Enkele problemen bij het op gelijkheid toetsen van een aantal lokatieparameters

Eindhoven University of Technology
Computing Centre Note 36

Enkele problemen bij het op gelijkheid toetsen van een aantal lokatieparameters.

Rudy Douven

Een stage statistische analyse
onder leiding van Prof.dr. R. Doornbos en
drs. J.B. Dijkstra.

Mei 1987.

Inhoud: bladzijde

1. Het gedrag van enkele rangtoetsen en normale verdelingen bij kleine steekproefgroottes.

1.1. Inleiding.

De in dit verslag voorkomende rangtoetsen zijn ontwikkeld om een toetsingsgrootheid te bepalen waarmee men de nulhypothese H_0: "de steekproeven zijn afkomstig uit dezelfde verdeling" al dan niet kan verwerpen. In het onderstaande wordt onderzocht hoe de toetsen zich gedragen bij de alternatieve nulhypothese H_0' van gelijke populatiegemiddelden μ_i. Dit wordt onderzocht voor normale verdelingen bij kleine steekproefgroottes. De hier gebruikte rangtoetsen hebben als gemeenschappelijke eigenschap dat alle waarnemingen, bestaande uit k steekproeven, als één geheel worden behandeld. Dat wil zeggen dat de rangnummerstoekenning onafhankelijk gebeurt van de steekproef waaruit de waarneming komt.

Laat $x_1 \ldots x_N$ een combinatie zijn van K steekproeven. R_i geeft het rangnummer van elke waarneming x_i aan en S_j geeft de indices aan van steekproef j. n_j is de bijbehorende steekproefgrootte $(i = 1, \ldots, N$ en $j = 1, \ldots, k)$.

De drie onderzochte rangtoetsen zijn:

1) De toets van Kruskal & Wallis (1952):

$$Q_1 = \frac{12}{N(N+1)} \sum_{j=1}^{K} n_j(\bar{R}_j - \bar{R})^2 .$$

Hierin is $\bar{R} = \frac{N+1}{2}$ en $\bar{R}_j$ is het gemiddelde van de rangnummers in de j-de groep.

2) De toets van Van der Waerden:

$$Q_2 = \frac{N-1}{h} \sum_{j=1}^{k} \frac{1}{n_j} \left[\sum_{i \in S_j} \Phi^{-1}\left(\frac{R_i}{N+1}\right)\right]^2 .$$

Hierin is $h = \sum_{i=1}^{N} [\Phi^{-1} (\frac{i}{N+1})]^2$ en Φ^{-1} is de inverse functie van de standaard normale verdeling.

3) De toets van Mood & Brown (1950):

$$Q_3 = 4 \sum_{j=1}^{k} \frac{1}{n_j} A_j^2 - N .$$

Hierin is A_j het aantal waarnemingen in steekproef j met een rangnummer groter dan $\frac{1}{2}(N+1)$.

Hun gedrag onder H_0 is voor alle drie de toetsen dat ze asymptotisch verdeeld zijn als χ^2_{k-1}. Het gedrag van de drie toetsen als H_0 niet geldt is echter verschillend. Kruskal & Wallis is asymptotisch optimaal voor een logistische verdeling, Van der Waerden asymptotisch optimaal voor een normale verdeling en Mood & Brown voor een dubbel exponentiële verdeling.
Deze informatie is echter vooral nuttig indien men te maken heeft met grote steekproefgroottes. In het komende gedeelte wordt het gedrag onderzocht voor kleine steekproefgroottes. Iedere steekproef had minimaal 6 waarnemingen en maximaal 15 waarnemingen uit een normale verdeling. Verder hebben we K = 3 genomen. Dit vanwege tijdgebrek maar ook omdat het waarschijnlijk is dat het gedrag voor grotere K's niet wezenlijk zal verschillen.

1.2. Het gedrag onder H_0'.

Met simulatie hebben we de kans van het aantal verwerpingen van H'_0, indien waar, geschat met 2500 herhalingen.
De steekproeven werden gegenereerd uit normale verdelingen met $\mu_i = 0$ en $\sigma_i^2 = 1$. H_0' wordt verworpen indien $Q_i > \chi_2^2(\alpha)$ $(i = 1, 2, 3)$. Voor de nominale onbetrouwbaarheid α kozen we 0.05. In tabel 1 staan 12 verschillende patronen voor $n_1 = n_2 = n_3$ en in tabel 2 staan 12 verschillende patronen voor $n_1 > n_2 > n_3$. De geschatte standaarddeviatie van de gerealiseerde onbetrouwbaarheid is $[(0.05 \times 0.95)/2500]^{\frac{1}{2}} = 0.004359 = 0.4359\%$.
Laat d nu het percentage van verworpen hypothesen zijn min 5, gedeeld door de standaarddeviatie. In tabel 1.2 is voor elk patroon de waarde van d weergegeven. In de tabel worden drie categorieën onderscheiden door stippellijnen: $d < -2$ (conservatief), $-2 < d < 2$ (accuraat), $2 < d$ (progressief).

K&W	VDW	M&B		K&W	VDW	M&B
		2	$d < -4$			1
		1	$-4 < d < -3$	1	2	
1	4	2	$-3 < d < -2$	1		4
5	4	2	$-2 < d < -1$	3	3	1
6	4	2	$-1 < d < 0$	5	6	2
		1	$0 < d < 1$	2	1	1
			$1 < d < 2$			1
			$2 < d < 3$			1
		1	$3 < d < 4$			1
		1	$4 < d$			

<u>tabel 1</u>: $n_1=n_2=n_3$, $\sigma_1=\sigma_2=\sigma_3$

<u>tabel 2</u>: $n_1>n_2>n_3$, $\sigma_1=\sigma_2=\sigma_3$

Tabel 1 en 2 suggereren de volgende conclusies:

- De toetsen van Kruskal en Wallis en Van der Waerden zijn ietwat conservatief voor kleine n_i bij gebruik van de χ^2-verdeling.
- De toets van Mood & Brown geeft slechte resultaten en lijkt ongeschikt voor steekproefgroottes kleiner dan 15.
- De invloed van verschillende steekproefgroottes ten opzichte van gelijke steekproefgroottes is bijzonder gering.

Tabel 3, 4 en 5 geven weer voor elke toets 12 patronen weer. Nu echter met $\sigma_1 < \sigma_2 < \sigma_3$. We kozen voor de combinaties $\sigma_1 = 1$, $\sigma_2 = 2$, $\sigma_3 = 3$ en $\sigma_1 = 1$, $\sigma_2 = 3$, $\sigma_3 = 5$.

	tabel 3 K&W	tabel 3 VDW	tabel 3 M&B	tabel 4 K&W	tabel 4 VDW	tabel 4 M&B	tabel 5 K&W	tabel 5 VDW	tabel 5 M&B
$d < -4$	5	7							
$-4 < d < -3$	2								
$-3 < d < -2$		4				1			
$-2 < d < -1$	2	1				2			1
$-1 < d < 0$	2		1		4				1
$0 < d < 1$	1		1	1	4	1			1
$1 < d < 2$			1	2	3			1	1
$2 < d < 3$			2	5	1	1	1		1
$3 < d < 4$			2	1		1			2
$4 < d$			5	3		6	11	11	5

<u>tabel 3</u>: $n_1<n_2<n_3$, $\sigma_1<\sigma_2<\sigma_3$ <u>tabel 4</u>: $n_1=n_2=n_3$, $\sigma_1<\sigma_2<\sigma_3$ <u>tabel 5</u>: $n_1>n_2>n_3$, $\sigma_1<\sigma_2<\sigma_3$

Tabel 3, 4 en 5 suggereren de volgende conclusies:

- De toetsen verwerpen de ware nulhypothese H_0' te vaak. Slecht de Van der Waerden toets bij gelijke groepgroottes lijkt enigszins geschikt. Bij alle andere toetsingen ligt niet eens 50% van de d's in het betrouwbaarheidsgebied.
- Mood & Brown is progressief bij dit gebruik.
- Zowel voor de toets van Kruskal & Wallis als voor de toets van Van der Waerden geldt dat de toets conservatie is als kleine varianties optreden bij kleine steekproefgroottes en progressief is als grote varianties optreden bij kleine steekproefgroottes.

1.3. <u>Verklaringen.</u>

Dat de toetsen de ware nulhypothese H_0' bij ongelijke varianties nogal vaak verwerpt, komt omdat de toetsen ontworpen zijn voor de nulhypothese H_0: alle K steekproeven komen uit dezelfde verdeling. In een eerdere publikatie van J.B. Dijkstra en P.S.P.J. Werter: "Het gebruik van de toets van Kruskal & Wallis bij normale verdelingen met ongelijke varianties" wordt het gebruik van de nulhypothese H_0' verklaard. Bij symmetrische verdelingen lijken verschillen in de varianties hoegenaamd geen invloed te hebben op Q1. Hun onderzoekingen toonden echter het slechte gedrag van de toets van Kruskal & Wallis, bij het gebruik van H_0', al aan voor kleine steekproefgroottes. De toetsen van Van der Waerden en Mood & Brown blijken H_0' ook vaak te verwerpen.

De verklaring voor de laatste conclusie gebeurt in stapjes. Keren we terug naar de toets van Van der Waerden:

$$Q_2 = \frac{N-1}{h} \sum_{j=1}^{k} \frac{1}{n_j} \left[\sum_{i \in s_j} \Phi^{-1} \left(\frac{R_i}{N+1}\right)\right]^2$$ en stel nu:

$$A_j^2 = \left[\sum_{i \in s_j} \Phi^{-1} \left(\frac{R_i}{N+1}\right)\right]^2$$ dan is het variabele gedeelte in Q_2, de som:

$\sum_{j=1}^{k} \frac{1}{n_j} A_j^2$. In tabel 6 staan enkele verwerpingspercentages van Q_2 zoals we hebben gevonden in het simulatie-onderzoek.

n_1,	n_2,	n_3	$\sigma_1=1$, $\sigma_2=1$, $\sigma_3=1$	$\sigma_1=1$, $\sigma_2=2$, $\sigma_3=3$
8	8	8	4.60	5.16
10	10	10	4.40	4.68
12	12	12	4.44	5.04
14	14	14	4.64	5.20

<u>tabel 6</u>: de Van der Waerden toets.

Bij gelijke n_i's wordt het variabele gedeelte in Q_2: $A_1^2 + A_2^2 + A_3^2$. Uit tabel 6 is af te lezen dat deze som kleiner is bij gelijke varianties dan in het geval van ongelijke varianties. Dit heeft te maken met het gewicht dat elk rangnummer, in de A_i, krijgt. Rangnummers dicht in de buurt van ½(N+1) krijgen een lichter gewicht mee dan rangnummers die hier ver vanaf liggen. In het geval van gelijke varianties is de verwachting dat rangnummers gelijkmatiger over de drie groepen verspreid liggen dan in het geval van ongelijke varianties. Dan zullen verhoudingsgewijs in groep 3 meer rangnummers met een zwaarder gewicht zitten. De A_i's zullen zich dan ook als volgt verhouden: $A_1^2 < A_2^2 < A_3^2$. Met deze wetenschap is tabel 7 ook niet erg verrassend. Door een kleiner aantal waarnemingen in groep 3 krijgt $\frac{1}{n_3} A_3^2$ een extra groot aandeel in de Q_2.

Het gegeven grote variantie bij kleine groepsgrootte lijkt elkaar

n_1, n_2, n_3	$\sigma_1=1$, $\sigma_2=2$, $\sigma_3=3$	n_1, n_2, n_3	$\sigma_1=1$, $\sigma_2=3$, $\sigma_3=5$
6 8 10	3.00	6 8 10	2.92
8 8 8	5.16	8 8 8	5.68
10 8 6	7.52	10 8 6	8.40

tabel 7: de Van der Waerden toets

in de Q_2 te versterken. Omdat $\mathbb{E}A_i = 0$ ($i = 1, 2, 3$) zal de afwijking van A_i van 0 afhangen van de groepsgrootte n_i. Deze afwijking zal groter zijn naarmate n_i kleiner is. Combineren we dit met het gegeven $A_1^2 < A_2^2 < A_3^2$ dan zal in het geval $n_1 < n_2 < n_3$, Q_2 dus versterkt worden ten opzichte van Q_2 in het geval $n_1 = n_2 = n_3$. En andersom zal Q_2 verzwakt worden bij $n_1 > n_2 > n_3$ ten opzichte van het geval $n_1 = n_2 = n_3$.
De Kruskal & Wallis toets vertoont een identiek gedrag. Dit is al eens eerder aangetoond en wordt bewezen in een te verschijnen publikatie van J.B. Dijkstra.
De Mood & Brown toets kent zo'n gedrag niet. Hier krijgen de rangnummers in een groep ook slechts twee soorten gewichten mee: 1 als rangnummer $> \frac{1}{2}(N+1)$ is of 0 als het rangnummer $\leq \frac{1}{2}(N+1)$ is.

1.4. Conclusie.

Voor de nulhypothese H_0' van gelijke populatiegemiddelden μ_i lijken de onderzochte toetsen ongeschikt. Dit had te maken met de kleine steekproefgroottes. Slechts de Van der Waerden toets bij gelijke steekproefgroottes is het waard om verdere onderzoekingen te ondergaan.

2. Enkelvoudige variantie-analyse in de aanwezigheid van uitschieters.

2.1. Inleiding.

In een eerdere publikatie van J.B. Dijkstra en H. Linders: "Comparison of several mean values in the presence of outliers" kwam Huber's methode voor enkelvoudige variantie-analyse in de aanwezigheid van uitschieters als beste uit de bus. De Huber toets werd hier vergeleken met 4 andere toetsen waaronder één rangtoets; de Van der Waerden toets. De toetsen werden toegepast op data waarbij de fractie uitschieters klein is maar de waarde van de uitschieter zelf groot is ten opzichte van de overige waarnemingen.

In dit verslag wordt bekeken hoe de Huber toets in deze hoedanigheid zich gedraagt ten opzichte van twee andere rangtoetsen; de Mood & Brown toets en de Kruskal & Wallis toets. Het simulatie-onderzoek van J.B. Dijkstra en H. Linders wordt hier voor deze twee rangtoetsen herhaald. Hoewel dat onderzoek plaatsvond voor 3 en 6 groepen, bleek al na het simulatie-onderzoek bij 3 groepen dat de Huber toets ook betere resultaten geeft dan de Kruskal & Wallis toets en de Mood & Brown toets. Het onderzoek bij 6 groepen hebben we hier achterwege gelaten omdat het nu voor de hand ligt dat ook hier Huber een betere performance zal geven.

De gebruikte toetsen zijn:

1) De methode van Huber:

Het model in de klassieke enkelvoudige variantie-analyse is $\underline{y}_{ij} = \mu_i + \underline{e}_{ij}$ waarin de fouten $\underline{e}_{ij}$ verondersteld worden $N(0, \sigma^2)$ verdeeld te zijn met onbekende σ.

Index i staat voor het groepsnummer ($i = 1, \ldots, k$) en j voor het aantal waarnemingen binnen een groep ($j = 1, \ldots, n_i$). De onderzochte hypothese is H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$.

Zonder uitschieters wordt de hypothese als volgt getoetst:

$$F = \frac{\sum_{i=1}^{k} n_i \, (\bar{y}_i - \bar{y})^2/(k-1)}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2/(N-k)} ,$$

met $N = \sum_{i=1}^{k} n_i$, $\bar{y}_i$ is het gemiddelde binnen een groep en $\bar{y}$ is het gemiddelde van alle waarnemingen. H_0 wordt nu verworpen indien $F > F^{k-1}_{N-k}(\alpha)$ voor een willekeurige onbetrouwbaarheid α.

In het geval met uitschieters werd F door Huber als volgt aangepast: allereerst wordt in iedere groep een beginschatting voor de verwachting bepaald volgens de methode ROBUST MEAN. Deze methode is te vinden in de RC-Informatie PP-4.20. Daarna wordt een robuuste schatting σ voor de standaardafwijking van de residuen bepaald. Met behulp van deze σ wordt via een speciale iteratieve kleinste kwadratenmethode van elke waarneming y_{ij} een pseudowaarneming bepaald. Met deze pseudowaarnemingen wordt nu op klassieke wijze bovenstaande F bepaald. Deze gehele methode staat uitvoeriger beschreven in de RC-Informatie PP-4.20.:

§10 ROBUST ANOVA.

2) De rangtoetsen:

Laat $x_1 \ldots x_N$ een combinatie zijn van K steekproeven. R_i geeft het rangnummer van elke waarneming x_i aan en n_j is de bijbehorende steekproefgrootte dan is de toets van Kruskal & Wallis (1952):

$$K = \frac{12}{N(N+1)} \sum_{j=1}^{k} \Big[\sum_{i \in s_j} R_i \Big]^2 - 3(N+1),$$

met $N = \sum_{i=1}^{k} n_i$ en S_j een verzameling die de indices bevat van steekproef j.

De toets van Mood & Brown (1950):

$$M = 4 \sum_{j=1}^{k} \frac{1}{n_j} A_j^2 - N,$$

hierin is A_j het aantal waarnemingen in steekproef j met een rangnummer groter dan $\frac{1}{2}(N+1)$.

Voor beide rangtoetsen wordt H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ verworpen indien K, respectievelijk M, > $\chi^2_{k-1}(\alpha)$ voor een willekeurige onbetrouwbaarheid α.

2.2. De gerealiseerde onbetrouwbaarheid.

Voor een vergelijking van de drie toetsen is een simulatie-onderzoek gedaan naar de gerealiseerde onbetrouwbaarheid. We schatten de kans op het verwerpen van H_0: $\mu_1 = \ldots = \mu_k$, indien waar, door gebruik te maken van simulatie die berust op 2000 herhalingen. Dit is gedaan voor drie groepen. Voor de groepsgroottes werd 10, 25 en 40 genomen. De steekproeven werden gegenereerd uit normale verdelingen met $\mu_i = 0$ en $\sigma^2 = 1$. Symmetrische contaminatie werd gesimuleerd door $\sigma^2 = 50$ te nemen met de kansen 0, 0.03 en 0.1. Eenzijdige contaminatie werd gesimuleerd door alleen $\sigma^2 = 50$ te nemen voor de positieve waarnemingen. Om hetzelfde aantal verwachte uitschieters te krijgen, als in het geval van de symmetrische contaminatie, werden hier de kansen op $\sigma^2 = 50$; 0, 0.06 en 0.2 genomen. In de tabellen staan deze kansen onder ε.

Bij iedere situatie is in tabel 1 en 2 het percentage van verwerpingen aangegeven voor een toets, met nominale onbetrouwbaarheid $\alpha = 0.05$. De waarden van Huber hebben we overgenomen uit de publikaties van J.B. Dijkstra en H. Linders. Zij gebruikten een andere procedure dan de ROBUST ANOVA die momenteel beschikbaar is in het Rekencentrum. De procedures verschillen hierin dat de ROBUST ANOVA nog wel eens de waarde "false" wil afleveren. In dit onderzoek was dit het geval indien het aantal uitschieters binnen een groep de pan uitrees. In dat geval wordt niet meer aan H_0 voldaan. Als we deze pogingen buiten beschouwing laten blijken de twee procedures vrijwel dezelfde resultaten op te leveren.

n_1, n_2, n_3			ε	K&W	M&B	Huber
10	10	10	0	4.70	3.65	5.45
			0.03	4.45	3.75	5.15
			0.1	5.40	3.55	5.25
25	25	25	0	4.55	5.45	5.35
			0.03	4.80	5.45	5.20
			0.1	5.25	5.70	5.40
40	40	40	0	4.80	5.40	5.00
			0.03	5.60	5.25	5.25
			0.1	5.40	5.25	4.60
10	25	40	0	4.80	4.00	4.40
			0.03	4.45	4.05	5.70
			0.1	4.65	4.35	5.25

tabel 1: symmetrische contaminatie

n_1, n_2, n_3			ε	K&W	M&B	Huber
10	10	10	0	4.35	3.80	6.35
			0.06	4.40	3.65	5.65
			0.2	4.40	3.60	5.65
25	25	25	0	5.00	5.50	6.00
			0.06	4.60	4.10	5.20
			0.2	4.40	4.45	5.15
40	40	40	0	4.40	4.60	4.60
			0.06	4.45	4.60	5.80
			0.2	5.15	5.60	5.55
10	25	40	0	4.40	4.00	5.65
			0.06	5.90	5.25	5.30
			0.2	4.25	4.15	5.35

tabel 2: eenzijdige contaminatie

De geschatte standaarddeviatie van de gerealiseerde onbetrouwbaarheid is $[(0.05 \times 0.95)/2000]^{\frac{1}{2}} = 0.004873 = 0.4873\%$. Het betrouwbaarheidsinterval voor de waarden in tabel 1 en 2 wordt nu (4.03, 5.97). Voor iedere toets kunnen we in de volgende tabel aflezen hoe vaak zijn verwerpingspercentage buiten het interval uitkwam.

<u>symmetrische verstoring</u>:		<u>eenzijdige verstoring</u>:	
Kruskal & Wallis	0	Kruskal & Wallis	0
Mood & Brown	4	Mood & Brown	4
Huber	0	Huber	2

In tabel 3 hebben we de gevallen symmetrische verstoring, eenzijdige verstoring en geen verstoring nog eens naast elkaar geplaatst. De waarden in tabel 3 zijn gebaseerd op 8×2000 herhalingen.

contaminatie	K&W	M&B	Huber
geen (ε = 0)	4.625	4.550	5.350
symmetrisch	5.000	4.669	5.225
eenzijdig	4.694	4.425	5.456

<u>tabel 3</u>

Hun standaarddeviatie is $0.004873/\sqrt{8} = 0.001723 = 0.1723\%$. Het betrouwbaarheidsinterval voor de verwerpingspercentages in tabel 3 is nu: (4.655, 5.345).

Uit het voorgaande kunnen we de volgende conclusies trekken:

- De Mood & Brown toets lijkt het minst geschikt. Dit heeft echter grotendeels te maken met de lage verwerpingspercentages bij de groepsgroottes: 10, 10, 10. Waarschijnlijk is hier het gebruik van de χ^2-verdeling dubieus, want de M&B toets heeft deze verdeling alleen asymptotisch. Bij de grotere groepsgroottes is de M&B toets niet aantoonbaar slechter dan de andere twee toetsen.
- Uit tabel 3 blijkt dat de verstoring vrijwel geen invloed heeft in de verwerpingspercentages. Dit is het grote voordeel bij het gebruik van rangtoetsen.
- In de aanwezigheid van uitschieters geeft de M&B toets een conservatieve tendens en de methode van Huber een progressieve tendens.
- Er is geen duidelijke voorkeur voor een bepaalde toets.

2.3. Het onderscheidend vermogen.

Hier wordt hetzelfde simulatie-onderzoek gedaan als in het vorige hoofdstuk, met dit verschil dat de lokatieparameters verschillend worden ingesteld. Er wordt dus nu niet meer aan H_0 voldaan. We verkregen de volgende resultaten:

n_1, n_2, n_3	ε	$\mu_i(\times 0.1)$	K&W	M&B	Huber
10, 10, 10	0	0, 8, 16	83.95	61.55	84.50
	0.03		77.30	56.55	80.75
	0.1		58.60	46.90	68.25
25, 25, 25	0	0, 5, 10	86.20	69.85	85.45
	0.03		81.50	65.70	82.45
	0.1		66.35	58.45	71.00
40, 40, 40	0	0, 4, 8	86.65	71.80	87.05
	0.03		83.05	68.15	83.65
	0.1		69.80	62.70	75.50
10, 25, 40	0	0, 8, 13	90.20	72.45	91.00
	0.03		84.40	68.80	87.35
	0.1		70.25	59.10	80.45

tabel 4: symmetrische verstoring met ongelijke lokatieparameters

Tabel 4 geeft al een goede indicatie: Kruskal & Wallis is in alle gevallen beter dan de Mood & Brown toets. De Huber toets is in vrijwel alle gevallen beter dan de andere twee toetsen. Alleen in de situatie $n_1 = 25$, $n_2 = 25$, $n_3 = 25$, $\varepsilon = 0$ heeft Kruskal & Wallis een iets hogere waarde. Het scheelt echter maar bijzonder weinig, namelijk 0.4. Tabel 5 geeft weer de eenzijdige contaminatie.

n_1, n_2, n_3	ε	$\mu_i(\times 0.1)$	K&W	M&B	Huber
10, 10, 10	0	0, 8, 16	83.90	60.60	84.55
	0.06		78.10	57.65	80.40
	0.2		60.30	47.85	69.25
25, 25, 25	0	0, 5, 10	85.90	68.45	85.05
	0.06		81.55	67.15	82.50
	0.2		67.80	57.30	73.80
40, 40, 40	0	0, 4, 8	87.35	71.15	87.25
	0.06		83.60	67.70	84.00
	0.2		71.55	63.05	75.35
10, 25, 40	0	0, 8, 13	90.90	72.95	89.60
	0.06		85.45	68.75	87.65
	0.2		70.55	54.50	78.15

tabel 5: eenzijdige verstoring met ongelijke lokatieparameters

Ook deze tabel geeft weer dezelfde verhoudingen weer. De Kruskal & Wallis toets is alleen in het geval $\varepsilon = 0$ marginaal beter dan de Huber toets. In de aanwezigheid van uitschieters is de Huber toets in elke situatie beter dan de Kruskal & Wallis toets. De Mood & Brown toets is asymptotisch optimaal voor dubbelexponentiële verdelingen. Deze kenmerken zich door stevige staarten. Het is dan ook niet verrassend dat bij het onderscheidend vermogen deze toets te wensen overlaat bij het gebruik van kleine ε.

2.4. Conclusie.

Het onderscheidend vermogen van de toets van Huber is veel beter dan die van Kruskal & Wallis en Mood Brown. Huber blijft daarom de meest geschikte toets.

3. Het gedrag van gewinsoriseerde en getrimde toetsen voor enkelvoudige variantie-analyse bij normale verdelingen met ongelijke varianties.

3.1. Inleiding.

In een publikatie uit 1981 van J.B. Dijkstra en P.S.P.J. Werter: "Testing the equality of several means when the populations variances are unequal" is al eerder onderzoek verricht naar deze vorm van enkelvoudige variantie-analyse. In deze publikatie werden de Welch toets, de James toets en de Brown-Forsyte toets met elkaar vergeleken. Deze toetsen zijn speciaal ontwikkeld om de gelijkheid an lokatieparameters te toetsen bij normale verdelingen met ongelijke varianties. In hun onderzoek kwam de James toets als beste naar voren.

We vergelijken hier de James toets met gewinsoriseerde en getrimde toetsen waarbij we de resultaten van de James toets overnemen uit bovenstaande publikatie. De gewinsoriseerde en getrimde toetsen zijn ontwikkeld voor gelijke varianties. Het ligt daarom niet meteen voor de hand dat deze toetsen ook werkelijk betere resultaten zullen opleveren.

1. Winsorizing en trimming.

Veel methoden, met name die waarin met kwadratensommen wordt gewerkt, verliezen aan betrouwbaarheid als er zich uitschieters in de waarnemingen bevinden. Voor waarnemingen uit normale verdelingen met gelijke groepsvarianties hebben we de klassieke F-toets om de nulhypothese te toetsen dat de groepen een gelijk groepsgemiddelde hebben.

$$F = \frac{\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2/(N-k)}$$

Hierin is $N = \sum_{i=1}^{k} n_i$ en $\bar{y}_i$ is het groepsgemiddelde van groep i en $\bar{y}$ is het gemiddelde van alle waarnemingen. De index i geeft de groepsnummers aan ($i = 1 \ldots k$) en de index j de waarneming binnen groep i ($j = 1 \ldots n_i$). In de aanwezigheid van uitschieters kunnen we deze F als volgt aanpassen:

a) Winsorizing.

Laat de waarneming $y_{i1}, \ldots, y_{i,n_i}$ geordend zijn. Dan definiëren we het gewinsoriseerde gemiddelde $\bar{y}_{iwg}$ bij tweezijdige winsorizing met parameter g als volgt:

$$\bar{y}_{iwg} = \frac{1}{n_i}\{(g+1)y_{i,g+1} + y_{i,g+2} + \ldots + y_{i,n_i-g-1} + (g+1)y_{i,n_i-g}\}$$

De restkwadraatsom SS_{iwg} wordt nu:

$$SS_{iwg} = (g+1)(y_{i,g+1} - \bar{y}_{iwg})^2 + (y_{i,g+2} - \bar{y}_{iwg})^2 + \ldots$$
$$\ldots + (y_{i,n_i-g-1} - \bar{y}_{iwg})^2 + (g+1)(y_{i,n_i-g} - \bar{y}_{iwg})^2.$$

De gewinsoriseerde F-toets definiëren we nu:

$$F_{wg} = \frac{\sum_{i=1}^{k} h_i(\bar{y}_{iwg} - \bar{y}_{wg})^2/(k-1)}{\sum_{i=1}^{k} SS_{iwg}/(H-K)}$$

Hierin is $h_i = n_i - 2g$ voor $i = 1, \ldots, k$ en $\bar{y}_{wg} = \sum_{i=1}^{k} h_i\bar{y}_{iwg}/H$ met $H = \sum_{i=1}^{k} h_i$.

Winsorizing wil dus zeggen dat de staartelementen in de geordende groep vervangen worden door de meest extreme waarnemingen die niet in de staart zitten. Bij trimming worden de staartelementen gewoon weggelaten.

b) Trimming.

De noemer in de getrimde F-toets is dezelfde als in de gewinsoriseerde F-toets. Alleen de teller verandert. In de teller worden de gewinsoriseerde gemiddelden in F_{wg} vervangen hun getrimde gemiddelden $\bar{y}_{itg}$.

$$\bar{y}_{itg} := \frac{1}{n_i}\{y_{i,g+1} + y_{i,g+2} + \ldots + y_{i,n_i-g}\}.$$

Verder wordt in F_{wg}, $\bar{y}_{wg}$ vervangen door $\bar{y}_{tg}$ met

$$\bar{y}_{tg} = \sum_{i=1}^{k} h_i \bar{y}_{itg}/H.$$ Dus de getrimde F-toets is

$$F_{tg} = \frac{\sum_{i=1}^{k} h_i(\bar{y}_{itg} - \bar{y}_{tg})^2/(K-1)}{\sum_{i=1}^{k} SS_{iwg}/(H-K)}$$

Als $g = 0$ dan $F_{tg} = F_{wg} = F$

2. De James-toets.

De beschrijving van deze toets is nogal bewerkelijk, vandaar dat hier verwezen wordt naar de publikatie van Dijkstra en Werter. De toets staat ook beschreven in de RC-Informatie PP4-14.

3.2. De gerealiseerde onbetrouwbaarheid.

Voor de verschillende sets (n_i, μ_i, σ_i, g_i) wordt de gerealiseerde onbetroubaarheid van de drie hierboven beschreven toetsen bepaald. De n_i, μ_i, σ_i hebben we gekozen als in het onderzoek van Dijkstra en Werter. Voor de g_i werden verschillende waarden geprobeerd. De g_i werd wel zo gekozen dat hij overeenkwam met de groepsgrootte. De nulhypothese H_0: $\mu_1 = \mu_2 = .. = \mu_k$ wordt verworpen als $F_{tg} > F^{k-1}_{H-k}(\alpha)$ respectievelijk $F_{wg} > F^{k-1}_{H-k}(\alpha)$. Voor α kozen we de waarden 0.1, 0.05, 0.01. We hebben de groepsgemiddelden gelijk aan nul gesteld en vonden in tabel 1 de percentages van verwerpingen van de nulhypothese, dat alle gemiddelden gelijk zijn. Het simulatie-onderzoek is gebaseerd op 10.000 herhalingen. We kunnen nu voor iedere kolom weer een betrouwbaarheidsinterval opstellen:

groeps-grootttes	standaard-afwijkingen	g_i	WINSORIZING 10%	WINSORIZING 5%	WINSORIZING 1%	TRIMMING 10%	TRIMMING 5%	TRIMMING 1%	JAMES 10%	JAMES 5%	JAMES 1%
4 4 4	1 1 1	1 1 1	12.81	7.02	1.64	12.81	7.02	1.64	9.40*	4.44	0.76
4 4 4	1 2 3	1 1 1	15.75	9.46*	2.59	15.75	9.46*	2.59	10.44*	5.31*	1.28
4 6 8	1 1 1	1 1 1	9.65*	4.64*	1.07*	9.54*	4.63*	1.08*	9.74*	4.99*	1.03*
		1 1 2	11.40	6.22	1.49	11.38	6.20	1.49			
		1 2 3	13.72	7.99	2.03	13.21	7.62	1.97			
4 6 8	1 2 3	1 1 1	5.36	2.62	0.71	5.23	2.64	0.72	9.29	4.33	0.87*
		1 1 2	8.44	4.48	1.04*	8.32	4.51	1.03*			
		1 2 3	15.81	9.40	2.43	15.34	9.04	2.27			
4 4 4 4	1 1 1 1	1 1 1 1	12.77	6.94	1.59	12.77	6.94	1.59	9.95*	5.23*	0.91*
4 4 4 4	1 2 2 3	1 1 1 1	15.46	9.13	2.62	15.46	9.13	2.62	10.79	5.68	1.21
4 6 8 10	1 1 1 1	1 1 1 1	8.71	4.13	0.87*	8.63	4.23	0.84*	10.02*	4.92*	1.14*
		1 1 2 2	10.26*	5.09*	1.35	10.19*	5.03*	1.31			
		1 2 2 3	12.34	6.67	1.45	12.17	6.74	1.46			
4 6 8 10	1 2 2 3	1 1 1 1	4.79	2.48	0.40	4.79	2.30	0.39	9.56*	4.39	0.99*
		1 1 2 2	6.32	3.49	0.69	6.33	3.45	0.71			
		1 2 2 3	8.91	4.80*	1.29	8.75	4.86*	1.28			
4 6 8 10	3 2 2 1	1 1 1 1	22.40	14.37	5.52	22.27	14.15	5.29	10.48*	5.35*	1.64
		1 1 2 2	23.30	15.54	6.39	23.33	15.61	6.35			
		0 1 2 3	14.03	8.81	3.06	13.20	8.08	2.75			
10 10 10 10	1 1 1 1	0 0 0 0	9.75*	4.74*	0.90*	9.75*	4.74*	0.90*	9.94*	5.01*	0.91*
		1 1 1 1	10.17*	5.32*	1.19*	10.26*	5.59	1.28			
		2 2 2 2	10.13*	5.32*	1.28	10.56*	5.48	1.25			
		3 3 3 3	11.24	5.89	1.39	11.34	6.01	1.42			
10 10 10 10	1 2 2 3	0 0 0 0	11.23	6.18	1.71	11.23	6.18	1.71	10.15*	4.97*	1.01*
		1 1 1 1	11.57	6.57	1.85	11.71	6.76	1.80			
		2 2 2 2	11.79	6.83	1.84	12.19	6.88	1.91			
		3 3 3 3	13.89	8.54	2.95	13.93	8.76	3.02			
10 14 16 20	1 1 1 1	1 2 2 3	9.85*	5.17*	0.97*	10.06*	5.41*	1.04*	9.83*	4.56	1.03*
10 14 16 20	1 1.5 2 3	1 2 2 3	7.27	3.86	1.03*	7.56	4.03	1.07*	9.85*	5.02*	1.15*
		2 2 3 3	6.48	3.27	0.93*	6.74	3.42	0.92*			
10 14 16 20	3 2 1.5 1	2 2 3 3	19.80	12.68	5.25	20.18	12.98	5.36	9.66*	4.63*	0.94*
		1 2 2 3	19.87	11.59	4.43	18.12	11.99	4.58			

Tabel 1: de gerealiseerde onbetrouwbaarheid

α	betrouwbaarheidsinterval
0.1	(9.40, 10.60)
0.05	(4.564, 5.436)
0.01	(0.801, 1.199)

In tabel 1 zijn met ster aangegeven die verwerpingspercentages die in de respectievelijke betrouwbaarheidsintervallen liggen. De tabel toont al de slechte resultaten voor winsorisering en trimming. Winsorisering en trimming leveren slechts behoorlijke resultaten op bij gelijke varianties. als we echter kijken naar de verwerpingspercentages waar $g_i = 0$ (i= 1, ..., 4) dan bestaat het vermoeden dat de resultaten slechter zijn dan de klassieke F-toets. Bij ongelijke varianties zijn de resultaten ronduit slecht, zeker als men ze vergelijkt met de James-toets. Voor de percentages van de gewinsoriseerde en getrimde toetsen geldt dat grote varianties met grote steekproefgrootte's een te gering verwerpingspercentage ten opzichte van de nominale onbetrouwbaarheid opleveren. Voor grote varianties met kleine groepsgrootte's geldt het omgekeerde. Dit was ook wel te verwachten op grond van de noemer in beide toetsen.

3.3. Conclusie.

James blijft de meest geschikte toets. Het verschil tussen winsorisering en trimming in deze hoedanigheid is niet groot.

4. Literatuur.

[1] Dijkstra, J.B. en Werter, P.S.P.J.
Het gebruik van de toets van Kruskal & Wallis bij normale verdelingen met ongelijke varianties
Kwantitatieve Methoden 5 (1982), 151-158

[2] Dijkstra, J.B. en Werter P.S.P.J.
Testing the equality of several means when the populations variances are unequal
Commun. Statist.-simula. Computa.B10(6) (1981), 557-569

[3] Dijkstra, J.B. en Linders H.
Comparison of several mean values in the presence of outliers
TUE-RC 67704a

[4] Dijkstra, J.B.
Nonparametric comparison of several mean values with mild adaption to some sample characteristics
CC-note 20 (1984)

[5] Hajek, J. en Sidak, Z.
Theory of rank tests
Academic Press, New York (1967)

[6] Hontelez, J.
Het robuuste toetsen van gelijkheid van populatiegemiddelden waarbij de varianties gelijk zijn
CC-note 21 (1984)

[7] Huber, Peter J.
Robust Statistics
Wiley, New York (1981)

[8] RC-Informatie PP-4.10.
Verdelingsvrije toetsen

[9] RC-Informatie PP-4.11.
Verdelingsfuncties

[10] RC-Informatie PP-4.14.
Variantie-analyse

[11] RC-Informatie PP-4.15.
Aselecte trekkingen

[12] RC-Informatie PP-4.20.
Robuuste statistische methoden