UNIVERSITY OF GÖTEBORG
Department of Statistics

RESEARCH REPORT 1986:1
ISSN 0349-8034

MULTIPLE COMPARISON TESTS BASED ON THE BOOTSTRAP

by

Tommy Johnsson

Statistiska institutionen
Göteborgs Universitet
Viktoriagatan 13
S 411 25 Göteborg
Sweden

LIST OF CONTENTS

ABSTRACT

A multiple test procedure for pairwise comparisons based on the bootstrap is presented. It is a stagewise test without any distributional assumptions. It is also very general according to the number and types of hypotheses to be tested. The procedure is evaluated and to some extent compared to existing procedures. A FORTRAN computer program is available for the practical performance of the procedure suggested.

1 Introduction

The problem to be treated here is that of testing a number of hypotheses which are connected with each other. Connection means most of the times that the hypotheses are involved in the answering of one single major question. However, the relations among hypotheses could be more loose and the choice between one multiple test procedure versus many univariate tests is not always obvious. This latter question is given a brief discussion in Miller (1981) but is not to be handled further in the following. The assumption from now on is that, if a multiple test is suggested, there are good reasons for treating the hypotheses simultaneously.

The general formulation of the multiple test situation is as follows. A number of null hypotheses, $H_1, H_2, \ldots, H_n$ is to be tested against the alternatives $H_1^*, H_2^*, \ldots, H_n^*$. When deciding which hypotheses are true and which are not, there are two possible mistakes to be made. Rejecting a hypothesis which in fact is true, type I error, and accepting a hypothesis which in fact is false, type II error. Errors of type I are usually considered more serious and thus the probability of doing such an error is kept at a predetermined low level. In the multiple test case this means that the probability of rejecting any true null hypothesis should be set to a low multiple level, α, that is

$$P(\underset{i\varepsilon T}{U} \text{Reject } H_i) = \alpha \qquad (1)$$

where T is the set of indices for true null hypotheses. The lowest possible level of α is of course reached if it is decided never to reject any null hypothesis. Such a rule would on the other hand give a probability of commiting a type II error, β, that equals unity if there is some false null hypotheses. Or in other words, the probability of detecting a false null hypothesis, the power, would be zero. Thus there is a necessary trade off between α and β when establishing the rule of rejecting or accepting the hypotheses. This trade off occurs in almost every test situation and is by no means special to multiple tests. In spite of the fact that there are situations when β ought to be predetermined and controlled, the common practise of using a predetermined α is followed in this paper. This forms also a basis for comparing the performances of different tests.

2 Multiple testing

2.1 The problem

The general formulation of the multiple test situation given in the previous section contains a wide range of different problems. For the matter of simplicity just one, however rather general, problem is to be discussed here. The problem is to compare a number of groups and decide if the expected value of some variable is the same in all groups. If not, it is a part of the problem to tell which groups are differing. The null hypotheses in this case can be formulated

$$H_{0ij} : \mu_i = \mu_j \qquad i,j = 1,2,\ldots,L, i\neq j \tag{2}$$

which forms the overall null hypothesis

$$H_0 : \bigwedge_{i,j=1}^{L} H_{0ij} \quad , \quad i\neq j \tag{3}$$

or

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_L \; . \tag{4}$$

Although (3) and (4) are equivalent,(3) seems to be more consistent with the general formulation of testing M hypotheses. Here M equals $\binom{L}{2}$. According to (3) the natural formulation of the alternative hypothesis is

$$H_A : \mu_i \neq \mu_j \quad \exists i,j, i,j=1,2,\ldots,L, i\neq j \tag{5}$$

which is a whole set of different alternative hypotheses. One alternative is that all groups except one are equal and another alternative is that all groups are differing. In between those two extremes there are, unless $L \leq 3$, a number of different alternative hypotheses which the test is supposed to discriminate among. The latter, of course, provided that H_0 is rejected.

The final result of the test could be looked upon as a kind of clustering. That is, forming clusters of groups which are not possible to separate on the predetermined level of significance. When doing this one should pay some attention to the logical structure in order to avoid putting one group in two different clusters or other similar contradictions. It is obvious that some of the existing procedures for solving the multiple test problem do not take care of the logical structure.

2.2 Solutions

There are many possible ways of solving the problem described above. The procedures could be divided into different types according to some important criteria.

First of all one method, that has not been mentioned yet, could be sorted cut; the construction of multiple confidence regions. As the confidence region and the test are two branches on the same tree it is to some extent possibly to use the former instead of the latter. Some of the techniques below may also be converted to give confidence regions. The construction of confidence regions will however not be discussed in the following.

The test procedures could than be classified according to if they require any assumptions on the underlaying distribution. Many procedures are based on the normal distribution. This is an often used assumption but nevertheless it is sometimes a rather dubious one. The procedure suggested in this paper does not require any distributional assumptions at all.

The test itself could be conducted in two different ways. Either all pairs, μ_i and μ_j, $i \neq j$, are tested and concluded to be equal or different,or all groups are ranked in the order of their assumed true means. If the final result of clustering is to be reached with the first technique the direction or sign of the difference has to be stated. Otherwise the clusters C_1 containing μ_i and μ_j and C_2 containing μ_k, μ_l and μ_m could merely be stated to differ, $C_1 \neq C_2$, but not in which way, $C_1 > C_2$ or $C_1 < C_2$. This problem is discussed by Shaffer (1980), Holm (1977) and Marcus, Peritz, Gabriel (1976). When using a procedure of the second type the directional problem reduces to that of ties. This does not necessarily mean that a ranking procedure is superior to pairwise testing. Other problems, such as unknown significance levels, occur and make the ranking procedures sometimes rather dubious, Miller (1981). The test in this paper is based on pairwise comparisons with predetermined significance levels.

Especially the pairwise comparisons tests could be further divided into two subgroups. Depending on if they are performed in one single or several stages some procedures could be labelled multi-stage or stagewise tests. The principle in most stagewise tests is simple enough. The $\binom{L}{2}$ differencies are ordered in descending order and the pair that shows the largest difference are being tested first, the second largest after that and so on. The significance level in each step is adjusted to give the predetermined multiple level α. If the required significance is not met in a step, the hypothesis being tested there, as well as the following ones, are accepted, Holm (1977). The most apparent advantage of a multi-stage procedure compared to a single-stage is that the power of the test is concentrated in order to find false null hypothesis where they are likely to appear. The general result of this is higher power but it could also be used to make more precise statements. An example of the latter is the possibility of making a two-sided test containing a directional statement without any loss of neither power nor significance level, Holm (1980). The test suggested in this paper is a multi-stage one.

2.3 Classical procedures

In this section a brief discussion of some existing procedures is given. Some of them apply just to the very problem presented above, others contain it as a special case. It is pointed out whether the procedures are based on distributional assumptions, ranking or pairwise testing, multi-stage testing and in some cases if the tests in fact are converted confidence regions.

Tukey's studentized range, Miller (1980), requires normally distributed variables and also the same number of observations in all of the L groups as well as common variance. By pairwise comparisons confidence intervals are constructed for the differencies. The tests of the hypotheses $\mu_i \neq \mu_j$, $i \neq j$ are then performed simply by examining the intervals for the inclusion of zero. The utility of this method is essentially the construction of confidence regions. When it comes to testing hypotheses the method is often inferior to other procedures.

Scheffé's F projections, Miller (1980), originates from Scheffé's method for handling contrasts in an analysis of variance. The normal distribution is assumed and the differencies $\mu_i - \mu_j$, $i \neq j$, are regarded as special cases of general linear combinations. Both confidence regions and tests could be given. The procedure is rather general and for different special cases there are often better methods to be used.

Bonferroni t statistics, Miller (1980), depends solely on the simple probability inequality,

$$P(\overset{n}{\underset{i=1}{U}} A_i) \leq \sum_{i=1}^{n} P(A_i), \qquad (5)$$

which in this case gives a conservative bound for the significance level when the multiple test is made up by several univariate t-tests. If M two-sided hypotheses are to be tested simultaneously the level α/M in each test gives an overall significance level that does not exceed α. It is obvious that this procedure requires normally distributed variables, compares the groups pairwise and is not multi-stage. The method is general and very simple, the latter perhaps its greatest advantage, together with its surprisingly good power, Bohrer et al (1981).

Newman-Keuls multiple range test, Miller (1980), is a multi-stage procedure. It is performed by first testing the range of all L means, in the second stage testing the range of the (L-1) smallest and the (L-1) largest means respectively, in the third stage testing ranges of (L-2) means and so on. The difference between two means are then said to be significant provided the range of each and every subset which contains the two means is significant according to an α-level studentized range test. Although the test ends up with statements concerning pairs of means, differing or not, the results may easily be translated into a way of clustering the L groups. Consider the following example. Let $\bar{y}_i$, i=1,2,3,4,5, be the ordered sample means from five groups that should be tested along with the null hypotheses (2) against the alternatives (4). Display the means in a row and underline all combinations whose range fails to meet the significance level. The testing procedure shown in table 1 gives the following result:

$$\begin{array}{ccccc} \mu_1 & \mu_2 & \mu_3 & \mu_4 & \mu_5 \\ \cline{2-5} \\ \cline{1-2} \end{array} \quad . \tag{6}$$

Table 1: Neuman-Keuls multiple range test

Stage	Test	Significance
1	$\bar{y}_5-\bar{y}_1$	Yes
2	$\bar{y}_4-\bar{y}_1$	Yes
	$\bar{y}_5-\bar{y}_2$	Yes
3	$\bar{y}_3-\bar{y}_1$	Yes
	$\bar{y}_4-\bar{y}_2$	No; underline $\bar{y}_2$ through $\bar{y}_4$
	$\bar{y}_3-\bar{y}_1$	Yes
4	$\bar{y}_2-\bar{y}_1$	No; underline $\bar{y}_1$ through $\bar{y}_2$
	$\bar{y}_3-\bar{y}_2$	Omitted because $\bar{y}_2$ through $\bar{y}_4$ has already been underlined
	$\bar{y}_4-\bar{y}_3$	
	$\bar{y}_5-\bar{y}_4$	Yes

The conclusions to be drawn from (6) are that μ_5 differs from the other four means, μ_1 differs from μ_3, μ_4 and μ_5 and that no other differences are significant. The restrictive assumptions that has to be met when performing this test are normally distributed variables, common variance and the same number of observations in each group. A further development of this procedure is made by Begun and Gabriel (1981) and the problem of interpreting patterns like (6) is discussed by Shaffer (1981).

Duncan's multiple range test, Duncan (1955), Miller (1980) differs from Newman-Keuls only in the choice of significance levels at the different stages. Let the predetermined overall level be α and p the number of means involved in the actual stage, then the significance level, according to Duncan should be

$$\alpha_p = 1 - (1-\alpha)^{p-1} \tag{7}$$

while according to Newman-Keuls it should remain unchanged independently on the number of means, that is

$$\alpha_p = \alpha \; . \tag{8}$$

As (7) is less conservative than (8) it increases the power of the test but gives also less protection against false rejections of the null hypothesis due to the large number of declarations required. The latter is rather vital, since the major idea behind simultaneous testing is to avoid that problem. As the actual multiple significance level of this test differs from α, it can not be compared to α of other tests.

Multiple F test, Duncan (1955), Miller (1980), has the same structure as the multiple range tests above. The only differencies is that F-tests are used instead of range tests and that the number of observations in each group does not have to be the same. As with the range tests the α_p-levels can be chosen in several ways, for instance (7) or (8).

Fisher's least significant difference test, Miller (1980), has two stages. In the first stage the null hypothesis,(3), is tested by an α-level F-test. If the F-value is nonsignificant, the null hypothesis is accepted and if it is significant the next stage is performed. In the second stage all of the $\binom{L}{2}$

pairs of groups are tested by α-level t-tests and for a significant t-value the comparison is judged significant. As both the t- and the F-distributions are involved it is obvious the test requires normally distributed variables. In the sense that the test contains more than one stage it could be called a multi-stage one. The test has same good qualities. It is simple and it is based on familiar distributions. A question mark should, however, be put for the significance level. The first stage F-test protects against false rejections if the null hypothesis is true in all parts. If the F-test shows to be significant, and the test proceeds to the second stage t-tests, this protection is gone for the part, if any, of the null hypothesis that remains true. This is so because the t-tests are performed as $\binom{L}{2}$ independent tests without the extra guard of a simultaneous testing procedure. This lack of protection could be serious. Let $L=6$, $\alpha=0.05$ and assume that the F-test is significant due to just one mean, differing from the rest. That leaves $\binom{6-1}{2}=10$ comparisons that ought to be judged insignificant by the t-tests. The probability of misjudging at least one of them is however as high as

$$1 - (1-0.05)^{10} \approx 0.40 \qquad (9)$$

For $L=10$ it gets even worse, the probability of rejecting at least one true null hypothesis is then 0.84.

The k-sample rank statistics test, Miller (1980), is the non-parametric analog to the studentized range test mentioned above. Thus it does not need the assumption of an underlaying distribution such as the normal one, which is required for the studentized range test. The limitation on the number of observations is however still left, it has to be the same in all groups. This is due to the difficulties in computing critical points. The test-statistic is the maximum Wilcoxon two-sample rank statistic which for small number of groups and few observations has been tabulated. For increasing number of groups and/or observations one is depending on the limiting distribution, the multivariate normal, for calculations. When the rank test is compared to the studentized range rest it is found to be speedy, independent of normality assumptions and hence more efficient for nonnormal situations while the range test has greater efficiency when the variables really are normally, or near-normally, distributed.

The Kruskal-Wallis rank statistics test, Miller (1980) is the nonparametric rank analog to Scheffé's F projections. Compared to the previous rank test it has one great advantage as it does not require equal sample sizes. This makes the test more applicable but apart from that it is second best to the previous rank test. If it is possible to use both tests, the former one should be choosen.

The sequentially rejective method proposed by Holm (1977) is not a statistical test in itself, it is rather a procedure for administrating any test when performed in a multiple way. Consider the testing of (2) by means of the Bonferroni t statistics at the significance level α. If there are $M=\binom{L}{2}$ different pairs to be tested, the significance levels for each test should be α/M. When applying the sequentially rejective procedure on this problem the M hypotheses are ordered in descending order after the actually observed values on any test-statistic. The test-statistics are assumed to take on greater values as the true means depart from the null hypothesis. The first hypothesis, that is the one with the gratest value on the corresponding test statistic, is then tested on the α/M-level. If it is accepted the rest of the hypotheses are accepted as well. If it is rejected the procedure moves on with the testing of the second ordered hypothesis. At this stage the level is $\alpha/(M-1)$. If that one is accepted, the rest, except the first, are accepted and if it is rejected the third stage follows with the level $\alpha/(M-2)$. As long as the hypotheses are rejected the procedure goes on until the last hypothesis has been tested at the level $\alpha/1=\alpha$. This procedure is shown to have the multiple level of significance α, Holm (1977), while it is easily seen that the power is substantially increased compared to the Bonferroni procedure.

There are of course several other multiple test procedures then those mentioned here, see for instance Duncan (1955) and Miller (1980). Some of them are inferior to a test described and others are unable to handle the testing situation concerned in this paper. The reasons for not discussing them further are thereby clear.

2.4 New procedures

The theory of multiple testing has been discussed further by authers other than the already mentioned, for instance Kendall (1955) and Lehmann and Schaffer (1977). Proposals on new procedures or variates on the old ones, has been discussed, Begun and Gabriel (1981) and old procedures has been improved Miller (1980), Schaffer (1981). The main ideas remain however the same.

In the next chapter a recently developed resampling technique, the Bootstrap, is discussed,Efron (1982),and in the fourth chapter this technique will be applied to the multiple test problem earlier described.

3 The Bootstrap

3.1 The basic idea

The Bootstrap is a resampling method invented and developed by Bradley Efron. It is presented in for instance Efron (1982). The basic idea is simple. We would like to know something about a population, finite or infinite. As it is impossible to investigate the whole population we have to do the best we can with a sample from that very population. With some functions of the sample we try to estimate what is interesting in the population. When it comes to estimating we always act under some degree of uncertainty and the statistical theory is called on to provide adequate measures of accuracy. The usual question is whether the estimate would be the same during an infinite number of repeated samples or rather with how much it would vary. A measurement of variation could be received in two ways. One way is to repeat the sampling procedure a number of times and thereby observe the actual variation of the estimate. This seems to be rather stupid as the final accuracy would be substantially increased if the observations from the repeated samples were added to the original one forming one large sample, and not split the observations into a number of equaly informative estimates. The second way is to deduce the proporties of the estimate in a theoretical way. This often implies that some distributional assumptions has to be made about the population, for instance that the variable investigated is normally distributed. As long as the population really behaves according to the assumptions the theory holds but if the conditions for the theory is not quite fulfilled the resulting postulates concerning the estimates could be seriously wrong. The principle of the Bootstrap is to act as if the sample were an image of the population and by sampling with replacement from that image getting a large number of simulated new samples, so called Bootstrap-samples. By recording the estimate from each Bootstrap-sample the picture of the estimates variation emerges. One advantage of the procedure is obvious, it does not call for any distributional assumptions. On the other hand one drawback is almost as obvious, the method is depending on massive calculations that hardly could be done without the assistance of a computer. The latter is nowadays a minor problem but explains why the Bootstrap and related methods has been developed just recently. In the following it is assumed that the capacity of a computer is available whenever calculations of the type mentioned above are to be performed. The advantage of the methods being distribution-free is of greater importance. It makes it possible to apply the method to problems where theoretical properties are unknown and where the number of observations and/or the complexity makes the normal distribution unjustified. And even if the accuracy of some simple estimates could be given theoretically the analysis could, by means of the bootstrap, be extended to further aspects on the problem at hand. In order to explain the method a few examples are given below.

3.2 Estimating the variance of a sample mean

Consider a sample of size n from an unknown probability distribution F on the real line,

$$x_1, x_2, \ldots, x_n \sim F \tag{10}$$

independently and identically. From the observed values $x_1, x_2, \ldots, x_n$ the sample mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{11}$$

is computed and used as an estimate of the expected value of F. From the sample it is also possible to get an estimate of the accuracy of $\bar{x}$. This could be measured by the variance

$$V(\bar{x}) = E(\bar{x} - E(\bar{x}))^2 \tag{12}$$

which is estimated by

$$\hat{V}(\bar{x}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2 \ . \tag{13}$$

The bootstrap estimate of (12) is received in the following way. Let $\hat{F}$ be the empirical probability distribution of the data, putting the probability mass of 1/n on each x_i. Use $\hat{F}$ for drawing samples with replacement of size n. That is sampling among the observed values $x_1, x_2, \ldots, x_n$ and hence

$$x_1^*, x_2^*, \ldots x_n^* \sim \hat{F} \tag{14}$$

where x_i^* is one observation in the bootstrap sample. The bootstrap sample mean

$$\bar{x}^* = \frac{1}{n}\sum_{i=1}^{n} x_i^* \qquad (15)$$

has the variance

$$V(\bar{x}^*) = \frac{1}{n^2}\sum_{i=1}^{n} (x_i-\bar{x})^2 \qquad (16)$$

By repeating this sampling procedure say B times and each time compute the mean (15) it is possible to estimate the variance (12) without using (13). The bootstrap estimate of (12) is then

$$\hat{V}(\bar{x})_{BOOT} = \frac{1}{B-1}\sum_{J=1}^{B} (\bar{x}_j^*-\bar{\bar{x}}^*)^2 \qquad (17)$$

where $\bar{x}_j^*$ is the mean of bootstrap sample j and

$$\bar{\bar{x}}^* = \frac{1}{B}\sum_{j=1}^{B} \bar{x}_j^* \qquad (18)$$

If the number of observations n, were small then the number of possible different bootstrap samples would also be small and in that case the different bootstrap samples could be enumerated and the true value of $V(\bar{x})_{BOOT}$ computed instead of its estimate (17). This could however be done only if n is very small. As soon as n becomes large enough to be realistic for real data one is depending on the estimate (17). The error in this estimation is however not the crucial point in the method. The precision of (17) is increased with the number of Monte Carlo simulated bootstrap samples, B, and as B → ∞ the true value is obtained. Thus by making B large enough, and that is just a matter of computational time and cost, the estimation error could be held at an acceptable level. The more serious problem is that of estimating F, the probability distribution, of the underlaying process or population. When F is estimated from the sample, in the way given above, it is very difficult to say anything about the error in that estimation. Only two facts are certain. If $\hat{F}$ is an inaccurate estimate of F the method goes wrong as the simulations are performed under inadequate conditions. As with all statistical inference the accuracy of $\hat{F}$ as an estimate of F increases with the number of observations in the original sample. This latter problem deserves to be treated more extensively than what is done here.

3.3 Estimating the variance of $\hat{\theta}$

The estimation of $V(\bar{x})$ in the previous section could of course be performed without the bootstrap technique, the theoretically deduced formula for that is given in (12). The trouble with (12) is that it doesn't, in any obvious way, extend to estimators other than $\bar{x}$. So does however the bootstrap estimate (17).

Let $\hat{\theta}$ be any function of the original sample

$$\hat{\theta} = \hat{\theta}(x_1,x_2,\dots x_n) \tag{19}$$

where as before

$$x_1,x_2,\dots x_n \sim F \tag{20}$$

Estimate F with $\hat{F}$, the empirical probability distribution, draw a bootstrap sample from $\hat{F}$ and calculate

$$\hat{\Theta}^{*} = \hat{\Theta}(x_1^{*}, x_2^{*}, \ldots, x_n^{*}) \tag{21}$$

Independently repeat this B times, obtaining the replications $\hat{\Theta}_1^{*}, \hat{\Theta}_2^{*}, \ldots, \hat{\Theta}_B^{*}$ and calculate

$$\hat{V}(\hat{\Theta})_{\mathrm{BOOT}} = \frac{1}{B-1} \sum_{j=1}^{B} (\hat{\Theta}_j^{*} - \hat{\Theta}_{\bullet}^{*})^2 \tag{22}$$

where $\hat{\Theta}_{\bullet}^{*} = \frac{1}{B} \sum_{j=1}^{B} \hat{\Theta}_j^{*}$.

The general notation in (19)-(22) reveals one of the most important advantages with the bootstrap. It can be applied to complicated situations were theoretical analysis is hopeless. The $\hat{\Theta}$ above could be any statistic as, for instance, the median, a trimmed mean or a correlation coefficient.

3.4 Other applications

There are many possible applications, beside the ones given above, for the bootstrap. Efron (1982) gives several examples where the bootstrap gives results that hardly could be reached with pure theoretical analysis. One of the most important is perhaps the suggestion to use the technique for estimating bias. Other applications to be metioned are estimation of parameters in regression models and the extension to finite sample spaces. The latter makes the rationale for the bootstrap even more evident.

A slightly different application is given in Efron (1981) where the bootstrap is used to set standard errors and confidence intervals for parameters of an unknown distribution when the data is subject to right censoring. The estimates derived closely approximate the answers given by Greenwood's formula. A formula which requires much more analysis then does the bootstrap. On the other hand the latter method requires more computation.

In the next chapter the bootstrap will be applied to the multiple test problem outlined in chapter two.

4 The Bootstrap multiple test procedure

4.1 The basic idea

The Bootstrap multiple test procedure is a new application of the bootstrap technique described in the previous chapter. It could be regarded as an alternative to the test-procedures mentioned in chapter two. The basic idea is to form all possible pairwise differencies among the L means and with a number of bootstrap samples determine whether the observed differencies are likely to occur just by chance or if they imply significant distinctions between the means. The test is performed in a stagewise way in order to test the differencies in descending order, beginning with the largest. As an additional stage at the end of the procedure, the logical structure is taken into account.

4.2 The preliminary procedure

Consider the overall null-hypothesis given in chapter one,

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_L \tag{23}$$

The alternative to (23) consists of a set of different statements of which one formulation is given in (4). As indicated in (2) it is also possible to give the null-hypothesis as a conjunction of hypotheses. Doing this and at the same time connecting each null-hypotheses with its alternative gives the following:

$$\begin{array}{cc}
\underline{H_0} & \underline{H_A} \\
\mu_1=\mu_2 & \mu_1\neq\mu_2 \\
\mu_1=\mu_3 & \mu_1\neq\mu_3 \\
\vdots & \vdots \\
\mu_{L-1}=\mu_L & \mu_{L-1}\neq\mu_L \ .
\end{array} \tag{24}$$

The testing of these $\binom{L}{2}$ hypotheses does not give the complete solution. For each H_0 rejected there is a directional statement missing. As mentioned earlier it is a part of the problem to tell in what way the groups differ, if they do. The answer is given by reformulating (24) according to the principles outlined in Holm (1977), giving

$$\begin{array}{cc}
\underline{H_0} & \underline{H_A} \\
\mu_1\leq\mu_2 & \mu_1>\mu_2 \\
\mu_2\leq\mu_1 & \mu_2>\mu_1 \\
\mu_1\leq\mu_3 & \mu_1>\mu_3 \\
\mu_3\leq\mu_1 & \mu_3>\mu_1 \\
\vdots & \vdots \\
\mu_{L-1}\leq\mu_L & \mu_{L-1}>\mu_L \\
\mu_L\leq\mu_{L-1} & \mu_L>\mu_{L-1}
\end{array} \tag{25}$$

It should be noted that (25) contains twice as many hypotheses as does (24). For each $\mu_i\geq\mu_j$ there is a $\mu_j\geq\mu_i$ under H_0. Unless $\alpha>0.5$ these two hypotheses could however not be rejected at the same time.

The basis for the inference are L randomly selected samples of sizes n_1, $n_2,\ldots,n_L$ from the probability distributions or populations, finite or infinite, having the expected values or true means μ_1, μ_2, ..., μ_L. Let the samples form the estimates $\bar{y}_1$, $\bar{y}_2$, ..., $\bar{y}_L$ for μ_1, μ_2, ..., μ_L and define the observed differencies, $d_{i,j}$, as

$$d_{i,j} = \bar{y}_i - \bar{y}_j \tag{26}$$

which are estimates of the true differencies

$$D_{i,j} = \mu_i - \mu_j \tag{27}$$

The observed differencies are now to be arranged in descending order, starting with the largest positive value. Denote the largest d with d^1, the second largest with d^2 and so on until the smallest of the $L(L-1) = k$ differencies which has to be d^k-d^1 and let I_1, J_1 be the indices of d^1, I_2, J_2 the indices of d^2 and so on until the last pair, being the first indices in opposite order.

The hypotheses in (25) could now be put in the same order as the observed differencies, which along with the order index k, gives

$$\begin{array}{ccc}
\underline{k} & \underline{H_0^k} & \underline{H_A^k} \\
1 & \mu_{I_1} \leq \mu_{J_1} & \mu_{I_1} > \mu_{J_1} \\
2 & \mu_{I_2} \leq \mu_{J_2} & \mu_{I_2} > \mu_{J_2} \\
\cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot \\
k & \mu_{I_k} \leq \mu_{J_k} & \mu_{I_k} > \mu_{J_k} \; ,
\end{array} \tag{28}$$

where $I_k=J_1$, $I_{k-1}=J_2$, $J_k=I_1$, $J_{k-1}=I_2$ etc. From (28) it is obvious that the second half of the hypotheses is just a mirror image of the first half. It is also obvious that it is the hypotheses on the first half that one is out to reject, the rest is just serving as a formal complement making it possible to make the desired directional statements.

The hypotheses in (28) is now to be tested in the following sequentially rejective manner, suggested by Holm (1977):

$$\begin{array}{l} \text{Test } H_0^1 \\ \text{If accepted, accept } H_0^i; i\geq 1 \\ \text{If rejected, test } H_0^2 \\ \quad\vdots \\ \text{Test } H_0^k \\ \text{If accepted, accept } H_0^i; i\geq k \\ \text{If rejected, test } H_0^{k+1} \\ \quad\vdots \\ \text{Test } H_0^K \end{array} \qquad (29)$$

The decision of accepting or rejecting in each stage of (29) is made by means of the bootstrap technique.

Let F_i be the probability distribution or population with mean μ_i and let $\hat{F}_i$ be the empirical probability distribution of the i:th sample with zero mean. That is, before $\hat{F}_i$ is computed by putting the probability mass of $1/n_i$ on each observation, the sample mean $\bar{y}_i$ is subtracted and thus giving the expected value zero of $\hat{F}_i$. This point is crucial for the following moments as we are now dealing with L distributions, $\hat{F}_i$, having the same mean. That is exactly what the original null-hypothesis (23) is saying and the general theory of tests is telling us to act as if the null-hypothesis were true until we have evidence enough to reject it. Acting like that makes it also possible to preassign and control the significance level, α.

Use the $\hat{F}_i$:s to draw L bootstrap samples of sizes n_1, n_2, ..., n_L giving the bootstrap sample means $\hat{\bar{y}}_1^*$, $\bar{y}_2^*$, ..., $\bar{y}_L^*$. Note that the expected value of each sample mean is zero. Compute the bootstrap differencies, d_{ij}^*, as

$$d_{ij}^* = \bar{y}_i^* - \bar{y}_j^* \qquad (30)$$

and put them in the same order as (28), $d_{I_1J_1}^*$, $d_{I_2J_2}^*$, ..., $d_{I_kJ_k}^*$. This does not necessarily mean that the k bootstrap differencies themselves are put in descending order, they are just arranged according to (28) and hence according to the differencies, d_{ij}^k, in the real sample. For each sample difference, d_{ij}^k, it is now recorded whether any of the bootstrap differencies, $d_{I_LJ_L}^*$, $L\geq k$, is greater than or equal to d_{ij}^k. If this happens, it indicates that the observed sample difference could have appeared by pure chance and thus is giving no evidence against the null-hypothesis, H_0^k. The bootstrap samples are drawn from distributions with the same mean, zero, and hence any $d_{ij}^*\neq 0$ is purely random. Comparing the bootstrap differencies with a sample difference is then indicating whether the observed sample differencies is just a random deviation likely to occur under the null-hypothesis. The following numerical example, Example 4.1, shows the procedure step by step. For simplicity, just three groups are being tested.

The original overall null-hypothesis is

$$\mu_1 = \mu_2 = \mu_3 \tag{31}$$

Data consists of three samples of sizes $n_1=10$, $n_2=20$, $n_3=15$ giving the sample means $\bar{y}_1=1$, $\bar{y}_2=2$, $\bar{y}_3=5$ standard deviations $s_1=3.3$, $s_2=2.2$, $s_3=30$. However convenient it is not necessary to arrange the sample means in any order. Computing the sample differencies and putting them in descending order gives

$$\begin{aligned}
d^1_{3,1} &= \bar{y}_3 - \bar{y}_1 = 5 - 1 = 4\\
d^2_{3,2} &= \bar{y}_3 - \bar{y}_2 = 5 - 2 = 3\\
d^3_{2,1} &= \bar{y}_2 - \bar{y}_1 = 2 - 1 = 1\\
d^4_{1,2} &= \bar{y}_1 - \bar{y}_2 = 1 - 2 = -1\\
d^5_{2,3} &= \bar{y}_2 - \bar{y}_3 = 2 - 5 = -3\\
d^6_{1,3} &= \bar{y}_1 - \bar{y}_3 = 5 - 1 = -4
\end{aligned} \tag{32}$$

Formulating the null-hypotheses along with the alternatives according to (28) now gives,

$$\begin{array}{cc} \underline{H_0} & \underline{H_A} \\ \mu_3 \le \mu_1 & \mu_3 > \mu_1 \\ \mu_3 \le \mu_2 & \mu_3 > \mu_2 \\ \mu_2 \le \mu_1 & \mu_2 > \mu_1 \\ \mu_1 \le \mu_2 & \mu_1 > \mu_2 \\ \mu_2 \le \mu_3 & \mu_2 > \mu_3 \\ \mu_1 \le \mu_3 & \mu_1 > \mu_3 \end{array} \tag{33}$$

Let us now assume that the bootstrap samples, drawn with replacement from the real samples transformed to zero means, produce the bootstrap means $\bar{y}^*_1=0$, $\bar{y}^*_2=-1$, $\bar{y}^*_3=1$. Computing the bootstrap differencies and putting them in the same order as the sample differencies (32) gives

$$\begin{array}{l} d^*_{3,1} = \bar{y}^*_3 - \bar{y}^*_1 = 1 - 0 = 1 \\ d^*_{3,2} = \bar{y}^*_3 - \bar{y}^*_2 = 1 - (-1) = 2 \\ d^*_{2,1} = \bar{y}^*_2 - \bar{y}^*_1 = (-1) - 0 = -1 \\ d^*_{1,2} = \bar{y}^*_1 - \bar{y}^*_2 = 0 - (-1) = 1 \\ d^*_{2,3} = \bar{y}^*_2 - \bar{y}^*_3 = (-1) - 1 = -2 \\ d^*_{1,3} = \bar{y}^*_1 - \bar{y}^*_3 = 0 - 1 = -1 \end{array} \tag{34}$$

Recording for each sample difference whether $d^*_{I_L,J_L} \ge d^k_{i,j}$, $L \ge k$, gives

Table 2 : The outcome of one bootstrap sample,
example 4.1

k	H_0^k	d^k	$d^*\geq d^k$
1	$\mu_3\leq\mu_1$	4	No
2	$\mu_3\leq\mu_2$	3	No
3	$\mu_2\leq\mu_1$	1	Yes, since $d^*_{1,2} = 1\geq 1 = d^3$
4	$\mu_1\leq\mu_2$	-1	Yes, since $d^*_{1,2}\geq d^4$
5	$\mu_2\leq\mu_3$	-3	Yes, since $d^*_{2,3}\geq d^5$
6	$\mu_1\leq\mu_3$	-4	Yes $d^*_{1,3}\geq d^6$

which in this case indicates that the sample differencies 3 and 4 did not occur just by chance in the bootstrap samples while the differencies 1, -1, -3 and -4 did. The condition $L\geq k$ above should perhaps be given a second thought. This condition is a consequence of the multi-stage natur of the test procedure. The null-hypotheses, H_0^1, H_0^2, ..., H_0^k, are tested one by one in descending order and the condition for testing H_0^k is that all preceding hypotheses, H_0^1, H_0^2, ..., H_0^{k-1} are being rejected. As they have been rejected, and thus stated to be false, any random deviation emerging from the corresponding bootstrap differencies are of no interest. The means are assumed to differ and doing so the corresponding null-hypotheses are no longer part of the hypotheses to be tested. This point is perhaps more obvious after the next step in the procedure.

Obviously the results in table 2 are not enough to accept or reject any hypothesis. Inference based on one single bootstrap indication,as how to act,would be similar to use just one observation for estimating a population parameter. In the latter case one needs several observations and for the problem at hand the answer is several bootstrap indications received from repeated drawings of bootstrap samples. For each new set of bootstrap samples of sizes n_1, n_2, ..., n_L the bootstrap differencies are being computed and compared to the observed sample differencies. The same recordings as those described for the first set of bootstrap samples, are made for each replication. When, say, B replications are made, there are, for each sample difference, B indications of whether that difference is likely to occur just by chance or not. The predetermined level of significance, α, is now used to decide if the null-hypothesis is to be rejected or accepted. Let B_A^k be the number of times when

$$d^*_{I_L,J_L} \geq d^k_{i,j} \quad , \quad L \geq k \qquad (35)$$

and let B_R^k=B-B_A^k. That is, the bootstrap samples indicate B_A^k times out of B, that the observed difference, $d^k_{i,j}$, has ocurred by pure chance. Such an indication speaks for accepting H_0^k. As the level of significance is the predetermined, maximum, probability of wrongly rejecting the null-hypothesis, is it obvious that H_0^k should be rejected if and only if

$$\frac{B_A^k}{B} \leq \alpha \quad . \qquad (36)$$

The comparisons of (36) are made stagewise according to (29) and thus resulting in the rejection of a number of null-hypotheses in the beginning of the ordered sequence of (28). The number of rejected null-hypotheses being anything from zero to K/2.

Returning to the numerical example above this means that a large number of bootstrap samples should be drawn. Let us assume that the number of replications, B, equals 1000. This is enough to show the necessity of a computer for using the bootstrap technique. For each of the 1000 replications the bootstrap differencies are being computed according to (32) and ordered according to (34). Table 2 has to be reworked as the number of times when the condition (35) is fulfilled, B_A^k, now has to be shown. The table below is one possible outcome of the 1000 bootstrap replications.

Table 3: Test based on 1000 bootstrap samples, example 4.1

k	H_0^k	d^k	B_A^k	B_A^k/B
1	$\mu_3 \leq \mu_1$	4	1	0.001
2	$\mu_3 \leq \mu_2$	3	12	0.012
3	$\mu_2 \leq \mu_1$	1	443	0.443
4	$\mu_1 \leq \mu_2$	-1	992	0.992
5	$\mu_2 \leq \mu_3$	-3	1000	1.000
6	$\mu_1 \leq \mu_3$	-4	1000	1.000

The number of null-hypotheses to be rejected according to the results of table 3 depends on the level of significance. For

$\alpha=0.05$ the two null-hypotheses $\mu_3 \leq \mu_1$ and $\mu_3 \leq \mu_2$ are rejected while their alternatives and the remaining null-hypotheses are accepted. For $\alpha=0.01$ just the first null-hypothesis is rejected. It is also possible to regard the ratios B_A^k/B as P-values or observed significance levels.

4.3 Logical structure

When table 3 is completed and evaluated it is possible to end the test procedure. A final step using the logical structure could however be added. By taking into account the logical structure the power of the test is increased without effecting the level of significance. The idea is to work with possible clusterings of the means being tested. If no information is given, as significant differencies between any two means, there are several possible patterns the clustering can follow. For simplicity regard the three means in the example above. They could be clustered in one of the five ways given in table 4.

Table 4: Possible patterns of three means, example 4.1

Nr	Pattern	Denoted
1	$\mu_1=\mu_2=\mu_3$	123
2	$\mu_1=\mu_2\neq\mu_3$	12-3
3	$\mu_1\neq\mu_2=\mu_3$	1-23
4	$\mu_1=\mu_3\neq\mu_2$	13-2
5	$\mu_1\neq\mu_2\neq\mu_3\Lambda\mu_1\neq\mu_3$	1-2-3

It is to be noted that in reality the distributions or populations generating the samples are clustered in one and only one of the five ways listed in table 4. The trouble is that we do not know which one. When the first null-hypothesis is tested, no prior information is given and hence any of the patterns in table 4 is possible. If the first null-hypothesis is rejected the second is tested. When performing that test we have already stated that H_0^1 is false, and thus $\mu_3>\mu_1$. From this it follows that the patterns 123 and 13-2 are not feasible as those patterns have the means μ_1 and μ_3 in the same cluster. When testing H_0^2 there are only three possible patterns, 2, 3 and 5 and the test procedure could be performed conditioned on one of them. As it is impossible to say a priori which one of the patterns that gives the largest B_A^2, the bootstrap testing procedure has to be performed once for each possible pattern, in this case three times. The final decision whether to reject H_0^2 or not must namely be made for the largest of the possibly different P-values appearing in the three performancies of the test. Otherwise the predetermined significance level is violated and the protection against the rejection of true null-hypotheses abandoned. For each of the three possible patterns the testing now proceeds assuming the actual pattern to be true. This assumption is similar to the protective one, acting as the null-hypothesis were true until it is rejected, made in almost every test. When testing the second hypothesis, H_0^2, under the condition of pattern number two, see table 4, it is assumed that the third mean differs from the other two. In the first stage it has already been shown that μ_1 and μ_3 differs and this leads to the conclusion that the means are divided into at least two different clusters. One possibility is then pattern number two. From table 3 it follows that 12 of the 1000 bootstrap samples gave the indications that the difference $d_{3,2}$ in the real sample did occur by pure chance. We must now find out which bootstrap differencies, d^*_{ij}, who turned out to be greater than or equal to $d_{3,2}$. The difference $d^*_{3,1}$ has already been excluded and now the differencies $d^*_{1,3}$, $d^*_{2,3}$ and $d^*_{3,2}$ have to be excluded as well. All of them excluded because the appearance of large values among them is just showing what has already been proved or assumed and does not contribute anything to the answering of the question at hand, namely whether H^2_0 is true or false. It is obvious that the number of indications after the exclusions is less than or equal to the number before.

This part of the procedure is repeated for each possible pattern and, as a protection against type I errors, the largest number of indications and hence the largest P-value is assigned to the solution. For the last pattern, 1-2-3, it has to be zero, and thus uninteresting, and for one of the patterns it perhaps equals the previous value. If the latter happens the clustering is not worth-while, if it does not,the additional stage improves the power of the test. One possible outcome of the testing of H^2_0 based on clustering is shown in table 5, from which it follows

Table 5: Test of H_0^2 under clustering assumptions, example 4.1

Pattern	B_A^2	B_A^2/B
$\mu_1=\mu_2\neq\mu_3$	4	0.004
$\mu_1\neq\mu_2=\mu_3$	0	0
$\mu_1\neq\mu_2\neq\mu_3\Lambda\mu_1\neq\mu_3$	0	0

max $B_A^2/B=0.004$

that the P-value has decreased from 1,2% to 0.4% by the clustering. Of course this is not always the case, but the possibility of any improvement makes the clustering stage worth trying.

When performing the clustering stage above one important point is not to violate the significance level. In order to see that the significance level is kept during this stage the following reasoning could be put forth.

The different patterns that are possible when testing a certain hypothesis could be regarded as an exhaustive and disjunctive partitioning of the parameter space. That is, each pattern represents a combination of true and false null-hypotheses, where one, and only one, combination is true. The trouble is, however, not knowing which combination being true. In spite of this lack of knowledge, assume that the null-hypotheses, e.g. from table 3, being true belong to the set

M* while the false ones do not. That is, i,j $\in$ M* means that the null-hypothesis $\mu_i \leq \mu_j$ is true. The possibility of committing a type I error could then be formulated as

$$P(\max_{i,j \in M^*} (d_{i,j}) > \gamma^*) = \alpha \qquad (37)$$

where γ^* is the critical value given by the predetermined level of significance, α, if the test statistic $\max_{i,j\in M^*} d_{i,j}$ is used for testing the null-hypotheses $\mu_i \leq \mu_j$ for all i,j $\in$ M*.

During the preliminary stage of the procedure, the search for significant differencies is performed on the whole parameter space, except among the null-hypotheses already rejected. Let M^k be the set of null-hypotheses not rejected before the preliminary stage k. If a type I error has not already been committed, then $M^* \subseteq M^k$ and the level of significance is $\leq \alpha$ since

$$M^* \subseteq M^k \quad \Rightarrow \quad \gamma^* \leq \gamma^k \quad . \qquad (38)$$

When it comes to the clustering stage, there is one set, M^c, for each pattern c, c=1,2,...,C, where C is the number of possible patterns. At least one of the M^c:s, say M^{c*}, must contain the true combination M*, that is $M^* \leq M^{c*}$, and then it follows from (38) that $\gamma^* \leq \gamma^{c*}$. If it were known which M^c to call M^{c*}, the test procedure could be performed for that pattern only; since it is not, the procedure has to be performed for every possible pattern. γ^{c*} is then one of the γ^c:s received and since $\gamma^{c*} \leq \max \gamma^{c}$, $\max \gamma^{c}$ serves as an upper bound for γ^{c*}. Taking $\max \gamma^{c} = \gamma^{c*} \geq \gamma^{*}$ gives, at least, the predetermined level of significance, since this means that the single null-hypothesis being tested, if it is rejected at all, would be so regardless of which pattern is containing the true combination M*.

In terms of P-values, the value for the pattern containing M* should be choosen. As this is unknown the largest P-value is selected as being an upper bound for the true value. If the single null-hypothesis could, at a predetermined level of significance, be rejected for any M^{c}, then it could certainly be rejected for $M^{*} \subseteq M^{c*}$.

The reasoning above holds for any true probability distributions generating the samples. In the Bootstrap procedure these distributions, and thus the corresponding γ:s and P-values, are estimated by simulations.

4.4 The final procedure

The combination of the preliminary procedure of section 3.2 and the logical structure inclusion of section 3.3 could be done in at least two ways. One way is to run the clustering procedure at all stages. The advantage of this would be to attain the lowest possible P-value for each null-hypothesis being tested. Doing so would however also cause unnecessary calculations as some P-values in the beginning of the test procedure, that is the most obvious rejections, are low enough without the clustering. This argument is negligible if the cost of computation is zero. If the cost and time of computation has to be taken into account, another combination is perhaps preferable. One suggestion is to run the test according to the preliminary procedure until the predetermined significance level is reached at one stage. For the null-hypotheses at the preceding stages there is no need for any clustering as they have been rejected anyway. At the first stage where the significance level, α, has been reached, the test procedure would stop and no further rejections were to be made. At this stage the clustering is introduced. If the clustering cause the null-hypothesis at this stage to be rejected the test is continued at the next stage, if not, the procedure ends. Once the clustering has been introduced it is performed at all the following stages, as long as the null-hypotheses are rejected. If the preliminary procedure is unable to reject at stage i, it is certainly unable at stage j, j>i.

One improper use of the clustering stage has to be mentioned as well. When using the clustering it is possible to get a P-value < α at a stage where one or even many of the preceding null-hypotheses has not been rejected. If this happens one has to remember that a condition for the rejection of H_0^i is that H_0^j , j<i, already has been rejected. See also Cox and Spjøtvoll (1982) where this condition is neglected.

In the examples of the following section the latter of the two possible combinations of main and clustering procedures is used. Although the computer program used is more general, this is the test procedure finally suggested.

5 Examples and evaluation

5.1 Comparisons between methods

In this chapter a few examples are given in order to see how the previously outlined method really works. Along with the solutions provided by the bootstrap procedure, there are results given according to other methods. The examples are therefore choosen to fit even other procedures then the one suggested in this paper. However, it is to be remembered that almost every other method is subject to restrictions and assumptions of which the bootstrap procedure is perfectly unaware. This makes it possible to line up a number of problems where the bootstrap method is the only alternative and thus the outstanding one. Such problems would however be uninteresting from a comparing point of view. The examples to follow are taken from papers concerned with other methods of multiple testing. This means that a comparison is possible, at least to one alternative method.

5.2 The bootstrap versus the Newman-Keuls procedure

The two examples given below serve both as an illustration of the bootstrap method and as a comparison between this procedure and the one known as the Newman-Keuls procedure.

Example 5.1 is taken from Miller (1980). It consists of 5 groups with 5 observations per group. As the actual observations are not given, 25 values are simulated following the means and standard deviations from Miller (1980) and assumed to be normally distributed. The complete data can be found in appendix B.

The original overall null-hypothesis is as usual

$$\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 \quad , \qquad (39)$$

with the alternative

$$\mu_i \neq \mu_i \quad \exists \quad i \neq j \qquad (40)$$

For the bootstrap procedure the hypotheses are reformulated according to (25). The result from the running of the main procedure with B=1000 replications is given in table 6. If the significance level were α=5%, the clustering procedure ought to be performed for H_0^8 and onwards but let us assume that we are interested in pressing the P-value downwards even for H_0^7 and start the clustering with that instead. The results from the clustering are given in tables 7-9. Before turning to the results of the clustering one remark should be made according to the describing of patterns. Since the μ_i:s are already ordered in descending order, the meaning of for instance $\mu_1 \neq \mu_2 \neq \mu_3$ is that all μ_i are differing even μ_1 and μ_3 though that is not explicitly said. Accordingly the sixth pattern of table 7 means that $\mu_1 = \mu_3$, $\mu_1 \neq \mu_4$, $\mu_1 \neq \mu_5$ and $\mu_2 \neq \mu_5$, the eighth pattern that $\mu_i \neq \mu_j$, $\forall_{i,j,i \neq j}$ etc.

Table 6: Example 5.1 solved with the boostrap method main procedure B=1000

k	H_0^k	P-value=$B_A^k/B\cdot 100$
1	$\mu_5\leq\mu_1$	0
2	$\mu_5\leq\mu_2$	0
3	$\mu_3\leq\mu_1$	0
4	$\mu_5\leq\mu_4$	0
5	$\mu_4\leq\mu_1$	0.8
6	$\mu_5\leq\mu_3$	1.6
7	$\mu_3\leq\mu_2$	4.5
8	$\mu_2\leq\mu_1$	7.6
9	$\mu_3\leq\mu_4$	27.3
10	$\mu_4\leq\mu_2$	57.1
11	$\mu_2\leq\mu_4$	100.0
12	$\mu_4\leq\mu_3$	100.0
13	$\mu_1\leq\mu_2$	100.0
14	$\mu_2\leq\mu_3$	100.0
15	$\mu_3\leq\mu_5$	100.0
16	$\mu_1\leq\mu_4$	100.0
17	$\mu_4\leq\mu_5$	100.0
18	$\mu_1\leq\mu_3$	100.0
19	$\mu_2\leq\mu_5$	100.0
20	$\mu_1\leq\mu_5$	100.0

Table 7: Example 5.1, testing H_0^7 $\mu_3 \leq \mu_2$ with clustering

Pattern	P-value=$B_A^7/B \cdot 100$
$\mu_1=\mu_2\neq\mu_3=\mu_4\neq\mu_5$	0.3
$\mu_1=\mu_2=\mu_3\neq\mu_4\neq\mu_5$	0.3
$\mu_1\neq\mu_2=\mu_3=\mu_4=\mu_5$	1.5 = max $(B_A^7/B \cdot 100)$
$\mu_1\neq\mu_2=\mu_3\neq\mu_4\neq\mu_5$	1.4
$\mu_1\neq\mu_2=\mu_4\neq\mu_3\neq\mu_5$	0.5
$\mu_1\neq\mu_2\neq\mu_3=\mu_4\neq\mu_5$	0
$\mu_1\neq\mu_2\neq\mu_3\neq\mu_4\neq\mu_5$	0

Table 8: Example 5.1, testing H_0^8 : $\mu_2 \leq \mu_1$ with clustering

Pattern	P-value=$B_A^8/B \cdot 100$
$\mu_1=\mu_2\neq\mu_3=\mu_4\neq\mu_5$	1.2
$\mu_1=\mu_2\neq\mu_2\neq\mu_3\neq\mu_5$	1.2
$\mu_1\neq\mu_2=\mu_4\neq\mu_3\neq\mu_5$	1.4 = max $(B_A^8/B \cdot 100)$
$\mu_1\neq\mu_2\neq\mu_3=\mu_4\neq\mu_5$	0
$\mu_1\neq\mu_2\neq\mu_3\neq\mu_4\neq\mu_5$	0

Table 9: Example 5.1, testing $H_0^9 : \mu_3 \leq \mu_4$ with clustering

Pattern	P-value$=B_A^9/B\cdot 100$
$\mu_1 \neq \mu_2 = \mu_4 \neq \mu_3 = \mu_5$	$10.9 = \max\ (B_A^9/B\cdot 100)$
$\mu_1 \neq \mu_2 \neq \mu_3 = \mu_4 \neq \mu_5$	0.2
$\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5$	0

As the P-value for H_0^9 is as high as 10.9% the procedure is not continued. For $1.6 \leq \alpha \leq 10.9$ the results could be interpreted as follows. The samples put the means in the order

$$\mu_1 < \mu_2 < \mu_4 < \mu_3 < \mu_5 \qquad (41)$$

where the significance level is met for all comparisons except for $\mu_2 - \mu_4$ and $\mu_4 - \mu_3$ which are not significantly differing. The results could also be given according to (6), the underlining technique mentioned in section 2.3,

$$\begin{array}{ccccc} \mu_1 & \mu_2 & \mu_4 & \mu_3 & \mu_5 \\ \cline{3-4} & & & & \\ \cline{2-3} \end{array} \qquad (42)$$

where two means underlined by the same unbroken line are not significantly differing, while the others are.

The Neuman-Keuls procedure, Miller (1980), is now used to test the very same null-hypothesis. It is to be noted the figures are not the same as in Miller (1980), due to the random simulation, but of course the same as above. The test procedure can be followed in table 10. The critical values for the test could be found in Miller (1980).

Table 10: Example 5.1, solved by the Newman-Keuls procedure $\alpha=5\%$

Stage	Test	Significance
1	26.4-14.6	Yes
2	21.9-14.6	Yes
	26.4-18.0	Yes
3	19.5-14.6	Yes
	21.9-18.0	Yes
	26.4-19.5	Yes
2	18.0-14.6	No, underline μ_1 to μ_2
	19.5-18.0	No, underline μ_2 to μ_4
	21.9-19.5	No, underline μ_3 to μ_4
	26.4-21.9	Yes

With the underlining technique table 10 gives

$$\begin{array}{ccccc} \mu_1 & \mu_2 & \mu_4 & \mu_3 & \mu_5 \\ \cline{1-2} \cline{3-4} & & & & \\ \cline{2-3} \end{array} \qquad (43)$$

which is to be compared to (42). The only difference between the two methods is that the bootstrap procedures succeeded in rejecting the null-hypothesis $\mu_2 \leq \mu_1$ while the Newman-Keuls procedure did not. Whether this difference is important or not could not be postulated at this stage but nevertheless are the results an indication of the bootstrap procedure being more powerful.

Example 5.2 is taken from Hartley (1955). The problem here is to compare 6 means on the basis of 5 observations from each group. As in the previous example the variates are simulated according to a normal distribution. The complete data can be found in appendix C while the hypotheses are stated as before. When the bootstrap procedure, with B=500 replications, where applied the results of table 11 was achieved.

Table 11: Example 5.2, solved with the bootstrap method, main procedure, not all null-hypotheses being listed B=500

k	H_0^k	P-value=$B_A^k/B\cdot 100$
1	$\mu_6 \leq \mu_1$	0
2	$\mu_6 \leq \mu_3$	0.2
3	$\mu_5 \leq \mu_1$	0.8
4	$\mu_6 \leq \mu_2$	1.4
5	$\mu_4 \leq \mu_1$	1.4
6	$\mu_5 \leq \mu_3$	5.6
7	$\mu_6 \leq \mu_4$	11.8
8	$\mu_4 \leq \mu_3$	14.6
9	$\mu_2 \leq \mu_1$	16.0
⋮	⋮	⋮

Table 11 indicates that the clustering procedure ought to be run for H_0^6. As this is done, the P-value decreases to 5.0% under the pattern shown in table 12.

Table 12: Example 5.2, testing $H_0^6 : \mu_5 \leq \mu_3$ with clustering

Pattern	P-value$=B_A^6/B \cdot 100$
$\mu_1=\mu_2=\mu_3\neq\mu_4\neq\mu_5=\mu_6$	
$\vdots$	
$\mu_1\neq\mu_2=\mu_3=\mu_4=\mu_5\neq\mu_6$	$5.0 = \max\ (B_A^6/B \cdot 100)$
$\vdots$	
$\mu_1\neq\mu_2\neq\mu_3\neq\mu_4\neq\mu_5\neq\mu_6$	

The testing of H_0^7 gave a P-values of 10%. If it is agreed to reject H_0^6, the combined result of tables 11 and 12 could be given in the following terms. The samples imply the order

$$\mu_1 < \mu_3 < \mu_2 < \mu_4 < \mu_5 < \mu_6 \qquad (44)$$

among the means. μ_1 differs significantly from μ_4, μ_5, and μ_6 but from μ_3 and μ_2. μ_2 differs significantly only from μ_5 and μ_6 and so does μ_3. μ_4 differs significantly only from μ_1. μ_5 differs significantly from μ_1 and μ_3 while μ_6 does it from μ_1, μ_2 and μ_3. With the underlining technique this becomes

$$\begin{array}{cccccc} \mu_1 & \mu_3 & \mu_2 & \mu_4 & \mu_5 & \mu_6 \\ \cline{2-4} \\ \cline{1-3} \\ \cline{3-6} \end{array} . \qquad (45)$$

Table 13 gives the result when the Newman-Keuls procedure is applied to the same data. Again it should be noted that the actual figures are not the same as in previous paper, Hartley (1955).

Table 13: Example 5.2, solved by the Newman-Keuls procedure $\alpha=5\%$

Stage	Test	Significance
1	1609-1442	Yes
2	1554-1442	Yes
	1609-1472	Yes
3	1539-1442	Yes
	1554-1472	No, underline μ_3 to μ_5
	1609-1506	Yes
4	1506-1442	No, underline μ_1 to μ_2
	1539-1472	Already underlined
	1554-1506	Already underlined
	1609-1539	No, underline μ_4 to μ_6
5	No groups of 2 means could be significant since no groups of 3 means are.	

Drawing the lines suggested in table 13 gives

$$\begin{array}{cccccc} \mu_1 & \mu_3 & \mu_2 & \mu_4 & \mu_5 & \mu_6 \\ \cline{2-5} \\ \cline{1-3} \cline{4-6} \end{array} \qquad (46)$$

which is almost the same as (45). The only difference is that the bootstrap procedure pointed out μ_3 to differ significantly from μ_5 while the Newman-Keuls procedure did not. On the other hand the latter method shows a significant difference between μ_2 and μ_6 while the former technique state them to be not significantly different.

5.3 Evaluating the significance level

When the bootstrap procedure was compared to the Newman-Keuls method it was found that the former possibly had higher power. This means that the probability of committing a type-II error, β, is lower. For the same test procedure and sample size β could only be decreased by increasing α, the probability of committing a type-I error. Therefore, in this section some attention will be paid to wheather the significance level really keep the predetermined value, in this case 5%, or not. Such an investigation is however rather heavy to put through and that is the main reason for giving just two simple examples on this subject.

Consider L populations having the same mean and variance. Draw from each of them a sample of size n and compare the sample means. Since the populations are identical, a test that rejects any part of the null-hypothesis, $\mu_1=\mu_2=\ldots=\mu_L$, is committing a type-I error. By repeating the procedure over and over again the significance level is estimated.

In example 5.3 the variables are uniformly distributed, $x_i \sim R(0,100)$, L=3 and n=10. The results of 20 repeated bootstrap tests are given in table 14, from which it follows

Table 14: Example 5.3

Sample no	P-values for the largest differencies
1	13.3
2	80.5
3	28.0
4	19.4
5	83.4
6	19.4
7	16.7
8	60.6
9	73.4
10	13.8
11	62.4
12	94.1
13	26.5
14	50.3
15	36.6
16	24.0
17	16.8
18	76.7
19	71.3
20	47.7

that the procedure in this case does not commit any type-I error at all. In the long run the significance level is of course not zero but the result indicates that α is perhaps as low as it is meant to be, namely 5%.

In example 5.4 the variables are normally distributed. $x_1 \sim N(0,1)$, L=3 and n=5. The results of 20 repeated bootstrap tests are given in table 15. From this it follows

Table 15: Example 5.4

Sample no	P-value for the largest differencies	
1	97.8	
2	21.1	
3	1.8	18.5
4	69.4	
5	73.7	
6	5.6	
7	34.4	
8	50.1	
9	5.5	
10	32.7	
11	34.1	
12	23.7	
13	55.1	
14	35.9	
15	44.4	
16	4.7	17.3
17	79.1	
18	90.2	
19	35.0	
20	39.9	

that the procedure in this case wrongly rejects at least one true null-hypothesis two times out of 20, if α is predetermined to equal 5%. Estimating the real significance level from the 20 tests gives $\hat{\alpha}=2120=10\%$. This is however not as serious as it looks. Some calculations on the binomial distribution give that $P(X\geq 2|X\sim \text{Binom}(20, 0.05)=0.26$, which in terms of testing says that the deviation from $\alpha=5\%$ is in no way significant, α is not shown to be greater than 5%.

The four examples given in this chapter are mainly pointing in the same direction. The bootstrap multiple test procedure seems to have a little bit higher power than does the Newman-Keuls method but is still maintaining the same significance level. To some extent the probability of committing an error of type-I is lower with the bootstrap method. This is due to the methods taking care of the direction or sign of the differencies. The Newman-Keuls method is based on two-tailed tests and the protection against directional errors is therefore rather weak.

6 An application

The method outlined in chapter four has been applied to a real problem. A researcher in micro biology had developed a method of classifying observations into one of four groups. The real values, which are rather difficult to obtain, where then measured. The problem was now to investigate whether there were significant differencies between the true means of the four groups. Assuming that the observations are normally distributed with common variance, this problem could be solved by analysis of variance. This would however only give indications on the existence of differencies without showing where they are. Since the latter is also wanted, a multiple comparison test is called for. The results, completed with some descriptive statistics of the Bootstrap multiple test is given below.

Let x_{ij} be observation number j, j=1,2,..., n_i, classified to group number i, i=1,2,3,4. The number of observations, sample means and standard deviations are given in table 16.

Table 16: Descriptive statistics from the samples

i	n_i	$\bar{x}_i$	s_i
1	11	658.909	733.718
2	24	1634.083	2215.504
3	24	4098.750	3258.095
4	14	10256.350	3860.779

If the true means of the four groups are denoted μ_i, i=1,2,3,4, then the overall null-hypothesis to be tested is

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \tag{47}$$

with the alternative, H_A, that some pairs, at least one, are differing. The results of the preliminary stage of the Bootstrap test is given in table 17. The number of replications, B, is here 10000.

Table 17: The preliminary stage of the Bootstrap test

k	H_0^k	P-value = $B_A^k/B \cdot 100$
1	$\mu_4 \leq \mu_1$	0.00
2	$\mu_4 \leq \mu_2$	0.00
3	$\mu_4 \leq \mu_3$	0.00
4	$\mu_3 \leq \mu_1$	0.14
5	$\mu_3 \leq \mu_2$	2.59
6	$\mu_2 \leq \mu_1$	39.35
7	$\mu_1 \leq \mu_2$	100.00
⋮	⋮	⋮

At the significance level $\alpha = 0.05$, the hypotheses one through five could be rejected. This means that all pairs except 1 and 2 are significantly differing. Performing the second stage shows however that even the latter pair is differing on the $\alpha = 0.05$ level. The results are given in table 18.

Table 18: Testing H_0^6 : $\mu_2 \leq \mu_1$ with clustering

Pattern	P-value = $B_A^6/B \cdot 100$
$\mu_1 = \mu_2 \neq \mu_2 \neq \mu_3$	4.22
$\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$	0.00

The conclusions drawn from this experiment is then that the classification method developed yields four groups which, according to their means, all are significantly differing at the level of significance $\alpha = 0.05$. It could also be stated that the order is $\mu_1 < \mu_2 < \mu_3 < \mu_4$.

7 Conclusions

The problem of multiple testing is familiar to most statisticians. One solution to that problem has been suggested in this paper. Compared to other methods it is rather general according to distributional assumptions etc. This is just natural since the bootstrap procedure substitutes theoretical distributions with their empirical counterparts. That is practise instead of theory. Two important remarks could be made according to this substitution. First, a method like the bootstrap is heavily depending on the computer. While the user of traditional techniques is doing with a table of critical values, some formulas and possibly a desk calculator, the bootstrapper needs a computer even if the problem is rather small. Nowadays this need can be met for most statisticians. Second, it is always difficult to prove the qualities of a method relaying on simulations. The properties of the traditional methods are often derived by means of distinguished mathematics which is believed to be true, even if not always understood, by everyone. When it comes to the bootstrap and related techniques the theoretical results are often too far away and the only proofs obtainable are the ones based on large simulations. That is simulating a method that consists of simulations itself. To the born theorist this must be rather dubious.

The method presented in this paper is not proved to be better than other methods. There are just a few hints indicating that it seems to be slightly better. The advantages of the bootstrap procedure are however to obvious to be neglected, no need for distributional assumptions, no limits for the number of hypotheses or the number of observations and no restrictions like, for example, equal sample sizes. All this are good reasons for further development and evaluation of the bootstrap multiple test procedure.

ACKNOWLEDGEMENTS

I am indebted to Professor, Techn Dr Sture Holm, who suggested the topic for this paper and guided me in a most stimulative way during the work. I would also like to thank Mrs Britt Echevarría who carefully typed the manuscript.

REFERENCES

Begun, J.M. and Gabriel, K.R. (1981): Closure of the Newman-Keuls Multiple Comparisons Procedure, J.A.S.A. 76.

Bohrer, Chow, Faith, Joshi, Wu (1981): Multiple Three-Decision Rules for Factorial Simple Effects: Bonferroni Wins Again!, J.A.S.A. 76.

Cox, D.R. and Spjøtvoll, E. (1982): On Partitioning Means into Groups, Scandinavian Journal of Statistics 9.

Duncan, D.B. (1955): Multiple Range and Multiple F tests. Biometrics, 11:1-42.

Efron, B. (1981): Censored Data and the Bootstrap. J.A.S.A. 76.

Efron, B. (1982): The Jackknife, the Bootstrap and Other Resampling Plans, Society for Industrial and Applied Mathematics, Philadelphia.

Hartley, H.O. (1955): Some Recent Developments in Analysis of Variance, Communications on Pure and Applied Mathematics, 8, 47-72.

Holm, S. (1977): Sequentially Rejective Multiple Test Procedures, Institute of Mathematics and Statistics, University of Umeå, Statistical Research Report 1977-1.

Holm, S. (1980): On Multiple Test Procedures, Mathematical Statistics, Banach Center Publications, vol. 6.

Kendall, M.G. (1955): Further Contributions to the Theory of Paired Comparisons, Biometrics, March.

Lehmann, E.L. and Schaffer, J.P. (1977): On a Fundamental Theorem in Multiple Comparisons. J.A.S.A. 72.

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.

Miller, R.G.Jr (1980): Simultaneous Statistical Inference, 2nd ed., Springer-Verlag, New York,

Schaffer, J.P. (1980): Control of Directional Errors with Stagewise Multiple Test Procedures, The Annals of Statistics, vol. 8, no 6.

Schaffer, J.P. (1981): Complexity: An Interpretability Criterion for Multiple Comparisons, J.A.S.A. 76.

APPENDIX A

Data to Example 4.1, see section 4.2

j	Y_{1j}	Y_{2j}	Y_{3j}
1	-3	0	5
2	-3	0	5
3	-3	-1	5
4	1	-1	5
5	1	-1	5
6	1	-1	4
7	1	2	4
8	5	2	6
9	5	2	6
10	5	2	0
11		2	0
12		2	10
13		2	10
14		2	2
15		5	8
16		5	
17		5	
18		5	
19		4	
20		4	

APPENDIX B

Data to Example 5.1, see section 5.2, rounded to one decimal place

Obs			Grupp		
j	1	2	3	4	5
1	13.9	13.4	21.6	19.8	27.4
2	15.1	21.7	22.4	19.7	23.1
3	13.2	20.7	19.7	20.9	30.7
4	15.0	15.7	24.2	18.7	22.6
5	15.7	18.5	21.4	18.4	28.3
$\bar{x}_1$	14.6	18.0	21.9	19.5	26.4
s_i	1.0	3.5	1.6	1.0	3.5
μ_i	16.1	17.0	20.7	21.1	26.5
σ_i	2.7	2.7	2.7	2.7	2.7

APPENDIX C

Data to Example 5.2, see section 5.2

Obs	Grupp					
j	1	2	3	4	5	6
1	1462	1477	1491	1516	1537	1571
2	1449	1528	1484	1451	1542	1623
3	1417	1583	1350	1552	1532	1584
4	1451	1455	1531	1569	1613	1681
5	1429	1489	1505	1610	1546	1587
$\bar{y}_i$	1442	1506	1472	1539	1554	1609
s_i	18.2	50.4	70.6	60.0	33.4	44.5
μ_i	1470	1498	1505	1528	1564	1600
σ_i	49.5	49.5	49.5	49.5	49.5	49.5

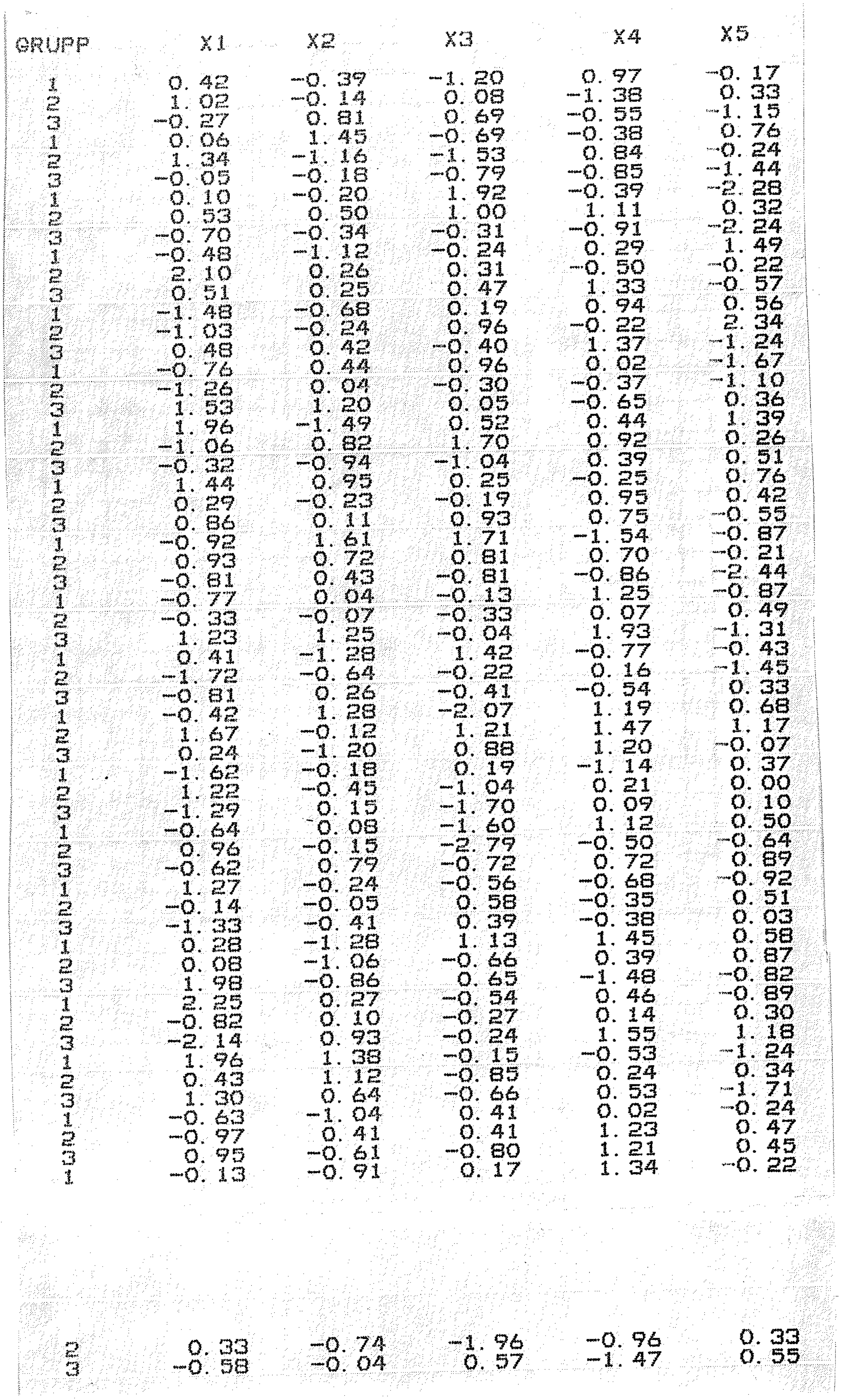
APPENDIX D

Data to Example 3, see section 5

GRUPP	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
1	40	56	21	12	3	18	57	51	50	96
2	42	71	10	67	62	61	82	71	36	19
3	64	32	50	7	51	39	84	27	43	78
1	50	17	10	7	77	9	0	79	48	25
2	1	26	85	34	52	76	83	48	3	39
3	0	78	28	28	91	33	84	35	93	71
1	23	62	62	5	52	93	60	49	43	23
2	86	24	99	35	71	4	58	24	3	25
3	60	78	24	32	52	97	55	79	55	80
1	92	85	71	41	33	33	44	93	73	22
2	74	72	31	68	74	12	43	56	85	75
3	92	42	33	24	0	5	41	53	60	47
1	93	61	23	44	41	78	31	52	68	46
2	11	53	98	39	1	59	58	45	43	66
3	88	91	16	50	2	18	23	24	1	10
1	25	37	29	68	29	24	89	38	24	12
2	3	89	66	67	20	37	14	98	79	22
3	16	9	22	35	44	99	49	96	84	15
1	37	79	66	50	94	87	53	0	72	83
2	80	64	31	62	32	90	46	81	88	86
3	69	32	3	63	42	72	44	41	54	98
1	77	2	55	75	2	87	17	9	39	49
2	44	77	66	58	32	3	1	49	82	83
3	41	22	59	62	48	1	4	64	84	27
1	96	62	95	13	60	11	51	49	30	21
2	29	46	61	83	51	47	86	49	66	68
3	33	56	91	9	78	2	94	1	64	74
1	31	6	57	26	31	73	58	0	31	36
2	63	10	67	67	78	64	2	79	74	40
3	20	74	4	14	56	47	57	39	84	37
1	77	67	54	30	17	27	77	7	53	56
2	60	49	46	67	94	88	51	83	88	7
3	77	12	56	93	1	70	53	36	27	62
1	60	86	99	4	20	49	57	5	27	41
2	53	91	56	90	11	78	55	85	35	5
3	77	81	74	6	68	49	29	38	2	78
1	18	58	66	24	84	50	98	15	56	98
2	73	2	62	19	36	58	86	32	92	32
3	86	91	36	52	46	58	48	70	94	30
1	19	73	61	17	44	26	89	67	56	78
2	96	45	9	14	85	53	20	55	3	18
3	48	7	58	6	34	70	74	7	89	81
1	2	60	23	22	86	26	38	68	69	20
2	61	89	57	82	28	5	28	7	24	81
3	93	23	65	85	9	12	38	86	38	92
1	83	7	57	41	32	26	23	75	91	25
2	79	71	81	99	19	49	22	98	96	94
3	67	37	85	36	88	42	93	73	17	52
1	18	51	25	75	65	87	63	85	3	50
2	45	75	75	83	85	58	63	84	85	21
3	96	27	66	25	11	46	59	53	65	53
1	38	17	96	21	7	10	33	11	35	0
2	57	40	44	91	13	20	41	18	70	25
3	9	100	32	20	41	44	0	39	21	94
1	81	13	70	34	51	42	22	38	64	25
2	66	73	14	35	30	51	47	46	16	28
3	19	10	88	58	10	85	7	95	84	6
1	17	41	97	19	80	51	64	52	63	84
2	65	44	73	4	6	21	46	30	33	22
3	78	40	91	79	14	24	35	36	85	80

APPENDIX E

Data to Example 4, see section 5



GRUPP	X1	X2	X3	X4	X5
1	0.42	-0.39	-1.20	0.97	-0.17
2	1.02	-0.14	0.08	-1.38	0.33
3	-0.27	0.81	0.69	-0.55	-1.15
1	0.06	1.45	-0.69	-0.38	0.76
2	1.34	-1.16	-1.53	0.84	-0.24
3	-0.05	-0.18	-0.79	-0.85	-1.44
1	0.10	-0.20	1.92	-0.39	-2.28
2	0.53	0.50	1.00	1.11	0.32
3	-0.70	-0.34	-0.31	-0.91	-2.24
1	-0.48	-1.12	-0.24	0.29	1.49
2	2.10	0.26	0.31	-0.50	-0.22
3	0.51	0.25	0.47	1.33	-0.57
1	-1.48	-0.68	0.19	0.94	0.56
2	-1.03	-0.24	0.96	-0.22	2.34
3	0.48	0.42	-0.40	1.37	-1.24
1	-0.76	0.44	0.96	0.02	-1.67
2	-1.26	0.04	-0.30	-0.37	-1.10
3	1.53	1.20	0.05	-0.65	0.36
1	1.96	-1.49	0.52	0.44	1.39
2	-1.06	0.82	1.70	0.92	0.26
3	-0.32	-0.94	-1.04	0.39	0.51
1	1.44	0.95	0.25	-0.25	0.76
2	0.29	-0.23	-0.19	0.95	0.42
3	0.86	0.11	0.93	0.75	-0.55
1	-0.92	1.61	1.71	-1.54	-0.87
2	0.93	0.72	0.81	0.70	-0.21
3	-0.81	0.43	-0.81	-0.86	-2.44
1	-0.77	0.04	-0.13	1.25	-0.87
2	-0.33	-0.07	-0.33	0.07	0.49
3	1.23	1.25	-0.04	1.93	-1.31
1	0.41	-1.28	1.42	-0.77	-0.43
2	-1.72	-0.64	-0.22	0.16	-1.45
3	-0.81	0.26	-0.41	-0.54	0.33
1	-0.42	1.28	-2.07	1.19	0.68
2	1.67	-0.12	1.21	1.47	1.17
3	0.24	-1.20	0.88	1.20	-0.07
1	-1.62	-0.18	0.19	-1.14	0.37
2	1.22	-0.45	-1.04	0.21	0.00
3	-1.29	0.15	-1.70	0.09	0.10
1	-0.64	0.08	-1.60	1.12	0.50
2	0.96	-0.15	-2.79	-0.50	-0.64
3	-0.62	0.79	-0.72	0.72	0.89
1	1.27	-0.24	-0.56	-0.68	-0.92
2	-0.14	-0.05	0.58	-0.35	0.51
3	-1.33	-0.41	0.39	-0.38	0.03
1	0.28	-1.28	1.13	1.45	0.58
2	0.08	-1.06	-0.66	0.39	0.87
3	1.98	-0.86	0.65	-1.48	-0.82
1	2.25	0.27	-0.54	0.46	-0.89
2	-0.82	0.10	-0.27	0.14	0.30
3	-2.14	0.93	-0.24	1.55	1.18
1	1.96	1.38	-0.15	-0.53	-1.24
2	0.43	1.12	-0.85	0.24	0.34
3	1.30	0.64	-0.66	0.53	-1.71
1	-0.63	-1.04	0.41	0.02	-0.24
2	-0.97	0.41	0.41	1.23	0.47
3	0.95	-0.61	-0.80	1.21	0.45
1	-0.13	-0.91	0.17	1.34	-0.22
2	0.33	-0.74	-1.96	-0.96	0.33
3	-0.58	-0.04	0.57	-1.47	0.55

APPENDIX F

Logical flowchart of the FORTRAN computer program performing the Bootstrap multiple comparison test.

1. Number of groups and observations and observed values are read into the program.
2. Calculations of means and standard deviations.
3. A table of means and standard deviations is printed.
4. All possible differencies among pair of means are calculated.
5. The differencies are sorted in descending order.
6. All observations are translated to give zero means.
7. All possible clustering patterns are enumerated.
8. A number of Bootstrap-samples are generated.
9. The observed differencies are tested in descending order.
10. The results of 9 is printed.
11. A number of Bootstrap-samples are generated (optional).
12. One difference is tested under clustering conditions (optional).
13. The result of 12 is printed (optional).
14. Go to 11 (optional).
15. End.

The steps 7, 8, 9, 11 and 12 are handled by two subroutines while the rest is performed by the main program.