

Robust One-Way ANOVA under Possibly Non-Regular Conditions

GUTTI JOGESH BABU*

Department of Statistics
Pennsylvania State University
USA

A. R. PADMANABHAN

Department of Mathematics
Monash University
Clayton
Australia
and
Department of Statistics
Pennsylvania State University
USA

MADAN L. PURI

Department of Mathematics
Indiana University
Bloomington
USA

Summary

Consider the one-way ANOVA problem of comparing the means m_1, m_2, \dots, m_c of c distributions $F_1(x) = F(x - m_1), \dots, F_c(x) = F(x - m_c)$. Solutions are available based on (i) normal-theory procedures, (ii) linear rank statistics and (iii) M -estimators.

The above model presupposes that F_1, F_2, \dots, F_c have equal variances (= homoscedasticity). However practising statisticians content that homoscedasticity is often violated in practice. Hence a more realistic problem to consider is $F_1(x) = F((x - m_1)/\sigma_1), \dots, F_c(x) = F((x - m_c)/\sigma_c)$, where F is symmetric about the origin and $\sigma_1, \dots, \sigma_c$ are unknown and possibly unequal (= heteroscedasticity). Now we have to compare m_1, m_2, \dots, m_c . At present, nonparametric tests of the equality of m_1, m_2, \dots, m_c are available. However, simultaneous tests for paired comparisons and contrasts and do not seem to be available.

This paper begins by proposing a solution applicable to both the homoscedastic and the heteroscedastic situations, assuming F to be symmetric. Then the assumptions of symmetry and the identical

* Research supported in part by NSF grants DMS-9007717 and DMS-9208066.

shapes of F_1, \dots, F_c are progressively relaxed and solutions are proposed for these cases as well. The procedures are all based on either the 15% trimmed means or the sample medians, whose quantiles are estimated by means of the bootstrap. Monte Carlo studies show that these procedures tend to be superior to the Wilcoxon procedure and Dunnett's normal theory procedure. A rigorous justification of the bootstrap is also presented. The methodology is illustrated by a comparison of mean effects of cocaine administration in pregnant female Sprague-Dawley rats, where skewness and heteroscedascity are known to be present.

Key words: Bias-corrected bootstrap; 15% trimmed means; Sample medians; Smooth bootstrap; Sprague-Dawley rats.

1. Introduction

Consider the one-way ANOVA problem of comparing the means m_1, m_2, \dots, m_c of c distributions $F_1(x) = F(x - m_1), \dots, F_c(x) = F(x - m_c)$. Solutions are available based on (i) normal-theory procedures (SCHEFFÉ, 1970), (ii) linear rank statistics (PURI and SEN, 1971) and (iii) M -estimators (RINGLAND, 1983). Note that in the foregoing model, F_1, \dots, F_c are assumed to be identical, except possibly for locations. This, in particular, implies that they have equal variances (= homoscedasticity). However, the homoscedasticity is often violated in practice (cf. BISHOP and DUDEWICZ, 1978, p. 419). This is because (a) heterogeneity of error may result from the erratic behaviour of the response to certain treatments (cf. STEEL and TORRIE, 1980, pp. 169–170) and (b) in using laboratory measurements to compare patients having a certain disease, with healthy subjects free of the disease, the disease may have a more pronounced effect on the laboratory values of some patients than some others, thereby tending to increase the dispersion (O'BRIEN, 1992, pp. 819–827).

Hence a more realistic model is $F_1(x) = F((x - m_1)/\sigma_1), \dots, F_c(x) = F((x - m_c)/\sigma_c)$, where F is an unknown distribution function, symmetric about the origin and $\sigma_1, \dots, \sigma_c$ are unknown and possibly unequal (= heteroscedasticity). Now we have to compare m_1, \dots, m_c .

For this problem, at present, there are three approaches available. The first is to assume that F is normal and develop a methodology accordingly (BISHOP and DUDEWICZ, 1978, pp. 419–424; DUNNETT, 1980, pp. 796–800). The second is to use a variance-stabilizing transformation such as the square-root or cube-root or logarithmic or inverse sine or inverse hyperbolic sine transformation (cf. DUDEWICZ, 1983, p. 613), all of which are highly non-linear. The third approach is to use location statistics, which are scale-invariant (COMPAGNONE and DENKER, 1966). The drawback of the first approach is its inapplicability to non-normal situation, which will clearly be the case, if the normal scores plot of the sample data is not linear (JOHNSON and BHATTACHARYA, 1986, pp. 218–220). As for the second ap-

proach, often it is not obvious which transformation is most appropriate. As regards the third approach there are some difficulties in applying it to our problem, as explained at the end of Section 4.

Section 2 contains some preliminaries. Section 3 proposes solutions to the above problem, which are applicable to both the homoscedastic and the heteroscedastic settings (thereby eliminating the need for a preliminary test of equality of variances). First the unknown distribution function F , is assumed to be symmetric about the origin. Then this problem is tackled under the weaker assumption that F is possibly skewed with median zero. Finally the restriction on the shapes of F_1, \dots, F_c are relaxed. In all these cases, the test statistics are based on either the 15%-trimmed means or the sample medians, and their quantiles are estimated using the bias-corrected bootstrap. Section 4 is devoted to Monte Carlo studies including a comparison of our procedures with their well-known competitors. Section 5 illustrates the theory by means of a comparison of the mean effects of cocaine on pregnant rats, where heteroscedasticity and skewness are present. The Appendix consists of asymptotics, including a rigorous justification of the bootstrap.

2. Some Preliminaries

Let $X_r = (X_{r1}, \dots, X_{rn_r})$ be a sample from an unknown distribution function F_r and \bar{X}_{rt} and $X_{r,med}$ denote respectively the 15%-trimmed mean and the sample median.

Let $w_r = \left(\frac{1}{n_r} \sum |X_{r0i} - X_{r,med}| \right)^2$ be the square of the mean absolute deviation of the X_r -sample from its median $X_{r,med}$ and $\hat{w}_r = w_r/n_r$ (cf. BABU and RAO, 1992). Next, the variance estimate \hat{v}_{rt} of \bar{X}_{rt} is defined as follows:

Let $\alpha = 0.15$, $n = n_r$, $[n\alpha]$ the integer part of $n\alpha$, $X_i = X_{ri}$ and $X_{(i)}$ the i th order statistic, $r = g - [n\alpha]$,

$$X_L = (1 - r) X_{(g+1)} + rX_{(g)},$$

$$X_U = (1 - r) X_{(n-g)} + rX_{(n+1-g)},$$

$$\bar{X}_w = n^{-1} \left[\sum_{g+1}^{n-g} (X_{(i)} + g(X_L + X_U)) \right],$$

and

$$\lambda = \left[\sum (X_{(i)} - \bar{X}_w)^2 + g((X_L - \bar{X}_w)^2 + (X_U - \bar{X}_w)^2) \right].$$

Then $\hat{v}_{rt} = n^{-1}(1 - 2\alpha)^{-1} (n - 2n\alpha - 1)^{-1} \lambda$ (cf. ROCKE, DOWNS, and ROCKE (1982)).

The statistics T_3, T_4, S_3, S_4 are defined later in this section. The issue which of the two families of statistics $\{T_1, T_2, T_3, T_4\}$ and $\{S_1, S_2, S_3, S_4\}$ is to be used,

will be clarified in later sections. In general T 's are used if symmetry is suggested by the preliminary analysis. Otherwise S 's are used.

If m_l is the location parameter corresponding to F_l ; $l = 1, 2, \dots, c$, then for testing $H_0 : m_1 = \dots = m_c$ against $H_1 : m_i \neq m_j$ for at least one pair (i, j) , the statistic to be used is either T_1 or S_1 . For paired comparisons, which involve the simultaneous testing of all pairs (m_i, m_j) for which $m_i - m_j \neq 0$, the statistic is T_2 or S_2 . For a simultaneous comparison of contrasts, the statistic is T_3 or S_3 . For example, let $c = 3$ and a simultaneous comparison of the contrasts $2m_1 - m_2 - m_3$, $2m_2 - m_1 - m_3$ and $2m_3 - m_1 - m_2$ be required. Then $T_3 = \max(A_1, A_2, A_3)$ and $S_3 = \max(B_1, B_2, B_3)$. Here

$$A_1 = (2\bar{X}_{1t} - \bar{X}_{2t} - \bar{X}_{3t})(4\hat{v}_{1t} + \hat{v}_{2t} + \hat{v}_{3t})^{-1/2},$$

$$A_2 = (2\bar{X}_{2t} - \bar{X}_{3t} - \bar{X}_{1t})(4\hat{v}_{2t} + \hat{v}_{3t} + \hat{v}_{1t})^{-1/2},$$

and

$$A_3 = (2\bar{X}_{3t} - \bar{X}_{1t} - \bar{X}_{2t})(4\hat{v}_{3t} + \hat{v}_{1t} + \hat{v}_{2t})^{-1/2}.$$

B_i is obtained from A_i by replacing \bar{X}_{rt} and \hat{v}_{rt} by $X_{r,\text{med}}$ and \hat{w}_r respectively.

For testing $H_0 : m_1 = \dots = m_c$ against the hypothesis of ordered alternatives

$$H_A : m_1 \leq m_2 \leq \dots \leq m_c$$

with strict inequality between m_a and m_{a+1} for at least one a , the choice is

$$T_4 = \sum_{1 \leq j < i \leq c} t_{ij} \quad \text{and} \quad S_4 = \sum_{1 \leq j < i \leq c} s_{ij}.$$

The null distributions of T_i and S_i , are intractable, and their asymptotic null distributions may not be of much use for practical sample sizes. On the other hand bootstrap method is known to give a better approximation than the one based on the normal approximation theory.

Therefore, bootstrap method is attractive, especially when the samples are of moderate size. In this connection, we shall use the bias-corrected refinement (cf. EFRON and TIBSHIRANI, 1986, p. 68), of the bootstrap percentile method. For brevity, it will be called the bias-corrected bootstrap.

One reason for the good performance of the bias-corrected bootstrap is that it leads to 'second-order correctness', which refers to rate of convergence, when there is a correction for the skewness of the distribution. Central Limit Theorem establishes only convergence, but does not provide any rate of convergence. Edgeworth expansions lead to approximation of a sampling distribution, by an expression involving normal distribution plus a skewness term plus error.

The skewness term is typically of the order $O(n^{-1/2})$ and the error term of the order $o(n^{-1/2})$. The second-order correctness has the potential of automatically correcting for the skewness term (cf. BABU and SINGH, 1983).

Let $X_{(l,1)} < X_{(l,2)} < \dots < X_{(l,n_l)}$ be the order statistics of $X_l = (X_{l1}, \dots, X_{ln_l})$. Write $X_{(l,0)} = 2X_{(l,1)} - X_{(l,2)}$ and $X_{(l,n_l+1)} = 2X_{(l,n_l)} - X_{(l,n_l-1)}$. Let F_l^* be the con-

tinuous distribution function, which places the probability $(n_l + 1)^{-1}$ uniformly on each of the intervals $(X_{(l,0)}, X_{(l,1)}, \dots, (X_{(l,n_l)}X_{(l,n_l+1)})$. Then bootstrap samples are drawn from F_l^* , $l = 1, 2, \dots, c$. This is the smooth bootstrap (cf. COLLINGS and HAMILTON, 1988).

Let \hat{G}_i be the resulting bootstrap distribution of T_i and $z_i = \Phi^{-1}(\hat{G}_i(T_i))$. Then the bias-corrected bootstrap quantile estimate of the 95% quantile of T_i is simply $\hat{G}_i^{-1}(\Phi(2z_i + 1.645))$, Φ being the cumulative standard normal distribution function.

Note that in the foregoing process the bootstrap is applied *separately* to each X_l -sample. Hence this methodology is applicable even when the underlying distributions are heteroscedastic or have different shapes.

Remark 2.1: The smooth bootstrap was used in preference to the classical bootstrap for the following reasons:

- (i) For small samples, say $n_i = 10$, the classical bootstrap may yield a sample, having only a single distinct observation, so that the corresponding scale estimate becomes zero. Then t_{ij}^* and s_{ij}^* (the bootstrap counterparts of t_{ij} and s_{ij}), become undefined.
- (ii) Even when the samples are not small, the classical bootstrap may yield tied observations, for whom the ranks are not well defined. Therefore, problems will arise with bootstrapping the Wilcoxon procedure (cf. Section 3). Of course ties can be resolved by the methods of AKRITAS, ARNOLD, and BRUNNER (1997) or BRUNNER, PURI, and SUN (1995). However, since we are already using bootstrapping, the smooth bootstrap is easier to apply.

The smooth bootstrap precludes possibilities (i) and (ii).

Remark 2.2: For estimating the q th quantile ($0 < q < 1$) of our test statistics, the smooth bootstrap quantile s_q is consistent (cf. Appendix). Moreover, for large samples, s_q and the bias-corrected bootstrap quantile b_q give nearly the same result. However, for practical sample sizes, the best results are obtained by using b_q . Hence, in all our Monte Carlo studies only b_q will be used.

Some other competitors to our procedures are the Wilcoxon procedure, Dunnett's normal theory procedure, and the procedures of BRUNNER and PURI (1996), HOLM (1979), SCHAFFER (1986) and MARCUS, PERITZ, and GABRIEL (1976). The first two are briefly described below while the others are described at the end of Section 4.

The Wilcoxon Procedure (cf. PURI and SEN, 1970, pp. 245–246).

For simplicity, we shall describe this only for the case of equal sample sizes (although the theory can be modified to cover unequal sample sizes). Let $n_1 = n_2 = \dots = n_c$. For $1 \leq i < j \leq c$, denote by $h(X_i, X_j)$ the Wilcoxon statistic based on the i th and j th samples. Let μ and V^2 be its null mean and variance respectively (as the sample sizes are equal, μ and V are the same for all i and j). Write $h_3(X_i, X_j) = n\sqrt{2}(h(X_i, X_j) - \mu)/V$. Denote by $R_{c,\alpha}$ the upper α th quantile

of the range of a sample of size c from the standard normal. As before, let m_i be the location parameter for F_i .

Simultaneous Test for Paired Comparisons

At level α , $m_i - m_j$ is regarded to be significantly different from zero, if $|h_S(X_i, X_j)| > R_{c,\alpha}$. Equivalently, H_0 is rejected if $\max_{1 \leq i < j \leq c} |h_S(X_i, X_j)| > R_{c,\alpha}$. A simultaneous test for contrasts can be performed as on page 250 of PURI and SEN (1971). The details, being messy, are omitted.

For the one-way ANOVA model involving normal distributions with unequal variances, Dunnett proposes four procedures, of which the GH procedure seems to be the best and is described below.

Dunnett’s GH Procedure (DUNNETT, 1980, p. 796).

Let \bar{X}_l and s_l^2 be the sample mean and sample variance of a sample of size n_l from F_l , $l = 1, 2, \dots, c$. Write $u_l = n_l - 1$ and

$$\hat{e}_{ij} = \left(\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j} \right) \left(\frac{s_i^4}{n_i^2 u_i} + \frac{s_j^4}{n_j^2 u_j} \right)^{-1}.$$

Denote by $SR_{\alpha, c\hat{e}_{ij}}$ the upper α quantile of the studentized range of c normal variates with \hat{e}_{ij} degrees of freedom.

Simultaneous Test for Paired Comparisons

For $i \neq j$, $m_i - m_j$ is considered to be significantly different from zero, if

$$\sqrt{2} |\bar{X}_i - \bar{X}_j| \left(\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j} \right)^{-1/2} > SR_{\alpha, c\hat{e}_{ij}}.$$

Remark 2.3: All Dunnett’s procedures including the GH procedure, are designed only for paired comparisons. They cannot handle contrasts, involving three or more treatments.

As there is no danger of confusion, we shall refer to Dunnett’s GH procedure as simply Dunnett’s procedure.

Finally we shall describe the statistic $Q_{1,ad}$ to be used for checking symmetry or skewness in Section 4, where F_1, \dots, F_c , differ if at all, only in location and scale. Define $U_{\gamma,i}$ (respectively $L_{\gamma,i}$) the mean of the γn_i upper (or lower) order statistics of the sample X_i (obtained by linear interpolation, if γn_i is not an integer). Set

$$Q_{2,i} = (U_{0.05,i} - L_{0.05,i}) / (U_{0.5,i} - L_{0.5,i}) \quad \text{and} \quad Q_2 = \left(\sum_1^c n_i Q_{2,i} \right) / \sum_1^c n_i.$$

In the case of the uniform, normal and double exponential distributions (which are standard examples of light-tailed, normal-tailed and heavy-tailed distributions), the population parameter q_2 corresponding to Q_2 is given by $q_2 = 1.9, 2.58$ and 3.3 respectively. The middle point of $(2.58, 3.3)$ is 2.94 . However, for computational simplicity, we replaced 2.94 by 3 (as a cut-off point). Based on empirical considerations, 5 was chosen as a lower cut-off point for very heavy tails. Thus F was classified as light- (or normal-) tailed or heavy-tailed or very heavy-tailed, according as $Q_2 < 3$ or $3 \leq Q_2 < 5$ or $5 \leq Q_2$ respectively.

Let M_i be the average of the middle 50% of the order statistics of the sample $X_i, i = 1, 2, \dots, c, Q_{1,i} = (U_{0.05,i} - M_i)/(M_i - L_{0.05,i})$ and $Q_1 = \sum_i n_i Q_{1,i} / \sum_i n_i$. According to HOGG, FISHER, and RANGLES (1975), the distributions may be assumed to be symmetric or right-skewed or left-skewed, according as $\frac{1}{2} \leq Q_1 \leq 2$, or $Q_1 > 2$ or $Q_1 < \frac{1}{2}$. However, the performance of Q_1 is adversely affected by outliers. Therefore, we introduce $Q_{1,ad}$, which is more effective than Q_1 . Let $Q_{1,0.10}$ and $Q_{1,0.20}$ denote Q_1 computed on the 10%- and 20%-trimmed samples respectively. Then

$$\begin{aligned} Q_{1,ad} &= Q_1, & \text{if } Q_2 < 3 \\ &= Q_{1,0.10}, & \text{if } 3 \leq Q_2 < 5 \\ &= Q_{1,0.20}, & \text{if } 5 \leq Q_2. \end{aligned}$$

Modifying the proposal of Hogg et al., we say that $\frac{1}{2} \leq Q_{1,ad} \leq 2$ indicates symmetry, while $Q_{1,ad} > 2$ and $Q_{1,ad} < \frac{1}{2}$ indicate right-skewness and left-skewness respectively. For practical sample sizes, the final level of the test will be affected by a preliminary test of symmetry (although asymptotically this effect will be zero). The extent to which it is affected, depends on the preliminary test statistic. In the case of $Q_1, Q_{1,ad}$ etc., the effect is very little. For, being based on order statistics, these have very good convergence properties, much better than those based on measures such as kurtosis (cf. HOGG, 1974, p. 913).

Remark 2.4: Bootstrap methods for testing and confidence intervals have been discussed in HINKLEY (1988 and 1989), DICICCIO and ROMANO (1988) and WU (1986). However, these do not apply to our problem, as briefly explained below:

1. The majority of theoretical results in HINKLEY (1988) and DICICCIO and ROMANO (1988, Section 2.4) deal with statistics, which are either functions of vector averages or representable by certain Volterra series (HINKLEY, 1988, p. 324). Nevertheless, our statistics T_2 and T_3 (for simultaneous testing of paired comparisons and contrasts respectively) do not satisfy the above conditions.
2. Assume for simplicity $c = 3$ and consider the heteroscedastic model $F_i(u) = F((u - m_i)/\sigma_i), i = 1 - 3$, where F is symmetric about zero and the σ_i 's are possibly unequal. For $i = 1 - 3$, let $\bar{X}_{i\alpha}$ denote the trimmed mean of the sample from the i th distribution. Using the pure significance tests in

HINKLEY, 1988, Section 5 or HINKLEY, 1989, Section 3, or the results in DICICCIO and ROMANO, 1988, Section 2.4, one can construct separate tests (or confidence intervals) for each m_1 , m_2 and m_3 . The resulting tests for $m_1 - m_2$, $m_2 - m_3$ and $m_3 - m_1$ will be based on $\bar{X}_{1\alpha} - \bar{X}_{2\alpha}$, $\bar{X}_{2\alpha} - \bar{X}_{3\alpha}$ and $\bar{X}_{3\alpha} - \bar{X}_{1\alpha}$ respectively. However, as these tests are not independent, the only way of combining them into an overall simultaneous test of the paired comparisons $\theta_1 - \theta_2$, $\theta_2 - \theta_3$ and $\theta_3 - \theta_1$ or the contrasts, say $2\theta_1 - \theta_2 - \theta_3$, $\theta_1 - 2\theta_2 + \theta_3$ and $\theta_1 + \theta_2 - 2\theta_3$ seems to use tests based on statistics of the type T_2 and T_3 introduced at the beginning of this section.

3. Using resampling methods, WU (1986) has proposed approximately unbiased variance estimators in the heteroscedastic regression model, including the special case of the c -sample location problem with unequal variances (WU, 1986, p. 1276). Nevertheless, he has not considered simultaneous testing of paired comparisons and contrasts. The simulation study in his paper considers confidence intervals for a single parametric function, assuming the normality of the underlying distribution.

Remark 2.5: The generalised two-sample Behrens-Fisher problem of comparing the medians of two unknown distributions with possibly different shapes was studied with and without the symmetry assumption in FLIGNER and POLICELLO (1981) and FLIGNER and RUST (1982) respectively. Our study of ANOVA for distributions with possibly different shapes can be regarded as a modified c -sample version of this problem. However, the techniques of these authors do not easily extend to the present setting.

3. Selection of the Statistics

(a) *Robustness to Possible Heteroscedasticity*

Let $X_r = (X_{r1}, \dots, X_{rm_r})$ be a sample from $F_r(u) = F((u - m_r)/\sigma_r)$, $r = 1, 2, \dots, c$ ($c > 2$), F an unknown distribution function symmetric about the origin and $\sigma_1, \dots, \sigma_c$ unknown and possibly unequal. The appropriate statistics to use in this case are T_1 , T_2 , T_3 and T_4 . See Section 2 for details on the choice of T_i .

(b) *Robustness to Possible Skewness as Well as Heteroscedasticity*

Suppose that the assumption of symmetry itself is suspect. Then one approach is to apply (to the data) the logarithmic (or some similar) transformation, hopefully achieving symmetry and to analyze the transformed data, as in (a) above. However, as explained in Section 1, it is preferable to avoid transformations altogether and work with the original data.

Note that the possible lack of symmetry implies that the error distribution has median zero, but may lack symmetry and the parameters of interest are

therefore the population medians m_1, m_2, \dots, m_c . This leaves us with two options.

I. Perform analysis using the family $\{S_1, S_2, S_3, S_4\}$.

II. Perform a preliminary test of symmetry, using $Q_{1,ad}$.

If that test supports symmetry; i.e. if $\frac{1}{2} \leq Q_{1,ad} \leq 2$, then proceed as (a) above by using $\{T_1, T_2, T_3, T_4\}$; otherwise proceed as in Option I by using $\{S_1, S_2, S_3, S_4\}$.

(c) *Robustness to Possible Differences in Distribution Shapes*

We have so far assumed that F_1, \dots, F_c have the same shape. This assumption is tenable, if the $Q - Q$ plot corresponding to any pair of samples is (at least approximately) linear. Otherwise, it becomes dubious, in which case, we may proceed as in Option I of (b). Here again the parameters of interest are population medians.

4. The Monte Carlo Studies

Due to restriction on computer time, of the four statistics T_1, T_2, T_3 and T_4 , only T_1, T_2 and T_3 were considered. Even here, only T_2 and T_3 were studied in detail, since preliminary studies found the performance of T_1 is similar to that of the other two. Similarly, in the family S_1, S_2, S_3 and S_4 only S_2 and S_3 were studied in detail.

The procedure (a) in Section 3 was studied for $c = 3, n_i = 10, 15$ and 20 and for $c = 5, n_i = 20, i = 1, 2, \dots, c$. The procedure (c) in Section 3 was studied for $c = 3$ and $n_i = 20, i = 1, 2, 3$. Then our procedure corresponding to Option II of (b) in Section 3 was compared, in terms of level and power, with the Wilcoxon procedure, and Dunnett's procedure, for $c = 3$ and $n_1 = n_2 = n_3 = 20$.

Random samples of sizes n_i ($i = 1, 2, \dots, c$) were drawn using IMSL subroutines on a VAX 23 computer, from a number of distributions to be specified shortly, and there were two thousand replications. Following the recommendations of EFRON and TIBSHIRANI (1986, p. 72), one thousand bootstrap samples were drawn from each sample. The nominal level was always kept at five percent. The quantiles of our test statistics were always estimated by the bias-corrected bootstrap quantiles (cf. Remark 2.2).

Let $N(0, 1)$ and $N(0, 4^2)$ denote respectively the standard normal and the normal distribution with mean 0 and variance 16. For a given integer λ , $\lambda\% cN$ will denote the mixture $(100 - \lambda)\% N(0, 1) + \lambda\% N(0, 4^2)$. The distributions studied were Beta (3, 3), shifted appropriately to the left so that its median became zero, $N(0, 1)$, $5\% cN$, $10\% cN$, $15\% cN$, and the standard exponential and Lognormal distributions, shifted appropriately to the left so that their medians become zero. These will be referred to as distributions, $B, N, 5\% cN, 10\% cN, 15\% cN, E$ and L respectively.

In connection with Tables 1, 2, 3, 5 and 6, heteroscedasticity was introduced as follows. When $c = 3$, the values in the second and third samples were multiplied by two and four respectively. When $c = 5$, the values in the second, third, fourth and fifth sampler were multiplied by two, three, four and five respectively.

In Table 4, the first, second and third samples were drawn respectively from distributions F_1, F_2 and F_3 having different shapes. Tables 5 and 6 compare our procedure (Option II of (b) in Section 3) with the Wilcoxon procedure and Dunnett's normal theory procedure.

Tables 1, 2 and 3 are concerned respectively with the procedure in Section 3, when $c = 3, n_i = 10$ and 15 and $c = 5, n_i = 20$. (Recall that now the T_2 -test is for paired comparisons, while the T_3 -test is for contrasts. When $c = 5, T_3 = \max(A_1, A_2, A_3, A_4, A_5)$ where $A_1 = (4\bar{X}_{1t} - \sum_{j=2}^5 \bar{X}_{jt}) (16\hat{v}_{jt} + \sum_{j=2}^5 \hat{v}_{jt})^{-1/2}$ and $A_i, i = 2, \dots, 5$ is obtained from A_1 by interchanging 1 and i).

Tables 1–6 contain the point estimates \hat{p} and also the 95% confidence intervals, i.e., the numbers correspond to the intervals $\hat{p} \pm 1.96 \sqrt{\hat{p}(1 - \hat{p})/2000}$.

Table 4 studies the procedure (c) in Section 3 for $c = 3$ and $n_i = 20$ for cases 1 and 2, F_1, F_2 and F_3 denoting respectively N, E and L in case 1 and B, E and $10\% cN$ in case 2.

Table 5 contains the empirical levels of our procedure in Option II of (b) of Section 3, and the Wilcoxon procedure and Dunnett's procedure. Recall that Option II chooses T_i or S_i according as $Q_{1,ad}$ suggests symmetry or skewness. Ta-

Table 1
Empirical levels (in percent) and confidence intervals ($c = 3, n_i = 10$)

Distributions	T_2 -test	T_3 -test
B	5.50 ± 0.99	5.70 ± 1.02
N	5.80 ± 1.02	5.95 ± 1.04
5% cN	6.00 ± 1.04	6.10 ± 1.05
10% cN	6.30 ± 1.06	6.60 ± 1.09
15% cN	6.70 ± 1.10	6.90 ± 1.11

Table 2
Empirical levels (in percent) and confidence intervals ($c = 3, n_i = 15$)

Distributions	T_2 -test	T_3 -test
B	4.40 ± 0.90	4.60 ± 0.92
N	4.70 ± 0.93	4.90 ± 0.95
5% cN	4.90 ± 0.95	5.05 ± 0.96
10% cN	5.09 ± 0.96	5.29 ± 0.98
15% cN	5.60 ± 1.01	5.90 ± 1.03

Table 3
Empirical levels (in percent) and confidence intervals ($c = 5, n_i = 20$)

Distributions	T_2 -test	T_3 -test
B	4.65 ± 0.92	4.85 ± 0.94
N	4.90 ± 0.95	5.20 ± 0.97
5% cN	5.00 ± 0.96	5.25 ± 0.98
10% cN	5.25 ± 0.98	5.35 ± 0.99
15% cN	5.80 ± 1.02	5.90 ± 1.03

Table 4
Empirical levels (in percent) and confidence intervals ($c = 5, n_i = 20$)

Cases	S_2 -test	S_3 -test
1	4.25 ± 0.88	4.1 ± 0.87
N	4.15 ± 0.87	4.1 ± 0.87

Table 5
Empirical levels (in percent) and confidence intervals ($c = 3, n_i = 20$)

Distribution	Option II		Wilcoxon		Dunnnett
<i>B</i>	4.9 ± 0.9	5.0 ± 0.9	48.0 ± 2.1	55.0 ± 2.1	4.9 ± 0.9*
<i>N</i>	4.9 ± 0.9	4.9 ± 0.9	54.7 ± 2.1	50.0 ± 2.1	4.6 ± 0.9*
5% <i>cN</i>	4.7 ± 0.9	4.5 ± 0.9	68.1 ± 2.0	7.3 ± 1.9	4.1 ± 0.8*
10% <i>cN</i>	4.9 ± 0.9	5.3 ± 0.9	78.1 ± 1.8	82.4 ± 1.6	3.6 ± 0.8*
15% <i>cN</i>	4.8 ± 0.9	4.7 ± 0.9	78.3 ± 1.8	83.0 ± 1.6	2.9 ± 0.7*
<i>E</i>	5.2 ± 0.9	5.2 ± 0.9	79.0 ± 1.7	83.5 ± 1.6	17.1 ± 1.6*
<i>L</i>	5.4 ± 0.9	5.4 ± 1.0	84.3 ± 1.5	89.5 ± 1.3	19.3 ± 1.7*

ble 6 contains a similar study of empirical powers, after the addition of 0.50 and 0.75 to the second and third sample values respectively.

In Table 5, the first and second entries under each procedure denote its empirical level for paired comparisons and contrasts respectively. The only exception is the * under Dunnnett’s procedure, denoting the inapplicability of the latter to contrasts (cf. Remark 2.3). A similar interpretation applies to Table 6 with ‘empirical level’ replaced by ‘empirical power’.

Table 5 shows that our procedure (Option II) is quite robust. Dunnnett’s procedure is good at the Beta and the Normal models, but becomes conservative (with the resulting loss of power) for heavy-tailed symmetric distributions and extremely liberal for skewed distributions. This resembles the performance of the two-sample *t*-test in such situations. The Wilcoxon procedure fares the worst, being unacceptably liberal. This is not surprising, in view of its unsatisfactory performance even in the simpler two-sample location model, with unequal variances (cf. PRATT, 1964, 665–680).

In Table 6, all the entries under the Wilcoxon procedure and the last two under the Dunnnett procedure, should be discounted, as the corresponding empirical levels are excessively high. Thus when both robustness of level and reasonably good power across distributions are taken into account, our procedure becomes the best performer. Moreover, the Dunnnett and the Wilcoxon procedures cannot be made robust by bootstrapping.

For the situation corresponding to distribution *L* in Table 5 (and nominal level 5 percent), the empirical level of Dunnnett’s procedure based on the bootstrap quantile was only 0.073 percent. One possible explanation for the failure of the bootstrap is the non-robustness of the sample means, on which Dunnnett’s statistic is based.

The appropriate quantile of the Wilcoxon procedure was estimated by bootstrapping separately from each sample $X_n - c_r X_{r,med}$, where c_r is the n_r -dimensional vector with all entries equal to 1. For the situation corresponding to distribution *L* in Table 5 and nominal level five percent, the empirical level of the Wilcoxon procedure for paired comparisons was 9.1 percent. Because of the unsatisfactory performance, the Wilcoxon procedure also was not considered any further.

Table 6a

Empirical power (in percent) and confidence intervals ($c = 3, n_i = 20$)

Distribution	Option II		Wilcoxon		Dunnett
<i>B</i>	13.1 ± 1.4	12.9 ± 1.4	66.0 ± 2.0	71.0 ± 1.9	15.2 ± 1.5*
<i>N</i>	12.2 ± 1.4	11.8 ± 1.4	60.0 ± 2.1	70.0 ± 2.0	13.9 ± 1.5*
5% <i>cN</i>	11.0 ± 1.3	10.6 ± 1.3	77.0 ± 1.8	81.0 ± 1.7	10.4 ± 1.3*
10% <i>cN</i>	10.8 ± 1.3	10.4 ± 1.3	92.0 ± 1.1	95.0 ± 0.9	8.4 ± 1.2*
15% <i>cN</i>	10.4 ± 1.3	10.2 ± 1.3	88.1 ± 1.4	94.2 ± 1.02	6.0 ± 1.0*
<i>E</i>	7.2 ± 1.1	10.0 ± 1.3	87.0 ± 1.4	92.2 ± 1.1	75.0 ± 1.9*
<i>L</i>	6.8 ± 1.1	9.4 ± 1.2	89.1 ± 1.3	91.5 ± 1.2	70.0 ± 2.0*

* not applicable to contrasts (cf. Remark 2.3)

Following a suggestion of the referee, our procedures were also studied for unequal sample sizes, when large variances appear with small sample sizes and small variances appear with large sample sizes. More precisely, samples of sizes $n_1 = 10, n_2 = 20$ and $n_3 = 30$ were drawn from the five symmetric distributions *B*, *N*, 5% *cN*, 10% *cN* and 15% *cN*. The values in the second sample were multiplied by two while those in the first sample were multiplied by four. The results are presented in Table 6b, where the first and second entries in each row, under the column ‘levels’ (and similarly under the column ‘powers’) correspond to paired comparisons and contrasts respectively. (The nominal level was 5%.)

We now compare our methodology with some existing procedures.

Statistics for the location-scale model have been proposed by COMPAGNONE and DENKER (1996) and DENKER and PURI (1992). One of them is based on the kernel function

$$\Psi(x_1, \dots, x_m) = \frac{\sum_1^m x_k}{\sqrt{\sum_{k=1}^m (x_k - \bar{x})^2}}$$

(COMPAGNONE and DENKER, 1996, p. 137). Because of its scale-invariance, this statistic is sensitive to location differences but unaffected by scale differences and

Table 6b

Empirical levels and powers (in percent)

Distribution	Levels	Powers
<i>B</i>	4.60, 4.82	8.24, 14.22
<i>N</i>	4.85, 4.97	7.70, 8.15
5% <i>cN</i>	4.74, 5.20	7.20, 7.26
10% <i>cN</i>	4.80, 5.30	6.32, 6.14
15% <i>cN</i>	5.10, 5.20	5.74, 5.92

therefore can be used for testing for location differences in presence of nuisance scale parameters. However, there are some difficulties in applying such statistics to our problems. As explained in the next paragraph, such tests may be, for practical sample sizes, liberal for some distributions and conservative for some others in the two-sample problem. The results are going to be similar for the c -sample problem. Moreover, inference (based on these statistics) seems to be limited to paired comparisons and not possible for contrasts.

Consider, for example, the much simpler statistic (for the two-sample scale problem) based on the kernel function $\sum_{k=1}^m (x_k - \bar{x})^2$ (COMPAGNONE and DENKER, p. 138). Even here, exact variance estimates could be tedious to compute for practical sample sizes (para. 3, page 145) and only an approximate variance estimate based on the jackknife has to be employed (page 146). Probably, as a result of this, the test is liberal for some distributions (Table III, entries under G_1 , G_2 and G_5 , pp. 151–152) and conservative for some others (Table IV, entry under G_4). Finally, the test based on a similar statistic in DENKER and PURI (1992) is somewhat less robust than the above test (cf. COMPAGNONE and DENKER, 1992, p. 142, second last paragraph).

Nonparametric methods for stratified two-sample designs were proposed in BRUNNER, PURI, and SUN (1995) even without the assumption of continuity of the underlying distribution function. Nonparametric hypotheses and rank statistics for unbalanced factorial designs (making allowance for heteroscedastic errors) were studied in AKRITAS, ARNOLD, and BRUNNER (1997), once again without assuming continuity of the underlying distribution function. For the one-way ANOVA model, the methodologies of these papers specialize to testing for the equality of the treatment effects and do not cover simultaneous testing for paired comparisons and contrasts.

BRUNNER and PURI (1996) deal with heteroscedasticity, making an assumption, which, in the case of our problem, means the following:

Let X_{1i} and X_{2j} be the observations from the i th and j th samples. Under the null hypothesis that the effects of the i th and j th treatments are the same, $P[X_{1i} > X_{2j}] = P[X_{2j} > X_{1i}]$. While this assumption automatically holds for symmetric distributions, it may not be valid for skewed distributions, which are also considered in our paper. Moreover, BRUNNER and PURI (1996) do not discuss simultaneous testing for paired comparisons and contrasts. When there are logical implications among the hypotheses and alternatives, the sequentially subjective Bonferroni (SRB) procedure due to HOLM (1979) and its modification MSRB due to SCHAFFER (1986), can, in conjugation with WELCH's test (1951), be used to increase the powers of tests for three or more treatments.

MARCUS, PERITZ, and GABRIEL (1976) have presented a method of discussing stepwise multiple testing procedures with fixed experimentwise error rate. However, in applications, these procedures require the existence of exact level alpha tests.

5. An Illustration

An experiment was conducted by ROBINSON et al. (1993) to determine the effects of cocaine administration to pregant female Sprague-Dawley rats. There were two forms of cocaine administration, namely acute and chronic. In the acute administration, cocaine (6 mg/kg) was injected on day 18 of pregnancy. In the chronic administration, cocaine (6 mg/kg) was injected on each of days 8 through 18 of pregnancy. In both groups of cocaine administration, cocaine levels in the blood-stream were measured on day 18 of pregnancy at 5, 20 or 60 minutes of injection.

The X_1 -, X_2 - and X_3 -samples in Table 7 are taken from the data sets corresponding to chronic 5, chronic 20 and acute 5 respectively.

Experience shows that such measurements are highly skewed to the right. The mean absolute deviations for the three samples were 240.75, 111.75 and 151.88 indicating heteroscedasticity.

However, our methodology (Option I of (b) in Section 3) based on the sample medians, was applicable. The absolute values of the studentised versions of the differences between the three pairs of sample medians were 0.43, 0.32 and 0.016 while the bootstrap critical value was 0.94. Therefore, the difference between no two population medians was statistically significant, thus leading to the acceptance of the null hypothesis.

Next the data were analyzed using the Dunnett procedure. Let d_{ij} be the absolute value of the studentised version of the difference between the means of the i th and j th samples, and c_{ij} the corresponding critical value based on the studentised range. For the given data set, $d_{12} = 0.86$, $c_{12} = 3.96$, $d_{23} = 7.64$, $c_{23} = 3.765$, $d_{13} = 3.79$ and $c_{13} = 3.804$. Hence the second and third location effects are significantly different, leading to the rejection of the null hypothesis.

Finally the data were analysed using the Wilcoxon procedure. Let r_{ij} denote the absolute value of the studentised version of the Wilcoxon statistic based on the i th and j th samples. Then $r_{12} = 4.21$, $r_{23} = 26.95$ and $r_{13} = 17.65$, while the critical value based on the studentised range of a normal sample of size 3 was 3.3144. Therefore, according to the Wilcoxon procedure, the difference between any two population medians was significant once leading to the rejection of null hypothesis. This is not suprising, since for skewed distributions, the Dunnett and the Wilcoxon procedures tended to reject the null hypothesis more often (than our procedure), as shown by the Monte Carlo studies. As a result, the conclusions yielded by the Dunnett and Wilcoxon procedures were at variance with the conclusions of our procedure.

Table 7

X_1	140	180	93	32	34	629	1169	33
X_2	92	66	141	35	230	163	372	463
X_3	1158	1076	612	672	538	561	650	630

Acknowledgements

We are deeply grateful to the referee for many valuable suggestions and having drawn our attention to the extensive literature on nonparametric methodology in ANOVA. We are also deeply grateful to Professor Chinchilli, of the Department of Health Sciences, Pennsylvania State University for valuable suggestions and having obtained permission (on our behalf) from Robinson et al. to use their data sets.

Appendix on Asymptotics

As $Q_{1,ad}$ converges in probability to a constant, it is enough to consider the trimmed mean and the sample median.

Justification of the bootstrap in the case of the trimmed mean

Let $n = n_1 + \dots + n_c$,

$$\hat{\theta}_i = \sqrt{n}(\bar{X}_{i,t} - m_i), \quad \theta_i^* = \sqrt{n}(\bar{X}_{i,t}^* - \bar{X}_{i,t}),$$

$$\tilde{t}_{ij} = (\hat{\theta}_i - \hat{\theta}_j) / \sqrt{n(\hat{v}_i + \hat{v}_j)}, \quad \text{and} \quad t_{ij}^b = (\theta_i^* - \theta_j^*) / \sqrt{n(v_i^* + v_j^*)},$$

where v_i^* denotes the bootstrap version of \hat{v}_i . For $k = 1, 2, 3, 4$, let \tilde{T}_k and T_k^b be obtained from T_k , by replacing t_{ij} respectively by \tilde{t}_{ij} and t_{ij}^b . Let $H_{k,n}$ and $H_{k,n}^*$ denote the distributions of \tilde{T}_k and T_k^b respectively. Under the null hypothesis, $\tilde{T}_k = T_k$ and its distribution is given by $H_{k,n}$. In the case of contrasts, we consider the case $c = 3$, since the proof easily extends to any $c > 0$.

Theorem 1: *Suppose $(n_i/n) \rightarrow p_i$ as $n \rightarrow \infty$ and $p_i > 0$ for $i = 1, \dots, c$. Then for almost all sample sequences and for $k = 1, 2, 3, 4$, we have*

$$\sup_x |H_{k,n}(x) - H_{k,n}^*(x)| \rightarrow 0. \tag{1}$$

If $t_{k,\beta}^*$ denote a β th quantile of $H_{k,n}^*$, then $P(\tilde{T}_k \leq t_{k,\beta}^*) \rightarrow \beta$.

Proof: By Theorem 3 and (P9) of BABU and SINGH (1984) and by Theorem 1. A of SINGH (1981), it follows that both $\hat{\theta}_i$ and for almost all sample sequences θ_i^* have the same limiting centered normal distribution. Further $n\hat{v}_i \rightarrow \gamma_i$ and $nv_i^* \rightarrow p\gamma_i$ for almost all sample sequences. As $\hat{\theta}_i$ are based on independent samples, the common limiting distribution of the vectors $(\hat{\theta}_1, \dots, \hat{\theta}_c)$ and $(\theta_1^*, \dots, \theta_c^*)$ is the same as that of $W = (\sqrt{\gamma_1} N_1, \dots, \sqrt{\gamma_c} N_c)$, where N_1, \dots, N_c are independent standard normal distributions. Consequently, both $H_{k,n}$ and $H_{k,n}^*$

have the same limiting distribution as that of $h_k(W)$, where

$$h_1(x_1, \dots, x_c) = \sum_{1 \leq i < j \leq c} (|x_i - x_j| / \sqrt{\gamma_i + \gamma_j}),$$

$$h_2(x_1, \dots, x_c) = \max_{1 \leq i < j \leq c} (|x_i - x_j| / \sqrt{\gamma_i + \gamma_j}),$$

$$h_3(x_1, x_2, x_3) = \max((2x_1 - x_2 - x_3) b_1, (2x_2 - x_1 - x_3) b_2, (2x_3 - x_1 - x_2) b_3),$$

$$h_4(x_1, \dots, x_c) = \sum_{1 \leq i < j \leq c} ((x_i - x_j) / \sqrt{\gamma_i + \gamma_j}),$$

and for $i = 1, 2, 3$, $b_i = (3\gamma_i + \gamma_1 + \gamma_2 + \gamma_3)^{-1/2}$.

It is easily seen that the maps h_k are continuous and that the distributions of $h_k(W)$ are continuous. This leads to (1). The limit second limit follows from Lemma 2.1 of BABU and BOSE (1988). This completes the proof of the theorem.

Justification of the bootstrap in the case of the sample median

Let $X_r = (X_{r,1}, \dots, X_{r,m_r})$ be independent samples from distributions F_r with median m_r . Let $X_{r,med}$ denote the sample median of X_r ,

$$M_i = \frac{1}{n_i} \sum_{j=1}^{n_i} |X_{ij} - X_{i,med}| \quad \text{and} \quad M_i^* = \frac{1}{n_i} \sum_{j=1}^{n_i} |X_{ij}^x - X_{i,med}^*|,$$

where $X_r^* = (X_{r1}^*, \dots, X_{r,m_r}^*)$ is a bootstrap sample (drawn from X_r), having median $X_{r,med}$. Let \tilde{S}_k be obtained from \tilde{T}_k by replacing $\hat{\theta}_i$ and \hat{v}_i respectively by $\sqrt{n} (X_{i,med}^* - m_i)$ and M_i^2/n_i . Let S_k^b be defined similarly by replacing θ_i^* and \hat{v}_i^* in T_k^b , respectively by $\sqrt{n} (X_{i,med}^* - X_{i,med})$ and M_i^{*2}/n_i . Let $R_{k,n} = R_{k,n}^*$ denote the distribution of \tilde{S}_k and S_k^b . Note that under the null hypothesis, $R_{k,n}$ denotes the distribution of S_k .

Theorem 2: Suppose F_i is differentiable in a neighborhood of its median m_i and $F_i'(m_i) > 0$. If $(n_i/n \rightarrow p_i$ as $n \rightarrow \infty)$ and $p_i > 0$ for $i = 1, \dots, c$, then for almost all sample sequences, we have $\sup_x |R_{k,n}(x) - R_{k,n}^*(x)| \rightarrow 0$ and for $k = 1, 2, 3, 4$. If $s_{k,\beta}^*$ denote a β th quantile of $R_{k,n}^*$, then $P(\tilde{S}_k \leq s_{k,\beta}^*) \rightarrow \beta$.

The proof will be given after establishing some preliminary lemmas.

Lemma 1: If $\{Z_i\}$ is a sequence of independent random variables from a common distribution, with $E(|Z_i|) < \infty$ then $n^{-2} \sum_{i=1}^n Z_i^2 \rightarrow 0$ a.e.

Proof: Note that $E(|Z|) < \infty$ implies $\sum_{i=1}^{\infty} P(|Z_i| \geq \epsilon) < \infty$, for all $\epsilon > 0$, see equation 21.9 on page 282 of BILLINGSLEY (1986). So by Borel-Cantelli lemma (BILLINGSLEY, 1986, p. 53), it follows that $n^{-1} \max_{1 \leq i \leq n} |Z_i| \rightarrow 0$ a.e. Hence by the strong law of large numbers (see BILLINGSLEY, 1986, p. 290), we have with prob-

ability one, that

$$\frac{1}{n^2} \sum_{i=1}^n Z_i^2 \leq \frac{1}{n} \sum_{i=1}^n |Z_i| \left(\frac{1}{n} 1 \leq j < n |Z_j| \right) \rightarrow 0.$$

Lemma 2: Let $\{Z_i\}$ be as in Lemma 1. Suppose in addition that the distribution G of Z_1 is strictly increasing and continuous in a neighbourhood of its median m . If m_n denotes the median of the sample Z_1, \dots, Z_n and m_n^* denotes the median of a corresponding bootstrap sample Z_1^*, \dots, Z_n^* , then

- (i) $m_n - m \rightarrow 0$ a.e.
- (ii) $m_n^* - m_n \rightarrow 0$ in probability for almost all sample sequences.
- (iii) $\frac{1}{n} \sum_{i=1}^n |Z_i - m_i| \rightarrow E(|Z_1 - m|)$ a.e.

and

- (iv) $\frac{1}{n} \sum_{i=1}^n |Z_i^* - m_i^*| - \frac{1}{n} \sum_{i=1}^n |Z_i - m_n| \rightarrow_p 0$ for almost all sample sequences.

Proof: Let G_n and G_n^* denote the empirical distribution functions of $\{Z_1, \dots, Z_n\}$ and $\{Z_1^*, \dots, Z_n^*\}$ respectively. Fix $\epsilon > 0$ and consider,

$$\begin{aligned} P(m_n > m + \epsilon) &\leq P\left(\frac{1}{2} \geq G_n(m + \epsilon)\right) \\ &\leq P\left(\frac{1}{2} - G(m + \epsilon) > G_n(m + \epsilon) - G(m + \epsilon)\right). \end{aligned}$$

As G is strictly increasing in the neighbourhood of m , $G(m + \epsilon) - \frac{1}{2} > \delta$ for some $\delta > 0$. Since

$$\sup_x |G_n(x) - G(x)| \rightarrow 0 \quad \text{a.e.} \tag{2}$$

(see BILLINGSLEY, 1986, p. 275), $\limsup_{n \rightarrow \infty} m_n \leq m + \epsilon$, a.e. Similarly $\liminf_{n \rightarrow \infty} m_n \geq m - \epsilon$, a.e. This proves (i). Now consider

$$\begin{aligned} P^*(m_n^* > m_n + \epsilon) &\leq P^*\left(\frac{1}{2} \geq G_n^*(m_n + \epsilon)\right) \\ &\leq P^*\left(\frac{1}{2} - G_n(m_n + \epsilon) > G_n^*(m_n + \epsilon) - G_n(m_n + \epsilon)\right). \end{aligned} \tag{3}$$

Note that by (i) and (2), $G_n(m_n + \epsilon) \rightarrow G(m + \epsilon) > \frac{1}{2} + \delta > 0$ a.e. Hence it follows that the left side of (3) is

$$\begin{aligned} &\leq P^*(|G_n(m_n + \epsilon) - G_n^*(m_n + \epsilon)| > \delta) \leq \delta^{-2} \text{Var}^*(G_n^*(m_n + \epsilon)) \\ &\leq n^{-1} \delta^{-2} \text{Var}^*(I(Z_1^* \leq m_n + \epsilon)) \leq n^{-1} \delta^{-2} \rightarrow 0. \end{aligned}$$

The result $P^*(m_n^* \leq m_n - \epsilon) \rightarrow 0$, can be established similarly. This proves (ii).

To prove (iii) and (iv) note that

$$\frac{1}{n} \sum_{i=1}^n (|z_i^* - m_n^*| - |z_i^* - m|) \leq |m_n^* - m| \leq |m_n^* - m_n| + |m_n - m|$$

and $\frac{1}{n} \sum_1^n (|Z_i - m_n| - |Z_i - m|) \leq |m_n - m|$. The strong law of large numbers and (i) give us (iii). For any $\eta > 0$, we have

$$\begin{aligned} & \eta^2 P^* \left(\left| \sum_1^n (|Z_i^* - m| - |Z_i - m|) \right| > \eta n \right) \\ & \leq n^{-1} \text{Var}^* |Z_1^* - m| \leq n^{-2} \sum_1^n (Z_i - m)^2 \rightarrow 0, \end{aligned}$$

for almost all samples by Lemma 1. This proves (iv).

Proof of Theorem 2: By Bahadur's representation of sample quantile (see GHOSH, 1971 for a proof), we have, for some θ_i , $X_{i,\text{med}} - m_i = (\hat{F}_i(m_i) - \frac{1}{2}) \theta_i + o_p(n^{-1/2})$, and for almost all samples, $X_{i,\text{med}}^* - X_{i,\text{med}} = (F_i^*(m_i) - \hat{F}_i(m_i)) \theta_i + o_p(n^{-1/2})$. Here \hat{F}_i and F_i^* respectively denote the empirical distribution functions of X_i and X_i^* . (The error terms above can be improved to $O_p(n^{-3/4} \log n)$ if F_i is twice differentiable in a neighborhood of m_i . See Theorem 5 of BABU and SINGH (1984).)

As $F_i'(m_i) > 0$, $E|X_{i1} - m_i| > 0$ and hence by lemma 2, $M_1/E(|X_{i1} - m_1|) \rightarrow 1$ a.e. and $M_i^*/M_i \rightarrow 1$ in probability for almost all sample sequences. Note that $\sqrt{2n}(\hat{F}_i(m_i) - \frac{1}{2})$ and for almost all sample sequences, $\sqrt{2n}(F_i^*(m_i) - \hat{F}_i(m_i))$ both converge in distribution to the standard normal distribution. The rest of the proof follows from the proof of Theorem 1.

References

- AKRITAS, M. G., ARNOLD, S. F., and BRUNNER, E., 1997: Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *J. Am. Statist. Ass.* **92**, 258–265.
- BRUNNER, E. and PURI, M. L., 1996: Nonparametric methods in design and analysis of experiments. In: C.-R. Rao and S. Ghosh (eds.): *Handbook of Statistics*. Vol. **13**, 631–673.
- BRUNNER, E., PURI, M. L., and SUN, S., 1995: Nonparametric methods for stratified two-sample designs with applications to multi-clinical trials. *J. Am. Statist. Assoc.* **90**, 1004–1014.
- BABU, G. J. and BOSE, A., 1988: Bootstrap confidence intervals. *Stat. Prob. Letters* **7**, 151–160.
- BABU, G. J. and RAO, C. R., 1992: Expansions for statistics involving the mean absolute deviations. *Ann. Inst. Statist. Math.* **44**, 387–403.
- BABU, G. J. and SINGH, K., 1983: Inference on means using the bootstrap. *Ann. Statist.* **11**, 999–1003.
- BABU, G. J. and SINGH, K., 1984: Asymptotic representations relating to jackknifing and bootstrapping L -statistics. *Sankhya A* **46**, 195–206.
- BILLINGSLEY, P., 1986: *Probability and Measure*, 2nd edition. John Wiley, New York.
- BISHOP, T. A. and DUDEWICZ, T. J., 1978: Exact analysis of variance with unequal variances. Test procedures and tables. *Technometrics* **20**, 419–424.
- COLLINGS, B. J. and HAMILTON, M. A., 1988: Estimating the power of the two-sample Wilcoxon test for shift. *Biometrics* **44**, 847–876.
- COMPAGNONE, D. and DENKER, M., 1996: Nonparametric tests for scale and location. *J. Nonpara. Statistic* **7**, 123–154.
- DENKER, M. and PURI, M. L., 1992: Asymptotic normality of two sample linear rank statistics under U -statistic structure. *J. Statist. Plann. Infer.* **32**, 89–110.
- DICICCIO, T. J. and ROMANO, R. P., 1988: A review of bootstrap confidence intervals. *J. R. Statist. Soc. B* **50**, 338–354.

- DUDEWICZ, E. J., 1983: Heteroscedasticity. In: S. Kotz and N. L. Johnson (eds.): *Encyclopedia of Statistical Sciences*. Vol. **3**, 611–619.
- DUNNETT, C. W., 1980: Pairwise multiple comparisons in the unequal variance case. *J. Am. Statist. Ass.* **75**, 796–800.
- EFRON, B. and TIBSHIRANI, R., 1986: Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. *Statistical Science* **1**, 54–75.
- FLIGNER, M. A. and POLICELLO, G. E., 1981: Robust rank procedures for the Behrens-Fisher problem. *J. Am. Statist. Ass.* **76**, 162–168.
- FLIGNER, M. A. and RUST, S. W., 1982: A modification of Mood's median test for the generalised Behrens-Fisher problem. *Biometrika* **69**, 221–226.
- GHOSH, J. K., 1971: A new proof of the Bahadur representation of quantiles and an application. *Ann. Math. Statist.* **42**, 1957–1961.
- HINKLEY, D. V., 1988: Bootstrap methods. *J. R. Statist. Soc. B* **50**, 321–337.
- HINKLEY, D. V., 1989: Bootstrap significance tests. *ISI Bulletin, 47th session* **53**, 65–74.
- HOGG, R. V., 1975: Adaptive robust procedures: a partial review and some suggestions for future applications and theory. *J. Am. Statist. Ass.* **69**, 909–925.
- HOGG, R. V., FISHER, D. M., and RANDLES, R. H., 1975: A two-sample adaptive distribution free test. *J. Am. Statist. Ass.* **70**, 656–661.
- HOLM, S., 1979: A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6**, 65–70.
- JOHNSON, R. and BHATTACHARYA, G., 1986: *STATISTICS: Principles and Methods*, 2nd edition. John Wiley, New York.
- MARCUS, R., PERITZ, E., and GABRIEL, K. R., 1976: On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–660.
- O'BRIEN, P. C., 1992: Robust procedures for testing equality of covariances. *Biometrics* **48**, 819–827.
- PRATT, J. W., 1964: Robustness of some procedure for the two-sample location problem. *J. Am. Statist. Ass.* **59**, 665–680.
- PURI, M. L. and SEN, P. K., 1971: *Nonparametric Methods in Multivariate Analysis*. John Wiley, New York.
- RINGLAND, J. T., 1983: Robust multiple comparisons. *J. Am. Statist. Ass.* **88**, 145–151.
- ROBINSON, S. E., ENTERS, E. K., JACKSON, G. F., CHINCHILLI, V. M., McDOWELL, K. P., PASCUA, J. R., ALLEN, H. M., and GUO, H., 1993: Maternal and fetal brain and plasma levels of cocaine and benzoylecgonine following acute or chronic maternal intravenous administration of cocaine. (Submitted).
- ROCKE, D. M., DOWNS, G. W., and ROCK, A. J., 1982: Are robust estimators really necessary? *Technometrics* **24**, 96–101.
- SCHAFFER, J. P., 1986: Modified sequentially rejective multiple test procedures. *J. Am. Statist. Ass.* **81**, 826–831.
- SCHEFFÉ, H. A., 1970: *The Analysis of Variance*, 6th edition. John Wiley, New York.
- SINGH, K., 1981: On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **9**, 1187–1195.
- STEEL, R. G. D. and TORRIE, J. H., 1980: *Principles and Procedures of Statistics: A Biomedical Approach*, 2nd edition. McGraw-Hill, New York.
- WELCH, B. L., 1951: On the comparison of several mean values: an alternative approach. *Biometrika* **38**, 330–336.
- WU, C. F. J., 1986: Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **4**, 1261–1295.

A. R. PADMANABHAN
 Department of Mathematics
 Monash University
 Clayton, Victoria 3168
 Australia

Received, January 1997
 Revised, July 1998
 Revised, December 1998
 Accepted, December 1998

