COMPARING RELATIVE AND TOTAL COST MULTIPLE COMPARISON PROCEDURES VIA JAMES-STEIN ESTIMATORS

Roy E. Welsch

WP 892-76 December 1976

COMPARING MULTIPLE COMPARISON PROCEDURES

1. INTRODUCTION

In a recent panel discussion on multiple comparison (MC) procedures [7], a question was raised about the advisability of journals requiring that certain specific MC procedures be used in the papers they published. The consensus seemed to be that while a journal might consider proposing standards it should in no way inhibit the use of other procedures which the author might consider more suitable to his problem.

In this paper we propose that instead of standardized procedures, the journals recommend that, where possible, the costs (losses) associated with a particular MC procedure be stated in at least two ways - total and relative.

The first way emphasizes the total cost of type 1 (1 and/or 3) errors (conversely type 2 errors) and is associated with control of the experiment-wise error rate (α) for all hypotheses containing groups of equal (or for type 3, nearly equal) population means. (Of course, this does not control the total cost of type 1 and 2 errors which increases with the number of populations considered.)

The second way focuses on the relative cost of type 1 and type 2 errors and has been most frequently associated with tests proposed by Waller and Duncan [8].

There are at least three basic approaches to inference in the MC situation; Bayesian (B) , James-Stein or empirical Bayes (JS), and maximum likelihood (ML). It is important, at least at the beginning, to separate discussions of cost from questions of inference. This has not often been done in the past and total cost seems to be limited to ML methods while relative costs are linked to the JS and B methods.

There are valid reasons to consider different inferential approaches. Sometimes B or JS procedures cannot be used because reasonable prior assumptions are not satisfied. (There are cases that we discuss later where it may pay to do some preliminary work in order to get the problem into a suitable form for such procedures.) On the other hand, the availability of prior information may make the strict use of ML methods very wastful.

The need to consider both types of cost is, perhaps, more controversial. Here, I think, we owe it to ourselves and our clients to be able to look at cost both ways, at least until there is a concensus on one approach or the other. One way to build a consensus is to have both α and k available in each problem and then see which is more useful.

This paper is organized in the following manner. The next three sections discuss how to compute total and relative costs for each of the three inferential methods. The concluding part notes some difficulties with the proposed methodology and discusses a few unsolved problems.

The author would like to acknowledge helpful discussions with John Tukey, Ray Faith, Bradley Efron and David Duncan.

2. BAYESIAN METHODS

We shall focus on the basic one-way ANOVA model with equal sample sizes, treatment means $\{\bar{x}_{i.}\}_{i=1}^{p}$, grand mean $\bar{x}$, and s^2 (standard error of the means) based on n-p degrees of freedom. Only the pairwise differences $\bar{x}_{i.}-\bar{x}_{j.}$, will be considered.

The three types of error are:

1. declare population means not equal when they actually are equal;
2. delcare equal, when actually not equal;
3. declare not equal, when actually not equal but reverse order.

Often type 3 errors are included with type 1 errors. We will let E stand for the experimentwise error rate and C for the comparisonwise error rate (for definitions see [6]).

In the Bayesian case assume that the population means $\mu_1,\ldots,\mu_p$ are independent samples from a population which is Gaussian with mean $\delta=0$ and variance σ_μ^2 (known). We also assume that the treatment means $\bar{x}_{i.}$ are independent Gaussian samples from $G(\mu_i, \sigma_e^2)$ where, to make our early discussion simple, we take σ_e^2 as known.

In this case the posterior distribution of the μ_i (which we will call $\hat{\mu}_i$) is Gaussian with mean $\bar{x}_{i.}/(1+r)$ and variance $\sigma_e^2/(1+r)$ where $r = \sigma_e^2/\sigma_\mu^2$. If we want to preserve the posterior E1 (the experimentwise type 1 error rate), we would compute $q^T_{\alpha,p,\infty}$ so that

$$P\left\{\underset{i-1,\ldots,p}{\text{range}}\left(\frac{\hat{\mu}_{i.}-\bar{x}_{i.}/(1+r)}{\sigma_e/(1+r)^{\frac{1}{2}}}\right) \geq q^T_{\alpha,p,\infty}\right\} = \alpha. \tag{2.1}$$

Thus we have a Bayesian range procedure with critical values

$$q^{T}_{\alpha,p,\infty}(1+r)^{1/2}\sigma_e \qquad (2.2)$$

where $q^{T}_{\alpha,p,\infty}$ is the standard studentized range statistic or Tukey HSD test ([6], p. 38). Since $r\geq 0$, it is clear that if we have any prior precision $(1/\sigma^2_\mu>0)$ and $\sigma^2_e>0$, then the B critical values will be larger than for the Tukey HSD test. This is reasonable because the more we feel the population means are clumped together, the more evidence the data must give us to separate them.

We could also have used a Scheffe´ F-test in place of the range. The critical values would then have been $\sigma_e[2(p-1)\ (1+r)\ q^{S}_{\alpha,p-1,\infty}]^{\frac{1}{2}}$ where $q^{S}_{\alpha,p-1,\infty}$ are the critical values for the usual Scheffe´ test.

The relative error B approach has been treated by Duncan [1]. The critical values are

$$z(k)\ (1+r)^{1/2}\ \sigma_e\ 2^{\frac{1}{2}} \qquad (2.3)$$

where z(k) satisfies

$$h(z) + zH(z)/(h(z) - zH(z)) = k \qquad (2.4)$$

with h(·) and H(·) denoting the p.d.f. and c.d.f. of a Gaussian (0,1) variable. We note that (2.3) is independent of p.

If α were given,we would find $k(\alpha)$ by finding the value k^* that satisfied $2^{\frac{1}{2}}z(k^*) = q^{T}_{\alpha,p,\infty}$ or $[2(p-1)\ q^{S}_{\alpha,p-1,\infty}]^{\frac{1}{2}}$ and conversely if k were given.

Tables 1 and 2 provide examples when p = 6 for α = .01, .05, .10 and for k = 50, 100 and 500. Extensive tables of the range are available in [4].

Tables 1 and 2 about here

3. JAMES-STEIN METHODS

The Bayesian case discussed in the previous section is difficult to implement because we do not often know r. The James-Stein methodology provides us with a way to estimate r.

If we again assume a Gaussian model for $\{\bar{x}_i\}_{i=1}^{p}$ then the James-Stein estimate for the treatment effects would be

$$\bar{y}_i = (1 - \frac{\lambda}{F}) (\bar{x}_i - \bar{x}) \tag{3.1}$$

where

$$F = \frac{\sum_{i=1}^{p} (x_i - \bar{x}_i)^2/(p-1)}{s^2} \tag{3.2}$$

and $\lambda = \lambda_F$ is often taken to be $(p-3)/(p-1)$. There does not appear to be a simple way to estimate the variance of $\bar{y}_i$ (see [3] for a recent discussion), so we used a quantity analogous to that for the pure Bayesian case, $s^2 (1 - \frac{\lambda}{F})$. We will exploit the idea that $(1+r)$ is like $(1 - \frac{\lambda}{F})^{-1}$ extensively.

To perform a Scheffe´ type MC procedure on the $\bar{y}_i$ we need to find critical numbers $b^S_{\alpha,p-1,n-p}$ such that

$$P\left\{\frac{2 \sum_{i=1}^{p} \bar{y}_i^2}{s^2(1-\frac{\lambda}{F})} \geq b^S_{\alpha,p-1,n-p}\right\} = \alpha$$

or

$$P\left\{2(p-1)F(1-\frac{\lambda}{F}) \geq b^S_{\alpha,p-1,n-p}\right\} = \alpha$$

or

$$P\left\{F \geq \lambda + b^S_{\alpha,p-1,n-p}/2(p-1)\right\} = \alpha.$$

Solving this gives critical values of

$$s\,[2(p-1)(q^S_{\alpha,p,n-p}-\lambda)(1-\frac{\lambda}{F})^{-1}]^{\frac{1}{2}} \tag{3.3}$$

which can be compared to the usual Scheffe´ values

$$s\,[2(p-1)\,q^S_{\alpha,p,n-p}\,]^{\frac{1}{2}}. \tag{3.4}$$

Thus the JS Scheffe´ test has smaller critical values whenever $F > q^S_{\alpha,p,n-p}$ and as a consequence is more powerful. In fact, this test can be viewed as an extension of Fisher's idea of using a preliminary F-test before applying a MC procedure.

It is formally possible to modify the Tukey HSD in a similar way. Let R denote the studentized range statistic and replace F by R^2 in (3.1). Then solving

$$P\left\{\underset{i}{\text{range}}\ \frac{(1-\frac{\lambda}{R^2})(x_i-\bar{x})}{s\,(1-\frac{\lambda}{R^2})^{1/2}} \geq b^T_{\alpha,p,n-p}\right\} = \alpha$$

gives the critical values

$$s\,[((q^T_{\alpha,p,n-p})^2-\lambda)\,(1-\frac{\lambda}{R^2})^{-1}]^{\frac{1}{2}}. \tag{3.5}$$

Since there is no James-Stein type result using R^2 , we do not know what $\lambda=\lambda_R$ should be. However, we can note that p-3 in the JS case is $E(\chi^2_{p-3})$ and use $E(W^2_{p-3})$, where W_n is the range of n independent G(0,1) variates, to replace p-3 in the range case. (The factor $(p-1)^{-1}$ in λ_F corrects for the fact that the numerator of the F statistic is divided by p-1.) Tables of the moments of the range are contained in [4]. This is only a heuristic approximation and developing a useful theory in this case remains an open problem.

For relative costs, Waller and Duncan [8] have again provided the necessary results. We shall take the liberty of calling their method a James-Stein approach, even though Waller and Duncan start from a pure Bayesian framework and require exchangeable priors in their development.

To compare total and relative costs in the JS Scheffe´ case we proceed as follows. First compute the F statistic and find λ_F. If α is given, compute (3.3) omitting the factor 2, and look in the Waller-Duncan [8] tables to find a k which gives approximately the same critical value. The procedure can be reversed if k is given first. We call the Waller and Duncan values t(k). Tables 3 and 4 provide examples when p=6, n-p=30, and F=3.

The procedure is similar for the JS Tukey case. First compute R, the studentized range statistic, and then look up $E(W^2_{p-3})=\lambda_R$. When α is given, compute (3.5), divide it by $2^{\frac{1}{2}}$ and again use the Waller and Duncan tables. Tables 3 and 4 provide examples when p=6, n-p=30, R=4.7, and λ_R=3.65. Extensive tables of the studentized range are contained in [5].

Tables 3 and 4 about here

4. MAXIMUM LIKELIHOOD

In this case the total cost approach has been discussed by many authors with a vast array of different techniques (for reviews, see [2], [6], and [9]). We shall continue to focus on the Tukey and Scheffé procedures which have widely available tables of critical values.

For maximum likelihood we do not, however, have available a relative cost approach. One way to fill this void is to let $r = 0$ (i.e., no prior precision) on the pure Bayesian case (section 2) and then use the initial values $z(k)\sigma_e(2)^{\frac{1}{2}}$ from (2.3). Of course, we do not know σ_e and it is natural to approximate it by s. We can find critical numbers for this by noticing that if we consider $(1 + r)$ analogous to $(1 - \frac{\lambda}{F})^{-1}$ then $r = 0$ corresponds to $F = \infty$. Thus we can use the Waller and Duncan tables for $F = \infty$ and the appropriate number of degrees of freedom for s.

When α is given, find the T or S statistic, divide by $2^{\frac{1}{2}}$ and find a k value (interpolation in log k is recommended) from the Waller and Duncan tables. The procedure is easily reversed using the tables in [5]. Examples for p=6, n-p=30 are given in Tables 5 and 6.

Tables 5 and 6 about here

5. CONCLUDING REMARKS

For Scheffe´ type MC procedures (and in some cases for the Tukey HSD) we have provided a way (there may be others) to choose a type of inference (Bayes, James-Stein, or maximum likelihood) and then compute a total cost value α for a given relative cost k and vice versa.

We have not mentioned comparisonwise (C1) error rates in the preceding derivations because for situations where MC procedures seem most relevant we do not think the C1 error rate is a useful measure. In fact, the results presented here free us from having to always compare the relative cost k with the C1 error rate associated with the least-significant difference test ([6], p. 90). For example, Waller and Duncan have argued ([8], p. 1486) that k = 100 corresponds to a C1 error rate of about 5%. But, depending on p and n-p we have seen how k = 100 can correspond to rather high experimentwise error rates. This fact should be reported when relative cost procedures are used. Conversely E1 = .05 can correspond to what some would consider an absurdly high relative cost ratio. This, too, should be reported.

Given the results of section 3, it is reasonable to ask why we should ever use a maximum likelihood procedure. First, the procedures discussed in this paper (especially B and JS) lean very heavily on Gaussian assumptions. The current theories of robust estimation can probably be more readily applied to maximum likelihood type procedures than to Bayesian methods. This advantage will keep ML methods in the picture for some time to come.

Second, it is important to remember that B and JS, in effect, shrink the ML (least-squares) estimates toward the grand mean. If we suspect that the population means may be in clumps (blocks) of equal or nearly equal means with some separation between the clumps, then can we argue that there is a common mean to shrink toward? This clumpiness of means is not consistent with a Gaussian prior and could lead to a large value of F and low critical values.

There is no simple way out of this dilemma. Perhaps we could first make a pass over the ordered sample means using a gap procedure [9], in order to break them into clumps. Within clumps it may seem more reasonable to use a B or JS procedure because the shrinking would be toward the mean of the clump, not the grand mean. Just how to compute error rates for such compound procedures is a subject for further research.

Table 1

α	.01	.05	.10
T	4.76	4.03	3.66
$T/2^{\frac{1}{2}}$	3.36	2.85	2.59
$k^*(T,\alpha)$	27800	4387	1719
S	5.50	4.70	4.30
$S/2^{\frac{1}{2}}$	3.89	3.32	3.04
$k^*(S,\alpha)$	189382	24052	8728

Table 2

k	500	100	50
$z(k)$	2.23	1.72	1.49
$\sqrt{2}\ z(k)$	3.15	2.43	2.11
$\alpha^*(T,k)$	.23	.52	.67
$z(k)/(p-1)^{\frac{1}{2}}$	1.00	.77	.67
$\alpha^*(S,k)$	.32	.57	.65

Table 3

α	.01	.05	.10
T	5.24	4.30	3.85
$T/2^{\frac{1}{2}}$	3.71	3.04	2.72
$k^*(T,\alpha)$	19486	3036	1250
S	6.08	5.03	4.52
$S/2^{\frac{1}{2}}$	4.30	3.56	3.20
$k^*(S,\alpha)$	100173	12852	4738

Table 4

k	500	100	50
$t(k)$	2.39	1.81	1.55
$\sqrt{2}\ t(k)$	3.38	2.56	2.13
$\alpha^*(T,k)$	.19	.48	.66
$t(k)/(p-1)^{\frac{1}{2}}$	1.07	.81	.69
$\alpha^*(S,k)$	.40	.55	.63

Table 5

α	.01	.05	.10
$q^{T}_{\alpha,5,30}$	5.24	4.30	3.85
$T/2^{\frac{1}{2}}$	3.71	3.04	2.72
(a)	3.79	2.99	2.59
$k^{*}(T,\alpha)$	1404	346	172
$q^{S}_{\alpha,5,30}$	3.70	2.53	2.05
$S/2^{\frac{1}{2}}$	4.30	3.56	3.20
(b)	4.40	3.47	3.01
$k^{*}(S,\alpha)$	4080	802	359

(a) $\left[.5((q^{T}_{\alpha,5,30})^{2} - \lambda_{R})\ (1-\lambda_{R}/R^{2})^{-1}\right]^{\frac{1}{2}}$

(b) $\left[(p-1)\ (q^{S}_{\alpha,5,30} - \lambda_{F})\ (1-\lambda_{F}/F)^{-1}\right]^{\frac{1}{2}}$

$F=3$, $\lambda_{F}=.6$, $\lambda_{F}/F=.2$; $R=4.7$, $\lambda_{R}=3.65$, $\lambda_{R}/R^{2}=.17$

Table 6

k	500	100	500
t(k)	3.20	2.28	1.92
(c)	4.54	3.50	3.13
α(T,k)	.04	.16	.26
(d)	2.24	1.43	1.19
α(S,k)	.08	.24	.34

(c) $[t^2(k)\ 2\ (1-\lambda_R/R^2) + \lambda_R]^{\frac{1}{2}}$

(d) $t^2(k)\ (1-\lambda_F/F) + \lambda_F$

$F=3$, $\lambda_F=.6$, $\lambda_F/F=.2$; $R=4.7$, $\lambda_R=3.65$, $\lambda_R/R^2=.17$

REFERENCES

[1] Duncan, D.B., "A Bayesian Approach to Multiple Comparisons," Technometrics, 7 (1965), 171-222.

[2] Einot, I. and Gabriel, K.R., "A Study of the Powers of Several Methods of Multiple Comparisons," Journal of the American Statistical Association, 70 (September 1975), 574-83.

[3] Faith, R.E., "Minimax Bayes Set and Point Estimators for a Multivariate Normal Mean, Technical Report 66, Department of Statistics, University of Michigan, Ann Arbor, Mich., February 1976.

[4] Harter, H.L., Clemm, D.S., and Guthrie, E.H., "The Probability Integrals of the Range and of the Studentized Range: Percentage Points and Moments of the Range," Wright Air Development Center Technical Report 58-484, Vol. I, 1959 (NTIS Document No. PB161037).

[5] Harter, H.L., Clemm, D.S., and Guthrie, E.G., "The Probability Integrals of the Range and of the Studentized Range-Probability Integral and Percentage Points of the Studentized Range: Critical Values For Duncan's New Multiple Range Test," Wright Air Development Center Technical Report 58-484, Vol. II, 1959 (NTIS Document No. AD231733).

[6] Miller, R.G., Simultaneous Statistical Inference, New York: McGraw-Hill, 1966.

[7] Panel Discussion on Multiple Comparison Procedures, American Statistical Association Annual Meeting, Atlanta, Georgia, August 1975.

[8] Waller, R.A. and Duncan, D.B., "A Bayes Rule for the Symmetric Multiple Comparisons Problem," Journal of the American Statistical Association, 64 (December 1969), 1484-1503, and Corrigenda 67 (1972), 253-5.

[9] Welsch, R.E., "Stepwise Multiple Comparison Procedures," to appear in the Journal of the American Statistical Association.