Circulation
JOURNAL OF THE AMERICAN HEART ASSOCIATION
American Heart Association®
Learn and Live℠

Multiple Comparisons Procedures

Howard J. Cabral

Circulation 2008;117;698-701
DOI: 10.1161/CIRCULATIONAHA.107.700971
Circulation is published by the American Heart Association. 7272 Greenville Avenue, Dallas, TX 72514
Copyright © 2008 American Heart Association. All rights reserved. Print ISSN: 0009-7322. Online ISSN: 1524-4539

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circ.ahajournals.org/cgi/content/full/117/5/698

Subscriptions: Information about subscribing to Circulation is online at
http://circ.ahajournals.org/subscriptions/

Permissions: Permissions & Rights Desk, Lippincott Williams & Wilkins, a division of Wolters Kluwer Health, 351 West Camden Street, Baltimore, MD 21202-2436. Phone: 410-528-4050. Fax: 410-528-8550. E-mail:
journalpermissions@lww.com

Reprints: Information about reprints can be found online at
http://www.lww.com/reprints

Downloaded from circ.ahajournals.org at FMRP BRAZ 35042 MEDIC RIB 80 on April 17, 2008

Statistical Primer for Cardiovascular Research

Multiple Comparisons Procedures

Howard J. Cabral, PhD, MPH

In biomedical research, a common question posed by investigators is whether or not an outcome of interest differs significantly between multiple independent groups of subjects in the study sample. For example, in a randomized clinical trial focusing on differences in a parameter of cardiovascular health such as systolic blood pressure or heart rate that is measured on a continuum, one might make the comparison of those who received a placebo, those who received a particular active drug, and those who received a different active drug. Another example of a multiple group comparison might arise in an observational study when comparisons between categories of race or ethnicity are of interest.

The statistical problem that arises from the use of multiple comparisons tests is that any subsequent tests of hypotheses will be performed on the outcome with the same data on which the global test was performed. This can result in an uncontrolled type I error rate (the rate of rejecting the null hypothesis when it should not be rejected). These tests can produce this statistical problem, which can be encountered in analyses of multiple treatment or exposure groups, multiple end points, or multiple interim analyses. This problem has been addressed from a broad perspective.[1] The present report, however, will focus on the statistical analysis strategies used when the global or omnibus test of differences on a continuous outcome across the multiple groups has been performed and statistical tests contrasting subgroups are then conducted. It serves as a follow-up to an earlier article[2] in the series of statistical tutorials in *Circulation* that addressed the use of the ANOVA in performing the global test of hypothesis for a continuous outcome. These statistical tests are often referred to as multiple comparisons procedures (MCPs). We will first present a brief review of the statistical foundations of 1-factor ANOVA and then will describe the 2 main types of MCPs with specific reference to the more commonly used MCPs. Finally, we will show a worked example of an analysis of data from a study of heart size in animals exposed to different conditions of physical exercise that will illustrate the use of 1-factor ANOVA with supplementary MCPs.

Review of 1-Factor ANOVA

The 1-factor ANOVA is used to compare mean values for a continuous outcome of interest across 2 or more independent groups, ie, groups in which subjects belong to only 1 group. In this analysis, the outcome, or dependent variable, is compared between the categories of the grouping, or independent, variable, which is referred to as the "factor." The categories of the factor are referred to as "levels" (whether or not they are ordered in some fashion). We will henceforth refer to these categories as "groups." In the population to which statistical inference is to be made, the outcome is assumed to be measured on a continuum and to follow a gaussian distribution for each group, with statistically independent values across individual subjects. Furthermore, the variances of the outcome are assumed to be equal across the groups. For k groups, the null hypothesis (H_0) is that the population means (μ) of the outcome are the same for all of the groups. This is commonly written in symbolic form as:

$$H_0: \mu_1=\mu_2=\ldots=\mu_k$$

versus the alternative hypothesis, H_1, that the k population means are not equal. This null hypothesis is referred to as the "global" hypothesis and its statistical testing as an "omnibus" test. Note that if the null hypothesis is rejected in the face of sufficient sample data, the question of where particular differences were present between the mean values is not addressed.

In situations such as tests of treatment efficacy in a phase-III clinical trial in which a placebo control is used, for example, it has been argued that rejection of the global null hypothesis is required before one proceeds with additional analyses to identify specific differences between subgroups of the factor of interest.[3] In contrast, it can also be argued that with a limited number of a priori (ie, preplanned) multiple comparisons (for example, in a confirmatory study), one should not have to reject the global null hypothesis to perform the preplanned MCPs.[4] We will adopt the former strategy and thus will next discuss the use of MCPs to answer the question of the significance of differences between specific subgroups assuming that the global test has been rejected at a given α, or "significance," level.

Multiple Comparisons Procedures

As noted above, the performance of multiple hypothesis tests subsequent to the global test can result in an uncontrolled type I error rate (the rate of rejecting the null hypothesis when it should not be rejected). MCPs are applied when the global null hypothesis of the study has been rejected at a given

From the Department of Biostatistics, Boston University School of Public Health, Boston, Mass.
Correspondence to Howard J. Cabral, PhD, MPH, Department of Biostatistics, Boston University School of Public Health, 715 Albany St, Crosstown Center, 3rd Floor, Boston, MA 02118. E-mail hjcab@bu.edu
(*Circulation*. 2008;117:698-701.)
© 2008 American Heart Association, Inc.

Circulation is available at http://circ.ahajournals.org DOI: 10.1161/CIRCULATIONAHA.107.700971

Downloaded from circ.ahajournals.org at FMRP-USP RAZ 35042 MEDIC RIB 80 on April 17, 2008

α-level either (1) in an a priori fashion when specific preselected comparisons are of interest or (2) in a post hoc fashion when the data suggest that specific groups be compared statistically. Commonly, this α-level is set at 0.05 for the experiment or study and is applied to the global test whether or not a priori or post hoc MCPs are conducted.

In determining how to maintain the overall α-level, an investigator must consider an MCP to reduce the α-level for each MCP test or to apply the same α-level for each MCP test as applied in the global test. In mathematical terms, each additional statistical test performed in addition to the global test in such a situation will actually increase the overall α-level for the study. For example, with k groups and interest in comparing each pair of the k mean values, $k(k-1)/2$ possible comparisons exist. If a separate α-level is applied here to each test of hypothesis, the actual α-level for the set of comparisons could be as large as $\alpha[k(k-1)/2]$. Thus, when $k=4$ and $\alpha=0.05$ for each test, the overall α-level for the set of tests could be as large as 0.30: ie, 0.05[4(3)/2]. Furthermore, if interest exists in comparisons beyond those between pairs of means, the number of multiple comparisons tests will increase. For example, in a study of 3 groups (A, B, and C), one might have interest in the following comparisons: A versus B, A versus C, B versus C, A and B versus C, A and C versus B, and B and C versus A (6 total, 3 pairwise comparisons). MCPs that maintain the overall α-level for the set of tests are said to control the "experimentwise" error rate; a related type, called "familywise" error rate control procedures, also effectively reduce the α-level for each post hoc test. We will simplify matters by referring to these 2 classes of procedures as "experimentwise," although technical differences between the 2 must be acknowledged and have been left to the reader to investigate independently.[4]

In contrast, MCPs that apply a separate α-level for each test are called "comparisonwise" error control procedures. In the case of the study with groups A, B, and C, the use of comparisonwise error control after the global null hypothesis has been rejected would entail the performance of 6 individual tests and application of an α-level of 0.05 for each test. Statistically oriented overviews of MCPs,[3] as well as general biostatistical and research texts,[5–8] cover in more detail the distinction between these classes of MCPs and their relative strengths and weaknesses. In addition, the use of MCPs in additional analytical frameworks, such as multiway ANOVA and repeated-measures analysis, is beyond the scope of this article.

Experimentwise Error Control Procedures

As noted above, when interest exists in maintaining the overall α-level for the experiment or study, an investigator may choose an MCP that controls the experimentwise error. Many MCPs have been developed to maintain this overall α-level (the term "experiment" stems from the early development of ANOVA in the context of experimental research). Most statistical software packages that offer applications for general linear models analyses such as ANOVA have also implemented MCPs with an array of choices for these tests. Among the more commonly used procedures in this class are the Tukey (John W. Tukey, PhD, unpublished data, 1953), Dunnett,[9] Scheffé,[10] and Bonferroni (Dunn)[11] tests. The options available for MCPs vary by software package. Investigators should consider their scientific question when choosing an MCP and not limit their choice by the availability of MCPs in the statistical software used for their global test.

The Tukey test is appropriate for the comparison of pairs of group means; originally developed for equal sample sizes per group, a modified version accommodates unequal sample sizes. A Dunnett test is applied in situations in which contrasts are limited to comparisons with a control group and not, for example, between the means of active treatment groups. The Scheffé test is applicable for more general comparisons than the comparison of pairs of group means and is more appropriate than the Tukey test if sample sizes per group differ markedly. It is considered to be more conservative than the Tukey test when pairs of group means are being compared. The Scheffé test can be computed for a specific contrast of group means by first determining the critical value of the F statistic for the α-level of interest, with $k-1$ degrees of freedom (df) for the numerator and $N-k$ degrees of freedom for the denominator when k groups are present and N subjects overall. This F value is then multiplied by $k-1$ to yield a new critical F value for the multiple comparisons contrast.

The Bonferroni (Dunn) procedure takes into account the number of comparisons to be made and is more conservative (less likely to find a significant difference) than the Tukey or Scheffé test in comparisons of pairs of group means, and it is considered to be the most conservative option among MCPs in most situations. It can be applied to general hypothesis tests in addition to ANOVA. The Bonferroni (Dunn) procedure is implemented by computing a new α-level for each multiple comparisons test based only on the overall α-level for the study and the number of comparisons to be made. In this approach, the new α, α", is equal to α/C, where C is the number of post hoc tests to be performed. Thus, in the previously discussed example with 3 groups, A, B, and C, and interest in all possible comparisons, the new alpha, α", would be equal to α/6. Modifications have been made to the Bonferroni procedure with the goal of improving statistical power and include the Holm[12] and Hochberg[13,14] procedures among the more prominent methods.

Comparisonwise Error Control Procedures

When an investigator has a limited number of comparisons to be made after the rejection of the global null hypothesis, especially if these were prespecified before this test was conducted, it may be of interest to employ an MCP that controls the comparisonwise error. As noted previously, MCPs that control the comparisonwise error rate typically use the same α-level for each test that is applied in the test of the global null hypothesis for the study. Thus, the likelihood of falsely rejecting each null hypothesis increases with additional tests. They also, however, provide the benefit to the investigator of being more powerful, ie, more likely to reject the null hypothesis of each test when it should be rejected. Examples of such procedures include the application of Fisher's LSD (least significant difference) test and linear contrasts.[6] In the LSD test, a variant of the standard 2-sample

Downloaded from circ.ahajournals.org at FMRP BRAZ 35042 MEDIC RIB 80 on April 17, 2008

Table 1. Heart Weight/Body Weight Ratio by Exercise-Duration Group

Exercise Duration						
10 min	2.5 d	1 wk	2 wk	3 wk	4 wk	4 wk, 1-wk Rest
4.29	4.49	5.38	5.44	5.50	5.54	4.66
4.43	4.54	5.18	5.59	6.47	5.70	4.90
4.17	4.65	4.83	5.64	6.03	5.47	4.91

Values are heart weight/body weight ratios.

t test is used in which the within-subjects mean square from the global test on the full data set is used as an estimate of the pooled variance, as opposed to using the variance estimates only from the 2 groups being compared. Linear contrasts can be used for more general comparisons, for example, when a set of groups is to be compared with another set of groups or when the study groups represent different dose levels of a treatment or exposure.

Worked Example

We now present an example of an analysis of data from a study of heart size in mice exposed to different conditions of physical exercise[15] that will illustrate the use of 1-factor ANOVA with supplementary MCPs. The design used 2 randomly assigned factors, long-term exercise versus no exercise (control), and exercise duration at 7 different ages after baseline. The sample included 30 mice in total, 21 assigned to the 7 different durations of long-term exercise (swimming) and 9 assigned to 3 durations in the control group (8, 12, and 13 weeks of age). The heart weight–to–body weight ratio at the time of euthanasia was examined as the outcome of interest in this study. We will limit our analyses here to the long-term exercise group to illustrate the use of 1-factor ANOVA. In Table 1, we show the data for this sample.

For these 21 animals, the global F test of differences in mean heart weight–to–body weight ratio for the 1-factor ANOVA was 20.59 (numerator $df=6$, denominator $df=14$; $R^2=0.90$) with $P<0.0001$. Thus, we reject the global null hypothesis of no difference in population mean heart weight–to–body weight ratio between groups using $\alpha=0.05$ and are next interested in where significant differences can be found between subgroups. Although one should in practice restrict the choice of multiple comparisons to those that are substantively meaningful, we will assume for the purpose of illustration in this case that all pairs of means are of interest and will examine results of contrasts between pairs of means using MCPs among those discussed earlier.

In Table 2, we present the means and SDs for the outcome of interest, heart weight–to–body weight ratio, together with a summary of the results of the application of 4 procedures that control the experimentwise type I error rate (Tukey test, Scheffé test, and the Bonferroni [Dunn] procedure) and one that controls the comparisonwise type I error rate (Fisher LSD test). To aid in the interpretation of differences between groups, means have been arranged so that the highest is presented at the top of the table and the lowest at the bottom the table (higher values indicating greater cardiac hypertrophy as a result of exercise).

In Table 2, we adopt a system used in the statistical software, SAS,[16] to identify statistically significant differences between groups. We apply a level of $\alpha=0.05$ to denote statistical significance. In this system, means of groups that share a letter are not significantly different, whereas the means of any groups that do not share a letter are significantly different. For example, for the results of the Tukey test, the "3 weeks" group is significantly different from all groups except "4 weeks" and "2 weeks," because they all share the letter "A" in the display. Likewise, the "10 minutes" group is significantly different from all groups except the "2.5 days" group, because it shares the letter "D" only with that group.

In the table overall, we see a similar set of results for comparisons of all pairs of means among the procedures that control the experimentwise error rate. These tests, however, are rejected at the 0.05 level less frequently than the Fisher's LSD test, which is expected given that the LSD test controls the comparisonwise error rate and should be more powerful. We note, however, that the actual type I error rate for the set of all pairwise comparisons here could be as large as $0.05[(7\times6)/2)]\leq1.00$ ($k=7$). In interpreting these results, however, one should keep in mind that the results of hypothesis tests are highly dependent on sample size, and only 3 mice in each of the 7 groups were examined in this sample. Greater distinction between the findings of these procedures may be observed in larger samples.

Table 2. Means and SDs of Heart Weight/Body Weight Ratio With Results of Multiple Comparisons Procedures (n=21 Mice, 3 per Group)

		Multiple Comparisons Procedure																			
Exercise-Duration Group	Mean (SD)	Tukey				Scheffé				Bonferroni (Dunn)				Fisher LSD							
3 wk	6.00 (0.49)	A				A				A				A							
4 wk	5.57 (0.12)	A	B			A	B			A	B				B						
2 wk	5.56 (0.10)	A	B			A	B			A	B				B						
1 wk	5.13 (0.28)		B	C			B	C			B	C				C					
4 wk, 1 wk of rest	4.82 (0.14)		B	C			B	C	D			C	D			C	D				
2.5 d	4.56 (0.08)			C	D			C	D			C	D					D	E		
10 min	4.30 (0.13)				D				D				D						E		

Means are sorted from largest to smallest, and means of groups that do not share a letter are significantly different.

Downloaded from circ.ahajournals.org at FMRP BRAZ 35042 MEDIC RIB 80 on April 17, 2008

Summary

We have presented information that should be helpful to investigators in biomedical research who have an interest in addressing study questions in which multiple comparisons of study groups are appropriate. We have reviewed the statistical framework for 1-factor ANOVA and have discussed how multiple comparisons after the rejection of the global null hypothesis are conceptually linked to the global test. We have described selected, commonly applied MCPs and have utilized them in the analysis of data from an animal study.

Several of the procedures discussed here can be applied in common statistical software in the context of ANCOVA models when variables are present that need to be controlled statistically to obtain valid inferences about differences between treatment or exposure groups (eg, Bonferroni, Scheffé, Tukey, and Dunnett tests in SAS). We have not covered this situation and leave this to the reader to investigate within their statistical software of choice. Researchers without extensive statistical analysis experience should be able to use this information to work with a professional statistician to better design and analyze data from studies in which multiple comparisons are of interest.

Disclosures

None.

References

1. Koch GG, Gansky SA. Statistical considerations for multiplicity in confirmatory protocols. *Drug Inf J*. 1996;30:523–533.
2. Larson MG. Analysis of variance. *Circulation*. 2008;117:115–121.
3. D'Agostino RB, Heeren TC. Multiple comparisons in over-the counter drug clinical trials with both positive and placebo controls. *Stat Med*. 1991;10:1–6.
4. D'Agostino RB, Massaro J, Kwan H, Cabral H. Strategies for dealing with multiple treatment comparisons in confirmatory clinical trials. *Drug Inf J*. 1993;27:625–641.
5. D'Agostino RB, Sullivan L, Beiser A. *Introductory Applied Biostatistics.* Belmont, Calif: Duxbury-Brooks/Cole; 2004.
6. Rosner B. *Fundamentals of Biostatistics.* 6th ed. Boston, Mass: Duxbury Press; 2005.
7. Kleinbaum DG, Kupper LL, Mueller KE, Nizam A. *Applied Regression Analysis and Multivariable Methods.* 5th ed. Boston, Mass: Duxbury Press; 1997.
8. Toothaker L. *Multiple Comparisons for Researchers.* New York, NY: Sage Publications; 1991.
9. Dunnett CW. New tables for multiple comparisons with a control. *Biometrics*. 1964;20:482–491.
10. Scheffé H. *The Analysis of Variance.* New York, NY: Wiley; 1959.
11. Dunn OJ. Multiple comparisons among means. *J Am Stat Assoc.* 1961; 56:52–64.
12. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat*. 1979;6:65–70.
13. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*. 1988;75:800–803.
14. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. *Stat Med*. 1990;9:811–818.
15. Available at: http://cardiogenomics.med.harvard.edu/groups/proj1/pages/swim_home.html. Accessed January 15, 2008.
16. *Statistical Analysis System (SAS), Version 9.1.* Cary, NC: SAS Institute; 2007.

KEY WORDS: analysis of variance ■ models, statistical ■ statistics

Downloaded from circ.ahajournals.org at FMRP BRAZ 35042 MEDIC RIB 80 on April 17, 2008