DOCUMENT RESUME

ED 319 754 TM 014 843

AUTHOR	Olejnik, Stephen; Lee, JaeShin
TITLE	Multiple Comparison Procedures when Population Variances Differ.
PUB DATE	Apr 90
NOTE	22p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE	Speeches/Conference Papers (150) -- Reports - Research/Technical (143) -- Information Analyses (070)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Comparative Analysis; *Research Methodology; *Statistical Analysis
IDENTIFIERS	*Multiple Comparisons; *Population Variances; Type I Errors; Variance (Statistical)

ABSTRACT

A review of the literature on multiple comparison procedures suggests several alternative approaches for comparing means when population variances differ. These include: (1) the approach of P. A. Games and J. F. Howell (1976); (2) C. W. Dunnett's C confidence interval (1980); and (3) Dunnett's T3 solution (1980). These procedures control the overall risk of a Type I error experimentwise at approximately the nominal significance level and have the best statistical power among alternative solutions. The two-stage multiple comparison procedures of Y. Hochberg and A. C. Tamhane (1987), and R. R. Wilcox (1987) are also discussed. These procedures and the Tukey-Kramer procedure were applied to data from a study of the effects of exercise on psychological and physiological variables with a total of 36 subjects. Textbooks and a sample of research studies were reviewed to determine the most frequently taught and used multiple comparison procedures. This review indicates that most applied researchers are not aware of the alternative solutions when variances differ. It is suggested that the Games-Howell procedure will provide a valid test for most purposes and should be included in statistical methods textbooks and classes. Four tables present data from the application and the reviews. (SLD)

Multiple Comparison Procedures when Population Variances Differ

Stephen Olejnik

JaeShin Lee

University of Georgia

Paper presented at the annual meeting of the American Educational Research Association in Boston, MA. April 19, 1990.

Multiple Comparison Procedures when Population Variances Differ

Over the past several years I've had an ongoing discussion with several of my colleagues regarding teaching statistical methods classes. One of the topics that we've debated has been the role multiple comparison procedures plays in our courses. It is well known that many procedures have been developed and each has its advantages and disadvantages. From an instructors point of view, choosing from among the many alternatives poses the problem of what to discuss in class. When asked, most of my colleagues say that they provide instruction on three or maybe four techniques. One friend feels strongly that only one procedure need be taught. Typically, the procedures mentioned have been Tukey's HSD, Bonferroni, Fisher's LSD and Scheffé. Many others have been mentioned but these seem to me to be the most popular among those to whom I've spoken. The four procedures I've mentioned differ in several important ways including: type I error rates, statistical power, types of contrasts to be examined, dealing with unequal sample sizes etc. But one characteristic that they all share is that they all assume that the populations sampled have equal variance. On reflection, I don't believe that anyone I have spoken to, has ever mentioned providing instruction on multiple comparison procedures when population variances differ. Several articles reviewing multiple comparison procedures have commented on the variance heterogeneity problem and have recommended approximate solutions. Statisticians certainly have been busy developing new and improved approaches to the problem but I wondered whether these solutions are taught and whether the applied researcher is familiar with the issues and the solutions.

The purpose of our paper is threefold: First, I wanted to briefly review the statistical literature on the alternative approaches for comparing means when population variances differ. Second, I wanted to demonstrate the approaches with a small data set. And third, I wanted to review textbooks and a sample of research studies to determine what are the most frequently taught multiple comparison procedures (as indicated by textbook coverage) and how often do appli· esearchers in educ ition and psychology consider the procedures that allow variances to differ. My comments are limited to situations where all pairwise contrasts are of interest and control of type I errors is set experimentwise.

Alternative Approximate Solutions

When population variances differ, several solutions have been suggested. Many of the proposed procedures control the overall risk of type I errors but have low statistical power. Three procedures that have often been recommended are those that have been developed by Games and Howell (1976) based on Welch's solution to the Behrens-Fisher problem, Dunnett C (1980) based on Cochran's (1964) solution to the Behrens-Fisher problem, and Dunnett T3 (1980) based on Sidak's (1967) uncorrelated-t inequality. These procedures control the overall risk of a type I error experimentwise at approximately the nominal significance level and have the best statistical power among the alternative solutions.

The Games-Howell procedure constructs a confidence interval as follows:

$$\bar{X}_j - \bar{X}_k \pm q_{a,w,1-\alpha}/\sqrt{2}\left[\sqrt{\frac{s_j^2}{n_j} + \frac{s_k^2}{n_k}}\right] \qquad (1)$$

Where $\bar{X}_j$, $\bar{X}_k$ are the sample means for groups j and k respectively,
q is the 1-α/2 centile of the studentized range distribution,
a is the number of levels of the independent variable,
s_j^2, s_k^2 are the unbiased estimates of the population variances for populations j and k respectively
n_j, n_k are the number of observations available from populations j and k respectively,

$$w = \frac{\left[\frac{s_j^2}{n_j} + \frac{s_k^2}{n_k}\right]^2}{\frac{\left[\frac{s_j^2}{n_j}\right]^2}{(n_j - 1)} + \frac{\left[\frac{s_k^2}{n_k}\right]^2}{(n_k - 1)}} \qquad (1b)$$

Dunnett's C confidence interval is constructed as:

$$\bar{X}_j - \bar{X}_k \pm A / \sqrt{2} \left[\sqrt{\frac{s_j^2}{n_j} + \frac{s_k^2}{n_k}}\right] \qquad (2)$$

Where $A = \dfrac{q_{a,v,1-\alpha}\dfrac{s_j^2}{n_j} + q_{a,v,1-\alpha}\dfrac{s_k^2}{n_k}}{\dfrac{s_j^2}{n_j} + \dfrac{s_k^2}{n_k}}$; (2b)

$v_j = n_j - 1$, $v_k = n_k - 1$;

the other terms are the same as those defined above.

Dunnett's T3 solution forms a confidence interval as:

$$\bar{X}_j - \bar{X}_k \pm M_{a,w,1-\alpha} \sqrt{\frac{s_j^2}{n_j} + \frac{s_k^2}{n_k}} \qquad (3)$$

Where M is the 1-α/2 centile of the Studentized Maximum-Modulus Distribution,

other terms are as defined previously.

All three procedures use the same estimate for the standard error and differ only in the identification of the critical values from their respective reference distributions.

Reviewers of the statistical literature differ in their recommendations regarding which procedure is "best". Keselman and Rogan (1977) and Jaccard, Becker and Wood (1984) both recommend that when variances differ that the Games-Howell procedure be used. Games, Keselman and Rogan (1981) concluded that all three approaches are acceptable. Stoline (1981) recommends the T3 or C procedures and Wilcox (1987) concurs with Stoline. The reason for the disagreement focuses on some research findings that indicate that the Games-Howell procedure can have an inflated type I error rate experimentwise for some conditions. Tamhane (1979) found inflated type I error rates for several situations but no clear pattern was identified. Dunnett (1980) concluded that the Games-Howell procedure maybe slightly liberal when population variances were equal and became conservative as the variances differed. Games and Howell (1976) reported similar evidence but only studied a small number of situations where variances were equal. Wilcox (1987b) replicated Dunnett's conditions for a 4x4 factorial structure and examined differences in cell means. His results were consistent with Dunnett's in that the Games-Howell was found to be liberal when variances were equal. He also found the Games-Howell procedure to be liberal when variances were unequal if cell sizes were small. All studies that considered the T3 and C procedures have consistently shown these procedures are conservative. The C procedure is more conservative than T3 when sample sizes are small but the opposite is true when sample sizes are large. With infinite degrees of freedom C and the Games-Howell procedures become identical (Dunnett, 1980). When statistical power has been considered the Games-Howell procedure consistently provides narrower confidence limits.

In all of the situations where the Games-Howell procedure has been found to be liberal the sample sizes were small ($n<15$). When the smallest group had at least 15 observations the type I error rate did not exceed the nominal significance level.

Exact Solutions

The procedures presented above are not exact tests. The actual confidence intervals only approximate the nominal confidence level. When population variances differ no single stage procedure can provide an exact solution. An alternative apprcach to multiple comparisons which has not received much attention in the social science literature is to estimate differences between population means using a two-step or two-stage sampling procedure. Basically these procedures require the researcher to select samples from the populations of interest, estimate the population variances, then depending on the acceptable margin of error and variance inequality, sample a second time from each population. Two two-stage multiple comparison procedures are discussed by Hochberg and Tamhane (1987) and by Wilcox (1987). Both texts present the two-stage procedures developed by Tamhane (1977) and Hochberg (1975).

Tamhane's two-stage procedure can be applied to any linear contrast but only pairwise differences are of interest here. In stage one random samples of n_o individuals from each population are selected and basic descriptive statistics are computed (means, standard deviations). Next, the researcher determines the total number of observations needed for each group (n_j) so that the margin of error is no greater than m units. This is determined from the following:

$$n_j = \max [\ n_o+1,\ (s_j^2/d)^* +1] \qquad (4)$$

Where n_o is the initial sample size,

s_j^2 is the sample variance of group j based on the initial sample of n_o,

and $d = (m/h)^2$. (4b)

Where m is the margin of error the researcher finu.· acceptable,

h i·· 1- α /2 centile point of J independent Student's t variates, based on = n_o-1 degrees of freedom,

$(\ )^*$ indicates the integer value.

At a minimum Tamhane's procedure require_ that at least one additional observation is needed from each population.

A weighted mean for the two samples is then computed as:

$$\tilde{X}_j = b_j k\ \bar{X}_{2j} + (1-b_j k)\bar{X}_{1j} \qquad (4c)$$

Where $\bar{X}_{1j}$, $\bar{X}_{2j}$ are the sample means from the first and second sampling stages for group j;

k is the number of additional observations made in stage two;

and b_j is computed as:

$$1/n_j\ \{\ 1+(1/s_j)[n_o(n_j{*}d - s_j^2)/k]^{1/2}\} \qquad (4d)$$

A confidence interval is then constructed as the difference between the weighted means with the margin of error equal..ing m units:

$$\tilde{X}_j - \tilde{X}_k \pm m.$$

An altern..tive two-stage approach suggested by Hochberg(1976) determines the number of additional observations needed in the second stage based on the following:

$$n_j = \max[n_o,(s^2/d)^* +1] \qquad (5)$$

The terms are the same as those defined by Tamhane above. The confidence interval is provided by:

$$\bar{X}_j - \bar{X}_k \pm h\ \ (\max[s_j/\sqrt{n_j},\ s_k/\sqrt{n_k}]) \qquad (5b)$$

The difference between means is estimated using all of the data from samples at both stage 1 and 2. The variance estimate however, is based on data from the first sampling stage.

An advantage for Hochberg's procedure is that additional observations may not be needed but a possible disadvantage is that the width of the confidence intervals can vary.

Example Problem

Since many applied researchers may not be familiar with these alternative procedures for contrasts when variances differ I thought it would be useful to demonstrate an application to a real data set. The data example I chose for the demonstration is taken from Moore and McCabe's (1989) new statistics textbook titled Introduction to the Practice of Statistics. One of the exercises in this text cites a dissertation study by Lobstein (1983). In this investigation the effects of exercise on psychological and physiological variables, were studied. Four groups were considered: A treatment group (T) who participated in an exercise program; a control group (C) who had volunteered to participate in the exercise program but for various reasons could not attend the treatment sessions, a group of joggers (J) and a group of sedentary people (S) who did not exercise regularly. One of the outcome measures used in the study was a physical fitness scale administered when the treatment was terminated. Descriptive statistics on the four groups are reported in table 1.

Insert Table 1 about here

For this demonstration I will focus on the contrast between the treatment (T) and control (C) groups.

First, consider the Games-Howell procedure. Using (1b) the degrees of freedom are:

$$w = \frac{\left[\frac{38.17}{10} + \frac{32.07}{5}\right]^2}{\frac{\left[\frac{38.17}{10}\right]^2}{10 - 1} + \frac{\left[\frac{32.07}{5}\right]^2}{5 - 1}} = 9.5$$

And the critical value from the Studentized range distribution is found using the truncated value of w as:

$$q_{4,9,.95} = 4.42$$

The .95 confidence interval from (1) would be estimated as:

$$291.91 - 308.97 + 4.42/\sqrt{2}\sqrt{\left[\frac{38.17^2}{10} + \frac{32.07^2}{5}\right]}$$

$$-17.06 \pm 58.58$$

The critical value for Dunnett's C procedure found from (2b) is determined as:

$$q_{4,9,.95} = 4.42$$

$$q_{4,4,.95} = 5.76;$$

and

$$A = \frac{4.42\,\frac{38.17^2}{10} + 5.76\,\frac{32.07^2}{5}}{\frac{38.17^2}{10} + \frac{32.07^2}{5}}$$

$$= 5.204$$

Using (2) the .95 confidence interval is estimated as:

$$291.91 - 308.97 \pm 5.20/\sqrt{2}\sqrt{\left[\frac{38.17^2}{10} + \frac{32.07^2}{5}\right]}$$

$$-17.06 \pm 68.99$$

For Dunnett's T3 procedure the critical value from the maximum-modulus distribution is found as:

$$M_{4,9,.95} = 3.27$$

And the .95 confidence interval from (3) is found as:

$$291.91 - 308.97 \pm 3.27\sqrt{\left(\frac{38.17^2}{10} + \frac{32.07^2}{5}\right)}$$

$$-17.06 \pm 61.30$$

For the two-stage procedures let's assume the data in table 1 report sample data from the first sampling stage than to determine the number of additional observations needed. The critical value h from the distribution of a independent Student t-variates has degrees of freedom equal to the integer value of $a/\Sigma(n_{oj}-1)$ (Wilcox, 1987).

$$\nu = 4/(.4909)=8$$

$$h_{4,8,.95} = 4.40$$

If we assume that a margin of error equalling 65 is acceptable, then using (4b) and (4) the sample sizes needed for the Treatment (T) and Control (C) groups are determined as:

$$d = (65/4.40)^2$$

$$= 218.23$$

$$n_T = \max\left[10+1, \left[(38.17)^2/218.23\right]^* + 1\right] = 11$$

$$n_C = \max\left[5+1, \left[(32.07)^2/218.23\right]^* + 1\right] = 6$$

Thus for the contrast between the Treatment and Control groups one additional individual (k=1) would be needed from each group.

The weighting factor from (4d) for the Treatment and Control groups would be determined as:

$$b_T = 1/11\{ 1+(1/38.17)[10(11*218.23-38.17^2)/(11-10)]^{1/2}\} = .3223;$$

$$b_C = 1/6\ \{ 1+(1/32.07)[5(6*218.23-32.07^2/(6-5)]^{1/2}\} =.3614.$$

Finally, the weighted group means are calculated as:

$$\tilde{X}_T = .3223\bar{X}_{2T} + .6777\bar{X}_{1T}$$
$$\tilde{X}_C = .3614\bar{X}_{2C} + .6386\bar{X}_{1C}$$

The confidence interval is estimated as:

$$\tilde{X}_T - \tilde{X}_C \pm 65.0.$$

For Hochberg's procedure using (5) with the same value for d as was determined above:

$$n_T = \max\ [10, \left[(38.17)^2/218.23\right]^* + 1] = 10$$
$$n_C = \max\ [\ 5, \left[(32.07)^2/218.23\right]^* + 1] = 5$$

In the present example additional observations would not be needed from either T or C groups.

The confidence interval from (5b) is found as:

$219.91-308.97 \pm 4.40(14.34)$

-17.06 ± 63.11

Table 2 present the results when these five procedures are used for the six possible pairwise contrasts for the problem. The Tukey-Kramer procedure was added for comparative purposes I chose the Tukey-Kramer procedure since it it fairly well known and is generally recommended for

---

Insert Table 2 about here

---

situations where sample sizes are unequal, the error-rate is set experimentwise, and it can be assumed that population variances are equal. For Tamhane's procedure one additional observation was needed for the T, C, and J groups and 9 additional observations would be needed from the S group. With Hochberg's procedure no additional observations would be needed for the T, C, and J groups but 9 additional observations would be needed from the S group. The point estimates for the contrasts between groups would likely be different with Tamhane's procedure but only contrasts involving the S group would have point estimates that would change with Hochberg's procedure.

Statistics Textbooks

For the third portion of this paper we were interested in examining textbooks that might be used for statistical methods classes in the behavioral sciences. To identify possible texts we went to the libraries at the University of Georgia and obtained a listing of all books that were listed under the key words: behavioral statistics, statistics, and social sciences-statistical methods. A list of 161 titles were printed. We then went through this list and identified only those books that might be used as a textbook and was published from 1980 to the present. Textbooks on topics such as factor analysis, multivariate analysis, regression analysis, sampling we excluded from consideration. Finally after examining these books we only included those texts that discussed analysis of variance. For our analysis we examined 48 texts. From this list, 9 (19%) of the texts had no presentation on multiple comparisons or contrast analysis. Table 3 summarizes the frequency with which the most popular multiple comparison procedures were presented. The percentages

---

Insert Table 3 about here

---

reported in table 3 are based on the number of books that discuss multiple comparison procedures (39). By a considerable margin the Tukey HSD and Scheffe's multiple comparison procedures were the most popular techniques taught. Only 6 (15%) of the textbooks we examined discussed the issue of variance heterogeneity in the context of multiple comparisons. All six of these texts had discussed the Games-Howell procedure but only two of them commented on Dunnett's T3 procedure.

Journal Articles

Finally, we were interested in examining empirical research studies that have tested hypotheses on the equality of means and have examined pairwise contrasts. For our review we examined five research journals: American Educational Research Journal, Journal of Educational Psychology, Reading Research Quarterly, Journal of Experimental Education, and Journal of Educational Research. We limited our review to the four year period between 1985 and 1988. Table 4 summarizes what we found. By far the most popular multiple comparison procedure used was the Newman-Keuls. This finding is consistent with the results reported by Jaccard, Becker and Wood (1982) who also found the Newman-Keuls procedure the most popular technique in a survey of articles published by the American Psychological Association in 1982. Contrary to what Jaccard et al found, our survey indicated that Tukey's HSD procedure was the second most popular procedure. In reviewing the five research journals we did not find a single article that used a technique that did not assume equal variances. Perhaps this is not surprising since so few texts discuss the alternative procedures and those that do have only recently been published. It is possible of course that in applied research studies the assumption of equal variances is generally met so the techniques I've be discussing are rarely needed. We thought we might be able address this possibility by examining the descriptive statistics reported in the research articles. Unfortunately we found that many studies do not report indices of spread. Of those studies that did report the sample standard deviations, many appeared to have variances that were homogeneous but several had variances that differed by more than a factor of 2. Unequal sample sizes were generally common in the studies we examined.

Conclusions

Based on what I have read and learned about the issue of contrast analysis with heterogeneous variances I have come the following two conclusions. First I am pretty much convinced that most applied researchers are unaware of the problem and probably are unaware of the alternative solutions when variances differ. Second, I think that for most research studies the Games-Howell procedure will provide a valid test and should be included in statistical methods textbooks and classes.

References

Cochran, W. G. (1964). Approximate significance levels of the Behrens-Fisher test. Biometrics, 20, 191-195.

Dunnett, C. W. (1980). Pairwise multiple comparisons in the unequal variance case. Journal of the American Statistical Association, 75, 796-800.

Games, P. A., & Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal N's and/or variances: A monte carlo study. Journal of Educational Statistics, 1, 113-125.

Games, P. A., Keselman, H. J., & Rogan, J. C. (1981). Simultaneous pairwise multiple comparison procedures for means when sample sizes are unequal. Psychological Bulletin, 90, 594-598.

Hochberg, Y. (1975). Simultaneous inference under Behrens-Fisher conditions: A tow sample approach. Communications in Statistics, 4, 1109-1119.

Hochberg, Y. & Tamhane, A. C. (1987). Multiple Comparison Procedures. New York: John Wiley & Sons.

Howell, J. F. & Games, P. A. (1974). The effects of variance heterogeneity on simultaneous multiple-comaprison procedures with equal sample sizes. British Journal of Mathematical and Statistical Psychology, 27, 72-81.

Jaccard, J., Becker, M. A., & Wood, G. (1984). Pairwise multiple comparison procedures: A review. Psychological Bulletin, 96, 589-596.

Keselman, H. J., Games, P. A., & Rogan, J. C. (1979). An addendum to "A comparison of the modified-Tukey and Scheffe methods of multiple comparisons for pairwise contrasts. Journal of the American Statistical Association, 74, 626-627.

Keselman, H. J. & Rogan, J. C. (1978). A comparison of the modified-Tukey and Scheffe methods of multiple comparisons for pairwise contrasts. Journal of the American Statistical Association, 73, 47-52.

Keselman, H. J. & Rogan, J. C. (1977). The Tukey multiple comparison test: 1953-1976. Psychological Bulletin, 84, 1050-1056.

Lobstein, D. (1983). A multivariate studey of exercise training effects on beta-endorphin and emotionality in psychologically normal, medically healthy men. Unpublished doctoral dissertation Purdue Universtiy, Indiana.

Moore, D. S. & McCabe, G. P. (1989). Introduction to the Practice of Statistics. New York: W. H. Freeman.

Sidak, Z. (1967). Rectangular confidence rregions for the means of multivariate normal distributions. Journal of the American Statistical Association, 62, 626-633.

Stoline, M. R. (1981). The status of multiple comparisons: Simultaneous estimation of all pairwise comparisons in one-way ANOVA designs. Journal of the American Statistical Association, 35, 134-141.

Tamhane, A. C. (1979). A comparison of procedures for multiple comparisons of means with unequal variances. Journal of the American Statistical Association, 74, 471-479.

Tamhane, A. C. (1977) Multiple comparisons in model I one-way ANOVA with unequal variances. Communications in Statistics-Theory and Methods, 1, 15-32.

Wilcox, R. R. (1987). Within row pairwise comparisons in a two-way ANOVA. Communications in Statistics-Simulation, 16, 939-955.

Wilcox, R. R. (1987). New Statistical Procedures for the Social Sciences: Nodern Solutions to Basic Problems. Hilsdale, New Jersey: Lawrence Erlbaum.

Table 1

Means, standard deviations, and sample sizes for four groups investigating the effect of exercise on fitness.

	T	C	J	S
Mean	291.91	308.97	366.87	226.07
Standard Deviations	38.17	32.07	41.19	63.53
Sample Size	10	5	11	10

T=Treatment, C=Control, J=Joggers, S=Sedentary.

Table 2

Summary of half width's for confidence intervals based alternative procedures.

Contrast[a]	Point Estimate	TK[b]	GH	C	T3	Tam	H
T - C	-17.06	69.73	58.58	68.99	61.30	65.00	63.11
T - J	-74.96	55.63	48.97	53.56	50.74	65.00	53.65
T - S	65.89	56.94	68.11	73.25	70.78	65.00	88.40
C - J	-57.90	68.67	59.29	69.05	62.04	65.00	63.11
C - S	82.90	69.73	73.14	85.04	76.27	65.00	88.40
J - S	140.80	55.63	68.14	73.40	70.62	65.00	88.40

[a]T=Treatment, C=Control, J=Joggers, S=Sedentary.

[b]TK=Tukey-Kramer, GH=Games-Howell, C=Dunnett C, T3=Dunnett T3, Tam=Tamhane, H=Hochberg.

Table 3.

Frequency with which the most popular multiple comparison procedures were discussed.

Multiple Comparison Procedure	Frequency	Percent
Tukey	27	69
Scheffe	24	62
Fisher LSD	11	28
Newman-Keuls	10	26
Bonferroni	8	21
Dunnett	6	15
Games and Howell	6	15
Duncan	4	10
Dunnett T3	2	5
Dunnett C	2	5

Table 4.

Frequency counts of the most popular multiple comparison procedures found in journal articles 1985-1988.

Journal[a]	Tukey HSD	Newman-Keuls	Dunn-Bonferroni	Scheffe	Duncan	Fisher LSD
	Procedure					
AERJ	3	6	2	1		
JEP	18	26	7	9	1	3
RRQ	6	11	1	2	1	
JEE	4	3		1		
JER	6	8	2	6		1
Total	37	54	12	19	2	4

[a] AERJ=American Educational Research Journal, JEP=Journal of Educational Psychology, RRQ=Reading Research Quarterly, JEE=Journal of Experimental Education, JER=Journal of Educational Research.