On group sequential tests based on robust location and scale estimators in the two-sample problem

Andreas Christmann
University of Dortmund, HRZ, D-44221 Dortmund, GERMANY

Summary

The behaviour of group sequential tests in the two-sample problem is investigated if one replaces the classical non-robust estimators in the t-test statistic by modern robust estimators of location and scale. Hampel's 3-part redescending M-estimator 25A used in the Princeton study and the robust scale estimators length of the shortest half proposed by Rousseeuw & Leroy and Q proposed by Rousseeuw & Croux are considered. Of special interest are level, power and the average sample size number of the tests. It is investigated, whether commerical software can be used to apply these tests.

Key words: Average sample size number; Group sequential test; Length of the shortest half; Outliers; Redescending M-estimator; Robustness; Scale estimator Q.

1. Introduction

The famous Princeton study (Andrews et al., 1972) had a strong influence for further research on robustness. Much recent research concentrates on robustness properties of estimators for a fixed sample size, e.g. the behaviour of the influence function, breakdown point, maximal bias curve, and efficiency considerations, e.g. Huber (1981), Hampel et al. (1986) and Davies (1993). There are different strategies based on robust estimators to identify outliers, c.f. Hampel (1985), Rousseeuw & van Zomeren (1990), and Davies & Gather (1993). In some areas of applied statistics, e.g. in planning and analyzing clinical trials, group sequential plans play an important role. Such plans can reduce the average sample size number (ASN), i.e. the expected sample size when the test stops, which is attractive from ethical, time and financial aspects, c.f. Pocock (1977) and Pocock (1983, p.142ff.). In contrast to the fixed sample size case, much less research has been published on the application of robust estimators to group sequential plans. However, already Pocock (1977) considers in his fundamental paper a group sequential Wilcoxon test. Mehta et al. (1994) investigate exact permutational tests for group sequential clinical trials with special emphasis on the non-parametric group sequential Wilcoxon test. Silvapulle & Sen (1993) propose robust tests based on a Wald-type statistic in group sequential plans for one- and two-sided hypotheses in the linear model. The authors demonstrate by simulating a two-way analysis of variance model that their test based on an M-estimator corresponding to Huber's Proposal 2 (Huber, 1981) is power robust in contrast to the test depending on the least squares estimator.

The aim of the present paper is to study the behaviour of group sequential two-sample tests for location difference if one replaces the classical non-robust estimators in the t-test statistic by modern robust estimators for location and scale. Four criteria will be considered: the actual level and power of the test, the average sample size number, and the bias of the naive estimated standardized treatment difference. It is investigated, whether commerical software, e.g. EaSt (1993), can be used to apply these tests.

2. Group sequential design

Consider the following group sequential plan for the two-sample situation. Denote the maximal sample size for each of both treatment groups by N, and the maximal number of interim tests by K, $K \geq 1$. Let n_j be the sample size at stage j for each treatment group, and $N_j = n_1 + ... + n_j$ be the sample size up to stage j for each treatment group, $1 \leq j \leq K$. Assume that there are independent random variables $X_1, ..., X_N$ each with distribution function $F((\cdot - \mu_1)/\sigma)$, and $Y_1, ..., Y_N$ each with distribution function $G((\cdot - \mu_2)/\sigma)$. The location parameters $\mu_1 \in \mathrm{I\!R}$, $\mu_2 = \mu_1 + \Delta^* \in \mathrm{I\!R}$, and the scale parameter $\sigma \in (0, \infty)$ are unknown. Let $\Delta = \Delta^*/\sigma \in (0, \infty)$ denote the standardized treatment difference. The usual distribution assumption is that F and G are Gaussian. We will consider the two-sided testing problem

$$\mathrm{H}_0 : \Delta = 0 \quad \text{vs.} \quad \mathrm{H}_1 : \Delta \neq 0. \tag{2.1}$$

Of course, one-sided tests can be treated in an analogous manner. Group sequential tests will be considered which can only reject the hypothesis H_0 early, c.f. Pocock (1977). However, other group sequential test procedures can also be used, e.g. procedures proposed by O'Brien & Fleming (1979), Wang & Tsiatis (1987), Lan & DeMets (1983), and DeMets & Lan (1994). Define the test statistic T_j at stage j by

$$T_j = \sqrt{N_j}\,\frac{\hat{\mu}_{1,N_j} - \hat{\mu}_{2,N_j}}{\sqrt{\hat{\sigma}^2_{1,N_j} + \hat{\sigma}^2_{2,N_j}}} \quad , 1 \leq j \leq K. \tag{2.2}$$

The test decision of the test at stage j is defined by :

$$\begin{array}{lcl} |T_j| > c(j) & : & \text{STOP; Decision for } \mathrm{H}_1 \\ |T_j| \leq c(j) \text{ and } j < K & : & \text{Continue with stage } j+1 \\ |T_K| \leq c(K) & : & \text{STOP; Decision for } \mathrm{H}_0, \end{array}$$

where $c(j)$ denotes the critical constant at stage $j \in \{1, \ldots, K\}$. In the simulation we will consider the case $c(j) = c\, j^{a-0.5}$ where a is some fixed constant, c.f. Wang & Tsiatis (1987). The probability for an error of type I is distributed on the different stages of the interim tests such that

$$P_{\Delta=0}(\exists\, j \in \{1,\ldots,K\}; |T_j| > c(j)) = \alpha.$$

It is well-known that the maximum likelihood estimator of Δ will often be biased even under the classical normality assumption if this estimator is computed after a group sequential test has stopped and different bias reduction methods have been proposed, e.g. Cox (1952), Whitehead (1986), and Kim (1988, 1989). In this paper it is investigated how different pairs of distributions (F, G) and different pairs of robust estimators for $(\mu_1, \mu_2, \sigma_1, \sigma_2)$ influence the bias of the naive estimator for Δ given by $\hat{\Delta} = (\hat{\mu}_2 - \hat{\mu}_1)/[(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)/2]^{1/2}$.

3. Estimators

Three pairs of estimators will be considered for the unknown location and scale parameters. Of course, the classical estimators 'mean' $\bar{X}$ and 'standard deviation' S are used. Hampel's three-part redescending M-estimator 25A (Andrews et al., 1972) is one of the best location estimators in the Princeton study because 25A is asymptotically normal distributed and it has good robustness and good efficiency properties. It is defined as solution of

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \mu}{\hat{\sigma}}\right) = 0\,, \tag{3.1}$$

where

$$\begin{aligned} \psi(r) &= r && \text{if} \quad 0 \le |r| \le a \\ &= a\,\mathrm{sign}(r) && \qquad a \le |r| \le b \\ &= a\,\tfrac{c-|r|}{c-b}\,\mathrm{sign}(r) && \qquad b \le |r| \le c \\ &= 0 && \qquad |r| > c\,, \end{aligned}$$

and $a = 1.645$, $b = 3.0$, $c = 6.5$, $\hat{\sigma}$ is a scale estimator. In the Princeton study 25A is based on the scale estimator $1.483 \times \mathrm{MAD}$, where MAD denotes the median of the absolute deviations from the median. Recent work shows that there are other robust scale estimators which are promising alternatives to $1.483 \times$ MAD.

Denote the order statistics of $Y_1, \ldots, Y_N$ by $Y_{1:N} \leq \ldots \leq Y_{N:N}$. Rousseeuw & Leroy (1988) propose the scale estimator 'length of the shortest half'

$$SH = 0.7413 \cdot \{Y_{h+k-1:N} - Y_{k:N}; k = 1, ..., [\![(n+1)/2]\!]\}, \tag{3.2}$$

where $h = [\![n/2]\!]+1$, and $[\![r]\!]$ is the greatest integer less than or equal to r, $r \geq 0$. The constant 0.7413 is a correction factor which yields Fisher-consistency for normally distributed errors. Rousseeuw & Leroy (1988) consider also a modification of SH, say SH*, where the constant 0.7413 is replaced by a constant c_n to give approximately unbiased estimation results for normally distributed errors. Rousseeuw & Leroy (1988) show that SH has a breakdown point of approximately 0.5 and that the bias of SH can be much lower than the bias of $1.483\times$MAD if there are many outliers, see also Martin and Zamar (1993). Grübel (1988) and Davies (1990) prove that SH is asymptotically normal. Some other properties of SH are given in Christmann, Gather & Scholz (1994).

Croux & Rousseeuw (1992a,b) and Rousseeuw & Croux (1993) consider other robust alternatives to the median absolute deviation. They propose a class of high breakdown point scale estimators based on subranges, e.g. Q, and gave time efficient algorithms to compute such estimates. The scale estimator Q based on random variables $Y_1, \ldots, Y_N$ is defined by

$$Q = 2.2219 \cdot d_N \cdot \{|Y_i - Y_j|; 1 \leq i < j \leq N\}_{L:(N(N-1)/2)}, \tag{3.3}$$

where $h = [\![N/2]\!] + 1$, $L = h(h-1)/2$, $d_N = N/(N+1.4)$ for N odd, and $d_N = N/(N+3.8)$ for N even. Croux & Rousseeuw (1992b) propose to use other values of d_N for sample sizes $N \leq 9$. However, in the present paper Q will only be used for sample sizes larger than 9. The estimator Q has a finite sample breakdown point of approximately 0.5 but in contrast to $1.483\times$MAD, Q has a smooth and bounded influence function at the standard normal distribution. Further, Q is asymptotically normal. The gaussian efficiency of Q

is 82% in contrast to 37% for $1.483 \times \mathrm{MAD}$, c.f. Rousseeuw & Croux (1993). In this article, the very robust estimator SH and the more efficient estimator Q are used as robust scale estimators.

4. Design of the simulations

The tests with test statistic (2.2) are based on three pairs of estimators: $(\bar{X}, S)$, (25A,SH), and (25A,Q). Three different distributions will be considered. Let $\mathrm{N}(0,1)$ be the standard normal distribution, t_3 Student's t-distribution with 3 degrees of freedom, and MIXN a mixture of two normal distributions, which is defined by $0.9\mathrm{N}(0,1) + 0.1\mathrm{N}(10,10^2)$. The standard normal distribution is choosen for two reasons. It is a symmetric distribution with thin tails and many papers and commercial software for planning and analyzing group sequential studies, e.g. EaSt (1993) and the SAS/IML functions SEQ, SEQSCALE and SEQSHIFT (SAS, 1995) assume normally distributed errors in the test problem given in (2.1). Student's distribution t_3 is symmetric with heavier tails than N(0,1) and is often a good approximation to the distribution of high quality data, c.f. Hampel et al. (1986, p. 23). The mixture distribution given above is asymmetric and produces extreme outliers. We consider seven different pairs of distributions (F,G): $(\mathrm{N}(0,1), \mathrm{N}(0,1))$, $(\mathrm{N}(0,1), t_3)$, $(\mathrm{N}(0,1), \mathrm{MIXN})$, (t_3, t_3), (t_3, MIXN), $(\mathrm{MIXN}, \mathrm{MIXN})$, and $(\mathrm{MIXN}, \mathrm{N}(0,1))$. Five values of Δ are considered: $\Delta \in \{0, 0.25, 0.5, 0.75, 1.0\}$. The number of simulations for each design point is $10,000$. In the simulations $\mu = 0$ and $\sigma = 1$ are used which is no limitation because the considered estimators have the usual invariance properties.

Three different group sequential plans are considered. The corresponding critical constants are determined from simulations based on $10,000$ replications. The parameters are listed in Table 1. We assume $n_1 = \ldots = n_K$. The values of $N_K = K \times n_1$ are chosen using the software package EaSt (1993) such that the group sequential t-test has a power of approximately $1 - \beta = 0.95$ at $(F,G) = (\mathrm{N}(0,1), \mathrm{N}(\Delta,1))$ and $\Delta = 0.5$.

Table 1

Parameters for group sequential plans

Plan	α	K	n_1	a	c for estimator pair		
					$(\bar{X}, S)$	$(25A, SH)$	$(25A, Q)$
Pocock	0.05	3	40	0.5	2.2934	2.6473	2.3355
Pocock	0.01	5	34	0.5	3.0442	3.5630	3.1241
Wang & Tsiatis	0.01	5	34	0.4695	3.1308	3.6210	3.1902

Further, an O'Brien-Fleming design with Lan and DeMets boundaries with $K = 5$ interim tests at the processing times 0.4, 0.7, 0.8, 0.9, 1.0 is used. For level $\alpha = 0.05$ and power $1 - \beta = 0.90$ the maximal total sample size of $36 + 27 + 9 + 9 + 9 = 90$ is used for each treatment group. This simulation is done to investigate, whether the same critical constants can be used for the group sequential test based on $(25A, Q)$ as the commerical software package EaSt from Cytel uses for the classical test assuming normality provided that the sample sizes are not too small. I.e. the critical values for both tests at the 5 stages are 3.28492, 2.43011, 2.32008, 2.19412, and 2.08368, respectively. In this simulation the test statistic for the group sequential test based on $(25A, Q)$ is divided by appropriate constants depending on the sample sizes N_k such that the standardized test statistic has an approximately standard normal distribution under H_0, $k = 1, ..., 5$. These constants are 1.0149, 1.0038, 1.0030, 1.0221, 1.0135, respectively. All simulations are based on 10000 replications.

5. Results

The results for all considered designs are very similar. Therefore, only the results for Pocock's plan with $K = 3$ interim tests (Tables 2, 3 and 4) and for O'Brien-Fleming's design with Lan and DeMets boundaries with $K = 5$ interim tests (Table 5) are shown in detail. The results for $\Delta = 0.75$ are not shown because they are intermediate to those for $\Delta = 0.5$ and $\Delta = 1.0$.

First, Pocock's plan is considered. If the distributions in both treatment groups are normal, the group sequential t-test has higher power and lower average sample size number than the other two tests under consideration, c.f. Tables 2 and 3. For most other situations considered in our simulations the group sequential tests based on the robust estimators (25A,Q) and (25A,SH) show a better behaviour than the t-test.

Please insert Table 2

If the distribution is not normal in at least one treatment group but t_3 or MIXN, the application of the t-test can be dangerous. It can happen that the probability of an error of type I is approximately equal to α, but that the power is drastically reduced and the average sample size number is markedly higher than for normally distributed data. In the simulations this happens for the pairs (F, G) equal to (N(0,1), t_3), (t_3, t_3), and (MIXN, MIXN).
An application of the t-test can be very dangerous, too, for the other three pairs of (F, G), i.e. (N(0,1), MIXN), (t_3, MIXN), and (MIXN, N(0,1)), but for other reasons. For these pairs of distributions the t-test can have a probability for an error of type I which is drastically higher than α, in our simulations even higher than 10α. In these cases, the average sample size number of the t-test can be lower or higher than for normally distributed data.
For normally distributed data the power of the test based on (25A,Q) is only a few percents lower than for the t-test, and the average sample size number is only slightly higher than for the t-test. If in at least one treatment group the distribution is not normal, but t_3 or MIXN, the application of a group sequential test based on (25A,Q) is much safer with respect to level, power and average sample size than the use of the t-test. For the robust test the estimated power values increase and the average sample size numbers decrease with increasing treatment differences $|\Delta|$. Both points are not always true for the t-test, c.f. the pair of distributions (MIXN, N(0,1)).

Please insert Table 3

The group sequential test based on (25A,SH) shows a similar behaviour than the one based on (25A,Q) in our simulations. In general, the test based on

(25A,SH) has a somewhat lower power than the one based on (25A,Q) for normally distributed data. There are pairs of distributions (F, G), where the test based on (25A,Q) has higher power than the one based on (25A,SH) and vice versa. A similar result holds for the average sample size numbers. This behaviour is plausible because the scale estimator SH has a lower efficiency for normally distributed data than Q, and SH is more robust than Q, c.f. Rousseeuw & Croux (1993).
The estimated values of Median($\hat{\Delta} - \Delta$) are given in Table 4. Non-parametric 95% confidence intervals based on the 4902−th and 5099−th order statistics for Median($\hat{\Delta} - \Delta$) are computed, c.f. Serfling (1980, p. 102f). The widest confidence interval, i.e. the greatest difference between these order statistics, has the length 0.021.

Please insert Table 4

For normally distributed data the estimated values of Median($\hat{\Delta}$) are approximately equal to zero under H_0 but tend to be greater than Δ for positive values of Δ. For such data, $\hat{\Delta}$ based on $(\bar{X}, S)$ and (25A,Q) have comparable biases, but - as could be expected - the bias is not negligible for $\Delta \neq 0$, especially for $\Delta = 0.5$. Note, that $\Delta = 0.5$ is the value, for which the power of the tests should be approximately 0.95. For normally distributed data, the application of (25A,SH) yields greater biases than the other pairs to estimators if $\Delta > 0$.
For all six considered situations with non-normal data, the classical pair $(\bar{X}, S)$ yields values of Median($\hat{\Delta}$) which can drastically differ from Δ in both directions. For such situations, the application of (25A,Q) or (25A,SH) allows a much more stable estimation of Δ. However, none pair of estimators which is considered dominates the others for all situations.
Overall, the pair (25A,Q) yields the best results in the simulations from two aspects. For normally distributed data the results based on (25A,Q) do not differ too much from those produced by $(\bar{X}, S)$, whereas (25A,Q) yields more robust results with respect to actual level and power of the

test, average sample size number, and bias of the estimated standardized treatment difference. In our simulations, the test based on (25A,SH) does not give much more robust results than the one based on (25A,Q) for non-normal data, but for normally distributed data the loss of power using (25A,SH) is greater than for (25A,Q).

Therefore, for O'Brien-Fleming's design with Lan and DeMets boundaries and $K = 5$ interim tests only the group sequential t-test and the test based on $(25A, Q)$ are considered. The results are very similar to those given before. The strategy is to use the same critical values as are used in EaSt (1993) and to divide the test statistic based on $(25A, Q)$ by an appropriate constant, such that the *standardized* test statistic has approximately a standard normal distribution under H_0. This follows from (2.2) and Slutzky's theorem, because the considered scale estimators are consistent and the M-estimator 25A is asymptotically normal. This strategy avoids to determine the exact critical values via simulations. Table 5 shows that this is successful for moderate sample sizes, because level, power, and averaged sample size number are very similar to those for the group sequential t-test under normality, but yield similar robustness properties under the considered alternatives. This is also true for the naive estimator for Δ.

Please insert Table 5

Table 2
Estimated values of level and power (in percent) for two-sided group sequential Pocock tests;
Plan 1: $K = 3$, $\alpha = 5\%$, $1 - \beta \approx 95\%$ for $\Delta = 0.5$.

Hypothesis	Distributions		Group sequential test		
	F	G	$(\bar{X}, S)$	$(25A, SH)$	$(25A, Q)$
$\Delta = 0$	$N(0,1)$	$N(0,1)$	5.0	5.0	5.0
		$t3$	4.9	6.2	5.5
		$MIXN$	59.2	4.6	3.6
	$t3$	$t3$	5.2	7.5	6.0
		$MIXN$	53.8	5.8	4.3
	$MIXN$	$MIXN$	4.1	3.9	2.8
		$N(0,1)$	59.3	4.5	3.8
$\Delta = 0.25$	$N(0,1)$	$N(0,1)$	42.3	37.9	40.7
		$t3$	26.5	35.1	34.3
		$MIXN$	88.7	35.7	36.5
	$t3$	$t3$	20.0	32.6	29.4
		$MIXN$	82.6	33.5	31.3
	$MIXN$	$MIXN$	6.4	28.3	24.9
		$N(0,1)$	26.5	29.7	27.0
$\Delta = 0.5$	$N(0,1)$	$N(0,1)$	95.2	93.1	94.5
		$t3$	76.1	87.9	87.9
		$MIXN$	98.7	91.3	92.0
	$t3$	$t3$	60.1	82.9	80.4
		$MIXN$	96.3	86.0	84.8
	$MIXN$	$MIXN$	13.7	85.7	83.0
		$N(0,1)$	10.8	88.7	87.2
$\Delta = 1$	$N(0,1)$	$N(0,1)$	100.0	100.0	100.0
		$t3$	99.7	100.0	100.0
		$MIXN$	100.0	100.0	100.0
	$t3$	$t3$	98.2	100.0	100.0
		$MIXN$	99.9	100.0	100.0
	$MIXN$	$MIXN$	39.6	100.0	100.0
		$N(0,1)$	14.8	100.0	100.0

Table 3

Estimated average sample size number (ASN) for two-sided group sequential Pocock tests; Plan 1

Hypothesis	Distributions		Group sequential test		
	F	G	$(\bar{X}, S)$	$(25A, SH)$	$(25A, Q)$
$\Delta = 0$	$N(0,1)$	$N(0,1)$	117.5	117.3	117.6
		$t3$	117.6	116.6	117.3
		$MIXN$	104.1	117.5	118.3
	$t3$	$t3$	117.5	116.1	117.1
		$MIXN$	105.1	116.8	117.8
	$MIXN$	$MIXN$	118.2	117.9	118.7
		$N(0,1)$	103.9	117.4	118.0
$\Delta = 0.25$	$N(0,1)$	$N(0,1)$	103.9	104.6	104.7
		$t3$	109.7	105.3	107.1
		$MIXN$	86.2	106.0	107.1
	$t3$	$t3$	112.3	106.5	108.9
		$MIXN$	89.1	106.3	108.5
	$MIXN$	$MIXN$	117.1	108.7	111.3
		$N(0,1)$	114.0	108.0	110.4
$\Delta = 0.5$	$N(0,1)$	$N(0,1)$	67.8	69.7	69.6
		$t3$	85.2	74.6	77.2
		$MIXN$	66.6	72.5	74.1
	$t3$	$t3$	94.6	78.8	83.2
		$MIXN$	71.5	77.1	80.7
	$MIXN$	$MIXN$	114.2	78.5	83.6
		$N(0,1)$	116.1	75.7	79.7
$\Delta = 1$	$N(0,1)$	$N(0,1)$	40.6	40.8	40.8
		$t3$	47.2	42.1	42.7
		$MIXN$	44.9	41.3	41.6
	$t3$	$t3$	55.3	43.7	45.1
		$MIXN$	48.5	42.9	43.9
	$MIXN$	$MIXN$	102.8	42.3	43.8
		$N(0,1)$	110.4	41.8	42.7

Table 4

Estimated values for Median($\hat{\Delta} - \Delta$) for two-sided group sequential Pocock tests; Plan 1

Hypothesis	Distributions		Group sequential test		
	F	G	$(\bar{X}, S)$	$(25A, SH)$	$(25A, Q)$
$\Delta = 0$	$N(0,1)$	$N(0,1)$	-0.001	0.001	0.000
		$t3$	0.001	0.003	0.003
		$MIXN$	0.320	0.012	0.017
	$t3$	$t3$	0.002	0.002	0.002
		$MIXN$	0.307	0.013	0.016
	$MIXN$	$MIXN$	0.000	-0.003	-0.002
		$N(0,1)$	-0.320	-0.014	-0.018
$\Delta = 0.25$	$N(0,1)$	$N(0,1)$	0.011	0.030	0.010
		$t3$	-0.057	0.015	-0.018
		$MIXN$	0.159	0.024	-0.003
	$t3$	$t3$	-0.091	-0.004	-0.042
		$MIXN$	0.147	0.009	-0.026
	$MIXN$	$MIXN$	-0.192	-0.009	-0.045
		$N(0,1)$	-0.482	-0.006	-0.041
$\Delta = 0.5$	$N(0,1)$	$N(0,1)$	0.047	0.123	0.049
		$t3$	-0.079	0.096	-0.004
		$MIXN$	0.019	0.101	0.017
	$t3$	$t3$	-0.145	0.046	-0.054
		$MIXN$	-0.008	0.060	-0.039
	$MIXN$	$MIXN$	-0.382	0.037	-0.061
		$N(0,1)$	-0.650	0.069	-0.029
$\Delta = 1$	$N(0,1)$	$N(0,1)$	0.001	0.140	-0.007
		$t3$	-0.223	0.066	-0.112
		$MIXN$	-0.327	0.075	-0.093
	$t3$	$t3$	-0.342	-0.006	-0.195
		$N(0,1)$	-0.358	0.006	-0.174
	$MIXN$	$MIXN$	-0.752	0.000	-0.191
		$N(0,1)$	-0.984	0.053	-0.127

Table 5

Estimated values of level, power (in percent) and ASN for two-sided group sequential O'Brien-Fleming tests with Lan and DeMets boundaries; $K = 5$, $\alpha = 5\%$, $1 - \beta \approx 90\%$ for $\Delta = 0.5$.

Hypothesis	Distributions		P_Δ(decision for H_1)		ASN	
	F	G	$(\bar{X}, S)$	$(25A, Q)$	$(\bar{X}, S)$	$(25A, Q)$
$\Delta = 0$	$N(0,1)$	$N(0,1)$	5.0	5.0	89.2	89.2
		$t3$	5.3	5.9	89.2	89.0
		$MIXN$	51.4	4.4	82.9	89.3
	$t3$	$t3$	5.5	6.1	89.2	89.0
		$MIXN$	47.5	4.6	83.5	89.3
	$MIXN$	$MIXN$	4.9	3.0	89.4	89.5
		$N(0,1)$	51.8	4.4	83.0	89.4
$\Delta = 0.5$	$N(0,1)$	$N(0,1)$	90.9	89.8	66.1	66.6
		$t3$	69.6	81.6	74.7	70.4
		$MIXN$	95.7	86.4	67.3	69.0
	$t3$	$t3$	53.5	73.5	79.4	73.5
		$MIXN$	92.4	78.6	69.7	72.4
	$MIXN$	$MIXN$	13.4	77.1	87.9	73.8
		$N(0,1)$	9.7	80.2	88.7	71.8

6. Discussion

The investigated group sequential tests based on modern robust location and scale estimators give much more stable results than the group sequential t-test under the distributions considered here. On the other hand, under normality one does not loose much information if one uses tests based on modern robust estimators instead of the t-test.

The group sequential test based on Hampel's 3-part M-estimator $25A$ and the scale estimator Q proposed by Rousseeuw and Croux (1993) is an attractive alternative to the group sequential t-test, at least if the subsample sizes n_k for each group are not too small at the beginning of the test procedure, i.e. for $k = 1$. This test behaves very similar to the group sequential t-test under normality, but the behaviour is more stable for all 4 criteria - level, power, averaged sample size number and naive estimated standardized treatment difference - under the model deviations considered here.

It can be argued that the mixture model considered here may be too pessimistic for 'real life data' although the percentage of outliers is only 10%. However, the dramatic impact of such outliers on the group sequential t-test shows that the behaviour of this test can be very unstable. As mentioned earlier, Student's distribution with 3 degrees of freedom is often a good approximation to the distribution of high quality data, c.f. Hampel et al. (1986, p. 23). But even under these circumstances the group sequential t-test looses much more power and the average sample size number is substantially higher than for the alternative tests in the situations considered here.

Rousseeuw & Leroy (1988) propose a small sample modification of SH, say SH^*, to reduce the bias of SH in a fixed sample size problem. Although the sample sizes we considered in the simulations are not very small, all simulation are repeated for the considered group sequential Pocock test with $k = 3$ groups. The results for (25A,SH*) are very similar to those for (25A,SH). E.g. consider the pair of distributions (t_3, t_3). The simulated values of Median$(\hat{\Delta} - \Delta)$ for $\Delta = 0, 0.25, 0.5$ for the GST based on (25A,SH*) are 0.002, -0.005, and 0.043, respectively. The corresponding values for the

GST based on (25A,SH) are 0.002, -0.004, and 0.046, respectively.
An investigation of more complex estimators to reduce the bias and confidence intervals for the standardized treatment difference Δ from a robustness point of view is beyond the scope of this paper.

Acknowledgements

The author thanks P. J. Rousseeuw for making available the program to compute Q efficiently, and U. Gather and C. Croux for helpful discussions.

References

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., & Tukey, J.W. (1972). *Robust Estimates of Location. Survey and Advances.* Princeton: Princeton University Press.

Christmann, A., Gather, U. & Scholz, G. (1994). Some properties of the length of the shortest half. *Statistica Neerlandandica* **48**, 209-213.

Cox, D.R. (1952). A note on the sequential estimation of means. *Proc. Camb. Phil. Soc.* **48**, 447-450.

Croux, C. & Rousseeuw, P.J. (1992a). A Class of High-Breakdown Scale Estimators Based on Subranges, *Communications in Statistics Theory and Methods* **21**, 1935-1951.

Croux, C. & Rousseeuw, P.J. (1992b). Time efficient algorithms for two highly robust estimators of scale. *Computational Statistics* **1**, 411-428.

Davies, P.L. (1990). The asymptotics of S-estimators in the linear regression model. *Annals of Statistics* **18**, 1651-1675.

Davies, P.L. (1993). Aspects of robust linear regression. *Annals of Statistics* **21**, 1843-1899.

Davies, P.L. & Gather, U. (1993). The identification of multiple outliers. With discussion. *Journal of the American Statistical Association* **88**, 782-801.

DeMets, D.L. and Lan, K.K.G. (1994). Interim analysis: the alpha spending function approach. *Statistics in Medicine* **13**, 1341-1352.

Lan, K.K.G. & DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659-663.

EaSt (1993). *A Software Package of the Design and Interim Monitoring of Group Sequential Clinical Trials.* Cambridge: Cytel Software Corporation.

Grübel, R. (1988). The length of the shorth. *Annals of Statistics* **16**, 619-628.

Hampel, F.R. (1985). The breakdown points of the mean combined with some rejection rules. *Technometrics* **27**, 95-107.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. & Stahel, W.A. (1986). *Robust statistics: The Approach Based on Influence Functions*, New York: Wiley.

Huber, P.J. (1981). *Robust Statistics.* New York: Wiley.

Kim, K. (1988). Improved approximation for estimation following closed sequential tests. *Biometrika* **75**, 121-128.

Kim, K. (1989). Point estimation following group sequential tests. *Biometrics* **45**, 613-617.

DeMets, D.L. & Lan, K.K.G. (1994). Interim analysis: The alpha spending function approach. *Statistics in Medicine* **13**, 1341-1352.

Martin, R.D. & Zamar, R.H. (1993). Bias Robust Estimation of Scale. *Annals of Statistics* **22**, 991-1017.

Mehta, C.R., Patel, N., Senchaudhuri, P. & Tsiatis, A. (1994). Exact permutational tests for group sequential clinical trials. *Biometrics* **50**, 1042-1053.

Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191-199.

Pocock, S.J. (1983). *Clinical Trials. A practical approach.* New York: Wiley.

O'Brien, P.C. & Fleming, T.R. (1979). A Multiple Testing Procedure for Clinical Trials. *Biometrics* **35**, 549-556.

Rousseeuw, P.J. & Croux, C. (1993). Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association* **88**, 1273-1283.

Rousseeuw, P.J. & Leroy, A.M. (1988). A robust scale estimator based on the shortest half. *Statistica Neerlandica* **42**, 103-116.

Rousseeuw, P.J. and van Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association* **85**, 633-651.

SAS Institute Inc. (1995), SAS/IML Software: Changes and Enhancements through Release 6.11, Cary, NC: SAS Institute Inc..

Silvapulle, M.J. & Sen, P.K. (1993). Robust tests in group sequential analysis: one- and two-sided hypotheses in the linear model. *Annals of the Institute of Statistical Mathematics* **45**, 159-171.

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics.* New York: Wiley.

Wang, S.K. & Tsiatis, A.A. (1987). Approximately Optimal One-Parameter Boundaries for Group Sequential Trials. *Biometrics* **43**, 193-199.

Whitehead, J. (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika* **73**, 573-581.

APPENDIX: ADDITIONAL TABLES

Table 6a

Estimated values of level and power (in percent) for two-sided group sequential Pocock tests; Plan 2: $K = 5$, $\alpha = 1\%$, $1-\beta \approx 95\%$ for $\Delta = 0.5$.

Hypothesis	Distributions		Group sequential test		
	F	G	$(\bar{X}, S)$	$(25A, SH)$	$(25A, Q)$
$\Delta = 0$	$N(0,1)$	$N(0,1)$	1.0	1.0	1.0
		$t3$	0.9	1.1	0.8
		$MIXN$	40.4	0.9	0.6
	$t3$	$t3$	0.8	1.6	1.2
		$MIXN$	35.2	1.2	0.8
	$MIXN$	$MIXN$	0.4	0.6	0.3
		$N(0,1)$	40.8	0.8	0.7
$\Delta = 0.25$	$N(0,1)$	$N(0,1)$	29.2	22.1	27.0
		$t3$	14.2	20.0	20.5
		$MIXN$	81.2	18.9	20.9
	$t3$	$t3$	8.5	17.7	15.8
		$MIXN$	73.6	17.5	16.6
	$MIXN$	$MIXN$	1.1	13.2	11.7
		$N(0,1)$	10.3	14.2	13.8
$\Delta = 0.5$	$N(0,1)$	$N(0,1)$	95.0	90.4	93.8
		$t3$	69.5	84.0	85.0
		$MIXN$	98.3	88.1	90.4
	$t3$	$t3$	47.1	76.8	75.2
		$MIXN$	95.1	80.9	80.5
	$MIXN$	$MIXN$	4.2	79.6	78.3
		$N(0,1)$	2.8	82.8	83.2
$\Delta = 0.75$	$N(0,1)$	$N(0,1)$	100.0	99.9	100.0
		$t3$	97.0	99.8	99.8
		$MIXN$	99.9	99.9	99.9
	$t3$	$t3$	87.6	99.1	99.0
		$MIXN$	99.6	99.6	99.6
	$MIXN$	$MIXN$	12.1	99.8	99.8
		$N(0,1)$	5.4	99.9	99.9
$\Delta = 1$	$N(0,1)$	$N(0,1)$	100.0	100.0	100.0
		$t3$	99.7	100.0	100.0
		$MIXN$	100.0	100.0	100.0
	$t3$	$t3$	98.4	100.0	100.0
		$MIXN$	99.9	100.0	100.0
	$MIXN$	$MIXN$	25.8	100.0	100.0
		$N(0,1)$	11.0	100.0	100.0

Table 6b

Estimated average sample size number (ASN) for two-sided group sequential Pocock tests; Plan 2

Hypothesis	Distributions		Group sequential test		
	F	G	$(\bar{X}, S)$	$(25A, SH)$	$(25A, Q)$
$\Delta = 0$	$N(0,1)$	$N(0,1)$	169.2	169.0	169.2
		$t3$	169.2	168.9	169.3
		$MIXN$	157.7	169.1	169.5
	$t3$	$t3$	169.3	168.3	168.9
		$MIXN$	159.0	168.8	169.3
	$MIXN$	$MIXN$	169.8	169.3	169.7
		$N(0,1)$	157.8	169.1	169.5
$\Delta = 0.25$	$N(0,1)$	$N(0,1)$	154.6	156.9	155.9
		$t3$	162.4	157.9	159.2
		$MIXN$	132.6	159.2	159.9
	$t3$	$t3$	165.5	159.1	161.7
		$MIXN$	137.8	159.6	161.7
	$MIXN$	$MIXN$	169.4	162.8	164.7
		$N(0,1)$	167.3	161.6	163.0
$\Delta = 0.5$	$N(0,1)$	$N(0,1)$	93.4	100.6	97.1
		$t3$	125.3	109.1	111.3
		$MIXN$	100.2	106.1	106.6
	$t3$	$t3$	143.1	117.1	122.8
		$MIXN$	108.4	114.2	119.1
	$MIXN$	$MIXN$	167.6	117.2	123.8
		$N(0,1)$	167.9	112.7	116.8
$\Delta = 0.75$	$N(0,1)$	$N(0,1)$	53.6	56.1	55.8
		$t3$	78.9	63.0	65.8
		$MIXN$	75.3	59.9	61.9
	$t3$	$t3$	102.2	69.7	75.8
		$MIXN$	83.0	66.9	71.7
	$MIXN$	$MIXN$	162.8	67.0	73.5
		$N(0,1)$	163.5	63.3	67.3
$\Delta = 1$	$N(0,1)$	$N(0,1)$	38.9	39.9	39.9
		$t3$	53.9	44.0	46.0
		$MIXN$	59.7	42.1	43.5
	$t3$	$t3$	70.5	47.8	51.6
		$MIXN$	65.8	46.0	49.3
	$MIXN$	$MIXN$	154.6	45.7	50.1
		$N(0,1)$	157.2	43.5	46.3

Table 6c

Estimated values for Median($\hat{\Delta} - \Delta$) for two-sided group sequential Pocock tests; Plan 2

Hypothesis	Distributions		Group sequential test		
	F	G	$(\bar{X}, S)$	$(25A, SH)$	$(25A, Q)$
$\Delta = 0$	$N(0,1)$	$N(0,1)$	0.002	0.003	0.003
		$t3$	0.001	0.002	0.002
		$MIXN$	0.311	0.013	0.016
	$t3$	$t3$	0.000	0.000	0.000
		$MIXN$	0.299	0.011	0.014
	$MIXN$	$MIXN$	0.001	0.001	0.002
		$N(0,1)$	-0.311	-0.012	-0.016
$\Delta = 0.25$	$N(0,1)$	$N(0,1)$	0.005	0.019	0.004
		$t3$	-0.063	0.002	-0.025
		$MIXN$	0.153	0.013	-0.009
	$t3$	$t3$	-0.098	-0.015	-0.047
		$MIXN$	0.137	-0.004	-0.033
	$MIXN$	$MIXN$	-0.192	-0.013	-0.045
		$N(0,1)$	-0.480	-0.012	-0.041
$\Delta = 0.5$	$N(0,1)$	$N(0,1)$	0.042	0.095	0.044
		$t3$	-0.086	0.052	-0.019
		$MIXN$	-0.010	0.062	-0.004
	$t3$	$t3$	-0.182	0.018	-0.059
		$MIXN$	-0.034	0.028	-0.047
	$MIXN$	$MIXN$	-0.385	0.017	-0.059
		$N(0,1)$	-0.651	0.033	-0.038
$\Delta = 0.75$	$N(0,1)$	$N(0,1)$	0.048	0.163	0.049
		$t3$	-0.139	0.118	-0.036
		$MIXN$	-0.171	0.129	-0.005
	$t3$	$t3$	-0.235	0.036	-0.114
		$MIXN$	-0.196	0.055	-0.093
	$MIXN$	$MIXN$	-0.577	0.044	-0.111
		$N(0,1)$	-0.820	0.098	-0.058
$\Delta = 1$	$N(0,1)$	$N(0,1)$	0.010	0.163	0.004
		$t3$	-0.184	0.084	-0.091
		$MIXN$	-0.321	0.105	-0.073
	$t3$	$t3$	-0.321	0.026	-0.157
		$MIXN$	-0.363	0.037	-0.145
	$MIXN$	$MIXN$	-0.767	0.036	-0.155
		$N(0,1)$	-0.987	0.082	-0.104

Table 7a

Estimated values of level and power (in percent) for two-sided group sequential Wang & Tsiatis tests; Plan 3: $K = 5$, $\alpha = 1\%$, $1 - \beta \approx 95\%$ for $\Delta = 0.5$.

Hypothesis	Distributions		Group sequential test		
	F	G	$(\bar{X}, S)$	$(25A, SH)$	$(25A, Q)$
$\Delta = 0$	$N(0,1)$	$N(0,1)$	1.0	1.0	1.0
		$t3$	0.9	1.1	0.9
		$MIXN$	44.1	1.0	0.7
	$t3$	$t3$	0.8	1.7	1.2
		$MIXN$	38.7	1.3	0.8
	$MIXN$	$MIXN$	0.5	0.7	0.3
		$N(0,1)$	44.2	0.9	0.7
$\Delta = 0.25$	$N(0,1)$	$N(0,1)$	30.9	22.1	29.2
		$t3$	15.1	21.8	22.3
		$MIXN$	83.6	21.0	23.0
	$t3$	$t3$	9.2	19.6	17.3
		$MIXN$	76.2	19.8	18.4
	$MIXN$	$MIXN$	1.3	15.2	13.3
		$N(0,1)$	12.1	15.8	15.2
$\Delta = 0.5$	$N(0,1)$	$N(0,1)$	95.5	90.4	94.8
		$t3$	71.0	86.0	86.7
		$MIXN$	98.6	89.9	91.7
	$t3$	$t3$	48.9	79.3	77.4
		$MIXN$	96.0	83.2	82.6
	$MIXN$	$MIXN$	4.5	82.3	80.7
		$N(0,1)$	3.0	85.1	85.1
$\Delta = 0.75$	$N(0,1)$	$N(0,1)$	100.0	99.9	100.0
		$t3$	97.2	99.9	99.9
		$MIXN$	100.0	99.9	100.0
	$t3$	$t3$	88.4	99.3	99.1
		$MIXN$	99.7	99.7	99.7
	$MIXN$	$MIXN$	12.7	99.8	99.8
		$N(0,1)$	5.2	99.9	99.9
$\Delta = 1$	$N(0,1)$	$N(0,1)$	100.0	100.0	100.0
		$t3$	99.7	100.0	100.0
		$MIXN$	100.0	100.0	100.0
	$t3$	$t3$	98.5	100.0	100.0
		$MIXN$	99.9	100.0	100.0
	$MIXN$	$MIXN$	26.8	100.0	100.0
		$N(0,1)$	10.8	100.0	100.0

Table 7b

Estimated average sample size number (ASN) for two-sided group sequential Wang & Tsiatis tests; Plan 3

Hypothesis	Distributions		Group sequential test		
	F	G	$(\bar{X}, S)$	$(25A, SH)$	$(25A, Q)$
$\Delta = 0$	$N(0,1)$	$N(0,1)$	169.3	169.1	169.3
		$t3$	169.3	168.9	169.3
		$MIXN$	157.0	169.1	169.5
	$t3$	$t3$	169.4	168.4	169.0
		$MIXN$	158.4	168.8	169.4
	$MIXN$	$MIXN$	169.8	169.3	169.7
		$N(0,1)$	157.0	169.2	169.6
$\Delta = 0.25$	$N(0,1)$	$N(0,1)$	154.7	156.4	155.6
		$t3$	162.6	157.4	158.8
		$MIXN$	131.9	158.8	159.6
	$t3$	$t3$	165.6	158.7	161.5
		$MIXN$	137.2	159.1	161.4
	$MIXN$	$MIXN$	169.4	162.5	164.5
		$N(0,1)$	167.1	161.3	162.8
$\Delta = 0.5$	$N(0,1)$	$N(0,1)$	94.2	99.6	97.0
		$t3$	125.7	108.1	110.8
		$MIXN$	100.6	105.1	106.1
	$t3$	$t3$	143.3	116.2	122.2
		$MIXN$	108.5	113.2	118.3
	$MIXN$	$MIXN$	167.7	116.0	123.0
		$N(0,1)$	167.9	111.5	116.1
$\Delta = 0.75$	$N(0,1)$	$N(0,1)$	54.9	56.7	56.7
		$t3$	80.0	63.2	66.4
		$MIXN$	76.6	60.2	62.7
	$t3$	$t3$	103.0	69.8	76.3
		$MIXN$	84.2	67.1	72.2
	$MIXN$	$MIXN$	163.0	67.2	74.0
		$N(0,1)$	163.8	63.5	67.9
$\Delta = 1$	$N(0,1)$	$N(0,1)$	39.7	40.3	40.5
		$t3$	55.1	44.5	46.7
		$MIXN$	61.4	42.5	44.2
	$t3$	$t3$	71.8	48.2	52.3
		$MIXN$	67.4	46.6	50.2
	$MIXN$	$MIXN$	154.9	46.3	50.9
		$N(0,1)$	157.5	44.0	47.1

Table 7c

Estimated values for Median$(\hat{\Delta} - \Delta)$ *for two-sided group sequential Wang & Tsiatis tests; Plan 3*

Hypothesis	Distributions		Group sequential test		
	F	G	$(\bar{X}, S)$	$(25A, SH)$	$(25A, Q)$
$\Delta = 0$	$N(0,1)$	$N(0,1)$	0.002	0.003	0.003
		$t3$	0.001	0.002	0.002
		$MIXN$	0.311	0.013	0.016
	$t3$	$t3$	0.000	0.000	0.000
		$MIXN$	0.299	0.011	0.014
	$MIXN$	$MIXN$	0.001	0.001	0.002
		$N(0,1)$	-0.311	-0.012	-0.016
$\Delta = 0.25$	$N(0,1)$	$N(0,1)$	0.005	0.019	0.004
		$t3$	-0.063	0.002	-0.025
		$MIXN$	0.152	0.013	-0.008
	$t3$	$t3$	-0.098	-0.015	-0.047
		$MIXN$	0.136	-0.004	-0.033
	$MIXN$	$MIXN$	-0.192	-0.013	-0.045
		$N(0,1)$	-0.480	-0.012	-0.041
$\Delta = 0.5$	$N(0,1)$	$N(0,1)$	0.041	0.095	0.043
		$t3$	-0.087	0.051	-0.021
		$MIXN$	-0.012	0.061	-0.006
	$t3$	$t3$	-0.181	0.014	-0.061
		$MIXN$	-0.035	0.026	-0.048
	$MIXN$	$MIXN$	-0.385	0.014	-0.062
		$N(0,1)$	-0.651	0.031	-0.040
$\Delta = 0.75$	$N(0,1)$	$N(0,1)$	0.053	0.166	0.054
		$t3$	-0.142	0.110	-0.042
		$MIXN$	-0.172	0.132	-0.016
	$t3$	$t3$	-0.240	0.029	-0.118
		$MIXN$	-0.197	0.048	-0.096
	$MIXN$	$MIXN$	-0.578	0.038	-0.115
		$N(0,1)$	-0.820	0.087	-0.065
$\Delta = 1$	$N(0,1)$	$N(0,1)$	0.011	0.164	0.005
		$t3$	-0.178	0.086	-0.087
		$MIXN$	-0.329	0.106	-0.071
	$t3$	$t3$	-0.330	0.028	-0.154
		$MIXN$	-0.367	0.040	-0.142
	$MIXN$	$MIXN$	-0.767	0.038	-0.152
		$N(0,1)$	-0.987	0.084	-0.101