A New Multiple Comparison Method Using Resampling Technique for Medical, Pharmaceutical, and Chemical Data

Tatsuya Takagi\textsuperscript{a,b}, Kousuke Okamoto\textsuperscript{a}, Yumiko Yokogawa\textsuperscript{c}, Masahiko Yokota\textsuperscript{a}, Ken Kurokawa\textsuperscript{b,d}, Teruo Yasunaga\textsuperscript{b}

\textsuperscript{a} Graduate School of Pharmaceutical Sciences, Osaka University, 1-6 Yamadaoka, Suita, Osaka 565-0871 Japan
\textsuperscript{b} Genome Information Research Center, Osaka University, 3-1 Yamadaoka, Suita, Osaka 565-0871 Japan
\textsuperscript{c} Faculty of Pharmaceutical Sciences, Osaka University, 1-6 Yamadaoka, Suita, Osaka 565-0871 Japan
\textsuperscript{d} Research Institute for Microbial Disease, 3-1 Yamadaoka, Suita, Osaka 565-0871 Japan

(Received March 19, 2003; Accepted June 3, 2003)

It is not easy for experimental biologists to select an appropriate statistical method, especially in cases in which a multiple comparison method is necessary. Since various types of multiple comparison procedures have been proposed, experimentalists have had to face this difficulty. We propose a new multiple comparison procedure which has a wide range of applicability and uses a resampling technique that is currently popular in biostatistical literature. This new procedure enables us to solve multiple comparison problems more easily than conventional methods such as the Dunnett method. The power of tests of the new procedure was compared with some conventional multiple comparison methods, and satisfactory results were obtained.

Key Words: multiple comparisons, test of significance, resampling method, bootstrap method

Introduction

In the fields of biology, pharmaceutical sciences, chemistry, and medical sciences, multiple comparisons are a common problem for experimental scientists. For example, when multivariate data of the efficacies of new antibiotics (let the number of drugs be $n$) for a type of bacteria are measured, the average values of the efficacies have to be tested using a statistical test of significance to elucidate whether the new antibiotics are significantly more effective than the conventional
antibiotic. The Student’s t-test is commonly used to test significance. If the Student’s t-test is used repeatedly \((n)\) times in this example, with a significant level equal to 5%, it is quite likely that some drugs would be judged by chance to be significantly more efficient, even though there are no differences in the efficacies. This misjudging occurs because the probability that at least one drug is judged by chance to be significantly more efficient can reach as much as 1-0.95\(^n\). When \(n\) equals 10, \(p\) can reach 40%; therefore, it would be an exaggeration to say that all of the drugs are significantly more efficient. This problem — known as multiple comparisons — is so well known that many procedures have been proposed to avoid it. In the fields of medical and pharmaceutical sciences especially, this problem is serious. In these fields, such an error (i.e., Type-I error) should not be allowed; otherwise, inefficient drugs with possible unknown side effects could be introduced. Thus, the importance of multiple comparisons has recently been recognized in these fields, as well as the importance of using the most appropriate method for statistical analyses. Nevertheless, because too many multiple comparison procedures have been proposed, experimental biologists, who may not very familiar with biostatistics, have difficulty determining the most appropriate procedure for multiple comparison problems. Even now, experimental biology papers occasionally mention incorrect procedures for multiple comparisons, partly because some of the popular statistical program packages include inappropriate multiple comparison methods, such as Duncan’s method.

As stated above, various kinds of procedures have been proposed to solve multiple comparison problems. For instance, Bonferroni-type corrections give a corrected significant level for keeping nominal significant levels according to Bonferroni’s inequality: when \(n\) independent statistical tests are carried out, an adjusted significant level is obtained by dividing it by \(n\) in order to prevent the chance of committing Type-I errors that exceed the nominal significant level. Because this procedure is simple and easy, it has been widely used. Nevertheless, it is well known that this correction produces too conservative results. Many experimental biologists have used other types of procedures, such as the Dunnett and Tukey methods, for multiple comparison problems. The Dunnett and Tukey methods, which are parametric multiple comparison methods, are less conservative than the Bonferroni correction. Whereas Bonferroni-type methods can be widely used, the parametric methods are powerful only when the data fit a normal distribution and sample sizes are equal. Although some nonparametric multiple comparison methods, such as Steel-Dwass and Steel, can prevent the problem of “normal distribution spell,” their power of test is considered to be less than parametric methods such as the Dunnett method.

Considering the abovementioned problems, some multiple comparison methods using a resampling technique [1-3] have been proposed very recently; these methods take into account the correlations between statistical tests for multiple comparisons. This type of multiple comparison method was first proposed by Westfall and Young [4] and its use is possible due to the development of computer technologies. Although a number of multiple comparison procedures using resampling techniques have been developed [5-9], most are based on the adjustment of p-values calculated for tests of significance. In addition, most of the multiple comparisons using resampling techniques control the probability of a Type-I error (Type-I family-wise error rate) using adjusted p-values. The advantage of these methods is that it is not necessary to presume the distribution function for a certain statistic. Thus, methods using the resampling method are considered superior to Bonferroni-type methods in power of test because they consider the correlation structures among statistical tests. Nevertheless, the advantage of Bonferroni-type multiple comparison methods is their wide range of applicability. Considering the abovementioned situation, we propose a new method with the advantages of both the Bonferroni-type and the resampling methods: a wide range of applicability and enough power of tests.

Although this new method can be classified in the same category as the Westfall and Young method [4], it can also be classified as a Bonferroni-type method, because it adjusts the significant level of each test of significance in order to maintain the nominal significant level. Thus, as stated above, our new method has the advantages of both types of methods: a wide range of applicability like Bonferroni-type methods, and the high power of tests like the Westfall and Young method. Furthermore, users do not need any special statistical knowledge of multiple comparisons, due to our method’s theoretical and algorithmic simplicity. Our new method enables experimental researchers in the fields of biology, medical sciences, pharmaceutical sciences, and chemistry to carry out multiple comparisons with the appropriate power of tests, and without considering what kind of multiple comparison method they have to adopt for their data set.

**Algorithm**

Assume that the case data, control data, and the significant level are vectors \(\mathbf{x} = (x_{1j}, x_{2j}, \ldots, x_{nj}) (j=1, 2, \ldots, m)\), \(\mathbf{x}_0 = (x_{10}, x_{20}, \ldots, x_{n0})\), and \(\alpha_0\), respectively. For instance, in the case comparing all treatments against the control data using the nominal significant level \(\alpha_0\) the procedure of our new method is as follows:
1. Compute the standardized values, $\xi_{ij}$, according to the equation $\xi_{ij} = (x_{ij} - \bar{x}_j)/s_j$ ($i=1,2,\ldots,N$; $j=1,2,\ldots,m$). Here, $s_j$ means the standard deviation of the $j$-th group.

2. Generate the resampling data, $\xi_{ij}^*$, with replacement from the standardized data, $\xi_{ij}$, and repeat resampling $B$ times (Figure 1).

3. Initialize the $k$ value: $k=1$.

4. Let $\alpha_0$ be the initial value of $\alpha(k)$ that is used for the significant level.

5. Carry out the t-test for the $B$ data sets of the control data, using $\alpha(k)$ as the significant level.

6. Compare the number of data sets ($n_k$) of which at least one or more are rejected $H_0: \Box \neq \Box_{a0}$ are rejected with $n_k = B\alpha_k$. Here, $\Box_0$ and $\Box_j$ indicate the averages of the populations of the control and the $j$-th group, respectively.

7. If $n_k \leq n_0$, then $k=k+1$, $\alpha(k) = \alpha(k-1)+\alpha_{inc}$ ($\alpha_{inc}$ is a small increment value), and return to Step 5. If $n_k > n_0$, then calculate $\alpha_t$, where $\alpha_t$ indicates $\alpha(k)$ when $n_k$ equals $n_0$.

8. Repeat Steps 4–7, $T$ times.

9. Obtain the average $\alpha$ of $\alpha_t$ ($t=1,2,\ldots,T$), and set as the adjusted significant level, $\tilde{\alpha}$.

In Steps 1–9 above, although a slightly different process of sampling might be necessary, depending on the type of significance tests, the framework is conserved which generates resampled data under the

Figure 1. Two procedures of generating resampled samples with replacement used in this study.
condition that no null hypotheses are rejected. For example, when using the above procedure in the case of parametric tests, the observed data of each group are standardized, and then the data are resampled with replacement within each group. In the case of nonparametric tests, the observed data are not standardized, and then the data are resampled with replacement from every group. The value $\alpha_{inc}$ has to be sufficiently small so that the linear interpolation for Step 7 above is efficient.

As stated above, the Westfall-Young [4, 5] step-down approach (abbreviated as WY below) is the resampling-based multiple comparison which is similar to our new method. In particular, the bootstrap method is used for generating resampled data from observed data under the condition that all null hypotheses are accepted. However, whereas WY generates resampled data in order to adjust the p-value, our new method resamples in order to adjust the significant level. Furthermore, while WY includes the testing procedure, our new method does not include it. Therefore, our new method can be applied more widely than WY.

**Simulation Trials**

First of all, we checked whether our new procedure exceeds the nominal significant level using a Monte Carlo-type simulation. Considering that multiple comparisons for the differences of mean values are frequently used in the fields of chemistry, biology, and pharmaceutical sciences, the data sets for the significance test of differences between mean values were used for this purpose. Since the pair-wise comparisons between a control group and some case groups occupy an especially important position in these fields, the two types of multiple comparisons were investigated: one compares the mean of a control group to the means of other groups (a Dunnett-type comparison), the other compares all possible pairs of means (a Tukey-type comparison).

Furthermore, multiple linear regression methods are also important in these fields for analyzing multivariable data, which also include a multiple comparison problem in principle. Although semi-empirical criteria such as “more than four observations per one prediction variable” has been widely used in the field of medicinal chemistry, based upon the study of simulations under a limited condition, our new procedure enables practitioners to solve multiple comparison problems of multiple regression analyses generally, theoretically, and accurately. Therefore, in order to show the versatility of our new method, we tried to apply our new method to multiple comparisons for the significance test of partial regression coefficients of multiple linear regressions. Conventional multiple comparison methods (Bonferroni, Šidák, Tukey, and Dunnett) were also adopted for the data sets in order to compare the power of tests between these conventional methods and our new method. Because these trials were carried out to show the applicability and power of tests of our new method, we used common conditions, such as the same sample size for each data set, and the same 95% ordinal level of significance.

1) **Pair-wise comparisons between a control group and case groups**

**Trial-1.**

Data sets consisting of one control group and six case groups were generated using a standard normal pseudo-random number generator. The sample size of all the data sets was set at 20. This trial was carried out using four methods: Bonferroni, Šidák, Dunnett, and our new method. Although the Bonferroni and Šidák methods, as well as our new method, have flexible algorithms for multiple comparisons, the Dunnett method needs to assume a normal distribution for the data sets. Because the three methods (Bonferroni, Šidák, and our new method, but not the Dunnett method) have to decide the method of significance test, the Wilcoxon rank-sum test was adopted. The Wilcoxon rank-sum test was also adopted for the other trials (Trials 2–4) in this study.

**Trial-2.**

In order to generate the data for the nonparametric procedures, all data for both the control and case groups were generated using a pseudo-random number generator which follows the Weibull distribution, so that the data followed a skewed distribution. Other conditions were the same as those of Trial-1, except that the Dunnett method was replaced by the nonparametric Steel method.

2) **All pair-wise comparisons among multiple groups**

**Trial-3.**

Each data set of all four groups was generated using a standard normal pseudo-random number generator. Other conditions were the same as those of Trial-1, except that the Dunnett method was replaced by the Tukey method, which carries out all pair-wise comparisons.

**Trial-4.**

For Trial-4, pseudo-random numbers were generated on the basis of the Weibull distribution. Other conditions were the same as Trial-3, except that the Tukey method was replaced by the nonparametric Steel-Dwass method.

3) **Multiple comparisons of partial regression coefficients of multiple linear regressions**
Trial-5.

The principal component scores of random number matrices were used in order to avoid multicollinearity. Normal random numbers (a=0,  b=0.5) were added to the principal component scores for the predictive variables. Here, a and b mean the average and standard deviations, respectively. Standard normal random numbers were used as the correspondence variable. The data set used in this trial consisted of six predictive variables and the one correspondence variable, which has 100 observations. The same test methods as Trial-3, except for the Tukey method, were used for this trial.

Trial (Supplement)

Since most of the multiple comparison procedures with the resampling method compare adjusted p-values with the nominal significant level, it is natural that the probability of Type-I errors equals the nominal significant level. Nevertheless, due to the abovementioned conditions of other methods using resampling techniques, their applicability is limited. Recently, Troendle [6] introduced an SR2 (SR algorithm 2) algorithm, which is a stepwise multiple comparison procedure with the resampling method. This method is based on the same idea (that is, Bonferroni-type adjustments) as our new procedure; however, the purpose is not to adjust the calculated statistics for multiple comparisons but to adjust the significant level, which is used for determining whether the null hypotheses are rejected for obtaining the correct probabilities of Type-I errors. In order to compare SR2 with our new procedure, we obtained the probabilities of Type-I errors using SR2 for Trials 1 and 2, to which SR2 can apply. Although SR2 is based on the same idea as our new procedure, it is less applicable than our method because SR2 is a stepwise method that includes the statistical significant test procedure itself.

For all the simulation trials (Trials 1–5 and the Supplement Trial) in this study, the process from data generation, including bootstrap resampling, to significance test was repeated 10,000 times. Each multiple comparison test was carried out using 0.05 as the level of significance. We checked whether the probability of Type-I errors using our new procedure was less than the nominal level, as well as whether the results of the procedure were close enough to 0.05. For our new procedure, bootstrap resampling was carried out 5,000 times (B=5,000), and the significant level was adjusted using the average value of the 10 calculation results (T=10). Since the Wilcoxon rank sum test was used in our new procedure for all trials except Trial-5, resampling with replacement was carried out for each group. For Trial-5, resampling with replacement was carried out by “Separate Resampling,” as shown in Figure 2.

![Figure 2. “Usual Resampling” and “Separate Resampling” for vector data sets.](image-url)
Clinical Data Trial

We applied our new procedure to the clinical data set. The data set was extracted from reference [3]. It involves litter weights of mice assigned to three different dosage groups (5, 50, 500) and a control group (0 dosage). In this trial, the average weights for males in week 2 were adopted. Our new method was compared to the two conventional multiple comparison methods, the Bonferroni, and Šidák methods, which are similar to our new method in applicability. Since the data set is a dose-response relationship, the Williams method, which has less applicability than our new method, was also adopted as the conventional multiple comparison method. Welch’s t-test was used for the three methods (Bonferroni, Šidák, and our new methods) for the significance test of each sample. The null hypotheses for all the methods are
\[ \mu_c = \mu_1, \quad \mu_c = \mu_2, \quad \text{and} \quad \mu_c = \mu_3. \]
The alternative hypotheses are
\[ \mu_c > \mu_1 > \mu_2 > \mu_3, \]
where \( \mu_c \) and \( \mu_i \) indicate the averages of the control group and the \( i \)-th dosage group, respectively. Other conditions for all of these procedures were the same as those for the simulation trials.

All the computations in this study were carried out on a Fujitsu Prime Power 800 UNIX Workstation at the Genome Information Research Center, Osaka University.

Results and Discussion

<table>
<thead>
<tr>
<th>Trial</th>
<th>New</th>
<th>#</th>
<th>Bonferroni</th>
<th>Šidák</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.87</td>
<td>4.87</td>
<td>3.32</td>
<td>3.68</td>
</tr>
<tr>
<td>2</td>
<td>4.93</td>
<td>4.86</td>
<td>3.52</td>
<td>3.85</td>
</tr>
<tr>
<td>3</td>
<td>4.98</td>
<td>4.99</td>
<td>3.60</td>
<td>3.91</td>
</tr>
<tr>
<td>4</td>
<td>4.46</td>
<td>4.42</td>
<td>3.22</td>
<td>3.48</td>
</tr>
<tr>
<td>5</td>
<td>4.81</td>
<td>-</td>
<td>4.71</td>
<td>4.80</td>
</tr>
</tbody>
</table>

*: Each row shows trial results and each column corresponds to the procedure used. “New” is our procedure. The methods in column “#” are trial-dependent: Dunnett test for Trial-1, Steel test for Trial-2, Tukey test for Trial-3, Steel-Dwass test for Trial-4. The underlined numbers in each row indicate the values closest to 5%.

Table 1. Probabilities of Type-I Errors in Trials*

First, we checked whether the probabilities of Type-I errors for all the methods in all the trials, without adjusting the nominal significant level, were 5%. The results confirm that all of the simulation trials are appropriate: the probabilities of Type-I errors are equal to 5% for all methods in the five trials.

Then, the probabilities that Type-I errors occurred at least once were calculated (Table 3). The probability of Type-I errors for all trials using our new method was less than 5%, which satisfies the nominal significant level, as shown in Table 3. These results show that our new method is appropriate as a multiple comparison procedure.

We compared our new method with the conventional ones (Dunnett, Steel, Tukey and Steel-Dwass multiple comparisons). The simulation results indicate that our new method shows a higher power of tests than other conventional multiple comparison methods, except for Trial-4. Although only Trial-4 shows a slightly lower power of tests than the Steel-Dwass method, the values are almost the same.

Furthermore, the Bonferroni and Šidák methods, which can be applied as widely as our new method, both show much lower probabilities of Type-I errors. Only in the case of Trial-5 do both conventional methods show similar results. The reason for such small differences between our new method and the two conventional methods for Trial-5 is thought to be that two or more regression coefficients were rarely rejected at the same time under the condition of the artificial data generated and used in this study. In this trial, we have to mention that the adjustment of our new method led to correct results, and that these excellent results were obtained because our new method is the one which adjusted the significant level more accurately than any other similar method, such as the Šidák method (Table 2).

SR2 [6] was applied to the data used in Trial-1 and Trial-2 (Table 3). The results indicate that the power of test of SR2 was less than that of our new method on average. The reason for this difference is thought to be due to the fact that SR2 is based on only one simulation result, whereas our method is based on many simulation results.

<table>
<thead>
<tr>
<th>Case*</th>
<th>New</th>
<th>Bonferroni</th>
<th>Šidák</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>481</td>
<td>471</td>
<td>480</td>
</tr>
<tr>
<td>B</td>
<td>500</td>
<td>488</td>
<td>497</td>
</tr>
</tbody>
</table>

*: Row A shows the number of samples in which at least one or more null hypotheses were rejected. Row B shows the number of tests in which the null hypothesis was rejected.

Table 2. Number of Rejected Data Sets

<table>
<thead>
<tr>
<th>Trial*</th>
<th>New</th>
<th>SR2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.87</td>
<td>4.93</td>
</tr>
<tr>
<td>2</td>
<td>4.93</td>
<td>4.74</td>
</tr>
</tbody>
</table>

*: Each row in the table corresponds to the case applying SR2 to the data used in Trial-1 and Trial-2. The results in the “New” column are the same as those shown Table 1.
Table 4. Sample Size, Average and Variance of Control and Dosage Groups

<table>
<thead>
<tr>
<th>Dose</th>
<th>Size</th>
<th>Average</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (control)</td>
<td>20</td>
<td>31.08</td>
<td>7.41</td>
</tr>
<tr>
<td>5</td>
<td>19</td>
<td>27.19</td>
<td>21.86</td>
</tr>
<tr>
<td>50</td>
<td>18</td>
<td>28.60</td>
<td>16.75</td>
</tr>
<tr>
<td>500</td>
<td>17</td>
<td>27.65</td>
<td>25.75</td>
</tr>
</tbody>
</table>

We checked the applicability of our new procedure to a clinical data set. The basic statistics of the data are shown in Table 4. Since these significance tests deal with dose-response relationships, the order of the dosage groups (0, 5, 50, and 500) has to be considered. Therefore, the rejections of the null hypotheses of Bonferroni, Šidák, and our new procedures were carried out from the 500 to the 5 dosage groups, successively. If the null hypothesis is not rejected, the remaining hypotheses cannot be rejected. Whereas the Bonferroni and Šidák methods rejected the null hypothesis only for 500 dosage group, the conventional multiple comparison method (Williams method) and our new procedure rejected all null hypotheses (Table 5). The results indicate that the power of tests of our new procedure is higher than both the Bonferroni and Šidák methods and as high as the Williams method.

Table 5. Adjusted Significant Level and Results of Multiple Comparisons

<table>
<thead>
<tr>
<th>Method</th>
<th>Adj. α</th>
<th>500</th>
<th>50</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Williams</td>
<td>-</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Bonferroni</td>
<td>0.0167</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Šidák</td>
<td>0.0170</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>New</td>
<td>0.0202</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

*: Each column shows the test results and the adjusted α. Since the Williams method does not adjust α, “Adj. α” in row “Williams” is not shown. “*” indicates that the null hypothesis was rejected.

Conclusions

A new procedure that adjusts significant levels for tests of multiple comparisons was developed using the resampling method, which simulates the probability of Type-I errors. Our new method adjusted the significant level so that the probability that Type-I errors occurring at least once or more is equal to the nominal significant level. Therefore, single comparisons can be repeated to solve multiple comparison problems using our new methods. This new procedure is not only as appropriate as any of the other multiple comparison methods, but it is also a procedure which shows stable power that is not affected by conditions such as the type of test, statistics, and distribution of data. Our new method is applicable to various types of multiple comparison problems and we believe that it can be used by many practitioners.

Acknowledgments

This research was partially supported by the Ministry of Education, Science, Sports and Culture Grant-in-Aid for Scientific Research (C), 15590042, 2003.

References

標本再抽出法を用いた医薬学、化学データ解析のための新規多重比較法の開発

高木達也1,2*、岡本晃典1、横川由美子3、横田雅彦1、黒川顕2,4、
安永照雄2

1 大阪大学大学院 薬学研究科、大阪府吹田市山田丘1-6
2 大阪大学 遺伝情報実験センター、大阪府吹田市山田丘3-1
3 大阪大学 薬学部、大阪府吹田市山田丘1-6
4 大阪大学 微生物病研究所、大阪府吹田市山田丘3-1

統計学を専門としない実験生物学者が適切な統計手法を選択することは、さほど容易ではない。とりわけ多重比較法が問題になるケースでは、多種多様な手法が提案されていることもあり、困難を極めることも少なくない。そのため、より汎用的で検出力に優れる多重比較法の開発は、長らく望まれてきた。今回我々は、標本再抽出法を利用し、公称の有意水準を維持するように実質の有意水準を調節する、エンドユーザーが複雑な選択を必要としなくても統一的に利用が可能な、汎用性の高い多重比較検定アルゴリズムを開発した。この手法を用いれば、エンドユーザーは、Dunnett法のような従来の多様の手法に惑わされることなく、容易に多重比較を行うことができるようになる。計算機シミュレーション及び疫学データへの応用により、今回のアルゴリズムの検出力を従来の幾つかの多重比較法と比較したところ、満足すべき結果を得た。

キーワード：多重比較、有意性検定、標本再抽出法、ブートストラップ法

*ttakagi@phs.osaka-u.ac.jp