Biometrical Journal **45** (2003) 3, 325–334

Pairwise Comparisons in Each of Several Groups with Heterogenous Group Variances

S. H. CHEUNG[1,*], K. H. WU[1], and A. L. QUEK[2]

[1] Department of Statistics, Chinese University of Hong Kong, Shatin, Hong Kong, China
[2] Department of Statistics and Applied Probability, National University of Singapore, Singapore, Science Drive 2, Singapore 117543

Summary

Pairwise comparison procedures are frequently applied to analyze experimental results. In particular, practitioners in the area of medical researches often encounter situations which require these statistical techniques to compare various treatments. In this article, we focus on pairwise comparison procedures in a two-factor design, where comparisons of one factor are made simultaneously for each level of another factor. For example, several new drugs to treat a certain cancer are being compared for both male and female patients. Previous research efforts were mainly devoted to models with homogeneous variances. The current paper is to address more common scenario where group variances are heterogeneous.

Key words: Heterogenous Group Variances; Familywise Type I error; Multiple Comparisons.

1. Introduction

TUKEY's (1953) procedure has been widely used to conduct pairwise multiple comparisons in an one-factor balanced design. For unbalanced designs, Tukey's procedure was extended by UUSIPAIKKA (1985), SPURRIER and ISHAM (1985), and HAYTER (1989).

In a two-factor design (say factors A and B), simultaneous pairwise comparison techniques were developed by CHEUNG and CHAN (1996). We refer to their procedure as the *CC* procedure. In such cases, pairwise comparisons of one factor are made simultaneously for each level of another factor, at the same time maintaining a designated overall Type I error rate. For illustration, let us consider the following example. Five new drugs (factor A) are available to treat a certain cancer. Both male and female patients (factor B: Sex) are selected as subjects for the experiment. If there is interaction effect between factors A and B, it is reasonable to perform pairwise comparisons of drugs separately for male and female patients, yielding altogether 20 comparisons. As argued by CHEUNG and HOLLAND (1994),

* Corresponding author: b1@sparc2.sta.cuhk.edu.hk

 0323-3847/03/0304-0325 $ 17.50+.50/0

from the perspective of a particular patient, the relevant family of hypotheses probably consists only of those relating to the patient's gender, but from the perspective of a medical researcher who is equally interested in treatment efficacy for both genders, the relevant family may consist of all pairwise comparisons within either gender.

The methodology given in CHEUNG and CHAN (1996) only applies to homogeneous variances models. That is, with the above example, we are assuming that the variances among all patients (regardless of their gender) are the same. However, this assumption is not valid in many practical situations. With reference to the earlier example, the variation of treatment responses may appear to be quite different for patients with opposite genders. Hence, we propose a procedure which is appropriate for pairwise comparisons of treatments simultaneously for groups with different variances.

Hereafter, pairwise comparison procedures for a two-way design are conducted based on the following two-factor (factors A and B, say) fixed effect model:

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk}; \qquad i = 1, \ldots, r, \quad j = 1, \ldots, c, \quad k = 1, \ldots, n_{ij}, \qquad (1.1)$$

where there are r groups (treatment levels) in factor A, each group consisting of c treatments in factor B. Let Y_{ijk} be the kth observation of the response variable at the jth level of factor B and the ith level of factor A. Furthermore, let μ_{ij} and n_{ij} be the corresponding mean and sample size. For the *CC* procedure, ϵ_{ijk} are assumed to have independent $N(0, \sigma^2)$ distributions. However, in the current article, we will assume that ϵ_{ijk} are distributed independently as $N(0, \sigma_i^2)$. In other words, the variances are allowed to be different for different groups.

In practice, to assure that the *CC* procedure is appropriate to analyze data from a two-factor design, we need to test the null hypothesis that all the group variances are the same (H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_r^2$). One suitable candidate to test the above null hypothesis is the popular Bartlett's test (see KOLOSKA and NEVISON, 1989). If the null hypothesis is rejected, the proposed procedure in this article will be appropriate.

In Section 2, we briefly describe the *CC* procedure, indicating the inappropriateness of this procedure for groups with heterogeneous variances. Then, in Section 3, the test statistics for our proposed procedure will be derived. Afterwards, the evaluation of critical constants required to implement pairwise comparisons will be outlined. An approximation method is provided. Power of the test is discussed in Section 4. Finally, a numerical example illustrating our proposed procedure is given in Section 5.

2. The *CC* Procedure

In this section, we first review the *CC* procedure. Then the problem of the *CC* procedure in heterogeneous-group-variance environments will be examined.

The *CC* procedure is as follows. Denote the sample mean of treatment j in group i by $\bar{Y}_{ij}$ which is equal to $\sum_{k=1}^{n_{ij}} Y_{ijk}/n_{ij}$ where $Y_{ijk} \sim N(\mu_{ij}, \sigma^2)$. Let $\hat{\sigma}^2$ be the unbiased estimates of σ^2, independent of $\bar{Y}_{ij}$ and $\nu\hat{\sigma}^2/\sigma^2 \sim \chi^2$ with ν degrees of freedom. In addition, let $\boldsymbol{N} = (\boldsymbol{n}_1, \ldots, \boldsymbol{n}_r)'$ where $\boldsymbol{n}_i = (n_{i1}, \ldots, n_{ic})'$, $i = 1, \ldots, r$. Thus, $\boldsymbol{n}_i$ contains all the information about the sample sizes in group i.

With reference to model (1.1), to conduct pairwise multiple comparisons of the c treatment levels in each of the r groups, the inference problem under consideration is the $rc(c-1)/2$ simultaneous interval estimation of:

$$\mu_{ij_1} - \mu_{ij_2} \tag{2.1}$$

for all $i = 1, \ldots, r$ and $1 \le j_1 \ne j_2 \le c$, or the simultaneous testing of the $rc(c-1)/2$ hypotheses:

$$H_{i(j_1,j_2)} : \mu_{ij_1} = \mu_{ij_2} \tag{2.2}$$

against the two-sided alternatives:

$$H'_{i(j_1,j_2)} : \mu_{ij_1} \ne \mu_{ij_2}$$

for all $i = 1, \ldots, r$ and $1 \le j_1 \ne j_2 \le c$. To test the null hypotheses (2.2), the test statistics are:

$$D_{i(j_1,j_2)} = \frac{\bar{Y}_{ij_1} - \bar{Y}_{ij_2}}{\hat{\sigma}\sqrt{1/n_{ij_1} + 1/n_{ij_2}}}$$

for all $i = 1, \ldots, r$ and $1 \le j_1 \ne j_2 \le c$. Note that $\mu_{ij_1} - \mu_{ij_2} = 0$ when the null hypotheses are true.

Performing the estimation and testing procedures for Model (1.1) requires the computation of critical value $d_{(\alpha, r, c, \nu, \boldsymbol{N})}$. If the familywise type I error rate (the probability of committing at least one type I error) is controlled at a predetermined value α, the critical value is computed by solving the following equation:

$$P(|D_{i(j_1,j_2)}| \le d_{(\alpha,r,c,\nu,\boldsymbol{N})}; 1 \le j_1 \ne j_2 \le c,\ i = 1, \ldots, r) = 1 - \alpha\,.$$

To compute the value of $d_{(\alpha, r, c, \nu, \boldsymbol{N})}$, we can employ the Fortran coded subroutine (TPMC) which is provided in STATLIB electronic bulletin board (Web-page address: http://lib.stat.cmu.edu/). Nevertheless, one can also apply the improved algorithm given by Bretz, Hayter and Genz (2001) even though the formal algorithm is sufficient for practical purposes.

There have been a couple of extensions of the *CC* procedure. Cheung and Zhu (1998) developed techniques to perform one-sided pairwise comparisons for two-way designs. The *CC* procedure is also generalized to cases where covariance is included in the model. (Wong and Cheung, 2000).

Next we examine the problem of the *CC* procedure when group variances are not homogeneous.

In order to demonstrate the inappropriateness of the *CC* procedure in heterogeneous group variances situations, a simulation study was performed. Assume that group variances are heterogeneous. Therefore, $Y_{ijk} \sim N(\mu_{ij}, \sigma_i^2)$. Familywise error rate was calculated based on simulations with 100,000 replications. For the entire simulation study, $\mu_{ij} = 0$ for all i, j. For group variances σ_i^2, let $\sigma_1^2 = 1$ and when $i = 2, \ldots, r$, let $\sigma_i = \delta\sigma_{i-1}$ with $\delta = 1(0.1)3$. Note that $\delta = 1$ corresponds to the homogeneous variances case. In addition, the simulation study was repeated for equal sample sizes and selected values of r and c, namely, $r = 2, 3$ and $c = 3, 4, 5$. All tests were performed with a pre-assigned familywise error rate 0.05. The simulated familywise error rates are plotted in Figure 1.

As shown in Figure 1, the error rate is always larger than the nominal value, indicating that the *CC* procedure is unable to control familywise error rate at the pre-assigned level (0.05) when group variances are heterogeneous ($\delta > 1$). The departure from 0.05 is quite alarming especially when δ is large (for instance, familywise error rate is 0.3649 when $\delta = 3,\ r = 3,\ c = 5$).

There is another important feature which can be observed in Figure 1. For heterogeneous group variances, we found that the simulated familywise error rate

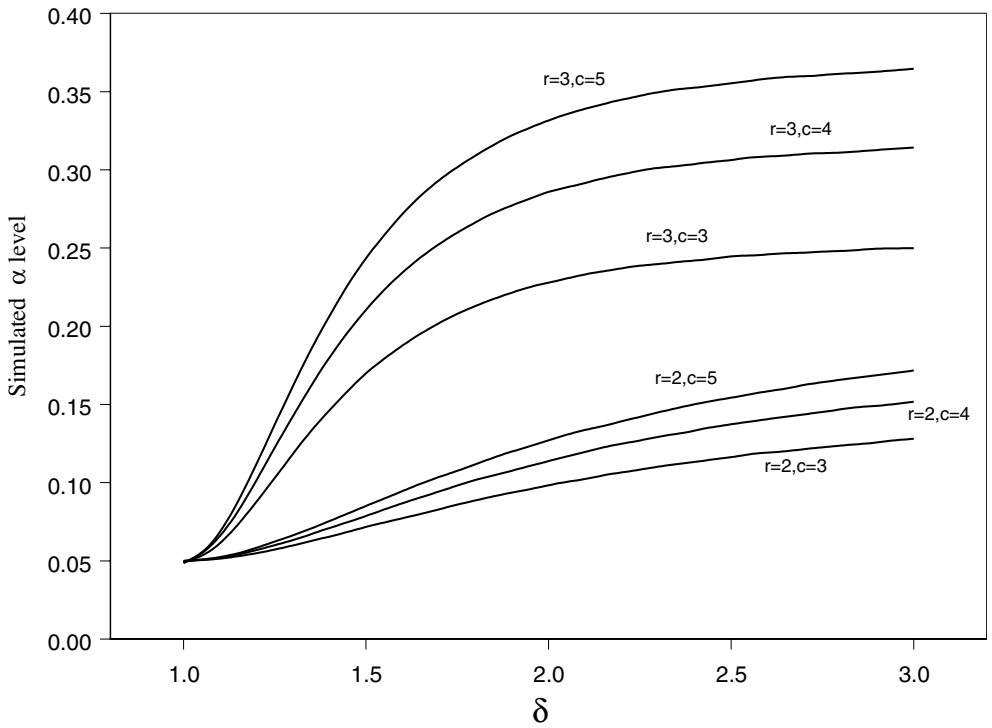

Fig. 1. Simulated familywise error rate for *CC* procedure with $\alpha = 0.05$

increases as either r or c increases. Therefore, the number of treatments and/or number of groups are large, the problem of inflated familywise type I error becomes more disturbing.

3. The New Procedure

In response to the incapability of the *CC* procedure to control familywise error rate under heterogeneous group variances models, we propose a new testing method.

First of all, let $Y_{ijk} \sim N(\mu_{ij}, \sigma_i^2)$. For $i = 1, \ldots, r$, let $\hat{\sigma}_i^2$ be the unbiased estimates of σ_i^2, independent of $\bar{Y}_{ij}$ and $\nu_i \hat{\sigma}_i^2 / \sigma_i^2 \sim \chi^2$ with ν_i degrees of freedom. Usually, $\hat{\sigma}_i^2$ is the pooled sample variance s_i^2 of group i.

The newly proposed test statistics are:

$$T_{i(j_1, j_2)} = \frac{\bar{Y}_{ij_1} - \bar{Y}_{ij_2}}{\hat{\sigma}_i \sqrt{1/n_{ij_1} + 1/n_{ij_2}}} \tag{3.1}$$

for all $i = 1, \ldots, r$ and $1 \le j_1 \ne j_2 \le c$.

Performing the estimation and testing procedures with the above statistics requires the computation of r critical values $t_{i,(\alpha, r, c, \nu_i, \boldsymbol{n}_i)}$. If the familywise type I error rate (the probability of committing at least one type I error) is controlled at a pre-determined value α, these r critical values are computed by solving the following equation:

$$P(|T_{i(j_1, j_2)}| \le t_{i(\alpha, r, c, \nu_i, \boldsymbol{n}_i)};\ 1 \le j_1 \ne j_2 \le c,\ i = 1, \ldots, r) = 1 - \alpha\,. \tag{3.2}$$

To compute the values of $t_{i(\alpha, r, c, \nu_i, \boldsymbol{n}_i)}$, we first note that since the overall pooled sample variance is not being used, the pivotal statistics $T_{i_1(j_1, j_2)}$ and $T_{i_2(j_3, j_4)}$ in (3.1) for $1 \le i_1 \ne i_2 \le r$; $1 \le j_1 \ne j_2$, $j_3 \ne j_4 \le c$, are independent from one another. Hence equation (3.2) can be rewritten as:

$$\prod_{i=1}^{r} P_i = 1 - \alpha\,,$$

where

$$P_i = P(|T_{i(j_1, j_2)}| \le t_{i(\alpha, r, c, \nu_i, \boldsymbol{n}_i)};\ 1 \le j_1 \ne j_2 \le c)\,. \tag{3.3}$$

Note that $1 - P_i$ represents the overall Type I error probability for testing all the hypotheses in group i. Without any prior information, we adopt the reasonable approach that the same Type I error rate is being assigned to each of the r groups. Consequently, from equation (3.3), we have for the ith group:

$$P_i = (1 - \alpha)^{1/r}\,.$$

Therefore, each $t_{i(\alpha, r, c, \nu_i, \boldsymbol{n}_i)}$ can be obtained by solving the equation

$$P(|T_{i(j_1, j_2)}| \le t_{i(\alpha, r, c, \nu_i, \boldsymbol{n}_i)};\ 1 \le j_1 \ne j_2 \le c) = (1 - \alpha)^{1/r}\,.$$

From the above two equations, we have

$$t_{i(\alpha_i, 1, c, \nu_i, \boldsymbol{n}_i)} = t_{i(\alpha, r, c, \nu_i, \boldsymbol{n}_i)}\,, \tag{3.4}$$

where

$$\alpha_i = 1 - (1 - \alpha)^{1/r} \tag{3.5}$$

for $i = 1, \ldots, r$.

With Equations (3.4) and (3.5), one can evaluate $t_{i(\alpha, r, c, \nu_i, \boldsymbol{n}_i)}$ as follows. With given α, compute α_i by Equation (3.5) and evaluate $t_{i(\alpha_i, 1, c, \nu_i, \boldsymbol{n}_i)}$ accordingly.

Now, let us consider subsequent simultaneous estimation and testing procedures. Following the computation of r critical values $t_{i(\alpha, r, c, \nu_i, \boldsymbol{n}_i)}$, the set of $100(1 - \alpha)\%$ simultaneous pairwise confidence intervals for (2.1) are:

$$(\bar{Y}_{ij_1} - \bar{Y}_{ij_2}) \pm t_{i(\alpha, r, c, \nu_i, \boldsymbol{n}_i)} \hat{\sigma}_i \sqrt{1/n_{ij_1} + 1/n_{ij_2}}$$

for all $i = 1, \ldots, r$ and $1 \le j_1 \ne j_2$, where the joint coverage probability of the $rc(c - 1)/2$ intervals is $1 - \alpha$.

With respect to simultaneous hypotheses testing, if the familywise type I error rate is α, each null hypothesis stated in (2.2) is rejected in favor of the two-sided alternative if and only if

$$\frac{|\bar{Y}_{ij_1} - \bar{Y}_{ij_2}|}{\hat{\sigma}_i \sqrt{1/n_{ij_1} + 1/n_{ij_2}}} > t_{i(\alpha, r, c, \nu_i, \boldsymbol{n}_i)}\,.$$

In many medical, biological and physiological researches, the disparities of sample sizes of the treatment levels in each group are usually small if they are not equal. In such cases, rather than applying the subroutine to compute the critical values, we can adopt the approximation method which normally yields very reasonable results (CHEUNG and CHAN, 1996).

In order to apply the approximation method, we need to derive the r approximate values for $t_{i(\alpha, r, c, \nu_i, \boldsymbol{n}_i)}$ in each group. By the Tukey-Kramer approximation technique (TUKEY, 1953; KRAMER, 1956), we have

$$t_{i(\alpha, r, c, \nu_i, \boldsymbol{n}_i)} \approx \frac{Q_{i(c, \nu_i, \alpha_i)}}{\sqrt{2}}$$

for $i = 1, \ldots, r$ where $Q_{i(c, \nu_i, \alpha_i)}$ is the upper α_i percentage point of the Studentized Range Distribution with parameters c, ν_i and α_i.

Hence, with the r approximate values, the approximate $100(1 - \alpha)\%$ level simultaneous pairwise confidence intervals for (2.1) are:

$$(\bar{Y}_{ij_1} - \bar{Y}_{ij_2}) \pm \frac{Q_{i(c, \nu_i, \alpha_i)}}{\sqrt{2}}\ \hat{\sigma}_i \sqrt{1/n_{ij_1} + 1/n_{ij_2}}$$

for all $i = 1, \ldots, r$ and $1 \le j_1 \ne j_2$, where the joint coverage probability of the $rc(c - 1)/2$ intervals is at least $1 - \alpha$. This conservative property of the approxi-

mation procedure was proved by HAYTER (1984). In accordance, with familywise type I the two-sided alternative in (2.2) will be rejected if and only if

$$\frac{|\bar{Y}_{ij_1} - \bar{Y}_{ij_2}|}{\hat{\sigma}_i \sqrt{1/n_{ij_1} + 1/n_{ij_2}}} > \frac{Q_{i(c,\nu_i,\alpha_i)}}{\sqrt{2}} .$$

4. Power

The evaluation of power is important not only for the assessment of the rejection capability of the testing procedure when the null is false, but also for the determination of sample size. In multiple comparisons, the definition of power is more complex. Here, two popular notions of power introduced by RAMSEY (1978) will be outlined. The first one is the any-pair power which is the probability of detecting at least one true difference among all pairs. The second one is the all-pairs power which is the probability of detecting all true differences among all pairs.

To compute power, we follow the ideas given in BIESHEUVEL and HOTHORN (2002). Let S be the subset of $\{i(j_1, j_2)\}$ such that the null hypotheses $H_{i(j_1,j_2)}$ are false. Then the any-pair power is given as

$$\begin{aligned}
\text{Power}_{\text{any-pair}} &= P(|T_{i(j_1,j_2)}| \geq t_{i(\alpha,r,c,\nu_i,\boldsymbol{n}_i)} \; \exists i(j_1, j_2) \in S) \\
&= 1 - P(|T_{i(j_1,j_2)}| < t_{i(\alpha,r,c,\nu_i,\boldsymbol{n}_i)} \; \forall i(j_1, j_2) \in S) \\
&= 1 - P\left[\left|\frac{(\bar{Y}_{ij_1} - \bar{Y}_{ij_2}) - (\mu_{ij_1} - \mu_{ij_2})}{\hat{\sigma}_i \sqrt{1/n_{ij_1} + 1/n_{ij_2}}} + \delta_{i(j_1,j_2)} \; \frac{\sigma_i}{\hat{\sigma}_i}\right| \right. \\
&\qquad \left. < t_{i(\alpha,r,c,\nu_i,\boldsymbol{n}_i)} \; \forall i(j_1, j_2) \in S\right],
\end{aligned}$$

where

$$\delta_{i(j_1,j_2)} = \frac{\mu_{ij_1} - \mu_{ij_2}}{\sigma_i \sqrt{1/n_{ij_1} + 1/n_{ij_2}}}$$

are the non-central parameters. On the other hand, the all-pair power is given as

$$\begin{aligned}
\text{Power}_{\text{all-pair}} = \; & P(|T_{i(j_1,j_2)}| \geq t_{i(\alpha,r,c,\nu_i,\boldsymbol{n}_i)} \; \forall i(j_1, j_2) \in S) \\
&= P\left[\left|\frac{(\bar{Y}_{ij_1} - \bar{Y}_{ij_2}) - (\mu_{ij_1} - \mu_{ij_2})}{\hat{\sigma}_i \sqrt{1/n_{ij_1} + 1/n_{ij_2}}} + \delta_{i(j_1,j_2)} \; \frac{\sigma_i}{\hat{\sigma}_i}\right| \right. \\
&\qquad \left. \geq t_{i(\alpha,r,c,\nu_i,\boldsymbol{n}_i)} \; \forall i(j_1, j_2) \in S\right].
\end{aligned}$$

The computation of the above probability involves the evaluation of non-central multivariate t-distributions. The algorithm given in BRETZ, HAYTER, and GENZ (2001) can be employed to complete the task.

5. Example

The example supplied in this section is extracted from the research undertaken by WARD, YUAN, CHEUNG and THOMPSON (2001). One objective of the study is to determine whether different diets of female rats affect their bone mineral content (BMC). The rats were divided into two groups $(r = 2)$. For the first group, the rats were exposed to secoisolariciresinol diglycoside (SDG) purified from flaxseed during suckling continuously to adolescence (postnatal day (PND) 50). While for the second group, the rats were exposed to SDG purified from flaxseed during suckling continuously to adult hood (PND 132). There were 5 different diets $(c = 5)$ for each group of rats. The treatments were:

1. BD-BD: Basal diet
2. 50S-DD: received 50S diet during lactation but from the end of lactation received BD
3. 50S-50S: received 50S diet during lactation through to PND 50 or 132
4. 100S-DD: received 100S diet during lactation but from the end of lactation received BD
5. 100S-100S: received 100S diet during lactation through to PND 50 or 132

where 50S and 100S were diets supplemented with different amount of SDG.

The data of the experiment are summarized in Table 1. A Bartlett's test was conducted to test for equality of error variances under the five diets in each of the two groups. After ascertaining that the variances are not significantly different at 0.05 level from one another within each group, we further conduct a test on the equality of group variances. With a p-value of less than 0.001, we conclude that the variances σ_i^2 for $i = 1, 2$ are significantly different. In fact, the sample variances $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are very different ($\hat{\sigma}_1^2 = 256.0$, $\hat{\sigma}_2^2 = 908.2$). Hence, the *CC* procedure is definitely not an appropriate tool for pairwise comparison.

From Table 1, it is obvious that the BMC of the BD-BD group is extremely high for the PND 132 group, suggesting that the existence of interaction effect between the two factors. In addition, a positive confirmation (p-value < 0.05) is also obtained from a test of interaction effect, assuming heterogeneous group variances (for the testing procedure, see for example: CHATTERJEE, HADI, and PRICE (2000)). Therefore pairwise multiple comparison of the diet effect should be con-

Table 1
Mean BMC of female rats fed on different diets at postnatal day (PND) 50 and 132

Group i	Diets				
	BD-BD	50S-BD	50S-50S	100S-BD	100S-100S
PND 50 ($i = 1$)	183 (8)	171 (10)	184 (8)	177 (9)	170 (11)
PND 132 ($i = 2$)	456 (7)	411 (7)	401 (10)	403 (9)	409 (8)

Number of female rats is given in parenthesis. The unit of BMC is mg.
$\hat{\sigma}_1^2 = 256.0$ and $\hat{\sigma}_2^2 = 908.2$.

Table 2
95% Joint confidence intervals for pairwise comparisons of mean BMC

	Group i	
	PND 50 ($i = 1$)	PND 132 ($i = 2$)
$\mu_1 - \mu_2$	(−11.73, 35.73)	(−5.74, 95.74)
$\mu_1 - \mu_3$	(−26.02, 24.02)	(8.22, 101.78)
$\mu_1 - \mu_4$	(−18.31, 30.31)	(5.16, 100.84)
$\mu_1 - \mu_5$	(−10.25, 36.25)	(−2.13, 96.13)
$\mu_2 - \mu_3$	(−36.73, 10.73)	(−36.78, 56.78)
$\mu_2 - \mu_4$	(−28.99, 16.99)	(−39.84, 55.84)
$\mu_2 - \mu_5$	(−20.86, 22.86)	(−47.13, 51.13)
$\mu_3 - \mu_4$	(−17.31, 31.31)	(−45.62, 41.62)
$\mu_3 - \mu_5$	(−9.25, 37.25)	(−53.03, 37.03)
$\mu_4 - \mu_5$	(−15.49, 29.49)	(−52.13, 40.13)

μ_{i1} = mean BMC of female rats fed on diet 'BD-BD' in group i.
μ_{i2} = mean BMC of female rats fed on diet '50S-BD' in group i.
μ_{i3} = mean BMC of female rats fed on diet '50S-50S' in group i.
μ_{i4} = mean BMC of female rats fed on diet '100S-BD' in group i.
μ_{i5} = mean BMC of female rats fed on diet '100S-100S' in group i.

ducted simultaneously for each group of rats, instead of using treatment means averaged across the groups.

For this numerical example, to construct simultaneous pairwise confidence intervals at $\alpha = 0.05$, one has to compute the four critical values $t_{1(0.05,2,5,41,\boldsymbol{n_1})}$ and $t_{2(0.05,2,5,36,\boldsymbol{n_2})}$ where $\boldsymbol{n}_1 = (8\ 10\ 8\ 9\ 11)$ and $\boldsymbol{n}_2 = (7\ 7\ 10\ 9\ 8)$. Furthermore, $\alpha_i = 0.0253,\ i = 1, 2$. Using our computer subroutine, the values computed are respectively 3.127 and 3.150. With the approximation method, these values change to 3.131 and 3.153. Note that the approximation values are very close to the exact values. However, they are all larger than the corresponding exact values, demonstrating the conservative properties of the approximation method.

With the exact values and follow the formula provided in Section 3, the set of 95% simultaneous pairwise confidence intervals are given in Table 2. We conclude that the mean BMC of female rats fed on the diet 'BD-BD' differs significantly from that of the diet '50S-50S' and '100S-BD' for Group 2 (PND 132) since their corresponding intervals do not contain zero. Conducting simultaneous hypotheses testing will lead to the same result because simultaneous pairwise intervals and testing procedures introduced in this article are in fact equivalent procedures.

Acknowledgements

The authors wish to thank two anonymous referees for several helpful comments and suggestions. This work was supported by a Hong Kong RGC Research Direct Grant 2060222.

References

BIESHEUVEL, E. and HOTHORN, L. A., 2002: Many-to-one comparisons in stratified designs. *Biometrical Journal* **44**, 101–116.

BRETZ, F., HAYTER, A. J., and GENZ, A., 2001: Critical point and power calculations for the studentized range test for generally correlated means. *Journal of Statistical Computation and Simulation* **71**, 85–99.

CHATTERJEE, S., HARDI, A. S., and PRICE, B., 2000: *Regression Analysis by Example*. Wiley: New York.

CHEUNG, S. H. and CHAN, W. S., 1996: Simultaneous confidence intervals for pairwise multiple comparisons in a two-way unbalanced design. *Biometrics* **52**, 463–472.

CHEUNG, S. H. and HOLLAND, B., 1994: A step-down procedure for multiple tests of treatment versus control in each of several groups. *Statistics in Medicine* **13**, 2261–2267.

CHEUNG, S. H. and ZHU, H. T., 1998: Simultaneous one-sided pairwise comparisons in a two-way design. *Biometrical Journal* **40**, 613–625.

HAYTER, A. J., 1984: A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative. *The Annals of Statistics* **12**, 61–75.

HAYTER, A. J., 1989: Pairwise comparisons of generally correlated means. *Journal of the American Statistical Association* **84**, 208–213.

KOKOSKA, S. and NEVISON, C., 1989: *Statistical Tables and Formulae*. Springer-Verlag: New York.

KRAMER, C. Y., 1956: Extension of multiple range test to group means with unequal numbers of replications. *Biometrics* **12**, 307–310.

RAMSEY, P. H., 1978: Power differences between pairwise multiple comparisons. *Journal of the American Statistical Association* **73**, 479–485.

SPURRIER, J. D. and ISHAM, S. P., 1985: Exact simultaneous confidence intervals for pairwise comparisons of three normal means. *Journal of the American Statistical Association* **80**, 438–442.

TUKEY, J. W., 1953: *The problem of multiple comparisons*. Unpublished report, Princeton University, New Jersey.

UUSIPAIKKA, E., 1985: Exact simultaneous confidence intervals for multiple comparisons among three or four mean values. *Journal of the American Statistical Association* **80**, 196–201.

WARD, W. E., YUAN, Y. V., CHEUNG, A. M., and THOMPSON, L. U., 2001: Exposure to purified lignan from flaxseed (Linum usitatissimum) alters bone development in female rats. *British Journal of Nutrition* **86**, 499–505.

WONG, Y. W. and CHEUNG, S. H., 2000: Simultaneous pairwise multiple comparisons in a two-way analysis of covariance model. *Journal of Applied Statistics* **27**, 281–291.

Received January 2002
Revised October 2002
Accepted November 2002