

A Comparative Study on Multiple Comparison Procedures

Rosa Arboretti¹, Stefano Bonnini², Livio Corain³, Francesca Solmi⁴

¹Dipartimento Territorio e Sistemi Agro-Forestali, Università di Padova

² Dipartimento di Matematica, Università di Ferrara

³ Dipartimento di Tecnica e Gestione dei sistemi industriali, Università di Padova

⁴ Dipartimento di Scienze Statistiche, Università di Padova

e-mail: rosa.arboretti@unipd.it, bnnsfn@unife.it; livio.corain@unipd.it; solmi@stat.unipd.it

Abstract

The topic of Multiple Comparison Procedures (MCPs) is receiving nowadays a growing theoretical and applicative attention in both industrial and scientific problems. This issue arises when we consider a response variable of interest and we wish to compare more than two treatments in order to find out possible significant differences among them. In this paper we propose a comparative simulation study in order to compare the power performance of the FWE controlling MCPs proposed in the literature in order to provide a practical contribute for the choice of the different techniques in different possible data scenarios.

Keywords: FWE - Familywise Error Rate, power comparison.

1. Multiple comparison procedures

The hypothesis testing problem connected with a multiple set of individual statistical tests, such as those of Multiple Comparison Procedures (MCPs), giving rise to the so called multiplicity issue, that is the problem to control the inferential errors of the whole set of null hypotheses of interest (Hochberg and Tamhane, 1987). The MCPs are commonly used to control the so called Familywise Error Rate (FWER). The strong form of FWE is the probability of rejecting any true null hypothesis H_i contained in a subset of true null hypothesis S ; that is formally:

$$\text{FWER}(S) = \Pr(\text{Reject at least one } H_{0i}, i \in S \mid H_{0i} \text{ is true for all } i \in S).$$

This paper aims at comparing some MCPs proposed in the literature in order to provide a practical contribute for the choice of the different techniques in different possible data scenarios: indeed it is well known that a MCP that works uniformly better than the others in all the situations does not exist (Montgomery, 2004), but it could be interesting to quantify the relative advantage of the different methods with respect to the other ones. In what follows we will refer to the one-way ANOVA problem in which n units are randomly assigned to C treatments or groups and a numeric response variable X is observed. Groups are supposed to have the same size n (balanced design). Units belonging to the j -th group are presumed to receive a treatment at the j -th level. We will refer to the so called fixed effect model, i.e.

$$Y_{ij} = \mu_j + \varepsilon_{ij} = \mu + \tau_j + \varepsilon_{ij}, \varepsilon_{ij} \sim \text{IID}(0, \sigma^2), i = 1, \dots, n, j = 1, \dots, C,$$

where τ_j and Y_{ij} are respectively the effect of the j -th treatment and the response variable for the i -th replicate and j -th treatment. The random term ε_{ij} represents the experimental error with zero mean, variance σ^2 and unknown continuous distribution P . The usual side-condition for effects are $\sum_j \tau_j = 0$.

In this framework we are interested in testing the individual $C \times (C-1)/2$ null hypotheses $H_{0(jh)}: \mu_j - \mu_h = 0, j, h = 1, \dots, C, j \neq h$, against $H_{1(jh)}: \mu_j - \mu_h \neq 0$, taking care of the multiplicity issue.

2. Simulation study and results

An analysis of recent literature has underlined some comparative studies of MCPs, in which no specific MCP is recognized as the best under different data scenarios (Rafter at al., 2002; Vallejo Seco at. al, 2001); anyway the advantage of procedures that take care of the dependence among the hypotheses of interest is clearly underlined.

In this work we performed a simulation study in order to compare the MCPs working at a fixed level for the FWE and looking for that procedure that maximizes the power, which in this framework can be defined in several slightly different ways. The MCP methods we have considered are: Bonferroni (B), Bonferroni-Holm (BH; Holm, 1979), Shaffer (two versions, called S1 and S2; Shaffer, 1986; 1995) and Chen-Seneta (CS; Seneta and Chen, 2005). As traditional parametric MCP we considered the Tukey's studentized range test (T; Tukey 1953) as well.

In the literature there are several different proposals to measure the power of MCPs (Horn and Dunnet, 2004). The most often used are the following:

- *any-pair power*: the probability of rejecting at least one false null hypothesis;
- *all-pair power*: the probability of rejecting all false null hypotheses (Ramsey, 1978);

In this work, we propose an alternative definition of power in order to evaluate the monotonicity of the power function for the whole procedure:

$$\text{Relative all-pair power}(\delta) = \Pr\{\text{Rejecting } H_{0(jh)} \mid \mu_j - \mu_h = \delta_{jh} = \delta\}.$$

The proposed simulation study considers different settings as regards the number of treatments to compared ($C=5$ and $C=10$), the sample size ($n=4$ and $n=8$) and the distribution of the random error (normal, exponential and Student's t with 2 d.f. – as examples of an asymmetric distribution and an heavy tailed distribution). Moreover, we have simulated some appropriate experimental settings (see Table 1), in order to consider different situations also in terms of the number of true null hypotheses.

Table 1: Setting details for the simulation study, in case of five treatments

No.	Setting	Means' values					No. of true H_0	Single true H_{ij}									
		μ_1	μ_2	μ_3	μ_4	μ_5		1-2	1-3	1-4	1-5	2-3	2-4	2-5	3-4	3-5	4-5
1	$\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5$	1	2	3	4	5	0	0	0	0	0	0	0	0	0	0	0
2	$\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 = \mu_5$	1	2	3	4/5	4/5	1	0	0	0	0	0	0	0	0	0	1
3	$\mu_1 \neq \mu_2 \neq \mu_3 = \mu_4 = \mu_5$	1	2	3/4/5	3/4/5	3/4/5	3	0	0	0	0	0	0	0	1	1	1
4	$\mu_1 \neq \mu_2 = \mu_3 = \mu_4 = \mu_5$	1	2/3/4/5	2/3/4/5	2/3/4/5	2/3/4/5	6	0	0	0	0	1	1	1	1	1	1

Interesting results have been obtained in this study. Table 2 summarizes the behavior of the procedures for the data scenario no. 1, reporting the relative all-pair power values at different values of the true difference δ between means (in terms of σ -units), the any and all-pair power of the procedures and the value of the FWER (under H_0).

Table 2: Rejection rates ($\alpha=0.05$) for $C=5$ treatments and sample size $n=4$ with different random error distributions

Normal random errors						
	B	B-H	S1	S2	T	CS
δ (delta)	<i>Relative all-pair</i>					
0	0,0030	0,0070	0,0070	0,0100	0,0050	0,0090
1	0,0555	0,0833	0,0907	0,1228	0,0765	0,0978
2	0,3727	0,4257	0,4567	0,5077	0,4357	0,4587
3	0,8075	0,8245	0,8500	0,8575	0,8545	0,8395
4	0,9770	0,9810	0,9830	0,9850	0,9845	0,9830
<i>any-pair</i>	0,9110	0,9110	0,9110	0,9110	0,9350	0,9230
<i>all-pair</i>	0,3283	0,3621	0,3814	0,4143	0,3691	0,3829
FWER	0,0030	0,0070	0,0070	0,0100	0,0050	0,0090
Exponential random errors						
	B	B-H	S1	S2	T	CS
δ (delta)	<i>Relative all-pair</i>					
0	0,0060	0,0120	0,0130	0,0260	0,0080	0,0170
1	0,0770	0,1160	0,1260	0,1593	0,0988	0,1345
2	0,4443	0,4953	0,5177	0,5627	0,5000	0,5203
3	0,8045	0,8185	0,8380	0,8435	0,8460	0,8305
4	0,9510	0,9580	0,9645	0,9660	0,9630	0,9620
<i>any-pair</i>	0,8930	0,8930	0,8930	0,8930	0,9270	0,9050
<i>all-pair</i>	0,3611	0,3986	0,4148	0,4458	0,3986	0,4178
FWER	0,0060	0,0120	0,0130	0,0260	0,0080	0,0170
Student's t random errors						
	B	B-H	S1	S2	T	CS
δ (delta)	<i>Relative all-pair</i>					
0	0,0010	0,0030	0,0030	0,0030	0,0030	0,0040
1	0,0135	0,0177	0,0220	0,0275	0,0182	0,0205
2	0,0777	0,0910	0,1020	0,1130	0,1053	0,0973
3	0,2445	0,2515	0,2685	0,2735	0,2870	0,2635
4	0,4270	0,4425	0,4585	0,4640	0,4845	0,4590
<i>any-pair</i>	0,3720	0,3720	0,3720	0,3720	0,4350	0,3900
<i>all-pair</i>	0,0862	0,0941	0,1034	0,1107	0,1070	0,1001
FWER	0,0010	0,0030	0,0030	0,0030	0,0030	0,0040

3. Conclusions

In general, the obtained results give a very strong indication for the advantage of using Shaffer S2 procedure in the considered data scenarios for the case of 5 treatments to be compared. With regard to the other considered procedures, an increasing power of Bonferroni, Bonferroni-Holm, and both CS and Shaffer S1 procedures can be noticed, and a similar behaviour of CS, S1 and T methods has been recorded for all the considered data situations. The power of all the considered procedures is shown to slowly decrease in the case of exponential distribution, and a bad behaviour in this sense can be seen in the case of the Student's t distribution, in which a strong decrease of power of all the procedures has been shown.

Moreover, one can notice how the use of the new definition of power (relative all-pair power), brings to a more complete information about the behavior of the procedures, showing a much more complete description of the differences among the procedures than considering only the standard all-pair and any-pair powers.

Acknowledgement: The research has been supported by the University of Padova CPDA088513 (Coordinator: Prof. Luigi Salmaso).

References

- Hochberg, Y., Tamhane, A. C. (1987) *Multiple comparison procedures*, Wiley.
- Holm, S. (1979) A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics*, 6: 65-70.
- Horn, M., Dunnett, C. W. (2004) Power and sample size comparisons of stepwise FWE and FDR controlling test procedures in the normal many-one case, *Recent Developments in Multiple comparison Procedures, Institute of Mathematical Statistics*, 47: 48-64.
- Montgomery, D. C. (2004) *Design and Analysis of Experiments*, Wiley.
- Rafter, J. A., Abell, M. L., Braselton, J. P. (2002) Multiple comparison methods for means, *Society for Industrial and Applied Mathematics*, 44: 259-278.
- Seneta, E., Chen, J. T. (2005) Simple stepwise tests of hypotheses and multiple comparisons, *International Statistical Review*, 73: 21-34.
- Shaffer, J. P. (1986) Modified sequentially rejective multiple test procedure, *Journal of the American Statistical Association*, 81: 826-831.
- Shaffer, J. P. (1995) Multiple hypothesis testing, *Annual Review of Psychology*, 46: 561-584.
- Tukey, J. W. (1953) The problem of multiple comparisons, Unpublished manuscript, Princeton University.
- Vallejo Seco, G., Menéndez de la Fuente, I. A., Escudero, J. R. (2001) Pairwise multiple comparisons under violation of the independence assumption, *Quality & Quantity*, 35: 71-76.
- Westfall, P. H., Tobias, R. D., Rom, D., Wolfinger, R. D., Hochberg Y. (1999) *Multiple comparisons and multiple tests using SAS*, SAS Institute Inc.