Analysis of Variance: The Fundamental Concepts

Steven F. Sawyer, PT, PhD

Analysis of variance (ANOVA) is a statistical tool used to detect differences between experimental group means. ANOVA is warranted in experimental designs with one dependent variable that is a continuous parametric numerical outcome measure, and multiple experimental groups within one or more independent (categorical) variables. In ANOVA terminology, independent variables are called factors, and groups within each factor are referred to as levels. The array of terms that are part and parcel of ANOVA can be intimidating to the uninitiated, such as: partitioning of variance, main effects, interactions, factors, sum of squares, mean squares, F scores, familywise alpha, multiple comparison procedures (or post hoc tests), effect size, statistical power, etc. How do these terms pertain to p values and statistical significance? What precisely is meant by a "statistically significant ANOVA"? How does analyzing variance result in an inferential decision about differences in group means? Can ANOVA be performed on non-parametric data? What are the virtues and potential pitfalls of ANOVA? These are the issues to be addressed in this primer on the use and interpretation of ANOVA. The intent is to provide the clinician reader, whose misspent youth did not include an enthusiastic reading of statistics textbooks, an understanding of the fundamentals of this widely used form of inferential statistical analysis.

ANOVA General Linear Models

ANOVA is based mathematically on linear regression and general linear models that quantify the relationship between the dependent variable and the independent variable(s). There are three different general linear models for ANOVA: (i) Fixed effects model (Model 1) makes inferences that are specific and valid only to the populations and treatments of the study. For example, if three treatments involve three different doses of a drug, inferential conclusions can only be drawn for those specific drug doses. The levels within each factor are fixed as defined by the experimental design. (ii) Random effects model (Model 2) makes inferences about levels of the factor that are not used in the study, such as a continuum of drug doses when the study only used three doses. This model pertains to random effects within levels, and makes inferences about a population's random variation. (iii) Mixed effects model (Model 3) contains both Fixed and Random effects.

In most types of orthopedic rehabilitation clinical research, the Fixed effects model is relevant since the statistical inferences being sought are fixed to the levels of the experimental design. For this reason, the Fixed effects model will be the focus of this article. Computer statistics programs typically default to the Fixed effects model for ANOVA analysis, but higher end programs can perform ANOVA with all three models.

ABSTRACT: Analysis of variance (ANOVA) is a statistical test for detecting differences in group means when there is one parametric dependent variable and one or more independent variables. This article summarizes the fundamentals of ANOVA for an intended benefit of the clinician reader of scientific literature who does not possess expertise in statistics. The emphasis is on conceptually-based perspectives regarding the use and interpretation of ANOVA, with minimal coverage of the mathematical foundations. Computational examples are provided. Assumptions underlying ANOVA include parametric data measures, normally distributed data, similar group variances, and independence of subjects. However, normality and variance assumptions can often be violated with impunity if sample sizes are sufficiently large and there are equal numbers of subjects in each group. A statistically significant ANOVA is typically followed up with a multiple comparison procedure to identify which group means differ from each other. The article concludes with a discussion of effect size and the important distinction between statistical significance and clinical significance.

KEYWORDS: Analysis of Variance, Interaction, Main Effects, Multiple Comparison Procedures
Assumptions of ANOVA

Assumptions for ANOVA pertain to the underlying mathematics of general linear models. Specifically, a data set should meet the following criteria before being subjected to ANOVA:

Parametric data: A parametric ANOVA, the topic of this article, requires parametric data (ratio or interval measures). There are non-parametric, one-factor versions of ANOVA for non-parametric ordinal (ranked) data, specifically the Kruskal-Wallis test for independent groups and the Friedman test for repeated measures analysis.

Normally distributed data within each group: ANOVA can be thought of as a way to infer whether the normal distribution curves of different data sets are best thought of as being from the same population or different populations (Figure 1). It follows that a fundamental assumption of parametric ANOVA is that each group of data (each level) be normally distributed. The Shapiro-Wilk test is commonly used to test for normality for group sample sizes (N) less than 50; D’Agostino’s modification is useful for larger samplings (N>50).

A normal distribution curve can be described by whether it has symmetry about the mean and the appropriate width and height (peakedness). These attributes are defined statistically by “skewness” and “kurtosis”, respectively. A normal distribution curve will have skewness = 0 and kurtosis = 3. (Note that an alternative definition of kurtosis subtracts 3 from the final value so that a normal distribution will have kurtosis = 0. This “minus 3” kurtosis value is sometimes referred to as “excess kurtosis” to distinguish it from the value obtained with the standard kurtosis function. The kurtosis value calculated by many statistical programs is the “minus 3” variant but is referred to, somewhat misleadingly, as “kurtosis.”). Normality of a data set can be assessed with a z-test in reference to the standard error of skewness (estimated as √[6 / N]) and the standard error of kurtosis (estimated as √[24 / N]). A conservative alpha of 0.01 (z ≥ ...
Homogeneity of variance within each group: Referring again to the notion that ANOVA compares normal distribution curves of data sets, these curves need to be similar to each other in shape and width for the comparison to be valid. In other words, the amount of data dispersion (variance) needs to be similar between groups. Two commonly invoked tests of homogeneity of variance are by Levene and Brown & Forsythe.

Independent Observations: A general assumption of parametric analysis is that the value of each observation for each subject is independent of (i.e., not related to or influenced by) the value of any other observation. For independent groups designs, this issue is addressed with random sampling, random assignment to groups, and experimental control of extraneous variables. This assumption is an inherent concern for repeated measures designs, in which an assumption of sphericity comes into play. When subjects are exposed to all levels of an independent variable (e.g., all treatments), it is conceivable that the effects of a treatment can persist and affect the response to subsequent treatments. For example, if a treatment effect for one level has a long half-time (analogous to a drug effect) and there is inadequate “wash out” time between exposure to different levels (treatments), there will be a carryover effect. A well designed and executed cross-over experimental design can mitigate carryover effects. Mauchly’s test of sphericity is commonly employed to test the assumption of independence in repeated measures designs. If the Mauchly test is statistically significant, corrections to the F score calculation are warranted. The two most commonly used correction methods are the Greenhouse-Geisser and Huynh-Feldt, which calculate a descriptive statistic called epsilon, which is a measure of the extent to which sphericity has been violated. The range of values for epsilon are 1 (no sphericity violation) to a lower boundary of 1 / (m−1), where m = number of levels. For example, with three groups, the range would be 1 to 0.50. The closer epsilon is to the lower boundary, the greater the degree of violation. There are three options for adjusting the ANOVA to account for the sphericity violation, all of which involve modifying degrees of freedom: use the lower boundary epsilon, which is the most conservative approach (least powerful) and will generate the largest p value, or use either the Greenhouse-Geisser epsilon or the Huynh-Feldt epsilon (most powerful) [statistical power is the ability of an inferential test to detect a difference that actually exists, i.e., a true positive].

Most commercially available statistics programs perform normality, homogeneity of variance and sphericity tests. Determination of the parametric nature of the data and soundness of the experimental design is the responsibility of the investigator, reviewers and critical readers of the literature.

Robustness of ANOVA to Violations of Normality and Variance Assumptions

ANOVA tests can handle moderate violations of normality and equal variance if there is a large enough sample size and a balanced design. As per the central limit theorem, the distribution of sample means approximates normality even with population distributions that are grossly skewed and non-normal, so long as the sample size of each group is large enough. There is no fixed definition of “large enough”, but a rule of thumb is N≥30. Thus, the mathematical validity of ANOVA is said to be “robust” in the face of violations of normality assumptions if there is an adequate sample size. ANOVA is more sensitive to violations of the homogeneity of variance assumption, but this is mitigated if sample sizes of factors and levels are equal or nearly so. If normality and homogeneity of variance violations are problematic, there are three options: (i) Mathematically transform (log, arcsin, etc.) the data to best mitigate the violation, with the cost of cognitive fog in understanding the meaning to the ANOVA results (e.g., “A statistically significant main effect was obtained for the arcsin transformation of degrees of ankle range of motion”). (ii) Use one of the non-parametric ANOVAs mentioned above, but at the cost of reduced power and being limited to one-factor analysis. (iii) Identify outliers in the data set using formal statistical criteria (not discussed here). Use caution in deleting outliers from the data set; such decisions need to be justified and explained in research reports. Removal of outliers will reduce deviations from normality and homogeneity of variance.

If You Understand t-Tests, You Already Know A Lot About ANOVA

As a starting point, the reader should understand that the familiar t-test is an ANOVA in abbreviated form. A t-test is used to infer on statistical grounds whether there are differences between group means for an experimental design with (i) one parametric dependent variable and (ii) one independent variable with two levels, i.e., there is one outcome measure and two groups. In clinical research, levels often correspond to different treatment groups; the term “level” does not imply any ordering of the groups.

The Null statistical hypothesis for a t-test is H0: μ1 = μ2, that is, the population means of the two groups are the same. Note that we are dealing with population means, which are almost always unknown and unknowable in clinical research. If the Null hypothesis involved sample means, there would be nothing to infer, since descriptive analysis provides this information. However, with inferential analysis using t-tests and ANOVA, the aim is to infer, without access to “the truth”; if the group population means differ from each other.

The Alternative hypothesis, which comes into play if the Null hypothesis is rejected, asserts that the group popula-
tion means differ. The Null hypothesis is rejected when the $p$ value yielded by the $t$-test is less than alpha. Alpha is the pre-determined upper limit risk for committing a Type I error, which is the statistical false positive of incorrectly rejecting the Null hypothesis and inferring the groups means differ when in fact the groups are from a single population. By convention, alpha is typically set to 0.05. The $p$ value generated by the $t$-test statistic is based on numerical analysis of the experimental data, and represents the probability of committing a Type I error if the Null hypothesis is rejected. When $p$ is less than alpha, there is a statistically significant result, i.e., the values in the two groups are inferred to differ from each other and to represent separate populations. The logic of statistical inference is analogous to a jury trial: at the outset of the trial (inferential analysis), the group data are presumed to be innocent of having different population means ( Null hypothesis) unless the differences in group means in the sampled data are sufficiently compelling to meet the standard of “beyond a reasonable doubt” ($p$ less than alpha), in which case a guilty verdict is rendered (reject Null hypothesis and accept Alternative hypothesis = statistical significance).

The test statistic for a $t$-test is the $t$ score. In conceptual terms, the calculation of a $t$ score for independent groups (i.e., not repeated measures) is as follows:

$$t = \frac{\text{statistical signal} / \text{statistical noise}}{\text{treatment effect} / \text{unexplained variance} (\text{"error variance"})}$$

$$t = \text{differences between sample means of the two groups} / \text{within-group variance}$$

The difference in group means represents the statistical signal since it is presumed to result from treatment effects of the different levels of the independent variable. The within-group variance is considered to be statistical noise and an “error” term because it is not explained by the influence of the independent variable on the dependent variable. The particulars of how the $t$ score is calculated depends on the experimental design (independent groups vs repeated measures) and whether variance between groups is equivalent; the reader is to referred to any number of statistics books for details about the formulae. The $t$ score is converted into a $p$ value based on the magnitude of the $t$ score (larger $t$ scores lead to smaller $p$ values) and the sample size (which relates to degrees of freedom).

**ANOVA Null Hypothesis and Alternative Hypothesis**

ANOVA is applicable when the aim is to infer differences in group values when there is one dependent variable and more than two groups, such as one independent variable with three or more levels, or when there are two or more independent variables. Since an independent variable is called a “factor”, ANOVAs are described in terms of the number of factors; if there are two independent variables, it is a two-factor ANOVA. In the simpler case of a one-factor ANOVA, the null hypothesis asserts that the population means for each level (group) of the independent variable are equal. Let’s use as an example a fictitious experiment with one dependent variable (pre/post changes in ankle range of motion in subjects who received one of three types of manual therapy treatment after surgical repair of a talus fracture). This constitutes a one-factor ANOVA with three levels (the three different types of treatment). The null hypothesis is: $H_0: \mu_1 = \mu_2 = \mu_3$. The Alternative hypothesis is that at least two of group means differ. Figure 1 provides a graphical presentation of this ANOVA statistical hypotheses: (i) the null hypothesis (left graph) asserts that the normal distribution curves of data for the three groups are identical in shape and position and therefore precisely overlap, whereas (ii) the Alternative hypothesis (right graph) asserts that these normal distribution curves are best described by the distribution indicated by the sample means, which represent an experimentally-derived estimate of the population means.

**The Mechanics of Calculating a One-factor ANOVA**

ANOVA evaluates differences in group means in a round-about fashion, and involves the “partitioning of variance” from calculations of “Sum of Squares” and “Mean Squares.” Three metrics are used in calculating the ANOVA test statistic, which is called the $F$ score (named after R.A. Fisher, the developer of ANOVA): (i) Grand Mean, which is the mean of all scores in all groups; (ii) Sum of Squares, which are of two kinds, the sum of all squared differences between group means and the Grand Mean (between-groups Sum of Squares) and the sum of squared differences between individual data scores and their respective group mean (within-groups Sum of Squares), and (iii) Mean Squares, also of two kinds (between-groups Mean Squares, within-groups Mean Squares), which are the average deviations of individual scores from their respective mean, calculated by dividing Sum of Squares by their appropriate degrees of freedom.

A key point to appreciate about ANOVA is that the data set variance is partitioned into statistical signal and statistical noise components to generate the $F$ score. The $F$ score for independent groups is calculated as:

$$F = \frac{\text{statistical signal} / \text{statistical noise}}{\text{treatment effect} / \text{unexplained variance} (\text{"error variance"})}$$

$$F = \frac{\text{Mean Squares Between Groups}}{\text{Mean Squares Within Groups (Error)}}$$

Note that the statistical signal, the MS$_{\text{Between Groups}}$, is an indirect measure of differences in group means. The MS$_{\text{Within Groups (Error)}}$ term is considered to represent statistical noise/error since this variance is not explained by the effect of the independent variable on the dependent variable. Here is the gist of the issue: as group means increasingly diverge from each other, there is increasingly more variance for between-group scores in relation to the Grand Mean, quantified as Sum of Squares$_{\text{Between Groups}}$, leading to a larger MS$_{\text{Between Groups}}$ term and a larger $F$ score. Conversely, as there is more variance within-group scores, quantified as Sum of Squares$_{\text{Within Groups (Error)}}$, the MS$_{\text{Within Groups (Error)}}$ term will increase, leading to a smaller $F$ score. Thus, for independent groups, large $F$ scores arise from large differences between group
means and/or small variances within groups. Larger F scores equate to lower p values, with the p value also influenced by the sample size and number of groups, each of which constitutes separate types of “degrees of freedom.”

ANOVA calculations are now the domain of computer software, but there is illustrative and heuristic value in manually performing the arithmetic calculation of the F score to garner insight into how analysis of data set variance generates a statistical inference about differences in group means. A numerical example is provided in Figure 1, in which the data set graphed in Figure 1 is listed and subjected to ANOVA, yielding a calculated F score and corresponding p value.

**Mathematical Equivalence of t-tests and ANOVA: t-tests are a Special Case of ANOVA**

Let’s briefly return to the notion that a t-test is a simplified version of ANOVA that is specific to the case of one independent variable with two groups. If we analyze the data in Figure 2 for the Type 1 treatment vs. Type 3 treatment group data (disregarding the Type 2 treatment group data to reduce the analysis to two groups), the t score for independent groups is 5.0 with a p value of 0.0025 (calculations not shown). For the same data assessed with ANOVA, the F score is 25.0 with a p value of 0.0025. The t-test and ANOVA generate identical p values. The mathematical relation between the two test statistics is: \( t^2 = F \).

**Repeated Measures ANOVA: Different Error Term, Greater Statistical Power**

The experimental designs emphasized thus far entail independent groups, in which each subject is “exposed” to only one level of an independent variable. In

---

**FIGURE 2.** The mechanics of calculating a F score for a one-factor ANOVA with independent groups by partitioning the data set variance as Sum of Squares and Mean Squares are shown below. This fictitious data set lists increased ankle range of motion pre/post for three different types of manual therapy treatments. For the sake of clarity and ease of calculation, a data set with an inappropriately small sample size is used.

<table>
<thead>
<tr>
<th>Subject Gender</th>
<th>Manual Therapy Type 1</th>
<th>Manual Therapy Type 2</th>
<th>Manual Therapy Type 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>14</td>
<td>16</td>
<td>20</td>
</tr>
<tr>
<td>Male</td>
<td>14</td>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td>Female</td>
<td>11</td>
<td>13</td>
<td>17</td>
</tr>
<tr>
<td>Female</td>
<td>13</td>
<td>13</td>
<td>17</td>
</tr>
<tr>
<td>Group Means</td>
<td>13</td>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td>Grand Mean</td>
<td>15</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the following, SS = Sum of Squares; MS = Mean Squares; df = degrees of freedom.

\[
SS_{\text{Total}} = SS_{\text{Between Groups}} + SS_{\text{Within Groups (Error)}}, \text{ and is calculated by summing the squares of differences between each data value vs. the Grand Mean. For this data set with a Grand Mean of 15:}
\]

\[
\]

\[
SS_{\text{Within Groups (Error)}} = SS_{\text{MT treatment Type 1 (Error)}} + SS_{\text{MT treatment Type 2 (Error)}} + SS_{\text{MT treatment Type 3 (Error)}}, \text{ in which the sum of squares within each group is calculated in reference to the group’s mean:}
\]

\[
SS_{\text{MT treatment Type 1 (Error)}} = (14-15)^2 + (14-15)^2 + (11-15)^2 + (13-15)^2 = 6
\]

\[
SS_{\text{MT treatment Type 2 (Error)}} = (16-15)^2 + (14-15)^2 + (13-14)^2 + (13-14)^2 = 6
\]

\[
SS_{\text{MT treatment Type 3 (Error)}} = (20-15)^2 + (18-18)^2 + (17-18)^2 + (17-18)^2 = 6
\]

\[
SS_{\text{Within Groups (Error)}} = 6 + 6 + 6 + 6 = 24. \text{ By subtraction, } SS_{\text{Between Groups}} = 74 - 18 = 56
\]

\( df \) refers to the number of independent measurements used in calculating a Sum of Squares.

\[
df_{\text{Between Groups}} = (\# \text{ of groups} - 1) = (3-1) = 2
\]

\[
df_{\text{Within Groups (Error)}} = (N - \# \text{ of groups}) = (12-1) = 9
\]

ANOVA test statistic, the F score, is calculated from Mean Squares \((SS/df)\):

\[
F = \frac{\text{Mean Squares}_{\text{Between Groups}}}{\text{Mean Squares}_{\text{Within Groups (Error)}}}
\]

\[
\text{Mean Squares}_{\text{Between Groups}} = \frac{SS_{\text{Between Groups}}}{df_{\text{Between Groups}}} = \frac{56}{2} = 28
\]

\[
\text{Mean Squares}_{\text{Within Groups (Error)}} = \frac{SS_{\text{Within Groups (Error)}}}{df_{\text{Within Groups (Error)}}} = \frac{18}{9} = 2
\]

So, \( F = \frac{28}{14} \approx 2 \)

With \( df_{\text{Between Groups}} = 2 \) and \( df_{\text{Within Groups (Error)}} = 9, \) this F score translates into \( p = 0.0017, \) a statistically significant result for alpha = 0.05.
the data set of Figure 2, this would involve each subject receiving only one of the three different treatments. If a subject is exposed to all levels of an independent variable, the mechanics of the ANOVA are altered to take into account that each subject serves as their own experimental control. Whereas the term for statistical signal, MS\textsubscript{Between Groups}, is unchanged, there is a new statistical noise term called MS\textsubscript{Within Subjects (Error)} that pertains to variance within each subject across all levels of the independent variable instead of between all subjects within one level. Since there is typically less variation within subjects than between subjects, the statistical error term is typically smaller in repeated measures designs. A smaller MS\textsubscript{Within Subjects (Error)} value leads to a larger F value and a smaller p value. As a result, repeated measures ANOVA typically have greater statistical power than independent groups ANOVA.

**Factorial ANOVA: Main Effects and Interactions**

An advantage of ANOVA is its ability to analyze an experimental design with multiple independent variables. When an ANOVA has two or more independent variables it is referred to as a factorial ANOVA, in contrast to the one-factor ANOVAs discussed thus far. This is efficient experimentally, because the effects of multiple independent variables on a dependent variable are tested on one cohort of subjects. Furthermore, factorial ANOVA permits, and requires, an evaluation of whether there is an interplay between different levels of the independent variables, which is called an interaction.

Definitions of terminology that is unique to factorial ANOVA are warranted: (i) **Main Effect** is the effect of an independent variable (a factor) on a dependent variable, determined separate from the effects of other independent variables. A main effect is a one-factor ANOVA that is performed on a factor that disregards the effects of other factors. In a two factor ANOVA, there are two main effects, one for each independent variable; a three-factor ANOVA has three main effects, and so on. (ii) **Interaction** describes an interplay between independent variables such that different levels of the independent variables have non-additive effects on the dependent variable. In formal terms, there is an interaction between two factors when the dependent variable response at levels of one factor differ from those produced at levels of the other factor(s). Interactions can be easily identified in graphs of group means. For example, again referring to the data set from Figure 2, let us now consider the effect of subject gender as a second independent variable. This would be a two factor ANOVA: one factor is the sex of subjects, called Gender, with two levels; the second factor is the type of manual therapy treatment, called Treatment, with three levels. A shorthand description of this design is 2x3 ANOVA (two factors with two and three levels, respectively). For this two-factor ANOVA, there are three Null hypotheses: (i) Main Effect for the Gender factor: Are there differences in the response (ankle range of motion) for males vs. females to manual therapy treatment (combining data for the three levels of the Treatment factor with respect to the two Gender factor levels)? (ii) Main Effect for the Treatment factor: Are there differences in the response for subjects in the three levels of the Treatment factor (combining data for males and females in the Gender factor with respect to the three Treatment factor levels)? (iii) Interaction: Are there differences due to neither the Gender or Treatment factors alone but to the combination of these factors? With respect to analysis of interactions, Figure 3 shows a table of group means for all levels the two independent variables, based on data from Figure 2. Note that the two independent variables are graphed in relation to the dependent variable. The two lines in the left graph are parallel, indicating the absence of an interaction between the levels of the two factors. An interaction would exist if the graphs were not parallel, such as in the right graph in which group means for males and females on the Type 2 treatment were switched for illustrative purposes. If the lines deviate from parallel to a sufficient degree, the interaction will be statistically significant. In this case with two factors, there is only one interaction to be evaluated. With three or more independent variables, there are multiple interactions that need to be considered. A statistically significant interaction complicates the interpretation of the Main Effects, since the factors are not independent of each other in their effects on the dependent variable. Interactions should be examined before Main Effects. If interactions are not statistically significant, then Main Effects can be easily evaluated as a series of one-factor ANOVAs.

**So There is a Statistically Significant ANOVA—Now What? Multiple Comparison Procedures**

If an ANOVA does not yield statistical significance on any main effects or interactions, the Null hypothesis (hypotheses) is (are) accepted, meaning that the different levels of independent variables did not have any differential effects on the dependent variable. The inferential statistical work is done (but see next section), unless confounding covariates are suspected, possibly warranting analysis of covariance (ANCOVA), which is beyond the scope of this article.

When statistical significance is obtained in an ANOVA, additional statistical tests are necessary to determine which of the group means differ from each other. These follow-up tests are referred to as multiple comparison procedures (MCPS) or post hoc tests. MCPS involve multiple pair-wise comparisons (or contrasts) in a fashion designed to maintain alpha for the family of comparisons to a specified level, typically 0.05. This is referred to as the familywise alpha. There are two general options for MCP tests: either perform multiple t-tests that require “manual” adjustment of the alpha for each pairwise test to maintain a familywise alpha of 0.05, or use a test such as the Tukey HSD (see below) that has built-in protection from alpha inflation. Multiple t-tests have their place, especially when only a subset of all possible pairwise comparisons are to be performed, but the special purpose MCPS are preferable when all pairwise comparisons are assessed.
Using the simple case of a statistically significant one-factor ANOVA, t-tests can be used for post hoc evaluation with the aim of identifying which levels differ from each. However, with multiple t-tests there is a need to adjust alpha for each t-test in such a way as to maintain the familywise alpha at 0.05. If all possible pairwise comparisons are performed, there will be a geometric increase in the number of t-tests as the number of levels increases, as defined by $C = m (m - 1) / 2$, where $C$ = number of pairwise comparisons, and $m$ = number of levels in a factor. For example, there are three pairwise comparisons for three levels; six comparisons for four levels; ten comparisons for five levels, and so forth. There is a need to maintain familywise alpha to 0.05 in these multiple comparisons to maintain the risk of Type I errors to no more than 5%. This is commonly accomplished with the Bonferroni (or Dunn) adjustment, in which alpha for each post hoc t-test is adjusted by dividing the familywise alpha (0.05) by the number of pairwise comparisons:

$$\alpha_{\text{Multiple t-tests}} = \alpha_{\text{Familywise}} / C$$
If there are two pairwise comparisons, alpha for each t-test is set to 0.05/2 = 0.25; for three comparisons, alpha is 0.05/3 = 0.0167, and so on. Any pairwise t-test with a p value less than the adjusted alpha would be considered statistically significant. The trade-off for preventing familywise alpha inflation is that as the number of comparisons increases, it becomes incrementally more difficult to attain statistical significance due to the lower alpha. Furthermore, the inflation of familywise alpha with multiple t-tests is not additive. As a result, Bonferroni adjustments overcompensate the alpha adjustment, making this the most conservative (least powerful) of all MCPs. For example, running two conservative (least powerful) t-tests is not additive. As a result, Bonferroni adjustment developed by Šidák12 to cut double familywise alpha to 0.10; it increases as the number of comparisons increases, thereby requiring greater differences between group means to attain statistical significance as the number of comparisons is increased. Some MCPs are better than others at balancing statistical power and Type I errors. By general consensus amongst statisticians, the Fisher Least Significant Difference (LSD) test and the Duncan's Multiple Range Test are considered to be overly powerful, with too high a likelihood of Type I errors (false positives). The Scheffé test is considered to be overly conservative, with too high a likelihood of Type 2 errors (false negatives), but is applicable when group sample sizes are markedly unequal17. The Tukey Honestly Significant Difference (HSD) is favored by many statisticians for its balance of statistical power and protection from Type I errors. It is worth noting that the power advantage of the Tukey HSD test obtains only when all possible pairwise comparisons are performed. The Student-Newman-Keuls (SNK) test statistic is computed identically to the Tukey HSD, however the critical value is determined differently using a step-wise approach, somewhat like the Holm method described above for t-tests. This makes the SNK test slightly more powerful than the Tukey HSD test. However, an advantage of the Tukey HSD test is that a variant called Tukey-Kramer HSD test can be used with unbalanced sample size designs, unlike the SNK test. The Dunnnett test is useful when planned pairwise tests are restricted to one group (e.g., a control group) being compared to all other groups (e.g., treatment groups).

In summary, (i) the Tukey HSD and Student-Newman-Keuls tests are recommended when performing all pairwise tests; (2) the Hochberg or Holm sequential alpha adjustments enhance the power of multiple post hoc t-tests while taking into account group mean differences, group variances, and group sample sizes in a fashion similar but not identical to the calculation of t. This q value is compared to a critical value generated from a q distribution (a distribution of differences in sample means).
maintaining control of familywise alpha; and (3) the Dunnett test is preferred when comparing one group to all other groups.

**Break from Tradition: Skip One-Factor ANOVA and Proceed Directly to a MCP**

Thus far, the conventional approach to ANOVA and MCPs has been presented, namely, run an ANOVA and if it is not significant, proceed no further; if the ANOVA is significant, run MCPs to determine which group means differ. However, it has long been held by some statisticians that in certain circumstances ANOVA can be skipped and that an appropriate MCP is the only necessary inferential test. To quote an influential paper by Wilkinson et al18, the ANOVA-followed-by-MCP approach "is usually wrong for several reasons. First, pairwise methods such as Tukey’s honestly significant difference procedure were designed to control a familywise error rate based on the sample size and number of comparisons. Preceding them with an omnibus F test in a stage-wise testing procedure defeats this design, making it unnecessarily conservative." Related to this perspective is the fact that inferential discrepancies are possible between ANOVA and MCPs, in which one is statistically significant and the other is not. This can occur when p values are near the boundary of alpha. Each MCP has slightly different criteria for statistical significance (based on either the t or q distribution), and all differ slightly from the criteria of F scores (based on the F distribution). An argument has also been put forth with respect to performing pre-planned MCPs without the need for a statistically significant ANOVA in clinical trials19. Nonetheless, the convention remains to perform ANOVA and then MCPs, but MCPs alone are a statistically valid option. ANOVA is especially warranted when there are multiple factors, due to the ability of ANOVA to detect interactions.

Wilkinson et al.18 also reminds researchers that it is rarely necessary to perform all pairwise comparisons. Selected pre-planned comparisons that are driven by the research hypothesis, and not a subcortical reflex to perform every conceivable pairwise comparison, will reduce the number of extraneous pairwise comparisons and false positives, and have the added benefit of increasing statistical power.

**ANOVA Effect Size**

Effect size is a unitless measure of the magnitude of treatment effects20. For ANOVA, there are two categories of effect size indices: (i) those based on proportions of sum of squares (η², partial η², ω²), and (ii) those based on a standardized difference between group means (such as Cohen’s d)21,22. The latter type of effect size index is useful for power analysis, and will be discussed briefly in the next section. To an ever increasing degree, peer review journals are requiring the presentation of effect sizes with descriptive summaries of data.

There are three commonly used effect size indices that are based on proportions of the familiar sum of squares values that form the foundation of ANOVA computations. The three indices are called eta squared (η²), partial eta squared (partial η²), and omega squared (ω²). These indices range in value from 0 (no effect) to 1 (maximal effect) because they are proportions of variance. These indices typically yield different values for effect size.

*Eta squared (η²)* is calculated as:

\[
\eta^2 = \frac{SS_{Between Groups}}{SS_{Total}}
\]

The SS_{Between Groups} term pertains to the independent variable of interest, whereas SS_{Total} is based on the entire data set. Specifically, for a factorial ANOVA, SS_{Total} = [SS_{Between Groups} for all factors + SS_{Error} + all SS_{Interaction}]. As such, the magnitude of η² for a given factor will be influenced by the number of other independent variables. For example, η² will tend to be larger in a one-factor design than in a two-factor design because in the latter the SS_{Error} term will be inflated to include sum of squares arising from the second factor.

*Partial eta squared (partial η²)* is calculated with respect to the sum of squares associated with the factor of interest, not the total sum of squares:

\[
\text{partial } \eta^2 = \frac{SS_{Between Groups}}{(SS_{Between Groups} + SS_{Error})}
\]

As with the η² calculation, the SS_{Between Group} numerator term for partial η² pertains to the independent variable of interest. However, the denominator differs from that of η². The denominator for partial η² is not based on the entire data set (SS_{Total}) but instead on only SS_{Between Group} and SS_{Error} for the factor being evaluated. For a one-factor ANOVA, the sum of square terms are identical for η² and partial η², so the values are identical; however, with factorial ANOVA the denominator for partial η² will always be smaller. For this reason, partial η² is always larger than η² with factorial ANOVA (unless a factor or interaction has absolutely no effect, as in the case of the interaction in Figure 4, for which both η² and partial η² equal 0).

*Omega squared (ω²)* is based on an estimation of the proportion of variance in the underlying population, in contrast to the η² and partial η² indices that are based on proportions of variance in the sample. For this reason, ω² will always be a smaller value than η² and partial η². Application of ω² is limited to between-subjects designs (i.e., not repeated measures) with equal samples sizes in all groups. Omega squared is calculated as follows:

\[
\omega^2 = \frac{SS_{Between Groups} - \frac{(df_{Between Groups})^2}{SS_{Total} + MS_{Error}}}{SS_{Total}}
\]

In contrast to η², which provides an upwardly biased estimate of effect size when the sample size is small, ω² calculates an unbiased estimate23.

The reader is cautioned that η² and partial η² are often misreported in the literature (e.g., η² incorrectly reported as partial η²)24,25. It is advisable to calculate these values by hand using the formulae shown above as a confirmation of the output of statistical software programs, to ensure accurate reporting. Refer to Figure 4 for sample calculations of these three effect size indices for a two-factor ANOVA.

The η² and partial η² indices have distinctly different attributes. Whether a given attribute is considered to be an advantage or disadvantage is a matter of perspective and context. Some authors24 argue the merits of eta squared, whereas others4 prefer partial eta squared. Nota-
An Analysis of Variance: The Fundamental Concepts

Analyzing issues pertaining to these indices include:

(i) Proportion of variance: When there is a statistically significant main effect or interaction, both $\eta^2$ and partial $\eta^2$ (and $\omega^2$) can be interpreted in terms of the percentage of variance accounted for by the corresponding independent variable, even though they will often yield different values for factorial ANOVAs. So if $\eta^2 = 0.20$ and partial $\eta^2 = 0.25$ for a given factor, these two effect size indices indicate that the factor accounts for 20% vs. 25%, respectively, of the total variability in the dependent variable scores.

(ii) Relative values: Since $\eta^2$ is either equal to (one-factor ANOVA) or less than (factorial ANOVA) partial $\eta^2$, the $\eta^2$ index is the more conservative measure of effect size. This can be viewed as a positive or negative attribute.

(iii) Additivity: $\eta^2$ is additive, but partial $\eta^2$ is not. Since $\eta^2$ for each factor is calculated in terms of the total sum of squares, all the $\eta^2$ for an ANOVA are additive and sum to 1 (i.e., they sum to equal the amount of variance in the dependent variable that arises from the effects of all the independent variables). In contrast, a factor’s partial $\eta^2$ is calculated in terms of that factor’s sum of squares (not the total sum of squares), so on mathematical grounds the individual partial $\eta^2$ from an ANOVA are not additive and do not necessarily sum to 1.

(iv) Effects of multiple factors: As the number of factors increases, the proportion of variance accounted for by each factor will necessarily decrease. Accordingly, $\eta^2$ decreases in an associated way. In contrast, partial $\eta^2$ for each factor is calculated within the sum of squares variance metrics of that particular factor, and is not influenced by the number of other factors.

How Many Subjects?

The aim of any experimental design is to have adequate statistical power to detect differences between groups that truly exist. There is no simple answer to the question of how many subjects are needed for statistical validity using ANOVA. Typical standards are to design a study with an alpha of 0.05 to have with statistical power of at least 0.80 (i.e., 80%
chance of detecting differences between group means that truly exists; alternatively, a 20% chance of committing a Type 2 error). Statistical power will be a function of effect size, sample size, and the number of independent variables and levels, among other things. Adequate sample size is a critical design consideration, and prospective (a priori) power analysis is performed to estimate the required sample size that will yield the desired level of power in the inferential analysis after data are collected. This entails a prediction of group mean differences and group standard deviations in the yet-to-be collected. Specifically, the effect size index used for prospective power analysis is based on a standardized measure such as Cohen’s $d$, which is based on predicted differences in group means (statistical signal) divided by standard deviation (statistical noise). Being based on differences instead of proportions, the $d$ effect size index is scaled differently than the $\eta^2$, partial $\eta^2$ and $\omega^2$ described above, and can exceed a value of 1.

The prediction of an experiment’s effect size that is part of a prospective power analysis is nothing more than an estimate. This estimate can be based on pilot study data, previously published findings, intuition or best guesses. A guiding principle should be to select an effect size that is deemed to be clinically relevant.

The approach used in a prospective power analysis is outlined below for the simple case of a $t$-test with independent groups and equal variance, in which the effect size index is defined as:

$$d = \frac{\text{difference in group means}}{\text{standard deviation of both groups}}$$

The estimate of the appropriate number of subjects in each group for the specified alpha and power is given by the following equation:

$$N_{\text{estimated}} = 2 \times \left[ \frac{(z_a + z_\beta)}{d} \right]^2$$

in which:

- $z_a$ is the $z$ value for the specified alpha. With an alpha of 0.05, $z_a = 1.96$ (2 tail).
- $z_\beta$ is the $z$ value for the specified beta (risk of Type 2 error). Power = 1 - $\beta$.

For $\beta = 0.20$ (power = 0.80), $z_\beta = 0.84$ (1 tail).

As a computational example, if the effect size $d$ is predicted to be 1.0 (which equates to a difference between group means of one standard deviation), then for alpha = 0.05 and power = 0.80 the appropriate sample size for both groups would be:

$$N_{\text{estimated}} = 2 \times \left[ \left( \frac{1.96 + 0.84}{1} \right) \right]^2 = 2 \times \frac{2.80^2}{1} = 2 \times 2.8^2 = 16$$

For a smaller effect size, a larger sample size is needed, e.g., $N = 63$ for an effect size of 0.5. The reader is cautioned that these sample sizes are estimates based on guesses about the predicted effect size; they do not guarantee statistical significance.

Prospective power analysis for ANOVA is more complex than outlined above for a simple $t$-test. ANOVAs can have numerous levels within a factor, multiple factors, and interactions, all of which need to be accounted for in a comprehensive power analysis. These complications raise the following cautionary note: ANOVA power analysis quickly devolves into a series of progressively more wild guesses (instead of “estimates”) of effect sizes as the number of independent variables and possible interactions increase. It is often advisable to focus a prospective power analysis for ANOVA on one factor that is of primary interest, so as simplify the power analysis and reduce the amount of unjustifiable guesses. The reader is referred to statistical textbooks (such as references 22, 26, 27) for different approaches that can be used for prospective power analysis for ANOVA designs. As a general guideline, it is desirable for group sample sizes to be large enough to invoke the central limit theorem in the statistical analysis ($>30$ or so) and for there to be a balanced design (equal sample sizes in each group).

Finally, a retrospective (post hoc) power analysis is warranted after data are collected. The aim is to determine the statistical power of the study, based on the effect size (not estimated, but calculated directly from the data) and sample size. This is particularly relevant for statistically non-significant findings, since the non-significance may have been the result of inadequate statistical power. The textbooks cited above, as well as many others, also discuss the mechanics of how to perform retrospective power analyses.

**Conclusion: Statistical Significance Should not be Confused with Clinical Significance**

ANOVA is a useful statistical tool for drawing inferential conclusions about how one or more independent variables influences a parametric dependent variable (outcome measure). It is imperative to keep in mind that statistical significance does not necessarily correspond to clinical significance. The much sought after statistically significant ANOVA $p$ value has only two purposes: to play a role in the inferential decision as to whether group means differ from each other (rejection of Null hypothesis), and to assign a probability of the risk of committing a Type 1 error if the Null hypothesis is rejected. Statistically significant ANOVA and MCPs say nothing about the magnitude of group mean differences, other than that a difference exists. A large sample size can produce statistical significance with small differences in group means; depending on the outcome measure, these small differences may have little clinical significance.

Assigning clinical significance is a judgment call that needs to take into account the magnitude of the differences between groups, which is best assessed by examination of effect sizes. Statistical significance plays the role of a searchlight to detect group differences, whereas effect size is useful for judging the clinical significance of these differences.

**REFERENCES**