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. 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 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. 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 (H₀) is that the population means (μᵢ) of the outcome are the same for all of the groups. This is commonly written in symbolic form as:

\[ H₀: \mu_1 = \mu_2 = \ldots = \mu_k \]

versus the alternative hypothesis, H₁, 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. 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. 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 α.

**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 (H₀) is that the population means (μᵢ) of the outcome are the same for all of the groups. This is commonly written in symbolic form as:

\[ H₀: \mu_1 = \mu_2 = \ldots = \mu_k \]
Experimentwise Error Control Procedures

As noted above, when interest exists in maintaining the overall \( \alpha \)-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 \( \alpha \)-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,\textsuperscript{9} Scheffé,\textsuperscript{10} and Bonferroni (Dunn)\textsuperscript{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 \( \alpha \)-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 test.

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 \( \alpha \)-level for each multiple comparisons test based only on the overall \( \alpha \)-level for the study and the number of comparisons to be made. In this approach, the new \( \alpha \), \( \alpha' \), is equal to \( \alpha/C \), 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 \( \alpha' \) would be equal to \( \alpha/6 \). Modifications have been made to the Bonferroni procedure with the goal of improving statistical power and include the Holm\textsuperscript{12} and Hochberg\textsuperscript{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 \( \alpha \)-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.\textsuperscript{6} In the LSD test, a variant of the standard 2-sample
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 that will illustrate the use of 1-factor ANOVA. In Table 1, we show the data for this outcome of interest in this study. We will limit our analyses here to the long-term exercise group to illustrate the hypothesis of no difference in population mean 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 of the table (higher values indicating greater cardiac hypertrophy as a result of exercise).

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 of 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, 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 $0.05[(7\times6)/2]=1.00 \ (k=7)$. 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 1. Heart Weight/Body Weight Ratio by Exercise-Duration Group

<table>
<thead>
<tr>
<th>Exercise Duration</th>
<th>Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 min</td>
<td>2.5 d</td>
</tr>
<tr>
<td>1 wk</td>
<td>2 wk</td>
</tr>
<tr>
<td>3 wk</td>
<td>4 wk</td>
</tr>
<tr>
<td>4 wk</td>
<td>1 wk Rest</td>
</tr>
</tbody>
</table>

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

Key Words: analysis of variance · models, statistical · statistics
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 75231
Copyright © 2008 American Heart Association, Inc. 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/content/117/5/698

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published
in Circulation can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial
Office. Once the online version of the published article for which permission is being requested is located,
click Request Permissions in the middle column of the Web page under Services. Further information about
this process is available in the Permissions and Rights Question and Answer document.

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

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