Journal of Modern Applied Statistical Methods

Volume 10 | Issue 1 Article 10

5-1-2011

Sample Size Considerations for Multiple Comparison Procedures in ANOVA

Gordon P. Brooks
Ohio University, brooksg@ohio.edu

George A. Johanson
Ohio University, johanson@ohio.edu

Recommended Citation
Brooks, Gordon P. and Johanson, George A. (2011) "Sample Size Considerations for Multiple Comparison Procedures in ANOVA," *Journal of Modern Applied Statistical Methods*: Vol. 10 : Iss. 1 , Article 10.
DOI: 10.22237/jmasm/1304222940

Journal of Modern Applied Statistical Methods
May 2011, Vol. 10, No. 1, 97-109

1538 – 9472/11/$95.00

Sample Size Considerations for Multiple Comparison Procedures in ANOVA

Gordon P. Brooks George A. Johanson
Ohio University,
Athens, Ohio USA

Adequate sample sizes for omnibus ANOVA tests do not necessarily provide sufficient statistical power for post hoc multiple comparisons typically performed following a significant omnibus *F* test. Results reported support a comparison-of-most-interest approach for sample size determination in ANOVA based on effect sizes for multiple comparisons.

Key words: Sample size, multiple comparison procedures, Tukey, ANOVA.

Introduction

The determination of an appropriate sample size is an often difficult, but critically important, element in the research design process. One of the chief functions of experimental design is to ensure that a study has adequate statistical power to detect meaningful differences, if indeed they exist (e.g., Hopkins & Hopkins, 1979). There is a very good reason why researchers should worry about statistical power a priori: If researchers are going to invest time and money in carrying out a study, then they would want to have a reasonable chance, perhaps 70% or 80%, to find a statistically significant difference between groups if it does exist in the population. Thus, a priori power, the probability of rejecting a null hypothesis that is indeed false, will inform researchers about how many subjects per group will be needed for adequate power (Light, Singer & Willett, 1990).

Among the most important matters impacting the choice of sample size is the particular statistical analysis that will be used to analyze data. For example, when a *t* test is used, the researcher commonly estimates an expected, standardized group mean difference effect size (such as Cohen's *d*) in order to determine an appropriate sample size. Sample sizes in analysis of variance (ANOVA) are often based on an effect size that represents an overall standardized difference in the means (such as Cohen's *f*), but these recommended sample sizes provide statistical power only for the omnibus null hypothesis (overall ANOVA) that no group means differ. Adequate sample size for the omnibus test does not necessarily provide sufficient statistical power for the post hoc multiple comparisons typically performed following a statistically significant (exploratory) omnibus test and in many cases the multiple comparisons are of most interest to a researcher.

The purpose of this study was to determine whether the knowledge that multiple comparison procedures will be used following a statistically significant omnibus ANOVA can be helpful in choosing a sample size for a given study. In particular, results using the Tukey HSD post hoc multiple comparison procedure (MCP) were examined to determine whether specific recommendations can be made about sample sizes when the Tukey MCP is used and three groups are compared. This evidence was used to reach conclusions about whether such an approach to sample size selection has merit. Note that this is a presentation of a new approach to sample size selection – specifically, a new way to think about effect sizes – for

Gordon P. Brooks is an Associate Professor of Educational Research and Evaluation. His research interests include statistics education, power and sample size analysis and Monte Carlo programming. Email him at: brooksg@ohio.edu. George A. Johanson is a Professor Emeritus of Educational Research and Evaluation. His research interests include survey research methods and differential item and person functioning. Email him at: johanson@ohio.edu.

exploratory ANOVA where post hoc comparisons are relevant. Other approaches are both more appropriate and more powerful when planned comparisons are made in a confirmatory analysis.

Theoretical Framework

Several factors play a role in sample size determination, including that after the statistical method and the directionality of the statistical alternative hypotheses have been decided, sample size, level of significance, effect size and statistical power are all functionally related. Other issues also impact statistical power, such as the reliability of measurements, unequal group sizes and unequal group variances. However, little consideration has been given to the role of post hoc multiple comparison tests in choosing adequate sample sizes.

In order to maintain reasonable experiment-wise Type I error rates when group means are compared, researchers often use ANOVA followed by an appropriate MCP. The overall ANOVA is tested using an omnibus test at a predetermined level of significance (e.g., 0.05). The post hoc tests that follow a statistically significant omnibus test are then often performed at an adjusted level of significance, based on the number of comparisons to be made.

For example, when comparing four groups, six pairwise group mean comparisons possible. If the researcher wishes to perform all six pairwise comparisons, the per comparison (i.e., per test) level of significance would be adjusted so that the entire set of follow-up tests does not exceed the experiment-wise alpha (e.g., if experiment-wise alpha is 0.05, the adjusted per comparison alpha might be 0.05/6 = 0.0083, using a Bonferroni approach). Each MCP performs this adjustment differently, resulting in different performance for each in terms of Type I error and statistical power (e.g., Carmer & Swanson, 1973; Einot & Gabriel, 1975; Toothaker, 1991).

Several methods exist for determining sample size for ANOVA. Most common are statistical power approaches based on Cohen's (1988) *f* effect size, which represents the standardized variability of the group means about the grand mean (Stevens, 2007). This method (and other similar methods) concentrates on the statistical power of the omnibus test in ANOVA. Others, Hinkle, Wiersma and Jurs (2003) and Levin (1975), for example, have recommended approaches based on how large the sample must be to detect a predetermined mean difference effect size between any two groups, or two extreme groups. Although Levin's approach is designed for use with the Scheffé multiple comparison procedure, Hinkle, et al. base their method on Cohen's *d* effect size for comparison between the two groups with the largest (most extreme) mean differences, and therefore do not consider the adjustments to alpha for multiple comparison procedures. Pan and Dayton (2005) provided sample size requirements for patterns of ordered means, but focused on an information criteria approach to pair-wise comparison procedures.

Comparison-of-Most-Interest

When determining sample sizes for a factorial ANOVA, researchers may choose the sample size that provides sufficient statistical power for all sources of variation (e.g., main effects and interactions). Alternatively, researchers may determine which effect is most important to them and select a sample size based on the expected effect size for that particular source of variation. For example, researchers may have most interest in the interaction effect or a particular main effect. Depending on the structure of the cell means, these effect sizes can vary and therefore result in different required sample sizes for the various main effects and interaction effects.

The approach presented in this study is based loosely on this effect-of-most-interest approach from factorial ANOVA as applied to one-way ANOVA: That is, beyond determining the sample size required for an omnibus test in one-way ANOVA, the new approach also determines the sample sizes required for the follow-up tests from a given set of population means.

For example, in a 3-group study the researcher may be able to estimate that a large effect exists between a control group and two types of treatment, but may expect a much smaller difference between two types of treatment. The comparison-of-most-interest may be the difference between the treatments and the control; however, the much smaller difference between the two treatments may be the most interesting. The researcher would use this information to determine an appropriate sample size for the study by selecting a sample size large enough for the smaller effect size between the types of treatment. This differs from an a priori set of planned comparisons in that the researcher may have a special interest in particular comparisons, but not have specific alternative research hypotheses to predict the direction of the mean differences. The procedure studied here is an adaptation of the Hinkle, et al. (2003) approach that looks at meaningful effect sizes between any groups rather than the Hinkle, et al. difference between only the two most extreme groups.

Even in an exploratory ANOVA, it is rarely satisfactory knowing only that a difference exists in the means (as given by the omnibus test); researchers typically also want to know between which groups the differences exist. Without consideration of the multiple comparison procedures during the sample size analysis, it is possible to find a statistically significant omnibus test with no pairwise group differences determined to be statistically significant in post hoc tests. Although other potential reasons for such a result exist, it may sometimes be an issue of statistical power.

An Example of the Problem

Suppose a researcher is analyzing the mean differences for three groups, where the means for groups 1 and 2 are both 0.0, but the third group mean is 0.8. This represents a relatively large pairwise difference between group 3 and both groups 1 and 2. Using the Cohen (1988) effect size, *f*, for ANOVA, this might be characterized as a relatively large effect: Cohen's large effect size is $f = 0.40$ and in this example $f = 0.38$. Cohen's sample size analysis, as implemented by the SPSS SamplePower program, indicates that 24 cases per group are required to achieve statistical power of 0.80 for the omnibus test in such a situation.

When performing a Monte Carlo analysis for this condition using the MC4G program (Brooks, 2008), approximately 80.8% of 100,000 samples resulted in statistically significant omnibus *F* statistics for the ANOVA among the three groups. However, the number of correct statistically significant Tukey HSD comparisons between groups 1 and 3 and between groups 2 and 3 (with a sample size of 24 in each group), was approximately 64.7%. At the adjusted alpha used by the Tukey HSD procedure, approximately 1.9% of the comparisons between groups 1 and 2 were statistically significant (and therefore Type I errors because both group 1 and 2 had the same mean).

These illustrative power analysis results imply that a number of samples from among the 100,000 had statistically significant omnibus *F* statistics while, at most, one of the non-null Tukey post hoc comparisons was statistically significant. The MC4G program reported that approximately 78.9% of samples had at least one significant Tukey comparison following a significant omnibus test. However, because only 64.7% of each non-null comparison were statistically significant, and because the group 1 versus group 2 comparison was significant as a Type I error in about 1.9% of the samples, this implies that - in many of those samples - only one of the two large, non-null comparisons was statistically significant.

From another perspective, in order to reach statistical power of 0.80 for the two non-null Tukey comparisons (i.e., group 1 vs. group 3 and group 2 vs. group 3), 32 cases are needed per group, for a total sample size of 96 (compared to 24 per group based solely on the omnibus test). With a total sample size of 96 the omnibus *F* test, however, had a power rate of approximately 0.91.

Methodology

An existing Monte Carlo program was modified so that it can ascertain appropriate sample sizes for pairwise comparisons calculated using the Tukey multiple comparison procedure. The MC4G: Monte Carlo Analyses for up to 4 Groups program was originally developed by one of the authors to perform Monte Carlo analyses for *t* tests and ANOVA in a Windows environment (Brooks, 2008). The current version of the program (MC4G version v2008) was upgraded to include the sample size analyses required for this study.

The MC4G program was compiled in Delphi 2007. The program uses the L'Ecuyer (1988) uniform pseudorandom number generator. Specifically, the FORTRAN code of Press, et al. (1992), was translated into Delphi Pascal. The L'Ecuyer generator was chosen due to its large period and because combined generators are recommended for use with the Box-Muller method for generating random normal deviates (Park & Miller, 1988), as is the case in MC4G. The computer algorithm for the Box-Muller method used in MC4G was adapted for Delphi Pascal from the standard Pascal code provided by Press, et al. (1989). Simulated samples were chosen randomly to test program function by comparison with results provided by SPSS.

Monte Carlo Design

In all simulations, normally distributed standardized data were generated to fit the given conditions for each simulation; that is, all variances were set to 1.0, while group means varied between 0.0 and 0.8, depending on the given effect size. A minimum of 10,000 replications were performed for the final sample size analysis in each condition. Specifically, a default value of 20,000 was used with the MC4G sample size analysis, which guaranteed that the final results would be based on at least 10,000 iterations (i.e., simulated samples). Samples sizes for all three groups were restricted to be equal. Some of the Monte Carlo simulations were run multiple times with different seeds to verify that the results were not an artifact of a poor seed choice.

Conditions included varying standardized mean differences among groups for a three-group ANOVA. In particular, groups varied such that all possible non-redundant patterns of pairwise mean differences were varied across groups from 0.0 to 0.8. The minimum non-null standardized mean difference between groups of 0.2 was chosen because of the very large sample sizes required for smaller effects; the maximum of 0.8 was chosen because of the very small sample sizes required when the mean differences are larger.

For example, whether the three group means were set at 0.2, 0.4 and 0.6 or at 0.3, 0.5 and 0.7, the pattern for both resulting standardized mean difference effect sizes (all standard deviations were 1.0) would be 0.2, 0.2 and 0.4, respectively. The mean differences - as effect sizes - are the key to the sample size analyses, not the absolute sizes of the means. Therefore, each pattern of mean differences was only included once. The result was 16 non-redundant comparison patterns that fit the mean difference conditions described (see Table 1).

Results

Three primary findings of interest were observed from this study. First, when the pattern of means resulted in a pattern where two of the three means are equal – and different from the third – there was a consistent pattern of sample sizes required for the comparison relative to the sample size required for the omnibus test. Second, when the pattern of means resulted in two of the three mean differences being equal – and different from the third – there was a consistent pattern of sample sizes required for the comparison relative to the sample size required for the omnibus test. Third, no matter what the pattern of means, a given absolute standardized mean difference effect size consistently required the same sample size to achieve the power desired.

Two Equal Means

In situations where two groups had the same mean and a third group mean differed, the non-null multiple comparisons required larger sample sizes than the omnibus ANOVA. For example, the condition where the pattern of standardized means was 0.0, 0.0 and 0.5 (therefore a pattern of mean differences of 0.0, 0.5 and 0.5) resulted in per group sample sizes of roughly 81 cases to achieve power of 0.80 for the two multiple comparisons with a standardized mean difference of 0.5 (see Table 2). This was compared to the 60 cases per group needed to achieve statistical power of 0.80 for the omnibus test.

All patterns with two similar means, regardless of the magnitude of the mean differences, resulted in a relative efficiency of sample sizes (omnibus per group sample size

Table 1: Patterns of Means Studied

Analysis	Group 1 Mean	Group 2 Mean	Group 3 Mean	Comparison Pattern[a]	Cohen *f* Effect Size	Cohen Total N	Cohen N Per Group
1	0.0	0.0	0.2	0.0, 0.2, 0.2	0.0943	1089	363
2	0.0	0.0	0.3	0.0, 0.3, 0.3	0.1414	486	162
3	0.0	0.0	0.4	0.0, 0.4, 0.4	0.1886	276	92
4	0.0	0.0	0.5	0.0, 0.5, 0.5	0.2357	177	59
5	0.0	0.0	0.6	0.0, 0.6, 0.6	0.2828	126	42
6	0.0	0.0	0.7	0.0, 0.7, 0.7	0.3300	93	31
7	0.0	0.0	0.8	0.0, 0.8, 0.8	0.3771	72	24
8	0.0	0.2	0.4	0.2, 0.2, 0.4	0.1633	366	122
9	0.0	0.2	0.5	0.2, 0.3, 0.5	0.2055	234	78
10	0.0	0.2	0.6	0.2, 0.4, 0.6	0.2494	159	53
11	0.0	0.2	0.7	0.2, 0.5, 0.7	0.2944	117	39
12	0.0	0.2	0.8	0.2, 0.6, 0.8	0.3399	87	29
13	0.0	0.3	0.6	0.3, 0.3, 0.6	0.2449	165	55
14	0.0	0.3	0.7	0.3, 0.4, 0.7	0.2867	123	41
15	0.0	0.3	0.8	0.3, 0.5, 0.8	0.3300	93	31
16	0.0	0.4	0.8	0.4, 0.4, 0.8	0.3266	96	32

[a] Comparison pattern indicates the standardized mean difference between Group 1 vs. Group 2, Group 2 vs. Group 3, and Group 1 vs. Group 3, respectively

divided by multiple comparison per group sample size) of approximately 0.70. Stated another way, in all cases where two groups had the same mean while a third group differed, the multiple comparisons required approximately 1.4 times more cases than the omnibus test did in order to achieve power of 0.80. For example, in the condition where the pattern of means was 0.0, 0.0 and 0.5, the multiple comparisons required 1.35 times more cases than did the overall test. For 0.0, 0.0 and 0.8, the multiple comparisons resulted in 1.38 times more cases. Complete relative efficiency results from the studied conditions can be reviewed in Table 2.

Two Equal Mean Differences

In conditions where two of the three mean differences were the same and the third mean difference was twice as large, the two smaller mean comparisons required a much larger sample size than the overall test, while the third comparison required roughly the same sample size as the omnibus test. For example, in the case where the pattern of means was 0.0, 0.3 and 0.6 (therefore a pattern of mean differences of 0.3, 0.3 and 0.6, respectively), the smaller mean comparisons required approximately 228 cases per group, while the third mean comparison required 57 cases per group. These values were compared to the omnibus test sample size of 55 cases per group for a power rate of 0.80.

Like the two similar means pattern described above, the relative efficiencies of the two similar mean differences pattern were consistent across results. In all cases where two mean differences were the same, the multiple comparison tests required approximately 4.2 times more cases than the omnibus test. For the third, different comparison, approximately 1.1 times more cases were needed. For example, in the 0.0, 0.4, 0.8 condition, the two equal multiple comparison tests (i.e., group 1 vs. group 2 and group 2 vs. group 3) required

Table 2: Sample Size Results for the Tukey HSD Multiple Comparison Procedure for the Primary Monte Carlo Design at Statistical Power of 0.80

Group 1 Mean	Group 2 Mean	Group 3 Mean	Comparison Tested	Total Sample Size	Sample Size per Group	Relative Efficiency[a]
0	0	0.2	Omnibus	1080	360	
			G1 v G2	*	*	
			G2 v G3	1521	507	1.41
			G3 v G1	1524	508	1.41
0	0	0.3	Omnibus	483	161	
			G1 v G2	*	*	
			G2 v G3	678	226	1.40
			G3 v G1	681	227	1.41
0	0	0.4	Omnibus	276	92	
			G1 v G2	*	*	
			G2 v G3	375	125	1.36
			G3 v G1	381	127	1.38
0	0	0.5	Omnibus	180	60	
			G1 v G2	*	*	
			G2 v G3	243	81	1.35
			G3 v G1	246	82	1.37
0	0	0.6	Omnibus	123	41	
			G1 v G2	*	*	
			G2 v G3	171	57	1.39
			G3 v G1	174	58	1.41
0	0	0.7	Omnibus	93	31	
			G1 v G2	*	*	
			G2 v G3	126	42	1.35
			G3 v G1	126	42	1.35
0	0	0.8	Omnibus	72	24	
			G1 v G2	*	*	
			G2 v G3	99	33	1.38
			G3 v G1	99	33	1.38

Notes: * indicates that the Null Hypothesis was true for the given comparison, thus no sample size analysis was performed; [a]Relative efficiency is calculated as the total sample size for the particular comparison divided by the total sample size for the omnibus test for the condition

Table 2 (continued): Sample Size Results for the Tukey HSD Multiple Comparison Procedure for the Primary Monte Carlo Design at Statistical Power of 0.80

Group 1 Mean	Group 2 Mean	Group 3 Mean	Comparison Tested	Total Sample Size	Sample Size per Group	Relative Efficiency[a]
0	0.2	0.4	Omnibus	366	122	
			G1 v G2	1524	508	4.16
			G2 v G3	1527	509	4.17
			G3 v G1	378	126	1.03
0	0.2	0.5	Omnibus	231	77	
			G1 v G2	1524	508	6.60
			G2 v G3	690	230	2.99
			G3 v G1	246	82	1.06
0	0.2	0.6	Omnibus	156	52	
			G1 v G2	1527	509	9.79
			G2 v G3	384	128	2.46
			G3 v G1	171	57	1.10
0	0.2	0.7	Omnibus	114	38	
			G1 v G2	1515	505	13.29
			G2 v G3	246	82	2.16
			G3 v G1	126	42	1.11
0	0.2	0.8	Omnibus	87	29	
			G1 v G2	1527	509	17.55
			G2 v G3	171	57	1.97
			G3 v G1	99	33	1.14
0	0.3	0.6	Omnibus	165	55	
			G1 v G2	684	228	4.15
			G2 v G3	684	228	4.15
			G3 v G1	171	57	1.04
0	0.3	0.7	Omnibus	120	40	
			G1 v G2	675	225	5.63
			G2 v G3	384	128	3.20
			G3 v G1	126	42	1.05

Notes: * indicates that the Null Hypothesis was true for the given comparison, thus no sample size analysis was performed; [a]Relative efficiency is calculated as the total sample size for the particular comparison divided by the total sample size for the omnibus test for the condition

Table 2 (continued): Sample Size Results for the Tukey HSD Multiple Comparison Procedure for the Primary Monte Carlo Design at Statistical Power of 0.80

Group 1 Mean	Group 2 Mean	Group 3 Mean	Comparison Tested	Total Sample Size	Sample Size per Group	Relative Efficiency[a]
0	0.3	0.8	Omnibus	93	31	
			G1 v G2	678	226	7.29
			G2 v G3	246	82	2.65
			G3 v G1	99	33	1.06
0	0.4	0.8	Omnibus	93	31	
			G1 v G2	381	127	4.10
			G2 v G3	378	126	4.06
			G3 v G1	99	33	1.06

Notes: * indicates that the Null Hypothesis was true for the given comparison, thus no sample size analysis was performed; [a]Relative efficiency is calculated as the total sample size for the particular comparison divided by the total sample size for the omnibus test for the condition

approximately 4.10 times more cases than the omnibus test (i.e., 127 vs. 31), while the third different mean comparison (i.e., group 1 vs. group 3) required just 33 cases, for a relative efficiency of 1.06. Very much the same results occurred for the (0.0, 0.2, and 0.4) and (0.0, 0.3, and 0.6) conditions of two similar mean differences (see Table 2).

Absolute Mean Difference Effect Sizes

There were also consistent required sample sizes for absolute standardized group mean difference effect sizes regardless of the pattern of means, that is, regardless of the pattern of means across the three groups, the same sample size was required for any given absolute mean difference (see Table 3). For example, when examining the specific results for a comparison-of-most-interest absolute standardized mean difference of 0.3, no matter whether the pattern of means was (0.0, 0.0, 0.3) or (0.0, 0.3, 0.6) or (0.0, 0.3, 0.8), results indicated that a total sample size of approximately 681 cases (227 per group) was required to achieve a statistical power rate of 0.80 for the comparison with a standardized mean difference effect size of 0.3. Thus, when researchers have a comparison-of-most-interest expected to be approximately 0.3, regardless of the expected effect sizes for the other possible comparisons, they would choose a total sample size of approximately 681 cases. Alternatively, if there are multiple comparisons-of-interest, then researchers in this example would choose 0.3 as the smallest among the set of most interesting comparisons and therefore choose sample sizes based on that smallest comparison-of-interest.

Conclusion

Perhaps even more important than the sample size tables produced for this study is the notion that when a researcher is considering sample size, it may not be sufficient to set sample size for the omnibus test being performed. Clearly, researchers should consider post hoc multiple comparisons in the same way they consider different sources of effects in factorial ANOVA: that is, the most important effects under study must be considered a priori so that adequate sample sizes may be obtained for the tests of those effects. With group comparison procedures such as ANOVA, these comparisons-of-most-interest are very frequently performed using post hoc comparison procedures.

Table 3: Sample Sizes Required for Statistical Power of .80 for the Tukey HSD Multiple Comparison Procedure Given Specific Absolute Standardized Mean Differences (regardless of the pattern of group means)

Standardized Mean Difference Effect Size	Total Sample Size	Per Group Sample Size
0.2	1521	507
0.3	681	227
0.4	381	127
0.5	246	82
0.6	171	57
0.7	126	42
0.8	99	33

These results clearly show that adequate statistical power for the omnibus ANOVA *F* test does not guarantee adequate statistical power for given pairwise MCPs performed post hoc. This condition may result in overall statistical significance for the omnibus *F* test, but no pairwise comparisons showing statistical significance. Although this will occur at times because the omnibus test is reflecting that a non-pairwise comparison is significant (e.g., one group compared to an average of two other groups in an experimental study where one control group is compared to an average of two experimental treatment groups), it will happen sometimes because there is not enough power for the adjusted-alpha MCP being performed by the researcher. In the end, researchers must determine whether they wish to have sufficient power for the overall test or for the often-more-informative post hoc pairwise comparisons. The comparison-of-most-interest approach to sample size selection may be useful for the latter situation.

Results of this study suggest that it may be inappropriate to select a sample size for ANOVA based only on the omnibus test. Clearly the expected pattern among the means has an impact on the usually important post hoc pairwise multiple comparisons. This may be analogous to situations involving other statistical methods, such as principal components analysis and MANOVA, where the pattern of correlations has an important impact on the power of the analyses, and therefore also sample size determination. Additionally, it is clear that the absolute size of the given comparison is also important. Both of these findings could be useful to researchers as they plan studies that will use ANOVA.

Sample Size Recommendations

Based on the results generated, certain specific recommendations can be made concerning sample sizes that researchers should use with ANOVA with three groups. It should be remembered that these results were limited to Tukey HSD comparisons performed using statistical power of 0.80. In particular, these recommendations follow from the three cases identified in the results.

Case 1: Two Equal Means

A researcher may be using two control groups and a single treatment group; alternatively, the researcher might expect two treatment groups each to be equally different from the single control group. In such cases, the researcher should determine the sample size required for the omnibus ANOVA test and then multiply that sample size by 1.4 to obtain the sample size required for the Tukey comparisons between the differing groups. For example, in a case where a single treatment group is expected to differ from two control groups by 0.6 (i.e., means of 0.6, 0.0 and 0.0 for the three groups, respectively), the researcher would determine that approximately 123 total cases are needed for the omnibus test to have statistical power of 0.80. If the researcher wants statistical power of 0.80 for the post hoc multiple comparisons, however, approximately (123 * 1.4) or 173 cases are needed.

Case 2: Two Equal Mean Differences

A researcher may expect one treatment to have twice the effect of the second treatment when each is compared to the third group (e.g., a control group). In such cases, the researcher should calculate the sample size required for the omnibus test and then multiply that sample size by 4.1 to obtain the sample size required for the Tukey comparisons between the differing groups. For example, in a case where the expected pattern of means across groups is 0.0, 0.3 and 0.6, the researcher would determine that approximately 165 total cases are needed for the omnibus test to have statistical power of 0.80. If the researcher wants statistical power close to 0.80 for the post hoc multiple comparisons, however, approximately (165 * 4.1) or 677 cases are needed.

Case 3: Absolute Mean Difference Effect Sizes

A researcher may expect that a certain pair of groups will differ by a given amount – no matter how they each differ from the third group. For example, a researcher may consider the comparison between group 1 and group 2 to be the most important and expect them to differ by a standardized mean difference of 0.5. In such a case, how much group 1 or group 2 differs from group 3 is irrelevant. Table 3 shows that 246 total cases are needed for the specific Tukey comparison between group 1 and group 2, given the expected mean difference of 0.5. In such a case, the sample size required for the omnibus test is also irrelevant, because in all cases the recommended sample sizes for the Tukey comparisons are larger than those required for the omnibus ANOVA test.

If however, the researcher expects a pattern of means that does not fit into Case 1 or Case 2 above, the absolute size of the expected mean differences can be used with Table 3. For example, if the means for group 1, group 2, and group 3 are expected to be 0.0, 0.3 and 0.8, respectively, then (a) 681 total cases are needed for the Tukey comparison between groups 1 and 2, where the standardized mean difference is expected to be 0.3, (b) a total sample size of 99 is needed for the expected standardized difference of 0.8 between group 1 and group 3, and (c) 246 total cases are needed for the Tukey comparison between group 2 and group 3. If all three comparisons are considered equally important, the researcher would choose 681 total cases in order to have statistical power of at least 0.80 for all comparisons. However, if the comparison-of-most-interest is the group 2 versus group 3 comparison, then the 246 total cases may be the sample size selected.

Pilot Studies and Monte Carlo Analyses

The results show that the sample size required for the omnibus *F* statistic to reach a given level of statistical power is frequently not sufficient for the non-null multiple comparisons to achieve the same power. In fact, it could be argued that using sample sizes chosen based on Cohen's *f* are inappropriate even when the study is completely exploratory and the researcher has absolutely no research hypothesis concerning the mean differences. When the work is completely exploratory, it may be even more critical to have enough statistical power to find non-null multiple comparisons, rather than simply finding that there is a difference among means somewhere.

An expected pattern of means might be available in relevant literature. However, when the relevant literature provides few clues about such effect sizes, another way to determine sample sizes for a multiple group comparison study might be to conduct a pilot study using a sampling strategy very similar to what will be used in the final study. That is, one cannot necessarily expect pilot study samples chosen conveniently to produce results similar to those obtained from representative random samples from a given population. A well-done pilot study sample, however, might provide clues to the pattern of means, the pattern of mean differences, or the absolute sizes of the mean differences the researcher might expect in the population, thereby helping to determine what sample sizes might be necessary to have sufficient power for the post hoc comparisons. These standardized mean difference effect sizes could then be used in a Monte Carlo analysis, much as was performed for this study, to determine the necessary sample sizes for the post hoc MCPs. Because the results presented here are limited to only a few specific conditions with statistical power of 0.80, the use of Monte Carlo analyses for other circumstances may be critical because sample size tables do not exist for most multiple comparison procedures.

Finally, it is important to note that with enough evidence or knowledge about the groups, exploratory ANOVA may not be a good choice, that is, there may be times there exists enough information to estimate a group mean difference without being able to predict a directional difference between those means. In such cases, the comparison-of-most-interest approach may be useful. However, when enough information is available to make such a prediction, statistical power would be gained by using directional tests and planned contrasts in the analyses described herein.

Future Research

A variety of questions, both philosophical and practical, exist that might be posed for future research based on the results presented. A few suggestions are:

Other Procedures Designed to Control Alpha-Inflation when Multiple Tests are Performed

Although several ad hoc analyses suggested that these results might hold also for Tukey comparisons at other statistical power levels, this would need to be confirmed by further study. Similarly, some analyses performed for Bonferroni revealed the same three cases of results reported here, but would need to be examined with further study. Future research might also investigate whether similar results occur for other multiple comparison procedures (e.g., Fisher LSD, Scheffé, Dunnett). Similarly, additional research should investigate the impact of unequal sample sizes and unequal variances across groups on the total sample sizes required to achieve target levels of statistical power for specialized MCPs (e.g., Games-Howell). Further, how this comparison-of-most-interest approach works within factorial ANOVA, as follow-up to statistically significant main effects, may also be worth investigating.

In light of other approaches that control the increase in Type I errors that occur when multiple null hypothesis tests are performed, it may be argued that perhaps MCPs should be abandoned altogether. For example, researchers could explore the effect on sample size when the Holm (1979) procedure is used (Green & Salkind, 2005; Lubrook, 1998) or when the Benjamini and Hochberg (1995) False Discovery Rate approach is used (Thissen, Steinberg & Kuang, 2002; Williams, Jones & Tukey, 1999), or perhaps no adjustment to alpha should be made for multiple comparison procedures, as is often the case when the statistical significance of regression coefficients is examined following a statistically significant regression model – this too, would impact sample size requirements.

Cross Validation

There are very different ways to think about how to determine required sample sizes for research; perhaps statistical power analyses are not the best way to determine sample size at all. Future research could investigate whether some adaptation of the cross-validity approaches recommended for multiple regression (e.g., Algina & Keselman, 2000; Brooks & Barcikowski, 1996; Park & Dudycha, 1974; Stevens, 2007) would be more useful for researchers in group comparison studies. The basic idea behind the cross-validation approaches is that researchers would be more likely to find results, especially effect sizes, that will replicate if sample sizes are large enough for cross-validation.

A Priori Contrasts and t Tests

Future researchers could compare these results to multiple individual *t* tests or other planned comparisons performed as a priori contrasts when using either an adjusted or unadjusted alpha. It may be that MCP sample sizes are functionally related to *t* test sample sizes using a relative efficiency approach similar to that done in this study. Future researchers might investigate whether the results change if only a subset of more important pairwise comparisons are performed (e.g., simple or repeated contrasts), instead of all possible pairwise comparisons. Similar analyses might also be performed for common non-pairwise comparisons, such as Helmert or polynomial contrasts.

Relative Efficiency

Although no function emerged for some mean difference patterns in the three-group analyses, there may be a less obvious function at work. One could study how well relative efficiency works with larger numbers of groups, with effect sizes larger or smaller than those investigated here, and with different statistical power targets than 0.80. A similar study with four or more groups would involve many more possible mean difference patterns, but could help to provide answers to some of these questions. Such a study would also verify whether such results occur with more than three groups. Finally, the present study can be modified to include non-normal data and different sample sizes in each group.

References

Algina, J., & Keselman, H. J. (2000). Cross-validation sample sizes. *Applied Psychological Measurement*, *24*, 173–179.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *Series B*, *57*, 289-300.

Brooks, G. P. (2008). *MC4G: Monte Carlo Analyses for up to 4 Groups*. [Computer software and manuals]. Retrieved from {http://www.ohio.edu/people/brooksg/software.htm}.

Brooks, G. P., & Barcikowski, R. S. (1999, April). *The precision efficacy analysis for regression sample size method*. Paper presented at the meeting of the American Educational Research Association, Montreal, Canada (ERIC Document Reproduction Service No. ED 449 177).

Carmer, S. G., & Swanson, M. R. (1973). An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. *Journal of the American Statistical Association*, *68*, 66-74.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (*2nd Ed.*). Hillsdale, NJ: Lawrence Erlbaum Associates.

Einot, I., & Gabriel, K. R. (1975). A study of the powers of several methods of multiple comparisons. *Journal of the American Statistical Association*, *70*, 574-583.

Green, S. B., & Salkind, N. J. (2005). *Using SPSS for Windows and Macintosh: Analyzing and understanding data* (*4th Ed.*). Upper Saddle River, NJ: Pearson Prentice Hall.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (*5th Ed.*). Boston: Houghton Mifflin.

Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65-70.

Hopkins, K. D., & Hopkins, B. R. (1979). The effect of the reliability of the dependent variable on power. *Journal of Special Education*, *13*, 463-466.

L'Ecuyer, P. (1988). Efficient and portable combined random number generators. *Communications of the ACM*, *31*, 742-749, 774.

Levin, J. R. (1975). Determining sample size for planned and post hoc analysis of variance comparisons. *Journal of Educational Measurement*, *12*, 99-108.

Light, R. J., Singer, J. D., & Willett, J. B. (1990). *By design: Planning research on higher education*. Cambridge, MA: Harvard University.

Ludbrook, J. (1998). Multiple comparison procedures updated. *Clinical and Experimental Pharmacology and Physiology*, *25*, 1032–1037.

Pan, X., & Dayton, C. M. (2005). Sample Size Selection for Pair-Wise Comparisons Using Information Criteria. *Journal of Modern Applied Statistical Methods*, *4*(*2*), 601-608.

Park, C. N., & Dudycha, A. L. (1974). A Cross-Validation Approach to Sample Size Determination for Regression Models. *Journal of the American Statistical Association*, *69*, 214-218.

Park, S. K., & Miller, K. W. (1988). Random number generators: Good ones are hard to find. *Communications of the ACM*, *31*, 1192-1201.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1989). *Numerical recipes in Pascal: The art of scientific computing*. New York: Cambridge University.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in FORTRAN: The art of scientific computing* (*2nd Ed.*). New York: Cambridge University.

Stevens, J. (2007). *Intermediate statistics: A modern approach* (*3rd Ed.*). New York: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics*, *27*, 77-83.

Toothaker, L. E. (1991). *Multiple comparisons for researchers*. Newbury Park: Sage.

Williams, V. S. L., Jones, L. V., & Tukey, J.W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics*, *24*, 42-69.