Journal of Korean
Data & Information Science Society
2007, Vol. 18, No. 2, pp. 525 ~ 534

Statistically Proper Multiple Range Tests for a Within Subject Factor in a Repeated Measures Design

Cheolyong Park[1] · Sangbum Park[2]

Abstract

It is a common practice in many research areas that multiple range tests for a between subject factor such as Tukey are applied to a within subject factor in a repeated measures design. Tukey procedure, however, sometimes detects no pairs with different means even when the hypothesis of all equal level means is rejected. This study attempts to provide a rationale for the proposition that Tukey is inappropriate post hoc procedure for a within subject factor in which the observations are correlated. We introduce two multiple range tests, Bonferroni and Scheffe, for a within subject factor and show that Bonferroni is more appropriate than Scheffe for pairwise multiple comparisons. Subsequent simulation study indicates that Tukey has significantly less power than Bonferroni in detecting actual difference between means of some pairs when the observations of a within subject factor are highly correlated.

Keywords : Bonferroni Procedure, Multiple Range Test, Repeated Measures Design, Tukey Procedure

1. Introduction

In many research areas, statistical methodologies are employed as a tool to provide an objective evidence for statistical hypotheses. However, some statistical methodologies are not easy to understand and are too complicated to apply in practice. Accordingly, researchers tend to apply easy-to-use alternatives to a situation where they might not be statistically correct. For example, while multivariate ANOVA (analysis of variance) is appropriate for testing equality of

[1] Associate Professor, Department of Statistics, Keimyung University, Taegu 704-701
E-mail: cypark1@kmu.ac.kr

[2] Associate Professor, Department of Physical Education, Keimyung University, Taegu 704-701. E-mail: sbpark@kmu.ac.kr

mean vectors, it is based on a complicated matrix algebra and does not lead to a unique procedure for the test (see sections 8.4 and 8.6 of Anderson (1984) for details). Consequently, a sequence of univariate ANOVA is often used for testing equality of mean values of each component since it is quite familar to most researchers and provides a unique F-test.

Multiple range tests for within subject factors in a repeated measures design provide another example of such applications. Since there has not been much study on post-hoc test procedures for within subject factors, multiple range tests developed for a between subject factor such as Tukey are often applied to a within subject factor. For example, in a study on the coupling of eye, finger, elbow, and shoulder movements during aiming task, Helsen et al. (2000) analyzed the effect of limb component (finger, elbow, or shoulder) on various dependent variables using ANOVA with repeated measures and Tukey's HSD post hoc procedure. Weir et al. (1998) also used Tukey's HSD as post hoc test for a significant within subject factor. Examples like this can be easily found in studies employing repeated measures design (see Barthelemy and Boulinguez (2002), Harvey et al. (2002), and Teixeira (2000) among others).

If Tukey procedure is applied to a significant within subject factor, however, it sometimes detects no pairs with different means even when the hypothesis of all equal level means is rejected. In other words, the sensitivity of Tukey to actual difference between level means decreases when sample means are (positively) correlated. In a repeated measures design, the observations of a within subject factor are usually (positively) correlated because they are measured from the same subject. Therefore, application of Tukey or other post hoc test procedures for between subject factors to a significant within subject factor may lead to an incorrect interpretation of the result.

In this paper, we attempt to provide a rationale for our proposition that Tukey is inappropriate post hoc procedure for a within subject factor in which the observations are correlated. First, we introduce two multiple range test procedures, Bonferroni and Scheffe, for a within subject factor, which are statistically correct and easy to use. We propose to use Bonferroni for pairwise comparisons, e.g. comparison between level 1 and level 2, and Scheffe for contrasts, e.g. comparison between the average of levels 1 and 2, and level 3. Since we are interested in pairwise multiple comparisons for a significant within subject factor with more than three levels, Bonferroni seems to be better than Scheffe and we provide a numerical evidence for it.

Although not being optimal, we suggest to use Bonferroni for pairwise multiple comparisons because it is quite easy to use in practice. To improve the power for detecting actual difference between means of some pairs, a stepwise procedure called the false discovery rate criterion, suggested by Benjamini and Hochberg (1995) and later developed by Benjamini and Liu (1999) for generally correlated test statistics, might be employed. However, this procedure is not considered in this study since it is not for the family of all possible pairs.

Next, we provide empirical evidence based on a simulation study as well as theoretical justification for a better performance of Bonferroni than Tukey in detecting pairs with different means. Although our comparison uses Tukey as a competitor, the same story holds for other methods for a between subject factor. Tukey has been a natural competitor since it is a well known multiple comparison method for comparing all possible pairs (see p. 574 of Neter et al. (1985) for details). We show that as the correlation among the observations of a within subject factor increases, Tukey performs worse whereas Bonferroni performs equally well without regard to the correlation.

2. Suggested Multiple Range Tests: Bonferroni and Scheffe

Consider a repeated measures design with one within subject factor having k levels. In this design, a vector $(X_1, X_2, \cdots, X_k)$ of k levels is observed from each subject and the observations $X_1, X_2, \cdots, X_k$ are correlated. Suppose that n subjects are randomly selected for an experiment and let $(X_{i1}, X_{i2}, \cdots, X_{ik})$ be the vector of observations from the i-th subject for $i = 1, 2, \cdots, n$. Let μ_j be the j-th level mean and let $\overline{X_j} = \sum_{i=1}^{n} X_{ij}/n$ be the j-th sample mean for $j = 1, 2, \cdots, k$.

The first step of analysis is to test if all level means are equal, i.e. $\mu_1 = \cdots = \mu_k$. If the number of levels is two, then the paired t-test is usually used for testing the hypothesis. In other words, we conclude that $\mu_1 \neq \mu_2$ at significance level α if

$$|t_{12}| = \frac{|\overline{X_1} - \overline{X_2}|}{\sqrt{S_d^2/n}} > t(\alpha/2, n-1),$$

where $S_d^2 = \sum_{k=1}^{n} (D_k - \overline{D})^2/(n-1)$ is the sample variance of $D_k = X_{k1} - X_{k2}$'s and $t(\alpha, df)$ is the upper α-th quantile of the t distribution with degrees of freedom df. When the number of levels is more than two, Hotelling T^2 is usually employed for testing the hypothesis (see p. 227 of Johnson and Wichern (1992) for details).

When the hypothesis of all equal level means is rejected, we apply multiple range tests in order to find some deviations from the hypothesis. A natural choice of deviations from the hypothesis is of form $\mu_i \neq \mu_j$ for some $i \neq j$. In this case, we search for pairs with different means and we call this process as pairwise multiple comparisons. In other words, we search for all possible combinations $\binom{k}{2}$ of pairs to find some with different means. A more specific form of deviations can be expressed as a nonzero contrast; a contrast is a linear combination of the means $\sum_{i=1}^{k} c_i \mu_i$ with $\sum_{i=1}^{k} c_i = 0$. In general, a pair with different means can be expressed as a contrast. For example $\mu_1 \neq \mu_2$ can be expressed as

$$\sum_{i=1}^{k} c_i \mu_i \neq 0 \quad \text{with} c_1 = 1, c_2 = -1, c_3 = \cdots = c_k = 0 .$$

Note that there are infinite number of nonzero contrasts.

We now explain how Bonferroni and Scheffe are conducted. Since Bonferroni is not applicable when the family of interest is the set of infinite number of contrasts, we will focus on pairwise multiple comparisons. In other words, we search for pairs with different means among $m \equiv \binom{k}{2}$ possible pairs. Bonferroni concludes that $\mu_i \neq \mu_j$ at significance level α if

$$|t_{ij}| > t(\alpha/(2^* m), n-1) \qquad (2\text{-}1)$$

for any $i \neq j$, where t_{ij} is the paired t-test statistic for comparing the i-th and j-th levels of the factor. The only difference between the paired t-test and Bonferroni is that the significance level α is adjusted to $\frac{\alpha}{m} = \frac{\alpha}{\text{number of comparisons}}$ in Bonferroni: The false discovery rate criterion adjusts the significance level somewhat differently to improve the power for detecting actual difference between means of some pairs (see Benjamini and Hochberg (1995) and Benjamini and Liu (1999) for details). Therefore Bonferroni can not be used when there are infinite number of comparisons. Bonferroni can be expressed in terms of the p-value of the paired t-test; we conclude that $\mu_i \neq \mu_j$ at significance level α if the p-value is less than α/m. This is a convenient way of applying Bonferroni in practice when we have the p-value of the paired t-test from computer output.

On the other hand, Scheffe concludes that $\mu_i \neq \mu_j$ at significance level α if

$$|t_{ij}| > \sqrt{\frac{(n-1)(k-1)}{(n-k+1)} F(\alpha, k-1, n-k+1)}$$

for any $i \neq j$, where $F(\alpha, df1, df2)$ is the upper α-th quantile of F distribution with degrees of freedom df1 and df2 (see p. 227 of Johnson and Wichern (1992) for details). Table 1 shows the ratio of the table values of Bonferroni and Scheffe, given by

$$t(\alpha/(2^*m),n-1)/\sqrt{\frac{(n-1)(k-1)}{n-k+1}F(\alpha,k-1,n-k+1)},$$

for various k, n, and $\alpha=0.05$. Since the ratio is 1 for k=2, this case is not reported.

<Table 1> The ratio of table values of Bonferroni and Scheffe
(n=10, 20, 40, 60; k=3, 4, 5, 6, 7; $\alpha=0.05$)

n	k				
	3	4	5	6	7
10	0.9261	0.8216	0.7074	0.5865	0.4581
20	0.9583	0.8992	0.8398	0.7824	0.7274
40	0.9693	0.9245	0.8808	0.8403	0.8029
60	0.9725	0.9315	0.8920	0.8558	0.8227

Scheffe can be used for searching infinite number of contrasts including pairwise comparisons, and thus the table value of Scheffe, as expected, is bigger than that of Bonferroni. This means that Bonferroni has more power for detecting pairs with different means than Scheffe. Since we focus on pairwise multiple comparisons for a significant within subject factor, we will use Bonferroni from now on.

We finally explain how Tukey is conducted. This method assumes that the observations $X_1,X_2,\cdots,X_k$ are independent and searches for pairs with different means after the hypothesis of all equal level means is rejected via the usual one-way ANOVA table. More specifically, Tukey concludes that $\mu_i\neq\mu_j$ at significance level α if

$$|T_{ij}|=\frac{|\overline{X_i}-\overline{X_j}|}{\sqrt{2MSE/n}}>\frac{q(\alpha,k,k(n-1))}{\sqrt{2}}, \qquad (2\text{-}2)$$

where $MSE=\sum_{i=1}^{n}\sum_{j=1}^{k}(X_{ij}-\overline{X_j})^2/(k(n-1))$ is the mean square error of the ANOVA table and $q(\alpha,df1,df2)$ is the upper α-th quantile of the studentized range distribution with degrees of freedom df1 and df2 (see p. 575 of Neter et al. (1985) for details). When the number of levels is two, T_{ij} is exactly the same as so called two sample t-test statistic. This means that Tukey employs the two sample t-test instead of the paired t-test for correlated paired observations and thus we suspect that Tukey performs well.

3. Justifications for a Better Performance of Bonferroni Than Tukey

In this section, we provide justifications why Bonferroni performs better than Tukey when there is dependence among the observations of a within subject factor. To make comparisons simple, we assume that the observations $X_1, X_2, \cdots, X_k$ of the within subject factor have the same variance σ^2 and that any two different observations have the same correlation ρ. Here we mainly focus on the case where $\rho > 0$ since the observations from the same subject tend to have positive correlations and since the actual significance level of Tukey becomes larger than the nominal α. In Section 2 and 3, we implicitly assume that the observations follow a multivariate normal distribution.

Theoretical justification is based on asymptotic arguments which hold for large n. Note that

$$t_{ij} \approx \frac{\overline{X_i} - \overline{X_j}}{\sqrt{E(S_d^2)/n}} = \frac{\overline{X_i} - \overline{X_j}}{\sqrt{2\sigma^2(1-\rho)/n}}$$

and

$$T_{ij} \approx \frac{\overline{X_i} - \overline{X_j}}{\sqrt{2 \cdot E(MSE)/n}} = \frac{\overline{X_i} - \overline{X_j}}{\sqrt{2\sigma^2/n}}$$

hold for large n, where t_{ij} and T_{ij} are defined in (2-1) and (2-2), respectively. Thus, for large n, $|t_{ij}| \approx |T_{ij}|/\sqrt{1-\rho}$ holds and $|t_{ij}|$ is approximately $1/\sqrt{1-\rho}$ times as big as $|T_{ij}|$. However, the table values of Bonferroni and Tukey are almost the same for large n for the case $k = 2,3,4,5,6,7$: See Table 2 for the ratio of the table values

$$\frac{t(\alpha/(2^*m), n-1)}{q(\alpha, k, k(n-1))/\sqrt{2}}$$

at $\alpha = 0.05$, where $m \equiv \binom{k}{2}$ is defined just prior to (2-1).

<Table 2> The ratio of the table values of Bonferroni and Tukey
($k = 2,3,4,5,6,7$; $\alpha = 0.05$; $n = \infty$)

n	k					
	2	3	4	5	6	7
∞	1.000	1.022	1.027	1.029	1.030	1.030

For large n, therefore, Tukey confronts with some problems if ρ is far away from

0. As ρ becomes close to 1, $|T_{ij}|$ tends to produce smaller values than $|t_{ij}|$ and thus Tukey will have less power than Bonferroni for detecting actual difference between means present in some pairs. As ρ becomes close to -1, $|T_{ij}|$ tends to produce larger values than $|t_{ij}|$ and thus Tukey will have actual significance level much larger than nominal α, i.e. Tukey will conclude more often than expected that some pairs have different means even when they actually have the same means.

We have conducted a simulation study to provide empirical evidence that Bonferroni performs better than Tukey in detecting actual difference between means of some pairs when the correlation ρ of the observations are positive but not too close to 0. When the correlation is negative, however, the actual significance level of Tukey becomes larger than nominal α and therefore we exclude the case of negative correlations from this simulation study. The setup for the simulation is given as follows. We take $\rho = 0.2, 0.4, 0.6, 0.8$, $n = 10, 20, 40, 60$, and $k = 2,3,4,5,6,7$. For each ρ, n, k,

1) We generate a random sample of size n from a multivariate normal distribution with $\mu_i = 0.5(i-1)/(k-1)\,(i = 1, \cdots, k;$ equally spaced between 0 and 0.5), $\sigma^2 = 1$ and ρ.
2) Bonferroni and Tukey are conducted separately to see if they detect some pairs with different means.
3) We repeat Steps 1-2 500 times and calculate the proportion of repetitions in which each method detects at least one pair with different means.

The proportion calculated in Step 3 is called 'empirical power' for detecting pairs with different means. There are a lot of simulation results but they show almost the same trend. Thus the simulation results for $k = 2, 4, 7$ are selected for presentation and shown in Tables 3, 4, and 5.

<Table 3> The empirical powers of Bonferroni and Tukey for k=2

n	Bonferroni				Tukey			
	ρ				ρ			
	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
10	0.170	0.166	0.184	0.198	0.136	0.068	0.040	0.006
20	0.322	0.350	0.326	0.306	0.258	0.200	0.070	0.008
40	0.598	0.580	0.542	0.588	0.522	0.352	0.168	0.032
60	0.746	0.770	0.772	0.794	0.656	0.598	0.374	0.052

<Table 4> The empirical powers of Bonferroni and Tukey for k=4

n	Bonferroni				Tukey			
	ρ				ρ			
	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
10	0.088	0.120	0.090	0.120	0.050	0.040	0.012	0.002
20	0.184	0.184	0.222	0.202	0.118	0.052	0.012	0.000
40	0.416	0.432	0.414	0.400	0.348	0.182	0.042	0.002
60	0.610	0.634	0.658	0.598	0.528	0.332	0.120	0.008

<Table 5> The empirical powers of Bonferroni and Tukey for k=7

n	Bonferroni				Tukey			
	ρ				ρ			
	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
10	0.060	0.086	0.082	0.088	0.044	0.024	0.006	0.000
20	0.176	0.154	0.152	0.136	0.106	0.038	0.002	0.000
40	0.392	0.392	0.380	0.382	0.292	0.154	0.026	0.000
60	0.558	0.572	0.614	0.596	0.478	0.258	0.046	0.000

These results can be summarized as the following:

1) The empirical power of Bonferroni is higher than Tukey for each configuration of our simulation study.
2) For fixed n and k, the empirical power of Bonferroni is almost the same for all ρ but that of Tukey decreases substantially as ρ becomes larger. As a result, the difference between empirical powers of Bonferroni and Tukey increases as ρ becomes larger for fixed n and k.
3) For fixed n and ρ, the empirical power of each method decreases as the number of k becomes larger. For fixed k and ρ, the empirical power of each method increases as the sample size n becomes larger.

Summary result 3 is well perceived in statistics and is not directly related to our issue of comparing Bonferroni with Tukey. Summary results 1 and 2 tell us that Bonferroni is better than Tukey for practically resonable ranges of values $\rho \geq 0.2, n \geq 10$, $k \leq 7$ and that, when the correlation ρ among the observations is high, Tukey can hardly detect actual difference between means present in some pairs.

4. Concluding Remarks

When a statistical methodology is proposed, there are assumptions under which the methodology is supposed to work well. Some assumptions do not produce serious errors in result unless they are violated in extreme form. Others, however, are vital to the methodology and violation of such assumptions may result in an absurd conclusion. For example, while violation of the assumptions such as normality and equal variances in one-way ANOVA does not cause a serious problem in F-test if all levels have almost the same sample sizes, the nonindependence of observations has serious effects on F-test (see pp. 624-25 of Neter et al. (1985) for details).

Similarly, application of Tukey to post-hoc test procedures in a repeated measures design may lead to a misleading conclusion on the relationship between the level means. Employing Tukey for pairwise multiple comparisons requires satisfaction of the assumption of independence. In a repeated measures design, however, this assumption is usually violated and Tukey performs badly when it is applied to the (positively) correlated observations of a within subject factor. When the correlation ρ is as high as 0.8, Tukey becomes similar to a procedure with actual significance level much smaller than nominal α and thus seldom detects actual difference between means present in some pairs.

On the other hand, Bonferroni is hardly affected by the correlation among the observations of a within subject factor, and we have provided empirical evidence and theoretical justification for it. Also, Bonferroni is easy to apply in practice since all we need to do is to perform paired t-tests for all pairs by any convenient statistical software; the pair with p-value less than the adjusted significance level α/m is declared to have different means. Considering its statistical power for detecting actual differences among level means and convenience for using in practice, Bonferroni seems to be a better approach than Tukey or other methods for a between subject factor as post hoc test procedures for a significant within subject factor.

References

1. Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*, second ed., John Wiley & Sons, Inc., New York.
2. Barthelemy, S., and Boulinguez, P. (2002). Orienting visuospatial attention generates manual reaction time asymmetries in target detection and pointing, *Behavioural Brain Research*, 133, 109-116.
3. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society Series B*, 57, 289-300.

4. Benjamini, Y., and Liu, W. (1999). A distribution-free multiple test procedure that controls the false discovery rate, *Tel Aviv. P-SOR-99-3*, Department of Statistics and O.R., Tel Aviv University.
5. Harvey, M., Olk, B., Muir, K., and Gilchrist, I. D. (2002). Manual responses and saccades in chronic and recovered hemispatial neglect: a study using visual search, *Neuropsychologia*, 40, 705-717.
6. Helsen, W. F., Elliott, D., Starkes J. L., and Ricker, K. L. (2000). Coupling of eye, finger, elbow, and shoulder movements during manual aiming, *Journal of Motor Behavior*, 32, 241-248.
7. Johnson, R. A., and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis*, third ed., Prentice Hall, New Jersey.
8. Neter, J., Wasserman, W., and Kutner, M. H. (1985). *Applied Linear Statistical Model*, second ed., Irwin, Illinois.
9. Teixeira, L. A. (2000). Timing and force components in bilateral transfer of learning, *Brain and Cognition*, 44, 455-469.
10. Weir, P. L., MacDonald, J. R., Mallat, B. J., Leavitt, J. L., and Roy, E. A. (1998). Age-related differences in prehension: the influence of task goals, *Journal of Motor Behavior*, 30, 79-89.

[received date : Apr. 2007, accepted date : Apr. 2007]