k = 4$, $n = 10$ P,Q en R toetsen met W of M scores beschouwd. In dit hoofdstuk zal eerst behandeld worden hoe de kritieke waarden bepaald of benaderd kunnen worden en hoe goed de benadering is.

3.1. Benaderde kritieke waarden

Het is bekend dat de volgende benaderde kritieke waarden gekozen kunnen worden (zie Miller, Hajek)

$$WQ(\alpha) = \frac{kn^2(kn+1)}{12} (\chi_{k-1}^2(\alpha) + 3(kn+1))$$

$$WP(\alpha) = \frac{n(2n+1)}{2} + q_{k,\infty}^\alpha n \left(\frac{2n+1}{24} \right)^{\frac{1}{2}}$$

$$WR(\alpha) = q_{k,\infty}^\alpha n \left(\frac{k(kn+1)}{12} \right)^{\frac{1}{2}}$$

$$MQ(\alpha) = \frac{kn^2}{4(kn-1)} (\chi_{k-1}^2(\alpha) + kn - 1)$$

$$MP(\alpha) = \frac{n}{2} + q_{k,\infty}^\alpha n(8(2n-1))^{-\frac{1}{2}} + \frac{1}{2}$$

$$MR(\alpha) = q_{k,\infty}^\alpha \frac{n}{2}$$

waarbij $q_{k,\infty}^\alpha$ het α -kritieke punt van de studentised range met k en ∞ vrijheidsgraden is.

Voor $P(\alpha)$ kan ook de Bonferroni afchatting gebruikt worden

$$P^{k,n}(\alpha) \geq \binom{k}{2} P^{2,n}(\alpha)$$

waarbij $P^{k,n}$ de P grootheid bij k steekproeven en n herhalingen is.

3.2. Exacte kritieke waarden

Uit de literatuur zijn bij $k = 4$, $n = 10$ geen exacte tabellen beschikbaar.

De verdeling van MQ of MR is nog tamelijk eenvoudig te berekenen. Er geldt nl. met

$$M_i := \sum_{j=1}^n a_{M(r_{ij})}$$

dat

$$P(M_i = m_i, i = 1, 2, 3, 4) = \frac{\binom{10}{m_1} \binom{10}{m_2} \binom{10}{m_3} \binom{10}{m_4}}{\binom{40}{20}}$$

$$\sum_{i=1}^4 m_i = 20$$

en het is eenvoudig in te zien door welke combinaties (m_1, m_2, m_3, m_4) een bepaalde waarde voor MQ of MR wordt geleverd.

Voor de overige testen zoals MP, WP, WQ en WR komt het berekenen van de verdeling onder H_0 neer op het opratelen van $(40!)/(10!)^4$ permutaties. Om ons deze rekentijd te besparen hebben we enkele duizenden trekkingen uit H_0 gesimuleerd en de empirische verdeling van de toetsingsgrootheden bepaald. Voor tabellen van exacte of empirische verdeling onder H_0 zie de appendix.

Tabel 3.2 toont de kwaliteit van de asymptotische benaderingen en Bonferroni afschattingen uit 3.1.

overschrijdingskansen (geïnterpoleerd)	k=4 , n=10		k=4 , n=20		k=8 , n=10	
	$\alpha=.05$	$\alpha=.01$	$\alpha=.05$	$\alpha=.01$	$\alpha=.05$	$\alpha=.01$
$P(WQ > WQ(\alpha))$	.044	.006	.051	.009		
$P(WR > WR(\alpha))$	.045	.007	.050	.009	.040	.008
$P(WP > WP(\alpha))$	.043	.006	.050	.005	.032	.003
$P(WP > \binom{k}{2} WP^{2,n}(\alpha))$ (Bonferroni)	.04	.008	.042	.009	.035	.008
$P(MQ > MQ(\alpha))$	.052	.010	.051	.009	.050	.008
$P(MR > MR(\alpha))$	.096	.018	.078	.016	.089	.016
$P(MP > MP(\alpha))$	.049	.005	.085	.012	.029	.012
$P(MP > \binom{k}{2} MP^{2,n}(\alpha))$ (Bonferroni)	.04	.009	.045	.010	.03	.007

tabel 3.2 Kwaliteit van asymptotische benadering en
Bonferroni afschatting voor kritieke waarden

We zien dat de benadering van $MR(\alpha)$ slecht is, maar deze is gelukkig nog relatief eenvoudig exact te genereren. De benadering van $MP(\alpha)$ is ruw, de Bonferroni afschatting heeft zeker bij kleine α en k de voorkeur.

3.3. Kritieke waarden bij de stapsgewijze Q toets

We zullen nu de bepaling van de kritieke waarden $Q_s(\alpha(s))$, $s = 2, 3, \dots, k$ van de parenvergelijkende Q toets behandelen.

De grootheden Q_s zijn discreet, we zoeken naar de maximale waarden $Q_s(\alpha(s))$ waarvoor

$$\beta_s := P(Q_s > Q_s(\alpha(s))) \leq \alpha(s) .$$

De onbetrouwbaarheid van de parenvergelijkende Q test is dan

$$\alpha_p = \max_{\substack{(p_1, p_2, \dots, p_m) \\ m \geq 1, \sum p_i = k}} \left\{ 1 - \prod_{i=1}^m (1 - \beta_{p_i}) \right\} \quad (*)$$

Het is nu gewenst dat de getallen waarover in (*) het maximum wordt genomen alle niet te ver van elkaar aflaggen, opdat op alle niveaus $s=2, 3, \dots, k$ met een ongeveer evengrote simultane onbetrouwbaarheid wordt getoetst.

Ter bepaling van de kritieke waarden van de stapsgewijze MQ toets is in de appendix een tabel van de exacte verdeling van MQ_s voor $s = 2, 3, 4$ en $n = 10(2)20(5)30$ gegeven. Voor $s = 2$ is deze bijvoorbeeld af te leiden uit Lieberman & Owen, voor $s > 2$ berekend zoals in 3.2 besproken. Voor $s > 4$ nemen we aan dat de verdeling voldoende continu is om een asymptotische benadering te gebruiken.

4. VERGELIJKINGEN IN ONDERSCHIEDINGSVERMOGEN

We zullen nu de resultaten van enkele simulaties bespreken. Het aantal simulaties per alternatief bedraagt minimaal 500. We onderzochten hoe de verschillen in onderscheidingsvermogen liggen bij

- gebruik van dezelfde scores, maar verschillende klasse van toets
- gebruik van dezelfde klasse, doch met verschillende scores
- bij uniforme, dubbele exponentiële of Cauchy verschuivingsalternatieven.

In principe zullen we $k = 4$, $n = 10$ beschouwen en zoals al gemeld voornamelijk de bruikbaarheid van mediaanscores onderzoeken, vandaar dat we meestal bij dubbelexponentiële of Cauchy alternatieven kijken. We beschouwen altijd het onderscheidingsvermogen als functie van de exacte onbetrouwbaarheid, zonodig geïnterpoleerd. We noteren

$H(VD; \mu_1, \mu_2, \dots, \mu_k)$: verschuivingsalternatief van een VD verdelingssoort met μ_i de verwachting van de i^e populatie.

We nemen voor de DE, CY en UFO verdeling respectievelijk $f(x) = \frac{1}{2} \ln 10 \cdot 10^{-|x|}$, $(\pi(1+x^2))^{-1}$ op $\mathbb{R}$ en 1 op $[0,1)$ en translaties daarvan.

Wanneer we twee toetsen A en B vergelijken dan bedoelen

we met $A \gtrsim B$ dat A qua onderscheidingsvermogen evengoed of beter is, met $A \sim B$ dat ze ongeveer even goed zijn, enzovoorts.

We zullen drie soorten spreiding van gemiddelde bekijken, nl:

- $\mu_i = \mu_1 + ia$, a constant: gelijkgespreide μ_i 's
- $\mu_1 = \mu_2 = \mu_3 \neq \mu_4$, spreiding met uitschieter
- $\mu_1 = \mu_2 \neq \mu_3 = \mu_4$, minimale range spreiding,

deze spreiding heeft minimale range in de klasse van μ_i 's waarbij

$$\sum_{i=1}^k (\mu_i - \sum_{j=1}^k \mu_j / k)^2 \text{ constant is.}$$

4.1. Het toetsen van H_0

4.1.1. Een vergelijking tussen P,Q en R toetsen bij gebruik van dezelfde scores

In tabel 4.1 staat het onderscheidingsvermogen van enkele Wilcoxon score toetsen bij verschillende parameters en alternatieven bij enkele overschrijdingskansen α .

Alternatief	WQ	WR	WP	WQ	WR	WP	WQ	WR	WP	WQ	WR	WP
k = 4 n = 10	$\alpha = .0063$			$\alpha = .019$			$\alpha = .044$			$\alpha = .096$		
H(DE;0,.2,.4,.6)	.25	.23	.20	.38	.37	.35	.51	.49	.47	.65	.63	.62
H(DE;0,0,0,.5)	.22	.20	.18	.35	.35	.34	.48	.47	.46	.62	.62	.60
H(DE;0,0,.5,.5)	.34	.24	.23	.49	.40	.41	.61	.53	.55	.74	.70	.69
H(UFO;0,.1,.2,.3)	.15	.13	.07	.27	.24	.21	.40	.38	.35	.56	.55	.50
H(UFO;0,0,0,.3)	.22	.21	.16	.37	.35	.32	.48	.48	.46	.63	.63	.60
H(UFO;0,0,.3,.3)	.37	.25	.17	.51	.42	.39	.63	.56	.54	.75	.71	.67
H(CY;0,.8,1.6,2.4)	.27	.26	.21	.39	.38	.36	.51	.51	.47	.66	.63	.61
H(CY;0,0,0,1.75)	.15	.16	.15	.28	.28	.28	.42	.40	.40	.56	.56	.55
H(CY;0,0,1.7,1.7)	.24	.16	.18	.36	.30	.33	.50	.43	.45	.62	.58	.59
k = 4 n = 20	$\alpha = .009$			$\alpha = .026$			$\alpha = .051$			$\alpha = .095$		
H(DE;0,.11,.22,.33)	.18	.16	.16	.29	.27	.25	.38	.37	.37	.53	.51	.49
H(DE;0,0,.3,.3)	.29	.22	.23	.42	.36	.36	.54	.51	.49	.68	.64	.63
H(CY;0,0,1,1)	.23	.18	.20	.35	.30	.33	.46	.42	.44	.59	.57	.57

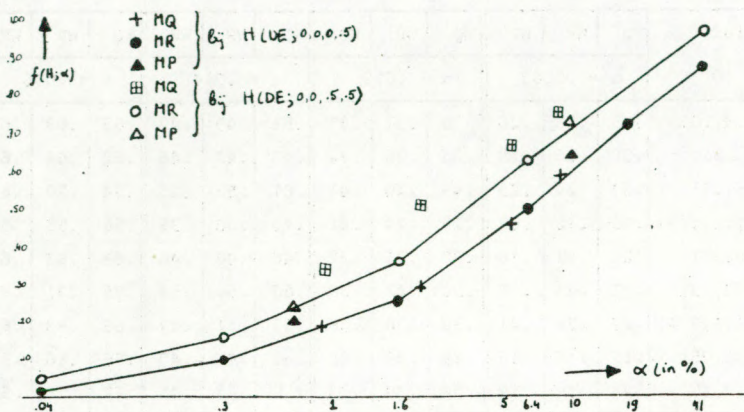
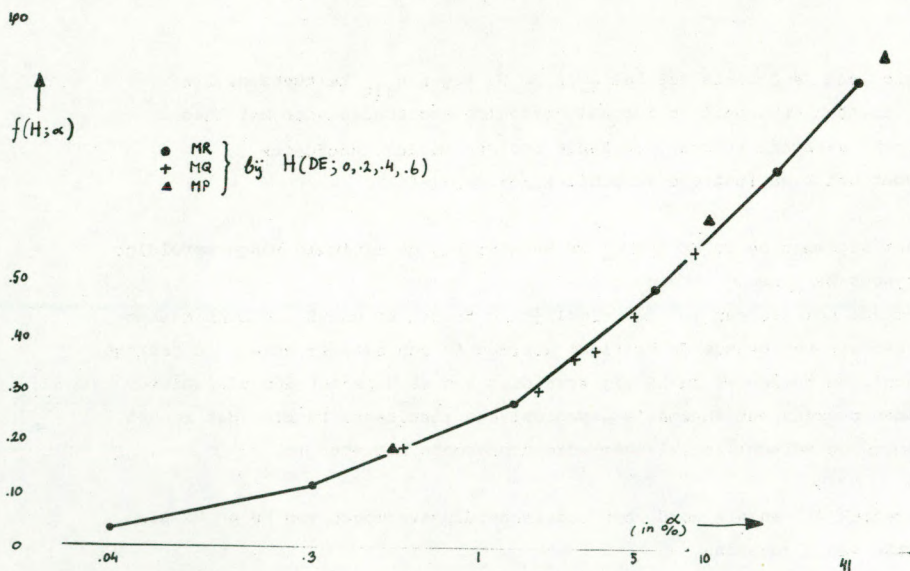
Tabel 4.1 $f(H; T(\alpha))$ bij verschillende H,T en α .

We zien dat de Q toets het beste is om H_0 tegen H_{alt} te toetsen. Zie voor toetsen gebaseerd op normaal verdeelde grootheden voor het theoretische verschil tussen Q en Range toetsen Miller, hoofdstuk 2, en voor het kwantitatieve verschil Einot/Gabriel.

In het algemeen geldt $WQ > WR \gtrsim WP$ behalve bij de minimale range spreiding waar soms $WR \lesssim WP$.

Dit stemt ook overeen met de vergelijking tussen WP en WR die Dunn maakte bij gebruik van benaderde kritieke waarden in een extreem geval. De geringe verschillen tussen WP en WR bij spreiding van de μ_i 's met één uitschieter stemmen overeen met Sherman's asymptotische resultaat. We zien dat in het algemeen de verschillen kleiner worden naarmate n groter is.

In grafiek 4.1 en 4.2 wordt het onderscheidingsvermogen van MP en MR als functie van α gegeven.

Grafiek 4.1. Onderscheidingsvermogen bij mediaanscores ($k=4$, $n=10$)

Grafiek 4.2. Onderscheidingsvermogen bij mediaanscores

Vergelijken we deze grafieken met tabel 4.1 dan zien we dat de verschillen tussen P,Q en R voor mediaanscores bijna hetzelfde zijn als bij Wilcoxon scores. De verschillen zijn gering, MQ is alleen bij een spreiding van de μ_1 's met minimale range beduidend beter dan MP of MR. Een verschil met Wilcoxon scores ligt hierin dat in het algemeen $MP \gtrsim MR$ is.

4.1.2. Een vergelijking tussen Wilcoxon en Mediaan scores bij dezelfde soort toetsingsgrootheid

Tabel 4.2 geeft een vergelijking tussen het onderscheidingsvermogen van MP en WP voor enkele gevallen.

Alternatief	α	WP	MP	α	WP	MP
k=2, n=10						
H(DE;0,.5)	.023	.34	.37	.179	.74	.72
H(CY;0,2)	.023	.37	.46	.179	.71	.80
n=20, k=2						
H(DE;0,.25)	.026	.22	.23	.112	.45	.47
H(CY;0,1.5)	.026	.48	.64	.112	.75	.84
k=4, n=10						
H(DE;0,.2,.4,.6)	.006	.19	.18	.101	.63	.63
H(DE;0,0,0,.5)		.17	.18		.63	.62
H(DE;0,0,.5,.5)		.22	.23		.70	.72
H(UFO;0,.1,.2,.3)		.07	.05		.50	.31
H(UFO;0,0,0,.3)		.13	.07		.63	.39
H(UFO;0,0,.3,3)		.17	.09		.68	.47
H(CY;0,.8,1.6,2.4)		.20	.26		.63	.72
H(CY;0,0,0,1.75)		.14	.21		.57	.65
H(CY;0,0,1.7,1.7)	.006	.17	.22	.101	.60	.73
k=4, n=20						
H(DE;0,.11,.22,.33)	.022	.24	.26	.125	.55	.56
H(DE;0,0,.3,.3)		.33	.38		.67	.69
H(CY;0,0,1,1)	.022	.31	.39	.125	.62	.70

Tabel 4.2 $f(H; P(\alpha))$ bij verschillende scores, alternatieven en onbetrouwbaarheid.

Uit de resultaten van Hajek volgt dat voor $k = 2$ en voor $\Delta \rightarrow 0$ waarbij Δ de verschuivingsparameter is $WP < MP$ is bij dubbelexponentiële of Cauchy verschuivingsalternatieven. Uit tabel 4.2 zien we dat

$WP \sim MP$ bij $n = 10$ en DE alternatieven
 $WP \sim MP$ bij $n = 20$ en DE alternatieven
 $WP < MP$ bij $n = 10, 20$ en CY alternatieven

geldt, en wel voor $k = 2$ en $k = 4$ in ongeveer dezelfde mate.

In 4.1.1. zagen we dat als het niet een spreiding van de μ_1 's met minimale range betreft dat

$$WQ \gtrsim WR \gtrsim WP \quad \text{en} \quad MP \gtrsim MR \sim MQ$$

geldt, dus in die gevallen is bij $n = 10$ en DE alternatieven

$$WQ \gtrsim WR \gtrsim WP \sim MP \gtrsim MR \sim MQ.$$

Voor een spreiding van de μ_1 's met minimale range blijkt dat $WQ \sim MQ$ en dus dat

$$WQ \sim MQ > WR \gtrsim WP \sim MP \sim MR$$

bij $n = 10$ en DE alternatieven. We concluderen dat bij $n = 10$ en DE alternatieven mediaanscores geen voorkeur verdienen. Voor $n = 20$ is het resultaat voor mediaanscores iets beter maar het verschil is zo gering dat de Wilcoxon scores nog de voorkeur genieten door hun betere asymptotische benadering en hun grotere keuze van in te stellen onbetrouwbaarheid. Bij Cauchy alternatieven geldt

$$MQ \gtrsim MP \gtrsim MR \gtrsim WQ \gtrsim WR \gtrsim WP$$

waarbij het verschil in onderscheidingsvermogen tussen MQ en WQ circa 15%, en tussen MP en WP circa 20% is.

Nog eenmaal naar tabel 4.2 kijkend zien we dat Wilcoxon scores veel beter dan mediaanscores zijn bij uniforme verschuivingsalternatieven, zoals te verwachten.

We zien dat de winst van mediaanscores in de daartoe geëigende gevallen gering is, terwijl verlies bij verkeerd gebruik groot is. Wilcoxon scores scoren nooit erg slecht.

4.2. Paarsgewijze populatietoetsen

Hieronder wordt het onderscheidend vermogen van parenvergelijkende toetsen vergeleken, in eerste instantie WP, WQ en WR en vervolgens worden enige opmerkingen over MP, MQ en MR gemaakt. We beperken ons tot het geval $k = 4$, $n = 10$ en een dubbele exponentieel verschuivingsalternatief.

4.2.1. Vergelijking van WP, WQ en WR

In tabel 4.3 zien we dat de stapsgewijze WQ over de gehele linie beter is dan WP of WR wat betreft per-pair power en mean-pair power. Voor een betere any-pair power van WQ zie de resultaten uit 4.1.

Over het algemeen is WP beter dan WR. Uitzondering hierop is dat de any-pair power (zie tabel 4.1) en de per-pair power van de verstuiteenliggende populaties bij WR iets beter is. Merk op dat de all-pair power van WR bij een alternatief met $\mu_i \neq \mu_j$ voor alle $i \neq j$, bij $k = 4$, $n = 10$ voor $\alpha_p \leq .20$ altijd nul is.

Samenvattend concluderen we dat de stapsgewijze WQ toets van de drie genoemden het beste is om populaties paarsgewijs verdelingsvrij te vergelijken. WP voldoet iets minder goed in onderscheidingsvermogen, maar maakt het mogelijk om simultane betrouwbaarheids intervallen te vinden. WR is in bijna alle opzichten inferieur aan WQ, met als enige uitzondering de geringe hoeveelheid benodigd rekenwerk.

4.2.2. Vergelijking van MP, MQ en MR

De vergelijking tussen MP, MQ en MR valt ongeveer hetzelfde uit als die bij WP, WQ en WR in de vorige paragraaf. We maken een tweetal opmerkingen. In de eerste plaats merken we op dat MR in nog sterkere mate inferieur is aan MP en MQ dan bij Wilcoxon scores. In geen enkel onderzocht geval heeft MR een beter soort power dan MP of MQ getoond. Vooral het onderscheidend vermogen van populatieparen aan dezelfde kant van de overall mediaan is bij MR zeer gering (zie grafiek 4.3) hetgeen begrijpelijk is omdat voor waarnemingen uit beide populaties de score vaak hetzelfde is. In plaats van

$$a_M(r) = 1 \text{ als } r \leq kn/2, \text{ 0 elders,}$$

Alternatief H	α_p	$f_{PP}(H; i, j; T_P(\alpha_p))$																		$f_{MP}(H; T_P(\alpha_p))$																				
		i=1			j=2			i=1			j=3			i=1			j=4			i=2			j=3			i=2			j=4			i=3			j=4					
		WQ	WR	WP	WQ	WR	WP	WQ	WR	WP	WQ	WR	WP	WQ	WR	WP	WQ	WR	WP	WQ	WR	WP	WQ	WR	WP	WQ	WR	WP	WQ	WR	WP	WQ	WR	WP						
H(DE;0,.2,.4,.6)	2	2	1	2	9	8	10	24	28	23	2	1	2	10	8	10	3	1	2	8	8	8																		
	5	6	2	4	19	15	17	39	42	36	7	3	4	20	14	17	7	2	4	16	13	14																		
	10	10	4	7	29	23	23	54	54	48	11	6	7	31	22	24	13	3	6	25	19	19																		
H(DE;0,.4,.8,1.2)	2	15	1	10	54	40	46	85	93	85	14	3	10	52	39	47	12	1	10	39	31	35																		
	5	26	3	17	72	53	62	95	96	92	25	6	17	72	52	60	24	4	18	52	36	44																		
	10	36	8	24	83	67	74	98	98	96	37	12	24	83	65	71	39	9	25	63	43	52																		
H(DE;0,.6,1.2,1.8)	2	34	1	24	87	71	81	99	100	98	32	3	24	85	69	79	31	1	21	61	41	55																		
	5	52	4	36	97	85	91	100	100	100	51	8	31	94	84	91	51	2	36	74	47	65																		
	10	66	9	49	98	93	96	100	100	100	64	17	49	98	91	94	66	7	48	82	53	73																		
H(DE;0,0,0,.5)	2	1	0	0	0	0	0	17	16	16	0	0	1	17	17	18	16	17	16	17	17	17																		
	5	1	1	1	2	1	1	31	26	28	1	0	1	31	26	27	30	27	27	31	26	27																		
	10	3	1	1	4	1	1	46	37	37	3	2	2	46	37	37	44	38	37	45	37	37																		
H(DE;0,0,.5,.5)	2	1	0	0	17	17	16	18	15	14	17	19	18	17	17	16	1	0	1	17	17	16																		
	5	2	0	1	32	27	29	32	25	24	31	27	29	33	26	26	2	0	2	32	26	27																		
	10	6	1	3	44	39	38	44	34	36	45	39	40	46	36	36	4	1	2	45	37	37																		
n = 20 H(DE;0,.4,.8,1.2)	1	31	4	20	91	78	85	100	100	100	31	6	19	91	80	87	28	3	19	62	45	55																		
	5	57	12	37	96	92	94	100	100	100	53	20	39	98	93	96	52	13	39	76	55	67																		
	10	66	23	49	99	96	97	100	100	100	66	30	50	100	97	98	66	24	52	83	61	75																		

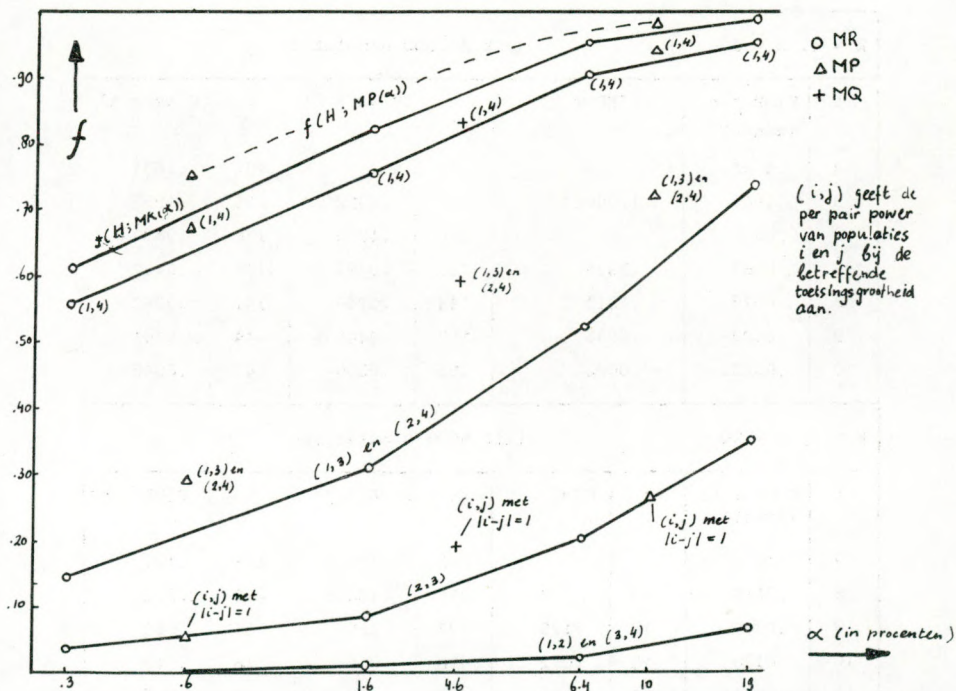
Tabel 4.3 Per Pair Power en Mean Pair Power bij verschillende toetsingsgrootheden T_p met Wilcoxon scores bij enkele alternatieven H en onbetrouwbaarheden α_p . $k=4$, $n=10$ (tenzij anders $n=20$ vermeld). f en α_p in procenten

zouden de scores

$$a_M(r) = k - i \text{ als } (i - 1) n < r \leq in$$

moeten worden gekozen.

In de tweede plaats merken we nog eens op dat het gedrag van grootheden met mediaanscores zeer discreet is. Dit had tot gevolg dat de stapsgewijze MQ enigszins conservatief is (zie 3.3) en hierdoor is noch MP noch MQ in het algemeen beter voor paarsgewijze populatietoetsen. De keuze tussen deze toetsen zal ook afhangen van de gewenste onbetrouwbaarheid α_p , als we tenminste het procedé van randomiseren willen vermijden. Bijvoorbeeld voor $k = 4$, $n = 20$ zagen we in 3.3.3. dat we bij MQ kunnen kiezen uit $\alpha_p = .051$ of $\alpha_p = .008$. Uit de tabel in de appendix zien we dat gebruik van MP alleen een α_p van .125 of .022 toelaat.



Grafiek 4.3. Onderscheidingsvermogen van paarsgewijze mediaantoetsen bij $H=H(DE; 0, .4, .8, 1.2)$

4.2.3. Slotopmerkingen

Tot nu toe is in 4.2 alleen het dubbelexponentiële verschuivingsalternatief beschouwd. We nemen aan dat bij alternatieven met andere verdelingen dezelfde tendens geldt als al in 4.1 geconstateerd werd bij het toetsen van H_0 . De vergelijkingen tussen Wilcoxon en mediaanscores bij DE alternatieven bevestigen dit, vergelijk de resultaten uit tabel 4.3 bij $H(DE; 0, .4, .8, .1.2)$, $k = 4$, $n = 10$ met grafiek 4.3.

Voor de parenvergelijkende toetsen houdt dit dan ook weer in dat het gebruik van mediaanscores slechts dan voordeel heeft als de gebruiker er sterk van overtuigd is dat hij te maken heeft met verdelingen met een zeer lange staart.

APPENDIXAppendix 1

De verdeling onder H_0 .

k = 4, n = 10		(uit 20.000 simulaties)					
t	P(MR ≥ t) (exact)	P(MP ≥ t)		t	P(WR ≥ t)	t	P(WP ≥ t)
4	.4102	1.0000		103	.2052	132	.1831
5	.1867			119	.1000	135	.1057
6	.0635			133	.0491	136	.0877
7	.0160	.5335		144	.0240	139	.0421
8	.0028	.1013		157	.0100	141	.0247
9	.0003	.0056		165	.0050	144	.0101
10	.00002	.0000				147	.0046

k = 4, n = 20		(uit 5000 simulaties)					
t	P(MR ≥ t) (exact)	t	P(MP ≥ t)	t	P(WR ≥ t)	t	P(WP ≥ t)
7	.1673			295	.193	485	.191
8	.0815	13	.399	335	.100	496	.093
9	.0346	14	.125	377	.050	505	.050
10	.0128	15	.022	428	.020	516	.019
11	.0041	16	.002	455	.010	523	.010
12	.0012	17	.000	480	.005	532	.005

Appendix 2

n:	10	12	14	16	18	20	25	30
k = 2	82 .11	122 .011	148 .042	178 .11	234 .015	270 .036	397 .054	548 .066
	68 2.30	104 .33	130 .70	160 1.21	212 .22	250 .38	373 .42	522 .41
	58 17.9	90 3.91	116 5.70	146 7.56	194 1.84	232 2.56	353 2.27	500 1.94
		80 22.0	106 25.7	136 28.9	180 9.43	218 11.3	337 8.87	482 6.98
k = 3			209 .011	264 .012	315 .039	374 .074	548 .49	771 .22
	117 .017	162 .009	203 .033	254 .040	305 .11	372 .096	546 .64	761 .36
	113 .060	158 .017	201 .049	248 .11	299 .24	362 .22	542 1.05	753 .61
	107 .16	150 .10	197 .081	246 .15	297 .33	356 .46	534 1.45	749 .93
	101 .89	146 .27	189 .30	242 .22	293 .47	354 .61	532 1.93	747 1.11
	99 1.44	140 .54	185 .71	234 .66	285 1.18	350 .83	530 2.48	737 1.83
	93 3.42	134 2.06	179 1.22	230 1.39	281 2.29	342 1.86	526 3.25	731 2.91
	89 12.2	132 3.13	173 3.66	224 2.18	275 3.39	338 3.40	524 3.69	729 3.53
	83 26.2	126 6.09	171 5.28	218 5.56	269 7.68	332 4.81	522 4.75	725 4.34
		122 17.5	165 9.11	216 7.74	267 10.4	326 9.93	516 6.47	717 7.10
k = 4			161 22.5	210 12.3	261 15.5	324 13.0	514 9.47	713 10.7
			250 .14	312 .23	386 .35	464 .48	708 .39	990 .77
		190 .12	248 .17	310 .40	382 .40	462 .70	706 .51	988 .81
	138 .16	188 .16	246 .28	308 .45	380 .52	458 .78	702 .67	986 1.05
	136 .21	186 .24	244 .29	306 .69	378 .83	456 .98	700 .78	984 1.16
	134 .34	184 .32	242 .41	304 .72	376 .93	454 1.48	698 .89	982 1.29
	132 .39	182 .62	240 .50	302 .96	374 1.36	452 1.63	694 1.21	980 1.36
	130 .79	180 .79	238 .71	300 1.12	372 1.40	450 2.27	692 1.31	978 1.53
	126 2.13	178 1.16	236 .88	298 1.52	370 1.80	448 2.35	690 1.74	976 1.62
	124 2.76	176 1.30	234 1.50	296 1.80	368 2.05	446 2.91	686 2.18	974 2.16
	122 3.68	174 2.15	232 1.84	294 2.82	366 2.66	444 3.26	684 2.72	972 2.35
	120 5.23	170 4.61	230 2.56	292 3.36	364 3.06	442 4.10	682 3.13	970 2.63
	118 8.72	168 5.72	228 2.80	290 4.46	362 4.52	440 4.63	678 4.00	968 2.95
	116 9.62	166 7.25	226 4.18	288 4.82	360 5.27	438 6.53	676 4.69	966 3.70
	114 19.6	164 9.51	222 7.79	286 6.73	358 6.77	436 7.49	674 5.58	964 3.75
	112 22.0	162 14.2	220 9.37	282 11.4	356 7.24	434 9.36	670 7.22	962 4.72
		160 15.3	218 11.4	280 13.4	354 9.64	432 9.93	668 7.85	958 5.04
			216 14.3	278 15.9	350 15.2	430 12.8	666 10.9	956 5.77

De verdeling van MQ onder H_0 voor enkele k en n . Getabelleerd is x en $P(MQ \geq x)$ (in procenten).

LITERATUUR:

- Carmen S.G & Swanson M.R.: An evaluation of ten pairwise multiple comparisons procedures by Monte Carlo methods
Comm. Stat. Simul. Computat. 1978, B7, pg. 117-128
- Dunn, Olive Jean: Multiple comparisons using rank sums,
Technometrics 6, 1964, pg 241-252

- Einot, Israel & Gabriel K.R.: A study of the powers of several methods for multiple comparisons,
JASA 70, 1975, pg 574-583
- Hajek, Jaroslav: A Course in nonparametric Statistics
Holden-day, 1969
- Lehmann, E.L.: Nonparametrics
Holden-day, San Francisco, 1975
- Lieberman, G.J. & Owen D.B.: Tables of the hypergeometrical probability distribution, Stanfor Univ. Press
- Lin, F.A. & Haseman J.K.: An evaluation of some nonparametric multiple comparisons procedures by Monte Carlo methods,
Comm. Stat-Simul Comput. 1978, B7, pg 117-128
- Miller, R.G.: Simultaneous Statistical inference,
McGraw-Hill, New York 1966
- Oude Voshaar, J.H.: Asymptotic $(k-1)$ -mean significance levels of a non-parametric method for multiple comparisons in the k -sample case,
Memorandum COSOR 77-18, Eindhoven, University of Technology
- Ramsey, Philip H: Power differences between pairwise multiple comparisons procedures by Monte Carlo methods,
JASA 68, pg 66-74
- Sherman, Ellen: A note on multiple comparisons using rank sums,
Technometrics 7, 1965, pg 255-256
- Steel, R.G.D.: A rank sum test for comparing all pairs of treatments,
Technometrics 2, 1960, pg 197-207
- Tobach, Smith, Rose & Richter: A table for making rank sum multiple paired comparisons,
Technometrics 9, 1967, pg 561 e.v.