ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 164, 2003

Czesław Domański, Dariusz Parys***

THE CLOSED BOOTSTRAP MULTIPLE TEST PROCEDURE

Abstract. In this paper we present how to apply the ideas of bootstrap and closed testing procedures on a multiple comparison test.

We consider L samples of size n_1, n_2, ..., n_l from L distributions, with expected values u_1, u_2, ..., u_L, are to be compared.

We consider the stepwise procedures introduced by Holm (1977). The test in each step is performed by means of the bootstrap technique. This procedures are always closed, but just the fact that a procedure is stepwise does not guarantee that it is closed. We discuss whether the appropriate conditions are met to make our bootstrap procedure closed. When the test is performed on very few observations the significance level is sometimes only approximately kept.

However, since the approximation are due to the bootstrap, and not to the test procedure itself, the multiple test discussed in this paper is likely to keep the multiple level of significance.

Key words: closed multiple tests, bootstrap technique, multiple comparisons.

1. THE PROBLEM

Let consider a situation where L samples, of sizes n_1, n_2, ..., n_L, from L distributions, with expected values μ_1, μ_2, ..., μ_L, are to be compared. The objective is to tell which μ_i's are different and eventually rank them in descending order. If $L > 2$ this is a problem of multiple testing, discussed in, e.g. Miller (1980), which according to the principles outlined in Holm (1977) and (1980), in terms of null and alternative hypotheses could be stated as

$$H_{0ij}\colon \mu_i \leqslant \mu_j; \quad i, j = 1, 2, ..., L \quad \text{and} \quad i \neq j \tag{1.1}$$

$$H_{Aij}\colon \mu_i > \mu_j; \quad i, j = 1, 2, ..., L \quad \text{and} \quad i \neq j \tag{1.2}$$

* Professor at the Department of the Statistical Methods, University of Lodz.

** Doctor at the Department of the Statistical Methods, University of Lodz.

An equivalent formulation of (1.1) is

$$H_0\colon \mu = \mu_2 = \mu_2 = \dots = \mu_L \tag{1.3}$$

which however is less suitable for the multiple test of pairwise comparisons to follow. It should be noted that (1.1) contains twise as many single hypotheses as does (1.3). For each $\mu_i \geqslant \mu_j$ there is also a $\mu_j \geqslant \mu_i$ under H_0. Within any such pair of hypotheses one is not supposed to reject more than one; at the most.

2. CLOSED TESTING PROCEDURE

Let $H = \{H_{01}, \dots, H_{0n}\}$ be a set of all null hypotheses. Assume that H is closed under intersection, that is $\bigwedge_{i \neq j} H_{0i} \cap H_{0j} \in H$.

If $H_{0i}\colon \theta \in \omega_{0i}$ and $H_{0j}\colon \theta \in \omega_{0j}$ than hypothesis $H_{0i} \cap H_{0j}$ can be written $\theta \in \omega_{0i} \cap \omega_{0j}$. Suppose that for each H_{0i} there is a test with P (reject $H_{0i} | H_{0j}$ is true) $\leqslant \alpha$.

Now, any $H_0\colon \theta \in \omega_0$, $H_0 \in H$ is rejected if and only if all hypothesis, H_{0i}, that are included in H_0 and belonging to H have been tested and rejected. Since a type I error is committed only if the intersection of all true hypothesis is tested and rejected, the significance level of this test is $\leqslant \alpha$. According to Marcus, Peritz and Gabriel (1976) a test procedure is closed and intersection if a multiple null hypothesis is rejected only if all hypotheses corresponding to smaller parameter sets are rejected at the same level.

3. THE BOOTSTRAP TECHNIQUE

The bootstrap, Efron (1982), is a resampling method for estimating the sampling distribution without knowing the distribution generating the sample. Consider a sample of size n from an unknown probability distribution F on the real line

$$x_1, x_2, \dots, x_n \sim F \tag{3.1}$$

independently and identically. Let $\hat{\theta}$ be a function of this sample

$$\hat{\theta} = \hat{\theta}(x_1, x_2, \dots, x_n) \tag{3.2}$$

and let $\hat{F}$ be the empirical probability distribution of the sample, putting the probability mass of $1/n$ on each x_i. Use $\hat{F}$ to draw a sample with replacement $x^1, x^2, \ldots, x^n$ of size n, that is, sampling among the observed values $x_1, x_2, \ldots, x_n$ and calculate

$$\hat{\theta}^* = \hat{\theta}(x^1, x^2, \ldots, x^n), \quad x^i \sim \hat{F} \tag{3.3}$$

from this bootstrap sample. This procedure is independently repeated B times giving the replications $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*$ and hence an image of the sampling distribution of $\hat{\theta}$. The replications could for instance be used to estimate the variance of $\hat{\theta}$,

$$\hat{V}(\hat{\theta})_{BOOT} = \frac{1}{B-1}\sum(\hat{\theta}_i^* - \hat{\theta}^*)^2 \tag{3.4}$$

where

$$\hat{\theta}^* = \frac{1}{B}\sum\hat{\theta}_i^*,$$

finding critical values for tests or constructing confidence limits. In the following the bootstrap technique is applied to the closed multiple test problem given above.

4. THE BOOTSTRAP MULTIPLE TEST PROCEDURE

The basic idea of the bootstrap multiple test procedure is to form all possible pairwise differences among the L sample means $\bar{y}_1, \ldots, \bar{y}_L$, and with a number of bootstrap samples determine whether the observed differences are likely to occur just by chance or if they imply differences between the corresponding true means.

Let $\delta_{ij} = \mu_i - \mu_j$, $i, j = 1, 2, \ldots, L$ $i \neq j$ be the true differences and $d_{ij} = \bar{y}_i - \bar{y}_j$ be the sample differences.

Let denote the largest difference as d_1, the second largest as d_2 and so on until d_k, where $K = L(L-1)$. The true differences δ_{ij} and hypotheses (1.1) and (1.2) are ordered in the same descending manner and since $\mu_i \leqslant \mu_j \Leftrightarrow \delta_{ij} \leqslant 0$ they could be stated as:

$$H_{0k}: \delta_k \leqslant 0; \quad k = 1, \ldots, K \tag{4.1}$$

$$H_{Ak}: \delta_k > 0; \quad k = 1, \ldots, K \tag{4.2}$$

These hypotheses (4.1) are now to be tested in the flowing sequentially rejective manner, suggested by Holm (1977). Test H_{0k}, if accepted, accept all $H_{0i}: i \geqslant k$, if rejected test H_{0k+1}, $k = 1, 2, \ldots, K-1$. The test in each step is performed by means of the bootstrap technique.

First, the L samples are translated to zero means by subtracting form each observation its sample mean, $\bar{y}_i$. A bootstrap sample is generated from these translated distributions and for each hypothesis in (4.1) the corresponding bootstrap difference d^k is calculated. Since $\mathrm{E}(\bar{y}^i) = 0$, where $\bar{y}^i$ is the mean of the bootstrap sample, $i = 1, 2, \ldots, L$, then $E(d^k) = 0$, $k = 1, 2, \ldots, K$, and hence any deviation from zero for d^k is random. By comparing the real sample differences with the bootstrap ones it is possible to conclude whether the former are likely to occur just be chance or if they indicate differences among the true means. Let ψ_k be the number of times the bootstrap differences. To begin with, all ψ_k equals zero and than they take on values according to the following:

$$\begin{aligned} &\text{if any } d^k < d_k \quad \text{for} \quad k = 1, \ldots, K \quad \text{and} \\ &k' = k,\ k+1,\ \ldots,\ K \quad \text{then} \quad \psi_{k'} = \psi_k + 1 \end{aligned} \tag{4.3}$$

That is, for the largest sample difference, d_1, it is noted whether any bootstrap difference is that large, if not, there is one indication of $\delta_1 > 0$, for d_2 it is noted whether any bootstrap difference, except d^1, is that large, and if not, there is one indication of $\delta_2 > 0$ and so on for the k sample differences.

However, since it is not enough with just one, possible, indication to reject a hypothesis, the whole procedure is repeated B times, where B is a rather large number e.g. 1000 to 10 000. After this, $0 \leqslant \psi_k \leqslant B$, $k = 1, 2, \ldots, K$, and it is easily seen that if the number of indications, ψ_k, for a hypothesis, H_{0k}, is large enough, then H_{0k} could be rejected. The probability of rejecting H_{0k} if it is true, the p-value, is namely ψ_k/B, that is, the fraction of times out of B when a difference as large as d_k occurred just by chance.

The decisions to accept or reject are now taken sequentially according to the order of (4.1) and if the overall significance level is pre-assigned to a, the rule is, starting with $k = 1$:

$$\begin{aligned} &\text{if } \psi_k/B \leqslant a \quad \text{reject} \quad H_{0k} \quad \text{and test } H_{0k+1} \\ &\text{if } \psi_k/B > a \quad \text{accept} \quad H_{0i},\ i > k \end{aligned} \tag{4.4}$$

It is to be observed that a hypothesis, H_{0k}, is not to be rejected, unless all hypotheses, H_{0i}, $i < k$, already have been so.

As a final step of the test procedure the logical structure is taken into account. Doing so, it is possible to increase the power without affecting the significance level. The idea is to not waste any power by counting both d^i and d^j and indications if it could be stated through the former rejections, that not both δ_i and δ_j could equal zero, $i, j = 1, 2, ..., K$.

The final step could be included in all stages of the procedure. However, this causes unnecessary calculations if the first hypothesis would be rejected anyway. Our suggestion is to reject as many hypotheses as possible without the final step and then include this step from the first hypothesis not being rejected and onwards. Of course the whole procedure stops when not even the final step is able to reject a certain hypothesis.

5. THE CLOSED BOOTSTRAP MULTIPLE TESTS

Let $\Delta = \{\delta_{ij}\}$ be the set of $K = L(L-1)$ true differences. Δ is possible to divide into Δ^+, including the positive elements and Δ^- including the negative and zero ones. Obviously $\Delta^+ \cup \Delta^- = \Delta$ and $\Delta^+ \cap \Delta^- = 0$. The test is supposed to tell whether $\delta_{ij} \in \Delta^+$ or $\delta_{ij} \in \Delta^-$, $i \neq j$.

In terms of δ_{ij} the null hypothesis (5.1) is

$$H_0 : \delta_{ij} \in \Delta^-; \quad \text{for all} \quad i \neq j \tag{5.1}$$

The set of null hypotheses $H = H\{H_0\}$, according to Marcus, Peritz, Gabriel (1976) is then the possible decisions of Δ^- and Δ^+. This set is obviously closed since the intersection between any two divisions results in a third one also included in H.

In terms of the sample differences, $d_{ij} = \bar{y}_i - \bar{y}_j$, the hypotheses are noted and tested in descending order.

Let $d^1 \geqslant d^2 \geqslant ... \geqslant d^K$ and $\delta^1 \geqslant \delta^2 \geqslant ... \geqslant \delta^K$ be the corresponding true differences. Then $\delta^k \in \Delta$, $k = 1, ..., K$ are tested only if all hypotheses $\delta^i \in \Delta^-$, $i < k$, have been tested and rejected.

Due to the inability to rejecting $\delta^{K^*+1} \in \Delta^-$, the test procedure stops and gives the final statement

$$\{\delta^1, \delta^2, ..., \delta^{K^*}\} \subset \Delta^+ \tag{5.2}$$

which is equivalent to rejecting the hypothesis that any δ^i, $i = 1, 2, ..., K^*$ belongs to Δ^-.

According to the stepwise character of the procedure the hypothesis rejected in step i, $i = 1, 2, ..., K^*$, is

$$\bigwedge_{j=1}^{K} (\delta^j \in \Delta^-) \tag{5.3}$$

In order to be a closed procedure all the hypotheses

$$\left(\bigwedge_{j=1}^{K} (\delta^j \in \Delta^-)\right) \wedge \left(\bigwedge_{j \in J} (\delta^j \in \Delta^-)\right) \tag{5.4}$$

should be rejected for all $J \subseteq \{K^* + 1,\ K^* + 2,\ ...,\ K\}$. Let for example $K = 6$ and $K^* = 3$. Then the hypotheses

$\{\delta^1, \delta^2, \delta^3, \delta^4, \delta^5, \delta^6\} \in \Delta^-$, $\{\delta^2, \delta^3, \delta^4, \delta^5, \delta^6\} \in \Delta^-$ and $\{\delta^3, \delta^4, \delta^5, \delta^6\} \in \Delta^-$ are rejected. To be a closed test the hypotheses $\{\delta^1, \delta^2, \delta^3\} \in \Delta^-$, $\{\delta^1, \delta^2, \delta^3, \delta^5\} \in \Delta^-$, $\{\delta^1, \delta^2, \delta^3, \delta^6\} \in \Delta^-$, $\{\delta^1, \delta^2, \delta^3, \delta^4, \delta^5\} \in \Delta^-$, $\{\delta^1, \delta^2, \delta^3, \delta^4, \delta^6\} \in \Delta^-$, $\{\delta^1, \delta^2, \delta^3, \delta^5, \delta^6] \in \Delta^-$, and $\{\delta^1, \delta^2, \delta^3, \delta^4, \delta^5, \delta^6\} \in \Delta^-$ should be rejected.

The nature of the bootstrap test is to simulate a large number of differences and records the number of times the difference d^k, $i = 1, 2, ..., K$ is exceeded, no matter where it did appear. When for example evaluating the hypothesis $\delta^1 \in \Delta^-$, all differences emerging from the bootstrap differences based on $d^1, d^2, ..., d^K$ are involved. If the proportion of differences exceeding d^1 is $\leqslant a$, $\delta^1 \varepsilon \Delta^-$ and hence $\{\delta^1, \delta^2, ..., \delta^k\} \in \Delta^-$ is rejected. If the proportion of defferences exceeding d^2 is $\leqslant a$, $\delta^1 \in \Delta^-$ and hence $\{\delta^2, \delta^2, ..., \delta^k\} \in \Delta^-$ is rejected and so on.

Let ψ_k be the number of times the bootstrap difference $i = 1, 2, ..., K$, exceeds d^k and B the total number of bootstrap replicates. Then

$$\{\delta^1, \delta^2, ..., \delta^K\} \in \Delta^- \tag{5.5}$$

rejected if $\sum_{k=1}^{K} \psi_k / B \leqslant a$.

When some δ^j is excluded from (5.5), the number of exceeding bootstrap differences will decrease, or possibly remain unchanged i.e. $\left(\sum_{i=1}^{K} \psi_i\right) - \psi_j \leqslant \sum_{i=1}^{K} \psi_i$ and thus the corresponding null hypothesis will be rejected. This holds that any $\{\delta\} \subset \{\delta^1, \delta^2, ..., \delta^K\}$ and especially

$$\left(\sum_{i=1}^{K^*} \psi_i\right) + \left(\sum_{i \in \{K^*+1, K^*+2, ..., K\}}^{K^*}\right) \leqslant \sum_{i=1}^{K} \psi_i \tag{5.6}$$

and hence the rejection of (5.3) implies the rejection of (5.4) which in turn is to say that the test procedure is closed one.

6. CONCLUSIONS

The problem of multiple comparisons is familiar to most statistications. One solution to that problem has been suggested in this paper. Compared to other methods it is rather general according to distributional assumptions. This is just natural since the bootstrap procedure substitutes theoretical distributions with their empirical counterparts. The are given some indications of higher power for the new method. As there as other advantages of bootstrap multiple test procedure, no need for distributional assumptions, no limits for the number of hypotheses or the number of observations and no restrictions like, e.g. equal sample sizes, there are good reasons for further development.

The test procedure discussed in this paper is shown to be closed. Hence it is likely to keep the multiple level of significance at the predetermined value. In theory this is so, but due to imperfections in the bootstrap estimations of the real distributions, the significance level is sometimes only approximately kept. Especially when the test performed on very few observations. However, since the approximations are due to the bootstrap, and not to test procedure itself, knowledge of the real distribution would give a procedure for multiple comparisons which exactly keeps the significance level.

REFERENCES

Efron B. (1982), *The Jacknife, the Bootstrap and Other Resamplin Plans*, Philadelfia, Society for Industrial and Applied Mathematics.

Holm S. (1977), *Sequentialy Rejective Multipe Procedures*, Statistical Research Report 1977–1, University of Umea, Institute of Mathematics and Statistics.

Marcus R., Peritz E., Gabriel K. R. (1976), *On Closed Testing Procedures with Special Reference to Orchred Analysis of Variance*, "Biometrica" **63**, 655–660.

Miller R. G. Jr (1980), *Simulations Statistical Inference*, 2^{nd} *edn.*, Springer Verlag, New York.

Czesław Domański, Dariusz Parys

WIELOKROTNY BOOTSTRAPOWY TEST DOMKNIĘCIA
(Streszczenie)

W pracy zaprezentowano zastosowania idei bootstrapowej i domkniętych procedur testowych we wnioskowaniu dotyczącym porównań wielokrotnych.

Rozważmy L prób o liczebnościach n_1, n_2, ..., n_L, odpowiednio pochodzących z L populacji. Porównuje się wartości oczekiwane $\mu_1, \mu_2, ..., \mu_L$. W procedurze kroczącej dla porównań

wielokrotnych zaproponowana przez Holma (1977) zastosowano technikę bootstrapową. Procedura jest niezależna od rozkładów badanych populacji, liczby badanych hipotez, równej liczby prób. Jest procedurą domkniętą, przez co utrzymany jest wielokrotny poziom istotności a na poziomie z góry ustalonym.