Een uitschieter-resistente procedure voor enkelvoudige klassieke variantie-analyse

THE-RC 59383

Eindhoven University of Technology

Computing Centre Note 21

Een uitschieter-resistente procedure voor enkelvoudige klassieke variantie-analyse

Jan Hontelez

Stage Statistische Analyse
o.l.v. Prof.dr. R. Doornbos
Drs. J.B. Dijkstra

oktober 1984

Inhoud: pagina

1. Inleiding 3
2. Methode 4
3. Een simulatie-onderzoek naar het gedrag van de toets 9
4. Conclusie 11
5. Suggesties voor nader onderzoek 12
6. Literatuur 13
Bijlage 1: Toelichting bij (3.1.)
Bijlage 2: Resultaten van de testen waarbij de H_0-hypothese waar is
Bijlage 3: Resultaten van de testen waarbij de H_0-hypothese niet waar is
Bijlage 4: Procedure voor het robuust toetsen van gelijkheid van populatiegemiddelden waarbij de varianties gelijk zijn:
ROBUST ANOVA

1. Inleiding.

Gegeven zijn k populaties die elk normaal verdeeld zijn met gemiddelden: $\mu_1, \mu_2, \ldots, \mu_k$ en allen met dezelfde (onbekende) variantie σ^2. We nemen nu een steekproef en willen de volgende hypothese toetsen:

$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$.

Dit geval is erg bekend en eenvoudig te toetsen. In de procedurebibliotheek is daarvoor een procedure "ONEWAYANOVA" (zie [1]) aanwezig. Het probleem is echter dat deze klassieke wijze van toetsen niet erg robuust is. We bedoelen hiermee dat de toets gevoelig is voor een kleine afwijking in de modelveronderstellingen, in dit geval bedoelen we: in de verdelingen van de populaties. Stel namelijk dat een klein percentage van het aantal waarnemingen 'uitschieter' is (bijvoorbeeld de lengten van personen in meter en een enkele keer is de decimale punt vergeten). De klassieke toets is hier niet ongevoelig voor. De toets is dus ten gevolge van enige foute waarnemingen onbetrouwbaar.

We zoeken nu een toets die wel robuust is. Peter J. Huber doet in [2] een suggestie. In deze stage gaan we na, althans een eerste aanzet daartoe, of dit werkelijk een robuuste toets voor bovenstaand geval is.

2. Methode.

Gegeven:

een steekproef $\underline{x} = (\underline{x}_m, \underline{x}_{m+1}, \cdots, \underline{x}_n)$

met $g_m, g_{m+1}, \cdots, g_n \qquad 1 \leq g_i \leq k;\ m \leq i \leq n$

waarbij g_i = populatie waartoe waarneming $\underline{x}_i$ behoort; de populaties zijn genummerd van 1 tot en met k.

k_j := aantal waarnemingen in groep j $(1 \leq j \leq k)$.

We nemen als model:

$\underline{x}_\ell = \beta_1 v_1 + \beta_2 v_2 + \cdots + \beta_{k-1} v_{k-1} + \beta_k + \underline{e}\ ;\ \underline{e} \sim N(0, 1)$

met $v_i = 1$ als $\underline{x}_\ell$ tot populatie i behoort, anders $v_i = 0$.

We krijgen dus nu voor de steekproef $\underline{x}$ in matrixnotatie:

$\underline{x} = V\beta + \underline{e}\ ;\ \underline{e} \sim N(0, \sigma^2 I)$.

Stel nu dat de waarnemingen geordend zijn, dus:

$\underline{x}_m, \cdots, \underline{x}_{m+k_1-1}$ is een steekproef uit groep 1

$\underline{x}_{m+k_1}, \cdots, \underline{x}_{m+k_1+k_2-1}$ is een steekproef uit groep 2, enz.

Dan geldt:

$$V = \begin{bmatrix} 1 & 0 & \cdots & 0 & 1 \\ \vdots\ (k_1) & \vdots & & \vdots & 1 \\ 1 & 0 & & \vdots & 1 \\ 0 & 1 & & \vdots & \vdots \\ \vdots & \vdots\ (k_2) & & \vdots & \vdots \\ 1 & 0 & & \vdots & \vdots \\ \vdots & \vdots & & 0 & \vdots \\ \vdots & \vdots & & 1\ (k_{k-1}) & \vdots \\ \vdots & \vdots & & 0 & \vdots \\ \vdots & \vdots & & \vdots & 1 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}$$

Met de kleinste kwadraten methode, dit is het minimaliseren van:
$\sum_{i=m}^{n} (\underline{x}_i - v_i.\beta)^2$, met v_i de i-de rij van de matrix V, vinden we een schatter voor β:

$$\underline{\hat{\beta}}_0 := (V^T V)^{-1} V^T \underline{x}$$

en een schatter voor $E\underline{x}$:

$$\underline{\hat{x}} := V(V^T V)^{-1} V^T \underline{x} =: H\underline{x}$$

H heet de 'hatmatrix' en heeft in dit geval de volgende vorm:

$$H = \begin{bmatrix} k_1^{-1} J_{k_1} & & \\ & k_2^{-1} J_{k_2} & \\ & & \ddots \\ & & & k_k^{-1} J_{k_k} \end{bmatrix} \qquad (2.1)$$

met $J_\ell = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \iddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$; $J_\ell \in M^1_{\ell,\ell}$

We definiëren: $\bar{\underline{x}} := \frac{1}{n-m+1} \sum_{i=m}^{n} \underline{x}_i$

H_0: $\beta_1 = \beta_2 = \ldots = \beta_{k-1} = 0$ wordt nu getoetst met:

$$\underline{F} := \frac{\frac{1}{k-1} \|\underline{\hat{x}} - \bar{\underline{x}}\|^2}{\frac{1}{n-m+1-k} \|\underline{x} - \underline{\hat{x}}\|^2} \qquad (2.2)$$

waarbij $\underline{F}$ onder H_0 een $F^{k-1}_{n-m+1-k}$-verdeling heeft.
Zie voor uitgebreidere behandeling bijvoorbeeld: [3].

Deze klassieke methode eist dat $\underline{e} \sim N(0, \sigma^2 I)$ en is dus gevoelig voor 'uitschieters' in de waarnemingen. Als nu $\underline{e}$ niet normaal verdeeld is, dan kunnen we $E\underline{x}$ robuust schatten. Dan zijn $\hat{\underline{x}}$ en $\bar{\underline{x}}$ tenminste asymptotisch normaal (zie [2], §7-4 & §7-5).

Robuust schatten van β (en dus ook van $E\underline{x}$) is het minimaliseren van:

$$\sum_{i=m}^{n} \rho\left(\frac{\underline{x}_i - v_i \cdot \beta}{\sigma}\right)$$

ofwel de oplossing van het stelsel:

$$\sum_{i=m}^{n} v_{ij} \psi\left(\frac{\underline{x}_i - v_i \cdot \beta}{\sigma}\right) = 0 \; ; \; 1 \leq j \leq k$$

met ρ een robuuste functie en $\psi = \rho'$. Als $\rho(x) = x^2$ dan zijn we terug bij het klassieke geval, de kleinste kwadratenmethode. Het is duidelijk dat we nu $\psi(x)$, of zoals vaak gedaan wordt: $\frac{\psi(x)}{x}$, verstandig moeten kiezen. In [4] staan er acht vermeld. We nemen hier de functie die de naam van Huber zelf draagt:

$$H(x) = \psi(x) = \begin{cases} x & |x| \leq A \\ \text{sign}(x).A & |x| > A \end{cases} \qquad (2.3)$$

Voor de constante A in $\psi(x)$ nemen we 1.35; deze keuze is aannemelijk gemaakt in [4], pagina 817 - 818.

Tenslotte hebben we nog een robuuste schatter voor σ nodig:

$$\hat{\underline{\sigma}} := 1.48 \underset{m \leq i \leq n}{\text{med}} \left| (\underline{x}_i - v_i \cdot \hat{\underline{\beta}}_0) - \left[\underset{m \leq i \leq n}{\text{med}} (\underline{x}_i - v_i \cdot \hat{\underline{\beta}}_0) \right] \right| \qquad (2.4)$$

met $\hat{\underline{\beta}}_0$ de kleinste kwadratenschatter voor β (zie [4], pagina 815). De factor 1.48 maakt deze schatter in benadering zuiver voor een normale verdeling, en het gebruik van de mediaan maakt hem ongevoelig voor uitschieters.

Hoe we dit stelsel op moeten lossen, gaan we hier niet op in. Er is een procedure "ROBUSTMULTIPLEREGRESSION" aanwezig (zie [5]), die β en dus ook $E\underline{x}$ robuust schat. Zie voor uitgebreidere behandeling bijvoorbeeld [2] en [4].

De suggestie van Huber, betreffende deze toets
H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ is nu (zie [2], pagina 195 - 198):
schat $\underline{Ex}$ en $\underline{E\bar{x}}$ robuust en vervang de noemer van (2.2) door:

$$\frac{1}{n-m+1-k} \cdot C^2 \cdot \sum_{i=m}^{n} \psi^2\left(\frac{\underline{x}_i - \hat{\underline{x}}_i}{\underline{\sigma}}\right) \cdot \hat{\underline{\sigma}}^2, \text{ met:}$$

$$C = 1 + \frac{k \, \mathrm{var}(\psi')}{(n-m+1).[E(\underline{\psi}')]^2} \tag{2.5}$$

waarbij $\underline{\sigma}$ zoals in (2.4) vermeld staat is geschat en $E(\underline{\psi}')$ en $\mathrm{var}(\underline{\psi}')$ als volgt (zie [2], pagina 174):

$$\hat{E}(\psi') = \frac{1}{n-m+1} \sum_{i=m}^{n} \psi'\left(\frac{\underline{x}_i - \hat{\underline{x}}_i}{\hat{\underline{\sigma}}}\right)$$

$$\hat{\mathrm{var}}(\underline{\psi}') = \frac{1}{n-m+1} \sum_{i=m}^{n} \left[\psi'\left(\frac{\underline{x}_i - \underline{x}_i}{\hat{\underline{\sigma}}}\right) - E(\underline{\psi}')\right]^2.$$

De robuuste-toets-procedure is dus als volgt:
Gegeven een steekproefrealisatie: x
- Schat $\underline{Ex}$ robuust: $\hat{x}$
- Bepaal $r^* := \begin{bmatrix} r_m^* \\ \\ r_n^* \end{bmatrix}$, met $r_j^* := \dfrac{C \cdot \psi\left(\dfrac{x_i - \hat{x}_i}{\hat{\sigma}}\right) \cdot \hat{\sigma}}{\dfrac{1}{n-m+1} \sum_{i=m}^{n} \psi'\left(\dfrac{x_i - \hat{x}_i}{\hat{\sigma}}\right) \cdot \hat{\sigma}}$
- $x^* := \hat{x} + r^*$
- Ga nu op de klassieke manier verder met x^* in plaats van x, dus $\bar{x}^* := \frac{1}{n-m+1} \sum_{i=m}^{n} x_i^*$ en:

$$F := \frac{\frac{1}{k-1} \| \hat{x} - \bar{x}^* \|^2}{\frac{1}{n-m+1-k} \| x^* - \hat{x} \|^2} \tag{2.6}$$

In het robuuste geval is F (2.2) slechts te benaderen door een $F^{k-1}_{n-m+1-k}$-verdeling.

Tenslotte merken we nog op dat Huber in [2], pagina 197 ook nog een voorwaarde aan het aantal waarnemingen heeft gesteld:

$$\frac{k}{n-m+1} \leq 0.2 \qquad (2.7)$$

In een eerder stadium echter (zie [2], pagina 162 en 165 - 170) is aangegeven dat alleen als $h = \max_{1\leq i\leq n-m+1} h_i \leq 0.2$ de resultaten betrouwbaar zijn, waarbij $h_1, h_2, \ldots, h_{n-m+1}$ de diagonaalelementen van de matrix H zijn. In dit geval is de eis dus (zie (2.1)):

$$\max_{1\leq i\leq k} \frac{1}{k_i} \leq 0.2, \text{ dus: } k_i \geq 5 \ (i = 1, 2, \ldots, k). \qquad (2.8)$$

Het aantal waarnemingen in een groep moet dus minstens 5 zijn; aan (2.7) is dan automatisch ook voldaan.

3. Een simulatie-onderzoek naar het gedrag van de toets.

We willen nu testen of de nieuwe methode werkelijk robuust is. Daartoe hebben we pseudo-random getallen gegenereerd die N(0, 1)-verdeeld zijn. Als we nu 10% 'slechte' waarnemingen, dus 10% uitschieters, wilden hebben, dan werd met een kans 0.1 het getal uit een $N(0, \sigma^2)$-verdeling gegenereerd en met 0.9 uit N(0, 1).
Voor σ hebben we de waarden 5 en 10 genomen.
Dus de distributiefunctie van de onderliggende verdeling van de getallen is:
$F_x = 0.9F_1 + 0.1F_\sigma$, waarbij:
F_1 de kansverdeling is bij een N(0, 1)-verdeling
F_σ de kansverdeling is bij een $N(0, \sigma^2)$-verdeling.
Voor k, het aantal groepen, hebben we 3, 5 en 8 genomen. De uitschieters hebben we zowel geconcentreerd in 1 of 2 populaties als over alle groepen verdeeld. Het percentage uitschieters hebben we vaak 0, 10 of 20% gekozen.

Elke test bestaat uit 500 simulaties. We hebben in alle gevallen de onbetrouwbaarheid $\alpha = 0.05$ genomen. De fractie p van het totaal aantal simulaties waarbij de H_0-hypothese verworpen wordt, is natuurlijk niet precies gelijk aan α. We vinden het een acceptabel verschil als voor p geldt (voor verantwoording: zie Bijlage 1):

$$\alpha - 2d \leq p \leq \alpha + 2d \qquad (3.1)$$

waarbij d de standaarddeviatie is van een binomiale verdeling:
$2d = 2\sqrt{\frac{\alpha(1-\alpha)}{N}}$ met N = aantal simulaties.
In dit geval is het acceptatiegebied [0.31, 0.69].
De opzet van dit simulatie-onderzoek is gehaald uit [6].

We hebben zowel robuust als klassiek getoetst. De resultaten staan in Bijlage 2 vermeld. Opmerkelijke verschillen treden pas op bij 5 en 8 populaties; bij 3 geven beide toetsen goede resultaten. Het is duidelijk dat de niet-robuuste toets niet goed is. De robuuste methode levert acceptabele resultaten op. Het gaat niet zo goed als het percentage uitschieters 20% is.
Blijkbaar kan de toets zo'n hoog percentage niet meer aan. Wat verder opvalt is, dat het percentage "H_0-hypothese wordt verworpen" vaker boven dan onder 5% ligt. Voor mogelijke oorzaken, zie: H5.

In Bijlage 3 staan de resultaten van eenzelfde simulatie, echter nu om na te gaan of de fout van de tweede soort niet te groot is. Ook hier is de robuuste methode beter dan de klassieke: bij 3 populaties is de fout van de tweede soort van de klassieke toets veel te groot; bij de robuuste methode varieert hij van 0.25 tot en met 0.65, zowel bij 3 als bij 5 groepen.

4. Conclusie.

De robuuste toets is betrouwbaar bij een percentage uitschieters van 10% of minder. Hoe groot de afwijkingen zijn, is niet belangrijk; zowel met $\sigma = 5$ als $\sigma = 10$, tegenover de niet-vervuilde waarnemingen met $\sigma = 1$, zijn de resultaten goed. De klassieke toets daarentegen is niet betrouwbaar bij een percentage van 10%, zowel met $\sigma = 5$ als $\sigma = 10$.

Om de twee volgende redenen zijn we nog niet helemaal tevreden met de toets:

1. De geschatte onbetrouwbaarheid ligt wat te hoog in onze testen; de meeste resultaten liggen boven 0.05.
2. Als het percentage uitschieters 20% is, dan is de geschatte onbetrouwbaarheid te groot, dus de H_0-hypothese wordt te vaak (ten onrechte) verworpen.

5. Suggesties voor nader onderzoek.

Tot slot nog enige suggesties voor nader onderzoek met betrekking tot verbetering van de robuuste toets.

Een opmerking van Huber (zie [2], pagina 197) is, dat F waarschijnlijk beter te benaderen is door een F_w^{k-1}-verdeling, waarbij w 'iets' kleiner is dan n-m+1-k. De grootte van w hangt af van de onderliggende kansverdeling van de steekproef. In onze testen was n-m+1-k groot (42 - 195), zodat het verschil tussen $F_w^{k-1}(0.05)$ en $F_{n-m+1-k}^{k-1}(0.05)$ klein zal zijn. Dit zou heel goed de reden kunnen zijn van het verschijnsel genoemd in 4 - 1. Suggestie is dan ook om na te gaan of de resultaten beter worden als de verhouding $\frac{w}{n-m+1-k}$ wat kleiner is dan 1, of het verschil n-m+1-k-w wat groter is dan 0.

Een andere mogelijkheid is de constante C (2.5) wat groter te nemen.

De constante A in de Huber-functie zou wat groter genomen kunnen worden.

Bovenstaande veronderstellingen kunnen wel het probleem genoemd in 4 - 1 maar zeker niet helemaal dat van 4 - 2 oplossen. Een reden van 'het niet robuust genoeg zijn van de toets' zou kunnen zijn dat de Huber-functie (2.3) niet robuust genoeg is voor deze toets. Een suggestie om de toets ook betrouwbaar te maken bij een percentage uitschieters van 20% is dan ook: probeer een robuuste functie die grote afwijkingen nog minder zwaar meeweegt dan de Huber-functie. Mogelijke functies staan, zoals al eerder opgemerkt, in [4].

6. Literatuur.

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

[2] Huber, Peter J.
Robust Statistics
Wiley, New York, 1981

[3] Bosch, A.J. en W.L.M.M. Senden
Lineaire Modellen
THE, Najaarssemester 1980

[4] Holland, Paul W. and Roy E. Welsch
Robust Regression Using Iteratively Reweighted Least-Squares
Communication in Statistics - Theor. Meth., A6(9), (1977),
813 - 827

[5] Algol procedures voor robuuste regressie (Voorl. uitgave)
THE-RC 58065

[6] Dijkstra, Jan B. and Paul S.P.J. Werter
Testing the equality of several means when the population variances are unequal
Communications in Statistics-Simula. Computa., B10(6), 1981,
557 - 569

Toelichting bij (3.1).

Onder H_0: $p = 0.05 = \alpha$ willen we een acceptatie-interval voor $\frac{y}{N}$ opstellen. y is hierbij het aantal maal dat de hypothese $\mu_1 = \ldots = \mu_k$ wordt verworpen van de in totaal N gedane simulaties.

Ook hier nemen we onbetrouwbaarheid 0.05, dus:

$$P(y_1 \leq \underline{y} \leq y_2 \mid H_0) = 0.95 \qquad (*)$$

Deze binomiale verdeling is te benaderen door een normale verdeling met $\mu = Np$ en $\sigma^2 = Np(1-p)$. Dus (*) wordt nu:

$$\Phi\left(\frac{y_2 - Np}{\sqrt{Np(1-p)}}\right) - \Phi\left(\frac{y_1 - Np}{\sqrt{Np(1-p)}}\right) = 0.95.$$

Dus (zie bijvoorbeeld Statistische Compendium):

$$\frac{y_i - Np}{\sqrt{Np(1-p)}} = \pm 1.960 \text{ ofwel: } \frac{y_i}{N} = p \pm 1.960 \sqrt{\frac{p(1-p)}{N}}.$$

Het acceptatie-interval wordt dus met een onbetrouwbaarheid van 0.05 = $[\alpha - 1.960d, \alpha + 1.960d]$, $d = \sqrt{\frac{\alpha(1-\alpha)}{N}}$.

Wij hebben in onze testen 1.960d afgerond naar 2d.

Resultaten van de testen waarbij H_0-hypothese waar is.

Onbetrouwbaarheid: 0.05; de waarnemingen zijn trekkingen uit een "vervuilde" N(0, 1)-verdeling;

Acceptatie-interval: 3.1% ≤ Perc. H_0 verworpen ≤ 6.9%

Aantal groepen		3	3	3	3	3	3	3
Aantal waarnemingen		15-15-15	15-15-15	20-50-50	15-15-15	40-40-40	40-40-40	10-50-50
Percentage uitschieters		10-10-10	0- 0-10	10-10-10	0- 0-20	0- 0-20	0- 0-20	0- 0-20
Variantie uitschieters		5	5	5	10	10	10	10
Perc. H_0 verworpen	Robuust	3.8	4.2	4.6	6.0	6.4	6.6	6.2
	Niet robuust	3.6	3.4	6.0	5.2	6.2	6.6	5.8

Aantal groepen		5	5	5	5	5	5	5
Aantal waarnemingen		15-15-15-15-15	15-15-15-15-15	15-20-17-15- 8	15- 8-17-15-20	15-15-15-15-15	40-40-40-40-40	40-40-40-40-40
Percentage uitschieters		10-10-10-10-10	0- 0- 0-10-10	0- 0- 0- 0-10	0- 0- 0- 0-10	0- 0- 0- 0-20	0- 0- 0-20-20	0- 0- 0- 0-30
Variantie uitschieters		5	5	10	10	10	10	10
Perc. H_0 verworpen	Robuust	5.6	5.8	6.0	4.6	5.4	7.8	7.2
	Niet robuust	3.0	7.2	10.6	2.6	6.8	10.0	10.0

Aantal groepen		5	5	5	5	5	5
Aantal waarnemingen		15-20-17-15- 8	15- 8-17-15-20	15-20-17-15- 8	15- 8-17-15-20	70-30-40-50-10	10-30-40-50-70
Percentage uitschieters		0- 0- 0- 0-20	0- 0- 0- 0-20	0- 0- 0- 0-20	0- 0- 0- 0-20	0- 0- 0- 0-20	0- 0- 0- 0-20
Variantie uitschieters		5	5	10	10	10	10
Perc. H_0 verworpen	Robuust	8.6	6.4	9.4	7.0	10.8	4.6
	Niet robuust	14.6	5.6	21.8	4.6	34.6	1.4

Resultaten van de testen waarbij de H_0-hypothese waar is (vervolg).

Aantal groepen		8	8	8	8	8
Aantal waarnemingen		12-12-12-12-12-12-12-12	12-12-12-12-12-12-12-12	12-12-12-12-12-12-12-12	15-15-15-15-15-15-15-15	15-15-15-15-15-15-15-15
Percentage uitschieters		10-10-10-10-10-10-10-10	0- 0- 0- 0- 0-10-10-10	0- 0- 0- 0- 0- 0- 0-10	10-10-10-10-10-10-10-10	0- 0- 0- 0- 0- 0- 0-10
Variantie uitschieters		5	5	5	5	5
Perc. H_0 verworpen	Robuust	6.0	5.6	5.4	5.8	4.6
	Niet robuust	3.4	5.6	5.2	4.0	4.8

Aantal groepen		8	8	8	8	8
Aantal waarnemingen		16-12-12-12-12-12-12- 8	8-12-12-12-12-12-12-16	16-12-12-12-12-12-12- 8	16-12-12-12-12-12-12- 8	16-12-12-12-12-12-12- 8
Percentage uitschieters		0- 0- 0- 0- 0- 0- 0-10	0- 0- 0- 0- 0- 0- 0-10	0- 0- 0- 0- 0- 0- 0-10	0- 0- 0- 0- 0- 0- 0-20	0- 0- 0- 0- 0- 0- 0-20
Variantie uitschieters		5	5	10	5	10
Perc. H_0 verworpen	Robuust	5.0	5.6	4.8	8.4	7.0
	Niet robuust	7.2	4.4	10.6	11.4	14.0

Aantal groepen		3	5
Aantal waarnemingen		15-15-15	15-15-15-15-15
Percentage uitschieters		0- 0- 0	0- 0- 0- 0- 0
Variantie uitschieters		-	-
Perc. H_0 verworpen	Robuust	5.6	6.0
	Niet robuust	6.0	5.6

Onderscheidingsvermogen van de toets.

Resultaten van de testen waarbij de H_0-hypothese niet waar is.

Onbetrouwbaarheid: 0.05; de waarnemingen zijn trekkingen uit een "vervuilde" N(μ, 1)-verdeling.

Aantal groepen		3	3	3	3	3
Aantal waarnemingen		15-15-15	15-15-15	15-15-15	15-15-15	15-15-15
Verwachting μ		0- 0- 1	0- 0- 1	0- 1- 0	1- 0- 0	0- 0- 1
Percentage uitschieters		10-10-10	0- 0-10	0- 0-10	0-10-10	0- 0- 0
Variantie uitschieters		5	5	5	5	-
Perc. H_0 verworpen	Robuust	60.8	71.4	74.4	66.8	77.2
	Niet robuust	35.8	55.2	59.6	46.0	79.8

Aantal groepen		5	5	5	5	5
Aantal waarnemingen		15-15-15-15-15	15-15-15-15-15	15-15-15-15-15	15-15-15-15-15	15-15-15-15-15
Verwachting μ		0- 0- 0- 0- 1	0- 0- 0- 0- 1	0- 0- 0- 1- 0	1- 0- 0- 0- 0	0 -0- 0- 0- 1
Percentage uitschieters		10-10-10-10-10	0- 0- 0- 0-10	0- 0- 0- 0-10	0-10-10-10-10	0- 0- 0- 0- 0
Variantie uitschieters		5	5	5	5	-
Perc. H_0 verworpen	Robuust	60.6	66.2	70.2	64.8	76.0
	Niet robuust	32.2	58.8	59.4	36.0	76.8

Bijlage 4 bij THE-RC 59383

Procedure voor het robuust toetsen van gelijkheid van populatiegemiddelden, waarbij de varianties gelijk zijn.
(Voorlopige uitgave).

ROBUST ANOVA

Korte functiebeschrijving.

Binnen een aantal groepen worden waarnemingen gegeven. Het is niet nodig dat het aantal waarnemingen per groep gelijk is; wel wordt geëist dat het aantal in elke groep minstens 5 is. Een volledige tabel voor variantie-analyse wordt berekend (zie Methode). Met behulp van een Fisher-verdeelde toetsingsgrootheid kan men de nulhypothese toetsen dat de populatiegemiddelden voor de verschillende groepen gelijk zijn. Deze toets vereist normale verdelingen en gelijke populatievarianties en is resistent tegen 10% uitschieters. Alles wordt dus zodanig bepaald dat grote fouten in de waarnemingen geen of weinig invloed hebben op het eindresultaat.

Procedure heading.

```
BOOLEAN PROCEDURE ROBUSTANOVA(X, GROUP, M, N, NUMBEROFGROUPS,
                              SUMSQUARES, DEGREESOFFREEDOM, MEANSQUARES,
                              FISHER);
VALUE M, N, NUMBEROFGROUPS;
INTEGER M, N, NUMBEROFGROUPS;
INTEGER ARRAY GROUP, DEGREESOFFREEDOM[*];
REAL FISHER;
REAL ARRAY X, SUMSQUARES, MEANSQUARES[*];
```

Formele parameters.

X	bevat bij aanroep de waarnemingen.
GROUP	bevat bij aanroep voor ieder element van X het nummer van de groep waartoe dit element behoort.
M, N	kleinste, respectievelijk grootste index voor X en GROUP.
NUMBEROFGROUPS	bevat bij aanroep het aantal groepen. Deze groepen zijn genummerd van 1 tot en met NUMBEROFGROUPS.
SUMSQUARES[1:3]	bevat na afloop de kwadraatsommen (zie Methode).
DEGREESOFFREEDOM[1:3]	bevat na afloop de aantallen vrijheidsgraden (zie Methode).
MEANSQUARES[1:3]	bevat na afloop de gemiddelde kwadraten (zie Methode).

FISHER	bevat na afloop de Fisher-verdeelde toetsingsgrootheid (zie Methode).
ROBUSTANOVA	TRUE, als het gelukt is om een goede robuuste schatting te maken met ROBUSTMULTIPLEREGRESSION (zie Methode). FALSE, als niet zo; dan is er dus geen variantie-analyse-tabel. SUMSQUARES, DEGREESOFFREEDOM, MEANSQUARES en FISHER hebben na afloop dan ook geen waarde van betekenis.

Methode.

Afkortingen: SS = SUMSQUARES

DF = DEGREESOFFREEDOM

MS = MEANSQUARES

F = FISHER

k = NUMBEROFGROUPS

k_i is de grootte van de steekproef uit de i-de groep.

$\hat{\underline{x}}_i$ is het robuust berekende steekproefgemiddelde van de i-de groep

x_{ij} is de j-de waarneming binnen de i-de groep

$n_j = x_{ij} - \hat{\underline{x}}_i$

n = N

m = M

Op robuuste wijze worden de gemiddelden van de groepen bepaald met behulp van ROBUSTMULTIPLEREGRESSION (zie [3]): $\hat{\underline{x}}_i$ $(1 \leq i \leq k)$. Als dit niet lukt, dan is ROBUSTANOVA = FALSE (ROBUSTANOVA := CONV, zie [3]). Vervolgens worden "nieuwe waarnemingen" gemaakt:

$$x^*_{ij} = \hat{\underline{x}}_i + r^*_{ij} \qquad 1 \leq i \leq k,\ 1 \leq j \leq k_i, \text{ met:}$$

$$r^*_{ij} = \frac{C \cdot \Psi\left(\frac{r_{ij}}{\sigma}\right)\sigma}{\frac{1}{n-m+1} \sum_{s=1}^{k} \sum_{t=1}^{s} \Psi'\left(\frac{r_{st}}{\sigma}\right)\sigma}$$

waarbij σ berekend wordt met ROBUSTSIGMA (zie [3]) en

$$C = 1 + \frac{k \cdot \mathrm{var}\left[\Psi'\left(\frac{r_{ij}}{\sigma}\right)\right]}{(n-m+1)\left(E\left[\Psi'\left(\frac{r_{ij}}{\sigma}\right)\right]\right)^2}$$

$$E\left[\Psi'\left(\frac{r_{ij}}{\sigma}\right)\right] = \frac{1}{n-m+1} \sum_{i=1}^{k} \sum_{j=1}^{k_i} \Psi'\left(\frac{r_{ij}}{\sigma}\right)$$

$$\mathrm{var}\left[\Psi'\left(\frac{r_{ij}}{\sigma}\right)\right] = \frac{1}{n-m+1} \sum_{i=1}^{k} \sum_{j=1}^{k_i} \left[\Psi'\left(\frac{r_{ij}}{\sigma}\right) - E\left[\Psi'\left(\frac{r_{ij}}{\sigma}\right)\right]\right]^2$$

en $\frac{\Psi(x)}{x}$ = WEIGHTFUNCTION (6, X, 1.35) = $\frac{1.35}{|x|}$ als $|x| > 1.35$, anders 1 (zie [3]).

Met deze nieuwe waarnemingen wordt nu de variantie-tabel gemaakt (zie tabel, waarbij $\bar{x}^* := \frac{1}{n-m+1} \sum_{i=1}^{k} \sum_{j=1}^{k_i} x^*_{ij}$).

Tabel:

Bron van variantie	kwadraatsom	vrijheids-graden	gemiddeld kwadraat	Fisher
tussen-groepen	$SS_1 = \sum_{i=1}^{k} k_i (\hat{\bar{x}}_i - \bar{x}^*)^2$	$DF_1 = k-1$	$MS_1 = \frac{SS_1}{DF_1}$	$F = \frac{MS_1}{MS_2}$
residu	$SS_2 = SS_3 - SS_1$	$DF_2 = n-m+1-k$	$MS_2 = \frac{SS_2}{DF_2}$	
totaal	$SS_3 = \sum_{i=1}^{k} \sum_{j=1}^{k_i} (x^*_{ij} - \bar{x}^*)^2$	$DF_3 = n-m$	$MS_3 = \frac{SS_3}{DF_3}$	

Externe relaties.

Uit de procedurebibliotheek worden aangeroepen:

ROBUSTMULTIPLEREGRESSION

ROBUSTSIGMA

WEIGHTFUNCTION

MULTIPLEREGRESSION

ABORT

Opmerkingen.

1. Aan de invoer worden de volgende eisen gesteld:
   - Er dienen tenminste twee groepen te zijn.
   - Alle groepen moeten minstens 5 waarnemingen bevatten.
   - De variantie mag niet de waarde 0 aannemen.

   Indien hieraan niet voldaan is, wordt de executie van het programma afgebroken en verschijnt een passende melding op de uitvoer.
2. Voor het toetsen van de nulhypothese kan gebruik gemaakt worden van FISHERSTATISTIC met DF_1 vrijheidsgraden in de teller en DF_2 in de noemer. Deze procedure staat beschreven in [2].

Literatuur.

[1] Huber, Peter J.
Robust Statistics
Wiley, New York, 1981

[2] RC-Informatie P-4.11.
Verdelingsfuncties

[3] Algol procedures voor robuuste regressie (Voorlopige uitgave)
THE-RC 58065

[4] Hontelez, Jan
Het robuuste toetsen van gelijkheid van populatiegemiddelden waarbij de varianties gelijk zijn
CC-Note 21