KM 25(1987)
pag 57 - 73

COMPARISON OF SEVERAL MEAN VALUES IN THE PRESENCE OF OUTLIERS

Jan B. Dijkstra and Hans Linders*

Abstract

The behaviour of several methods for one-way analysis of variance is examined for contaminated normal data. The test are: Classical Anova, Van der Waerden, Trimmed and Winsorized Anova and Huber's method. Two kinds of contamination are considered: symmetric and one-sided.

* Computing Centre, Eindhoven University of Technology, P.O. Box 513, 5600 MB EINDHOVEN, The Netherlands, tel. 040 - 474535/474089.

1. Introduction

The model in classical one-way anova is $\underline{y}_{ij} = \mu_i + \underline{e}_{ij}$ where the errors $\underline{e}_{ij}$ are supposed to be independently distributed as $N(0, \sigma^2)$ for unknown σ^2. The index i denotes the group-number ($i = 1, \ldots, k$) and j identifies the elements within the groups ($j = 1, \ldots, n_i$). The hypothesis of interest is H_0: $\mu_1 = \ldots = \mu_k$. According to the above conditions, this hypothesis can be tested with

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

where $N = \sum_{i=1}^{k} n_i$, $\bar{y}_i$ is the sample mean within the i-th group and $\bar{y}$ is the overall sample mean. The decision rule is to reject H_0 if $F > F_{N-k}^{k-1}(\alpha)$ for some chosen size α.

For contaminated normal data we consider the following modification: with (small) probability ε the distribution becomes $\underline{e}_{ij} \sim N(0, a\,\sigma^2)$, where $a >> 1$, and with probability $1-\varepsilon$ the distribution remains $N(0, \sigma^2)$. This contamination is symmetric; in the asymmetric case, multiplication by $\sqrt{a}$ is performed on the positive errors only, with probability 2ε. In both cases, the expected fraction of outliers is ε.
In practice, errors of this form with a high value of a may be caused by misplacing or forgetting a decimal point. And once fed into a computer file they are never seen by human eyes again.
Classical one-way anova is not designed for contaminated normal data. Using this test here might result in a probability of rejecting H_0 when true that differs from the chosen size α, or in a serious loss of power. In the next sections some alternatives are presented that seem to be more robust in these respects. A comparative study concerning the size and power of all the tests under consideration will be given, where the effect of symmetric and one-sided contamination is demonstrated by simulation.

2. Nonparametric Anova

In a nonparametric test the hypothesis is not the same as in the previous section, but it can be expressed as "all samples come from the same continuous distribution". Nonparametric anova has little power for the comparison of shapes, but it can be used to test the equality of location parameters. The density in case of symmetric contamination is given by:

$$f(x) = \varepsilon \frac{1}{\sigma\sqrt{a2\pi}} \exp\left[-\frac{x^2}{2a\sigma^2}\right] + (1-\varepsilon) \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{x^2}{2\sigma^2}\right]$$

and this represents a continuous distribution. Therefore the application of nonparametric anova is permitted. It is easily seen that this also holds for one-sided contamination.

Several nonparametric tests are available, but here we will use only the Van der Waerden test (1952). This test is based on the following statistic:

$$Q = \frac{N-1}{h} \sum_{i=1}^{k} \frac{1}{n_i} \left[\sum_{\ell \in S_i} \Phi^{-1}\left(\frac{R_\ell}{N+1}\right) \right]^2, \text{ where } h = \sum_{\ell=1}^{N} \left[\Phi^{-1}\left(\frac{\ell}{N+1}\right)\right]^2.$$

Here $y_1, \ldots, y_N$ represents the combined sample, where the groups are represented by sets of indices S_i for $i = 1, \ldots, k$. R_ℓ is the rank of y_ℓ and Φ denotes the standard normal distribution. Q is asymptotically distributed as χ^2_{k-1} and for small samples the critical values for Q are tabulated.

The reason for choosing the Van der Waerden test from the large collection of methods for nonparametric anova, lies in the fact that this is the only test that has for $\varepsilon = 0$ asymptotically the same efficiency as the classical test [Hajek (1969)]. By using this nonparametric method one is insured against the possible presence of outliers, and the premium one has to pay is the loss of power for small samples. For $k = 2$ this loss has already been shown to be moderate [Van der Laan and Oosterhoff (1967)].

3. Winsorizing and Trimming

Applications of these methods to the t-test for two samples have been published already. The t-test uses the statistic

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{SS_1 + SS_2}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{where } SS_i = \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

The hypothesis H_0: $\mu_1 = \mu_2$ is rejected if $|t| > t_{N-2}(\frac{1}{2}\alpha)$ for some chosen size α. This method is equivalent to classical one-way anova for $k = 2$: t^2 equals F and for the critical values the same relation holds ($t^2_\nu = F^1_\nu$).

Fung and Rahman (1980) Winsorized the t-test in an attempt to make it robust against the presence of outliers. This is done as follows: let $a_1, \ldots, a_n$ be an ordered sample. Then the mean and sum of squares of this sample, after two-sided Winsorizing with parameter g, are defined as:

$$\bar{a}_{wg} = \frac{1}{n}\{(g+1)a_{g+1} + a_{g+2} + \ldots + a_{n-g-1} + (g+1)a_{n-g}\}$$

$$SS_{wg} = (g+1)(a_{g+1} - \bar{a}_{wg})^2 + (a_{g+2} - \bar{a}_{wg})^2 + \ldots$$

$$\ldots + (a_{n-g-1} - \bar{a}_{wg})^2 + (g+1)(a_{n-g} - \bar{a}_{wg})^2.$$

The number of relevant observations hereby reduces to $h = n-2g$. Application of this technique to the t-test gives the following formula:

$$t_{wg} = \frac{\bar{y}_{1wg} - \bar{y}_{2wg}}{\sqrt{\frac{SS_{1wg} + SS_{2wg}}{h_1 + h_2 - 2}}\sqrt{\frac{1}{h_1} + \frac{1}{h_2}}}.$$

This statistic approximately follows a t-distribution with $h_1 + h_2 - 2$ degrees of freedom. Fung and Rahman used n_i instead of h_i under the second square-root sign, but that appears to be have been a typing error as can be concluded from a study by Yuen and Dixon (1973) on which they based their approach.

Winsorizing means replacing the tail-elements by the most extreme elements that are not considered to belong to the tails. Trimming is a different technique in which the tail-elements are simply deleted. Yuen and Dixon examined the behaviour of the trimmed t-test, where the numerator is based on trimmed means, but the denominator still contains

Winsorized sums of squares. In a simulation study with $n_i \geq 10$ both methods show the same qualities: The probability of rejecting H_0 when true is almost equal to the chosen size, and the power for normal distributions is only slightly below that of the classical t-test. For distributions with heavier tails the Winsorised and trimmed t-tests are even more powerful than the classical t-test for moderate values of g [Fung and Rahman (1980)].

Therefore it could be attractive to apply these techniques to classical one-way anova, which is nothing more than a generalisation of the t-test for $k > 2$. The Winsorized F-statistic is given by

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

where $\bar{y}_{wg} = \sum_{i=1}^{k} h_i \bar{y}_{iwg}/H$ and $H = \sum_{i=1}^{k} h_i$. For the trimmed F-statistic F_{tg} only the numerator of F_{wg} is modified; the Winsorized means are replaced by trimmed means $\bar{y}_{itg}$ and the trimmed overall sample mean is given by $\bar{y}_{tg} = \sum_{i=1}^{k} h_i \bar{y}_{itg}/H$. It is assumed that both F_{wg} and F_{tg} are approximately distributed as F_{H-k}^{k-1}. In a previous simulation [Dijkstra (1986)] it was found that the probability of rejecting H_0 when true differed too much from the chosen size α for Winsorised and trimmed anova. But after correction of the above mentioned typing error in the paper by Fung and Rahman the behaviour of these tests improved remarkably as will be shown later in this paper.

4. Robust Regression

The model for analysis of variance can be rewritten as a regression model:

$$\underline{y} = \beta_1 x_1 + \ldots + \beta_k x_k + \underline{e}.$$

The observations are represented by y and for every observation the group to which it belongs is identified by the dummy-variables $x_1, \ldots x_k$. This is done as follows: $x_i = 1$ if y belongs to group i and otherwise $x_i = 0$. If the errors were independently distributed as

$N(0, \sigma^2)$ then testing H_0: $\beta_1 = \ldots = \beta_k$ would be equivalent to testing H_0: $\mu_1 = \ldots = \mu_k$ in the model for classical one-way anova. The values of F and the corresponding numbers of degrees of freedom would be the same.

Several methods for dealing with outliers in regression models have already been published. Huber (1973) suggested a method with attractive properties that can be applied to the analysis of variance problem in this study.

The objective of classical regression is to minimize $\sum_{i=1}^{N} (y_i - x_i^T\beta)^2$ as a function of $\beta = (\beta_1, \ldots, \beta_k)^T$. Here $x_i = (x_{i1}, \ldots, x_{ik})^T$. It can easily be understood that outliers in y will have considerable influence on the estimation of β, because classical regression will square their residuals.

In robust regression a different objective is used:

$$\min_{\beta} \sum_{i=1}^{N} \rho\left(\frac{y_i - x_i^T\beta}{\sigma}\right).$$

In the classical case $\rho(r) = r^2$, but in robust regression one chooses a function that limits the influence of extreme residuals. Holland and Welsch (1977) mention eight different functions ρ with this desirable property. The objective will be at its minimum if

$$\sum_{i=1}^{N} x_{ij}\, \Psi\left(\frac{y_i - x_i^T\beta}{\sigma}\right) = 0$$

for $j = 1, \ldots, k$ and $\Psi = \rho'$. Several iterative methods for solving these equations can be considered. Initial estimates for $\beta_1, \ldots, \beta_k$ can be obtained by ordinary least squares, whereafter σ can be estimated as

$$\hat{\sigma} = 1.4826 \left[\operatorname{med}_j \left| (y_j - x_j^T\hat{\beta}) - \operatorname{med}_i (y_i - x_i^T\hat{\beta}) \right|\right].$$

Without restrictions on the weightfunction, convergence cannot in general be guaranteed if the estimation of σ is part of the iteration. Huber (1973) found a ρ that allows iteratively re-estimating of σ:

$$\rho(r) = \frac{r^2}{2} \text{ for } |r| \leq H$$

$$\rho(r) = H|r| - \frac{H^2}{2} \text{ for } |r| > H.$$

The sensitivity to outliers depends on the value of H. For H = 1.345 the efficiency is 95% for normal distributions. If the absolute value of a standardised residual exceeds H, its influence becomes linear instead of quadratic. Although Huber's ρ does not yield an extremely robust estimate (some authors prefer a ρ that becomes a constant for big values of r), this method is a considerable improvement on ordinary least squares in the presence of outliers.

In this case Newton's method yields a very efficient algorithm, because $\Psi = \rho'$ is a broken linear function.

For the construction of an outlier-resistant analysis of variance procedures we consider the above mentioned robust regression with Huber's ρ and H = 1.345. This approach results in fitted value $\hat{y}_i^{\cdot}$ and an estimate $\hat{\sigma}^{\cdot}$ for σ. Huber (1981) suggested a test for the hypothesis of equal population means that uses these estimates. His suggestion is the topic of the next section.

5. Huber's Method

In the classical situation (without outliers) the test statistic for $H_0: \mu_1 = \ldots = \mu_k$, is

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

Huber gave an $F^{\cdot}$ that is similar to F, but on which the outliers have less influence. In the numerator the first step is to replace $\bar{y}_i$ by $\hat{y}_i^{\cdot}$. In a more general model Huber suggests to replace $\bar{y}$ by an ordinary least squares fit using $\hat{y}^{\cdot}$ instead of y. In this case (without covariables) such a fit will yield the weighted mean

$$\bar{y}^{\cdot} = \frac{\sum_{i=1}^{k} n_i \hat{y}_i^{\cdot}}{N}.$$

After scaling this modified numerator follows under mild conditions

asymptotically a χ^2-distribution with the same number of degrees of freedom as in the classical test.

Dealing with the denominator is a bit more difficult: one single outlier can be the cause of an extremely high value, so that H_0 can be accepted although the location parameters are very different. Huber proposes to replace the denominator by the following expression (where the influence of the outliers is reduced considerably):

$$\frac{1}{N-k} \frac{c^2 \sum_{i=1}^{N} \Psi\left(\frac{r_i}{\hat{\sigma}^*}\right)^2 \hat{\sigma}^{*2}}{\left[\frac{1}{N} \sum_{i=1}^{N} \Psi'\left(\frac{r_i}{\hat{\sigma}^*}\right)\right]^2}$$

where $r_i = y_i - \hat{y}_i^*$

$$c = 1 + \frac{k \, \hat{var}(\Psi')}{N\left[\hat{E}(\Psi')\right]^2}$$

$$\hat{E}(\Psi') = \frac{1}{N} \sum_{i=1}^{N} \Psi'\left(\frac{r_i}{\hat{\sigma}^*}\right)$$

$$var(\Psi') = \frac{1}{N} \sum_{i=1}^{N} \left[\Psi'\left(\frac{r_i}{\hat{\sigma}^*}\right) - \hat{E}(\Psi')\right]^2.$$

Note that in case of Huber's ρ these formulae are simplified considerably, since Ψ' can only take values 0 and 1. In this case we have

$$c = 1 + k\frac{N-p}{Np},$$

where p is the number of observations for which $\Psi'\left(\frac{r_i}{\hat{\sigma}^*}\right) = 1$.

Just like in classical anova H_0: $\mu_1 = \ldots = \mu_k$ is to be rejected if $F^* > F_{N-k}^{k-1}(\alpha)$ for some chosen size α. Huber claims that the approximation of F^* by an F-distribution is reasonable if $n_i \geq 5$ for $i = 1, \ldots, k$. This is the same condition that is usually put forward for using nonparametric tests with a χ^2-distribution.

This approach is very sensitive to "leverage points" (Belsley, Kuh and Welsch, 1980); covariables can be included in the model, but they may not contain outliers. The test can be generalised to more complex designs, including interactions. In this respect Huber's methods seems more promising than its nonparametric alternatives, where the concept of rank-interaction is a complex matter, even for a simple two-way layout [De Kroon and Van der Laan (1981)].

6. The Actual Size of the Tests

We estimated the probability of rejecting H_0 when true by using a simulation with 2000 replications. This we did for 3 and 6 groups, symmetric and one-sided contamination and sample sizes of 10, 25 and 40. The samples were generated from normal populations with $\mu_i = 0$ and $\sigma^2 = 1$. Symmetric contamination was simulated by using $\sigma^2 = 50$ with probabilities 0, 0.03 and 0.1. For trimming and Winsorizing the constant g was chosen proportional to the sample sizes. The results of these simulations are presented in tables 1 and 2, where the estimated size for each situation is given as the percentage of rejections for a test with nominal size $\alpha = 0.05$.

n_i	ε	g	Anova	VdW	Trim	Wins	Huber
10	0	2	4.95	4.15	5.25	5.25	5.45
10	0.03	2	3.95	4.10	5.00	5.20	5.15
10	0.1	2	2.85	4.90	5.70	5.45	5.25
25	0	3	4.80	5.05	5.40	5.05	5.35
25	0.03	3	3.25	4.20	4.80	4.75	5.20
25	0.1	3	4.00	5.05	4.95	6.40	5.40
40	0	5	5.15	4.95	5.15	4.65	5.00
40	0.03	5	5.15	5.20	5.30	4.80	5.25
40	0.1	5	4.35	4.65	4.70	5.45	4.60
A	0	B	4.45	3.80	3.90	3.85	4.40
A	0.03	B	4.85	5.00	4.10	4.40	5.70
A	0.1	B	5.30	4.75	4.25	4.95	5.25

Table 1: Symmetric Contamination, k = 3

n_i	ε	g	Anova	VdW	Trim	Wins	Huber
10	0	2	5.25	4.50	5.65	4.95	6.05
10	0.03	2	3.10	3.50	5.35	4.70	5.45
10	0.1	2	3.20	3.85	5.20	5.25	5.20
25	0	3	4.40	4.20	4.50	3.95	5.05
25	0.03	3	4.25	4.95	4.95	4.70	5.30
25	0.1	3	3.95	5.00	4.35	7.20	5.15
40	0	5	5.90	5.65	5.90	5.30	6.05
40	0.03	5	4.20	4.70	5.30	5.05	5.30
40	0.1	5	4.30	4.35	4.20	6.10	4.15
C	0	D	4.75	4.65	3.95	3.40	5.25
C	0.03	D	4.45	4.55	4.75	4.20	5.75
C	0.1	D	6.20	5.15	4.00	5.85	5.55

Table 2: Symmetric Contamination, k = 6

code	meaning
A	10, 25, 40
B	2, 3, 5
C	10, 10, 25, 25, 40, 40
D	2, 2, 3, 3, 5, 5

Table 3: The codes used

In the case of one-sided contamination the use of $\sigma^2 = 50$ was restricted to positive observations. At the same time, the probability of a multiplication by $\sqrt{50}$ was doubled to 2ε, in order to get the same expected number of outliers as with symmetric contamination. The results of this simulation are presented in tables 4 and 5.

n_i	ε	g	Anova	VdW	Trim	Wins	Huber
10	0	2	4.75	4.15	5.60	5.45	6.35
10	0.03	2	4.00	4.75	5.75	5.70	5.65
10	0.1	2	3.75	5.00	5.20	5.35	5.65
25	0	3	5.65	5.20	5.30	5.20	6.00
25	0.03	3	3.80	4.40	4.75	4.65	5.20
25	0.1	3	3.50	4.90	3.90	7.50	5.15
40	0	5	4.85	4.70	4.60	4.10	4.60
40	0.03	5	4.75	5.25	5.45	5.15	5.80
40	0.1	5	4.95	5.60	4.75	9.40	5.55
A	0	B	5.10	5.35	4.60	4.50	5.65
A	0.03	B	4.35	4.80	4.40	4.50	5.30
A	0.1	B	4.90	4.55	3.60	6.05	5.35

Table 4: One-sided Contamination, k = 3

n_i	ε	g	Anova	VdW	Trim	Wins	Huber
10	0	2	5.80	4.65	4.85	5.10	6.10
10	0.03	2	3.95	4.60	5.90	5.75	6.15
10	0.1	2	3.15	5.05	4.85	5.20	6.35
25	0	3	4.55	4.15	5.00	4.55	5.20
25	0.03	3	5.55	5.20	5.35	6.00	5.90
25	0.1	3	4.15	4.90	4.25	11.80	5.65
40	0	5	4.15	4.35	4.30	3.55	4.30
40	0.03	5	4.40	4.90	4.55	4.90	4.95
40	0.1	5	4.35	4.10	3.55	12.70	4.10
C	0	D	5.55	4.95	4.60	4.55	6.20
C	0.03	D	5.20	4.80	3.95	4.30	5.55
C	0.1	D	5.90	5.15	4.05	9.15	6.15

Table 5: One-sided Contamination, k = 6

The tables are not very clear if one wants to compare these tests. The standard deviation of the estimated size is $(0.05 \times 0.95/2000)^{\frac{1}{2}} = 0.004873$ or 0.4873%. Let d be the percentage of rejected hypotheses minus 5, divided by this standard deviation. Tables 6 and 7 show the

values of d for each test. Three categories have been separated by dotted lines: $d < -2$ (conservative), $-2 \leq d < 2$ (accurate) and $2 \leq d$ (progressive).

	Anova	VdW	Trim	Wins	Huber
$d < -3$	4	1	0	1	0
$-3 \leq d < -2$	3	2	3	2	0
$-2 \leq d < -1$	7	6	5	2	2
$-1 \leq d < 1$	8	14	13	15	17
$1 \leq d < 2$	1	1	3	1	3
$2 \leq d < 3$	1	0	0	2	2
$3 \leq d < 4$	0	0	0	0	0
$4 \leq d < 5$	0	0	0	1	0
$5 \leq d$	0	0	0	0	0

Table 6: Symmetric Contamination

	Anova	VdW	Trim	Wins	Huber
$d < -3$	2	0	0	0	0
$-3 \leq d < -2$	4	0	4	1	0
$-2 \leq d < -1$	5	5	4	4	2
$-1 \leq d < 1$	8	18	13	10	7
$1 \leq d < 2$	5	1	3	2	8
$2 \leq d < 3$	0	0	0	2	7
$3 \leq d < 4$	0	0	0	0	0
$4 \leq d < 5$	0	0	0	0	0
$5 \leq d$	0	0	0	5	0

Table 7: One-sided Contamination

Tables 6 and 7 suggest the following conclusions:

- Classical anova tends to be conservative in the presence of outliers.

- The method of Van der Waerden is unaffected concerning the size by this kind of non-normality, which is just what might be expected from a nonparametric test.
- The trimmed test seems slightly conservative in this situation, but less than classical anova.
- Symmetric contamination does not seem to affect the Winsorized test very much, but this method is clearly not robust against one-sided contamination. Tables 4 and 5 show that the cases where $5 \leq d$ have a very high proportion of outliers: $\varepsilon = 0.1$. Such values of ε make it possible that outliers are found in the body of a sample and not only in its tails (as defined by g). It would be unreasonable to expect robustness against this situation in a Winsorized test, because a tail consisting of outliers can enter the computation. This problem can not occur in a trimmed test.
- Huber's method seems the best for symmetric contamination, although the differences with the other tests are not convincing (only classical anova is too convervative). Against one-sided contamination the suggestion of a slight progressiveness exists. Values of d between 2 and 3 occured in 7 cases. It is interesting to note that 4 of these cases contained no outliers ($\varepsilon = 0$), so that the results for these rows in the tables for symmetric and one-sided contamination should be similar. An examination of all the results for Huber's method shows that indeed a very slight progressiveness exists, but that the contamination has almost no influence (see table 8).

contamination	estimated size in %
none ($\varepsilon = 0$)	5.437
symmetric	5.228
one-sided	5.528

Table 8: Huber's Method

The estimated sizes in table 8 are based on 16 × 2000 replications, so that their standard deviation is 0.4873/4 = 0.1218. Two of the three sizes differ significantly from 5%, and it is clear that the approximation of Huber's test statistic by an F-distribution can be improved. But for practical purposes these results are acceptable.

7. A Comparison of Powers

Here we present a simulation study that differs from the one in the previous section in only one respect: the samples were generated with unequal location parameters. Table 9 is based on symmetric contamination with three samples.

n_i	ε	g	μ_i(× 0.1)	Anova	VdW	Trim	Wins	Huber
10	0	2	0, 8, 16	88.05	85.50	77.85	78.05	84.50
10	0.03	2	0, 8, 16	64.55	75.95	74.05	73.95	80.75
10	0.1	2	0, 8, 16	36.75	59.20	64.70	65.05	68.25
25	0	3	0, 5, 10	88.20	87.25	84.45	84.50	85.45
25	0.03	3	0, 5, 10	59.50	80.90	80.95	81.25	82.45
25	0.1	3	0, 5, 10	29.00	63.25	69.05	69.15	71.00
40	0	5	0, 4, 8	89.55	89.30	87.15	87.30	87.05
40	0.03	5	0, 4, 8	57.30	82.15	82.95	82.90	83.65
40	0.1	5	0, 4, 8	27.45	66.15	73.65	74.10	75.50
A	0	B	0, 8, 13	92.65	92.10	86.95	86.85	91.00
A	0.03	B	0, 8, 13	64.75	86.20	83.10	83.10	87.35
A	0.1	B	0, 8, 13	31.25	72.00	74.80	75.55	80.45

Table 9: Symmetric Contamination, k = 3

We also generated tables for symmetric contamination with k = 6 and one-sided contamination with k = 3 and k = 6, but the results were very similar and therefore they will not be presented here. A summary of these results is given in table 10, where the powers for uncontaminated data (ε = 0) are the means of 16 separate simulations with 2000 replications each. The other results are based on 8 simulations with the same number of replications.

This table suggests the following conclusions:

- Classical Anova is the most powerful test for normal data, but contamination reduces the power of this method considerably. It does not matter whether the contamination is symmetric or one-sided; only the number of outliers (for some chosen variance) appears to have any influence.

	ε	Anova	VdW	Trim	Wins	Huber
	0	90.50	89.44	85.25	85.41	88.19
symmetric	0.03	59.63	82.06	81.70	81.70	84.68
	0.1	28.55	65.54	71.43	72.54	75.12
one-sided	0.03	59.59	82.99	81.71	81.78	85.19
	0.1	29.20	68.71	65.48	68.88	75.08

Table 10: Comparison of Powers

- Table 9, as well as the tables that we did not include in this paper, show that the difference in power for normal data ($\varepsilon = 0$) between classical anova and the test of Van der Waerden almost disappears as the sample size increases from 10 to 40. Even for small samples ($n_i = 10$) the difference is only marginal. The influence of outliers on Van der Waerdens test is considerably smaller than on classical anova, especially as their number increases.
- Trimming and Winsorizing give similar results, except for one-sided contamination with $\varepsilon = 0.1$, where Winsorizing seems to provide a more powerful test. But that is just the situation where Winsorizing should not be trusted because outliers can occur between the tails of a sample (as defined by g) resulting in a probability of rejecting H_0 when true that exceeds the chosen size α considerably. Table 7 shows that trimming is insensitive to this problem, at least with our values of g. For the smaller values of ε, the values of g could be lowered, which might result in a somewhat higher power.
- Huber's method yields the most powerful test, except when the data come from uncontaminated normal distributions in which case classical anova and Van der Waerden's test have slightly more power.

The aim of this study was to select a test for outlier-resistant one-way anova that could be added to the local collection of statistical software at Eindhoven University of Technology. Considering the accuracy of the actual size, and the superior power of Huber's method, we reached the conclusion that this test was the appropriate choice. However, the differences with Van der Waerden's test and trimming are small, and Huber's greater power may be partly attributed to its greater size. So Van der Waerden's test and trimming are equally acceptable choices. On the other hand, Winsorizing is not to be recommended.

We like to thank prof.dr. R. Doornbos and prof.dr. P.J.M. Rousseeuw for their helpful comments on an earlier version of this paper.

8. Literature

[1] Beaton, A.E. and J.W. Tukey (1974)
The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data
Technometrics (16) 147-185

[2] Belsley, D.A., E. Kuh and R.E. Welsch (1980)
Regression Diagnostics
John Wiley & Sons, New York

[3] Dijkstra, Jan B. (1986)
Robuuste Variantie-analyse
Computing Centre Note 30, Eindhoven University of Technology

[4] Fung, K.Y. and S.M. Rahman (1980)
The two-sample Winsorized t
Communications in Statistics (B9. no. 4), 337-347

[5] Hajek, J. (1969)
A course in nonparametric statistics
Holden-Day, San Francisco

[6] Holland, P.W. and R.E. Welsch (1977)
Robust regression using iteratively reweighted least-squares
Communications in Statistics A6(9), 813-827

[7] Hontelez, J. (1984)
Een uitschieter-resistente procedure voor enkelvoudige klassieke variantie-analyse
Computing Centre Note 21, Eindhoven University of Technology

[8] Huber, P.J. (1972)
Robust statistics, a review
The Annals of Math. Stat. (43, no. 4), 1041-1067

[9] Huber, P.J. (1973)
Robust statistics: asymptotics, conjectures, and Monte Carlo
Ann. Statist. (1), 799-821

[10] Huber, P.J. (1981)
Robust Statistics
John Wiley & Sons, New York

[11] De Kroon, J. and P. van der Laan (1981)
Distribution-free test procedures in two-way layouts; a concept of rank-interaction
Statistica Neerlandica (35, no. 4), 189-213

[12] Van der Laan, P. and J. Oosterhoff (1967)
Experimental determination of the power functions of the two-sample rank tests of Wilcoxon, Van der Waerden and Terry by Monte Carlo techniques
Statistica Neerlandica (21, no. 1), 55-68

[13] Van der Waerden, B.L. (1952)
Order tests for the two-sample problem and their power
Indagationes Math. (14), 453-458

[14] Yuen, K.K. and W.J. Dixon (1973)
The approximate behaviour and performane of the two-sample trimmed t
Biometrika (60), 369-374

Ontvangen 02-12-1987
Geaccepteerd: 22-06-1987