ONE-STAGE MULTIPLE COMPARISONS WITH THE BEST FOR UNEQUAL VARIANCES

Frank J. Matejcik, South Dakota School of Mines and Barry L. Nelson, The Ohio State University
Frank J. Matejcik, Industrial Engineering, SD School of Mines, Rapid City, SD 57701

KEY WORDS: Heteroscedastic, simultaneous confidence intervals

Abstract

Multiple comparisons with the best (MCB) procedures are a class of procedures that provide simultaneous confidence intervals for the difference between each treatment mean and the best of the other treatment means. We introduce four new procedures to find MCB intervals under unequal variances. These procedures exploit Welch's approximate solution to the Behrens-Fisher problem. Two of the procedures can be shown to be conservative, and the others are approximate-conservative. We present computer simulation results that indicate the circumstances when each procedure is preferred.

1 Introduction

We consider the problem of comparing a small number of competitors. Although our work involves one-way designs used in ANOVA, our goal is to find the competitor with the largest expected performance, μ_i, for $i = 1, 2, \ldots, k$. We achieve this goal by using a class of multiple comparisons procedures known as Multiple Comparisons with the Best (MCB). We define MCB procedures as procedures that provide joint confidence intervals for $\mu_i - \max_{j \neq i} \mu_j$, for $i = 1, 2, \ldots, k$. We will propose, illustrate, and experimentally study one-stage MCB procedures that may be applied under mild assumptions; specifically, we permit unequal variances across competitors, thereby allowing MCB to be used more readily.

Describing MCB procedures requires some notation, which is defined next. Many of these terms are vectors, so we display vectors in boldface type.

Y is a single observation.

i is a subscript corresponding to a competitor.

k is the number of competitors, $i = 1, 2, \ldots, k$.

$\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_k]'$ is the objective quantity, a $k \times 1$ vector.

$\hat{\boldsymbol{\mu}}$ is an estimator of $\boldsymbol{\mu}$.

$\bar{\mathbf{Y}}$ is the overall sample mean, the estimator of $\boldsymbol{\mu}$ in our procedures.

Table 1: An extreme example

	Plan A	Plan B	Current Choice
mean	1.0	0.0	-1000
standard deviation	0.1	0.1	100

n_i is the sample size from competitor i.

ν is the degrees of freedom.

σ_i^2 is the variance from competitor i.

S_i^2 is the variance estimator from competitor i.

$t_{q,\nu}$ is the q quantile of the univariate Student's t distribution with ν degrees of freedom.

$T_{q,k,\nu,\mathbf{M}}$ is the q quantile of the multivariate t distribution of dimensions k with ν degrees of freedom, and correlation matrix $\mathbf{M}$.

$T_{q,k,\nu,\rho}$ is the q quantile of the multivariate t distribution of dimension k with ν degrees of freedom, and a correlation matrix having common off diagonal elements, ρ.

The paper is organized around four sections: MCB preliminaries, procedures, experiments, and conclusions.

We use normal-theory methods and compute one-stage MCB intervals. Alternatively, we could follow the recommendation of Matejcik and Nelson (1995) and design a second stage sample to allow inferences with a prespecified precision. Our procedures extend a rich body of literature on inference when variances are unequal and unknown. Welch (1947) is an early paper; see Bishop (1976) for an extensive survey.

When variances across competitors differ substantially, MCB procedures designed for equal variances can be made useless. Consider the problem of experimentally determining the competitor with the largest mean when the competitors have the parameters in Table 1. In situations of this nature MCB for equal variances fails to determine the best because the large variance of the "Current Choice" masks the small variance of Plans A and B. This is an extreme example, but where current choices are woefully inadequate and responsible alternatives exist, similar stituations can occur.

We will present four procedures which solve our problem. These procedures are applications of statistical methods for unequal variances specialized to MCB.

2 MCB Preliminaries

Before we present our procedures for one-stage MCB we present some general MCB results. When samples are sufficiently large, MCB correctly selects the competitor with the largest mean among k competitors with probability $1-\alpha$. This *optimization* is done by forming k joint confidence intervals for $\mu_i - \max_{j\neq i}\mu_j$, where i and j designate populations, treatments, systems, or competitors. Methods for finding and using these confidence intervals were first published in Hsu (1981).

Our presentation is structured around a result, which we call Hsu's lemma. Although Hsu did not present "Hsu's lemma" in the form that we will present it, his work provided the proof of the lemma and the foundations of MCB.

Lemma 1 (Hsu's multiple-bound lemma) *Let $\mu_{(1)} \leq \mu_{(2)} \leq \cdots \leq \mu_{(k)}$ be the ordered performance parameters of k competitors, and let $\hat{\mu}_{(1)}, \hat{\mu}_{(2)}, \cdots, \hat{\mu}_{(k)}$ be any estimates of the parameters. If for each i,*

$$\Pr\{\hat{\mu}_i - \hat{\mu}_j - (\mu_i - \mu_j) > -w_{ij}, \text{ for all } j \neq i\} = 1-\alpha, \tag{1}$$

then the following joint confidence intervals may be formed :

$$1-\alpha \leq \Pr\left\{ \begin{array}{r} \mu_i - \max_{j\neq i}\mu_j \in [D_i^-, D_i^+], \\ \text{for } i = 1,2,\ldots,k \end{array} \right\}, \tag{2}$$

where

$$D_i^+ = \left(\min_{j\neq i}[\hat{\mu}_i - \hat{\mu}_j + w_{ij}]\right)^+,$$

$$\mathcal{G} = \{\ell : D_\ell^+ > 0\},$$

and

$$D_i^- = \begin{cases} 0 & \text{if } \mathcal{G} = \{i\} \\ \left(\min_{\substack{j\in\mathcal{G}\\ j\neq i}}[\hat{\mu}_i - \hat{\mu}_j - w_{ji}]\right)^- & \text{otherwise .} \end{cases}$$

If we replace the $=$ in (1) with $\geq$, then (2) still holds.

To simplify notation we define the event

$$\{\mathrm{CS}(\hat{\boldsymbol{\mu}}, \mathbf{w})\} = \left\{ \begin{array}{r} \mu_i - \max_{j\neq i}\mu_j \in [D_i^-, D_i^+] \\ \text{for } i = 1,2,\ldots,k \end{array} \right\}.$$

Hsu's multiple-bound lemma is essentially proved in Chang and Hsu (1992). Informally, this lemma says that if we can do Multiple Comparisons with a Control (MCC), then we can do MCB. More precisely, the premise (1) is one-sided MCC (see e.g. Hochberg and Tamhane (1987) p.33, where the statistic $\hat{\mu}_i - \hat{\mu}_j$ is $\bar{Y}_i - \bar{Y}_j$, and $\mu_i - \mu_j$ is denoted $\theta_i - \theta_j$). The "bounds" w_{ij} have a helpful graphical interpretation. MCB intervals are often presented as whisker diagrams, so the w_{ij} are interpreted as "whisker lengths" in these diagrams. Accordingly, the w_{ij} are measures of the precision, which help define the length the MCB interval. All of our one-stage procedures are based on Hsu's Multiple Bound lemma.

3 MCB Procedures

We begin by introducing an MCB procedure in general form. Next, we describe and justify five specific procedures. The procedures differ only in the computation of the whisker lengths, w_{ij}. Finally, we discuss how we can determine which procedure should be recommended.

Procedure 1 (General one-stage procedure)

1. *Take samples from each competitor $(Y_{1,1},\ldots,Y_{1,n_1})$, $(Y_{2,1},\ldots,Y_{2,n_2}),\ldots,$ $(Y_{k,1},\ldots,Y_{k,n_k})$. We assume the competitors are independent.*

2. *Compute the following intermediate statistics for each competitor, $i = 1,\ldots,k$.*

$$\bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$$

$$S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2.$$

3. *Compute the whisker lengths, w_{ij}.*

4. *Finally, compute $\mathrm{CS}(\bar{\mathbf{Y}}, \mathbf{w})$.*

The usual homoscedastic MCB procedure fits the general form above, and we shall refer to it as procedure ST (for standard). For procedure ST we compute the whisker lengths using

$$w_{ij} = T_{1-\alpha,k-1,\nu,\mathbf{R}^{(i)}} S\left(\frac{1}{n_i} + \frac{1}{n_j}\right)^{1/2}$$

where the empirical variance, S^2, and degrees of freedom, ν, are the usual pooled variance estimator and its degrees of freedom, and the correlation matrix, $\mathbf{R}^{(i)}$, has off-diagonal elements given by

$$\rho_{\ell\ell'}^{(i)} = \left(\frac{n_\ell}{n_\ell + n_i}\right)^{1/2} \left(\frac{n_{\ell'}}{n_{\ell'} + n_i}\right)^{1/2}$$

$$(\ell \neq \ell' \neq i, 1 \leq i, \ell, \ell' \leq k)$$

(see e.g., Hochberg and Tamhane 1987). Notice that $\rho_{\ell\ell'}^{(i)} = 1/2$ when sample sizes are equal.

Hsu's multiple-bound lemma tells us that if we can do MCC, then we can do MCB. So we only need to find MCC procedures that work with unequal variances in order to derive heteroscedastic MCB procedures. Tamhane (1977) and Dunnett (1980) provide tools for developing heteroscedastic MCC procedures.

Tamhane (1977) gives a conservative MCC procedure based on Banerjee's Inequality. He also gives an approximate procedure using the Welch (1947) moment-based approximation to the Behrens-Fisher problem, which is derived in a similar manner, but uses Welch's (1947) moment-based approximation instead of Banerjee's Inequality. Applying Hsu's multiple bound lemma to these MCC procedures gives us MCB procedures B (for Banerjee) and T2, respectively, which both fit our general procedure.

Let $\beta = 1 - (1-\alpha)^{\frac{1}{(k-1)}}$.
For procedure B,

$$w_{ij} = \left(\frac{t^2_{\beta,n_i-1} S_i^2}{n_i} + \frac{t^2_{\beta,n_j-1} S_j^2}{n_i}\right)^{1/2}.$$

Similarly for procedure T2,

$$w_{ij} = t_{\beta,\hat{\nu}_{i,j}} \left(\frac{S_i^2}{n_i} + \frac{S_j^2}{n_j}\right)^{1/2},$$

where

$$\hat{\nu}_{i,j} = \frac{\left(S_i^2/n_i + S_j^2/n_j\right)^2}{\left[\frac{S_i^4}{n_i^2(n_i-1)} + \frac{S_j^4}{n_j^2(n_j-1)}\right]}.$$

Dunnett (1980) gives an approximate all-pairwise procedure related to the T2 procedure. Dunnett (1980) improves upon the all-pairwise procedure by using Sidak's Uncorrelated t Inequality instead of Slepian's Inequality. Hochberg and Tamhane (1987) suggest that an MCC version of Dunnett's procedure may be better than the approximate procedure of Tamhane (1977). Applying Hsu's multiple bound lemma to this MCC procedure gives us MCB procedure T3, which fits our general procedure. For procedure T3, $\hat{\nu}_{i,j}$ is the same as in T2, but

$$w_{ij} = T_{1-\alpha,k-1,\hat{\nu}_{i,j},0} \left(\frac{S_i^2}{n_i} + \frac{S_j^2}{n_j}\right)^{1/2}$$

Finally, Matejcik and Nelson (1995) give a one-stage MCB procedure (MN) of the form of our general procedure. Procedure MN is related to a two-stage heteroscedastic selection procedure (Rinott 1978). Both Rinott's (1978) procedure and procedure MN are derived by applying Slepian's Inequality to a standardized difference of the overall system means. For procedure MN,

$$w_{ij} = h_i \max\left\{\frac{S_i}{\sqrt{n_i}}, \frac{S_j}{\sqrt{n_j}}\right\},$$

where h_i is the solution to

$$\int_0^\infty \prod_{j\neq i}^{k-1} \left[\int_0^\infty \Phi\left(\frac{h_i}{\sqrt{\frac{n_j-1}{x} + \frac{n_i-1}{y}}}\right) \mathrm{f}_{n_j-1}(x)dx\right] \times \mathrm{f}_{n_i-1}(y)dy = 1-\alpha,$$

$\Phi(x)$ is the standard normal distribution function, and $\mathrm{f}_\nu(x)$ is the Chi-squared density function with ν degrees of freedom.

We must now decide which of these procedures to recommend. For MCC, the simulation results in Tamhane (1977) indicate that procedure T2 should perform adequately, and procedure B is needlessly conservative. Also, Dunnett (1980) notes (by an analytic observation) that T3 provides tighter intervals than T2. However, we have no comparisons for MCB of procedure MN to the others. Procedure ST is known to be generally conservative, so it may be able function adequately under some unequal-variance situations. Further, we should consider how well the unequal variance procedures perform under equal variances. Our next section uses simulation experiments to address these issues of performance of the procedures.

4 Experiments

We measured the results of our experiments by estimating three probabilities: the probability of coverage ($\mathcal{C}$), the probability of declaring correctly ($\mathcal{U}$), and the probability of correct and useful ($\mathcal{C}$ & $\mathcal{U}$) inference. By "coverage" we mean that the

Table 2: A test of an extreme example

2500 macro replications, $n = 3, k = 3$
$\boldsymbol{\mu} = (1, 0, -1000)'$,
$\sigma_1 = \sigma_2 = 0.1$, $\sigma_3 = 100$

$\alpha = 0.05$ Procedure	$\mathcal{C}\%$	$\mathcal{C}$ & $\mathcal{U}\%$	$\mathcal{U}\%$
ST	89.96	0.00	0.00
B	97.04	97.04	100.00
T2	96.00	96.00	100.00
T3	95.80	95.80	100.00
MN	98.52	98.40	99.88

Table 3: Usual MCB is exact, k=6, n=10

2500 macro replications, $\alpha = 0.05$
$\boldsymbol{\mu} = (1.897367, 0, 0, 0, 0, 0)'$,
$\sigma_i = 1$, $\forall i$

Proc.	$\mathcal{C}\%$	$\mathcal{C}\&\mathcal{U}\%$	$\mathcal{U}\%$	$\bar{w}$
ST	94.28	89.52	90.48	1.02
B	97.00	72.80	73.12	1.24
T2	95.60	81.20	82.20	1.13
T3	95.52	81.32	82.36	1.13
MN	98.00	61.12	61.20	1.35

MCB confidence intervals cover $\mu_i - \max_{i \neq j} \mu_j$, $i = 1, 2, \ldots, k$. By "declaring correctly" we mean that the lower bound of $\mu_{(k)} - \max_{j \neq (k)} \mu_j$ is zero. By "correct and useful" inference we mean that the procedure both declares correctly and covers the parameters. Later, we included $\bar{w}$; $\bar{w}$ is computed by first computing the average of the whisker lengths within each macro replication, then computing the average of those averages over all the replications. If we have no problems with coverage, the procedure with the shortest whisker length is the best.

We will first experimentally justify the need for our methods by demonstrating our claim in the Introduction (see Table 1 and related discussion) that use of procedure ST is imprudent in some situations. Our test simulated 2500 macro replications of an MCB computation for a case where $n = 3$ observations were taken from each of the $k = 3$ three competitors. All the observations were random variates generated to be independent and normally distributed with the parameters indicated in Tables 1 and 2. Table 2 summarizes the experiment's parameters and results. (Please notice that Table 2 reports four digits for each estimated probability, thus a reader may calculate exactly how many of the macro experiments met the criteria: $\mathcal{C}$, $\mathcal{U}$, and $\mathcal{C}$ & $\mathcal{U}$. This violates the common practice of reporting significant digits, but allows us to provide our raw data to readers without including additional tables. We continue with this practice throughout the tables.)

Clearly, procedure ST does not provide the nominal coverage for this extreme example, but all the other procedures do. More glaringly, we notice that procedure ST never declares correctly, but all the other procedures do declare correctly with high probability. So we have found procedure ST less useful dealing with this extreme example of unequal variances, and we have found that each of our other procedures perform well.

As we anticipated, procedure T3 performs better than procedure T2, and T2 performs better than procedure B. We have no references testing procedure MN with the others. Because procedure MN involves Slepian's inequality and a whisker length based on the greatest of a pair of variances, which is generally inferior to the approaches of other procedures, we expect that procedure MN will be worst.

4.1 Equal Variance Experiments

We consider that procedure MN may be particularly suspect under equal variances, and we know that procedure ST should perform well under equal variances. Our studies included experiments with k = 3,4,5,6,10 competitors; $1 - \alpha$ = 0.90,0.95,0.99 confidence levels; n = 3,10,20,30 observations; $\boldsymbol{\mu} = (4/\sqrt{n},0)$ & $(6/\sqrt{n},0)$; $\sigma_i = 1$ for $i = 1, \ldots, k$; and we used 2500 macro replication of each situation. Table 3 is typical of our results. More extensive results are available in Matejcik (1992).

Procedures B, T2, T3, and MN are all conservative. Overall, procedure MN was always inferior to the other procedures for the choices of $\boldsymbol{\mu}$ and σ_i where the probability of coverage of procedure ST is known to be 1-α. Moreover, T2 and T3 appear to be competitive to ST with regard to coverage, declaring correctly, being correct and useful, and whisker length. However, we have tested only one case with unequal variances, the extreme example.

4.2 Unequal Variance Experiments

We have not done extensive tests of unequal variance cases such as those in Dunnett (1980), Tamhane (1979), or Bishop (1976). Our procedures B and MN can be shown to be conservative, so there is no need to demonstrate experimentally that they are conservative. Our procedures T2 and T3 are both approximate-conservative procedures in the sense defined by Dunnett (1980); they are formed using conservative probability inequalities and use Welch's (1947) moment-based approximation for unequal variances. The all-pairwise procedures analogous to our T2 and T3 procedures, also approximate-conservative procedures, were extensively tested by Dunnett (1980) and Tamhane (1979). Dunnett (1980) did not find any situation where these procedures failed to be conservative. Tamhane (1979) observed a few situations where the all-pairwise analog to T2 appeared to be liberal, but these were disputed by the more sophisticated tests of Dunnett (1980). Based on these experiments, Hochberg and Tamhane (1987 p. 193) suggest procedures identical to the computation of w_{ij} in our procedures T2 and T3 for the MCC problem (1), which is in the premise of Hsu's multiple-bound lemma (lemma 1). Further, our T2 and T3 MCB procedures will generally be more conservative than the related MCC procedures, because MCB procedures based on Hsu's multiple-bound lemma (lemma 1) are generally more conservative than the related MCC procedures. For these reasons we anticipated that extensive tests would show our procedures T2 and T3 to be conservative. However, our experiments showed T2 and T3 to be liberal in some cases.

We consider extensive tests of T2 and T3 are of value to describe the limitations of approximate-conservative procedures. However, we have not yet performed an extensive study. But, we have studied some cases of unequal variances for normal distributions. These cases are chosen to illustrate that our procedures would be preferred to procedure ST under reasonable cases of unequal variances, and to test some conditions that may make procedures T2 and T3 liberal. Our studies included experiments with $k = 3,4,10$ competitors; $1-\alpha = 0.90, 0.95, 0.99$ confidence levels; $\boldsymbol{\mu} = (a, \mathbf{0})$($a$ was chosen to give Pr $\mathcal{C\&U}$ reasonable values); and we used 2500 macro replication of each situation. We were not able to test procedure MN in situations where the number of observations across competitors was not constant. Table 4 is one of our more curious results. More extensive results are available in Matejcik (1992).

Procedure ST is not generally conservative for unequal variances, so we must recommend other procedures for unequal variances. Procedure T3 was often the top performer, and was closely followed by procedure T2. Although, at times T2 and T3 are sometimes liberal. Procedure B was less conservative than procedure MN, but both procedures B and MN can be shown to be conservative. For this case of moderately unequal variances each of the procedures performed fairly well.

Table 4: k=4, n=(13,11,9,7)
2500 macro replications, $\alpha = 0.05$,
$\boldsymbol{\mu}'$=(3.5082,0,0,0)
$\sigma_1^2 = 1,\ \sigma_2^2 = 4,\ \sigma_3^2 = 7,\ \sigma_4^2 = 10$

Proc.	$\mathcal{C}\%$	$\mathcal{C\&U}\%$	$\mathcal{U}\%$	$\bar{w}$
ST	93.40	80.52	81.56	2.085
B	95.40	49.24	50.24	2.850
T2	94.36	51.44	53.00	2.689
T3	94.40	50.96	52.52	2.693

5 Conclusions

The experiments and discussion of the previous section allow us to recommend procedures. Under equal variances procedure ST is a clearly best, but not by much over T3. For unequal variances procedure T3 is the top performer and our recommendation. However, for some cases with unbalanced sample sizes procedures T2 and T3 are liberal. If software for T3 is not available or if repeated computations are needed such as in design of experiments, then procedure T2 is a good substitute. Procedure MN is not comparable with the other one-stage procedures, but it may be valuable in settings when both one-stage and two-stage procedures are frequently used.

6 References

Bishop, T. A. (1976). Heteroscedastic Anova, Manova, and Multiple-Comparisons. Ph.D. Disertation, Dept. of Statistics, The Ohio State University.

Chang, J. Y. and J. C. Hsu (1992). Optimal Designs for Multiple Comparisons with the Best. *Journal of Statistical Planning and Inference* **30,** 45-62.

Dunnett, C. W. (1980). Pairwise Multiple Comparisons in the Unequal Variance Case. *Journal of the American Statistical Association* **75,** 796-800.

Hochberg, Y., and A. C. Tamhane (1987). *Multiple Comparisons Procedures,* John Wiley and Sons, New York.

Hsu, J. C. (1981). Simultaneous Confidence Intervals for all Distances from the Best. *The Annals of Statistics* **9**, 1026-1034.

Matejcik, F. J. (1992). Heteroscedastic Multiple Comparison Procedures for Computer Simulation. Ph.D. Dissertation, Department of Industrial and Systems Engineering, The Ohio State University.

Matejcik, F. J., and B. L. Nelson (1995). Two-stage Multiple Comparisons with the Best for Computer Simulation. *Operations Research,* **43,** in press

Rinott, Y. (1978). On Two-Stage Selection Procedures and Related Probability-Inequalities. *Communications in Statistics, Series A* **7**, 799-811.

Tamhane, A. C. (1977). Multiple Comparisons in Model I: One-way ANOVA with Unequal Variances. *Communications in Statistics, Series A* **6,** 15-32.

Tamhane, A. C. (1979). A Comparison of Procedures for Multiple Comparisons of Means with Unequal Variances. *Journal of the American Statistical Association* **74,** 471-480.

Welch, B. L. (1947). The Generalization of Students' Problem When Several Different Population Variances are Involved. *Biometrika* **34,** 28-35.

AUTHOR BIOGRAPHIES

FRANK J. MATEJCIK is an Assistant Professor of Industrial Engineering at the South Dakota School of Mines and Technology. He received his Ph.D. in Industrial and Systems Engineering from The Ohio State University in 1992. His research interests include simulation experimental design and sequential methods. He is a member of ORSA, ASA, and ASQC.

BARRY L. NELSON is an Associate Professor in the Department of Industrial, Welding, and Systems Engineering at The Ohio State University. His research interests are experiment design and analysis of stochastic simulations. He is a past President of the TIMS College on Simulation and an Area Editor for *Operations Research* and an Associate Editor *TOMACS*.