Published in final edited form as:
J Surg Res. 2008 January ; 144(1): 158–170. doi:10.1016/j.jss.2007.02.053.

Analysis of Variance: Is There a Difference in Means and What Does It Mean?

Lillian S. Kao, M.D., M.S.[*,†,1] and **Charles E. Green, Ph.D.**[†]
[*]*Department of Surgery, University of Texas Health Science Center at Houston, Houston, Texas, USA*

[†]*Center for Clinical Research and Evidence-Based Medicine, University of Texas Health Science Center at Houston, Houston, Texas, USA*

Abstract

To critically evaluate the literature and to design valid studies, surgeons require an understanding of basic statistics. Despite the increasing complexity of reported statistical analyses in surgical journals and the decreasing use of inappropriate statistical methods, errors such as in the comparison of multiple groups still persist. This review introduces the statistical issues relating to multiple comparisons, describes the theoretical basis behind analysis of variance (ANOVA), discusses the essential differences between ANOVA and multiple *t*-tests, and provides an example of the computations and computer programming used in performing ANOVA.

Keywords

research/statistics and numerical data; data interpretation/statistical; models; statistical; review

INTRODUCTION

Suppose that a researcher performs an experiment to assess the effects of an antibiotic on interleukin-6 (IL-6) levels in a cecal ligation and puncture rat model. He randomizes 40 rats to one of four equally sized groups: placebo with sham laparotomy, antibiotic with sham laparotomy, placebo with cecal ligation and puncture, and antibiotic with cecal ligation and puncture. He measures IL-6 levels in all four groups and wishes to determine whether a difference exists between the levels in the control rats (placebo with sham laparotomy) and the other groups. He performs two-tailed student’s *t*-tests on all of the possible pairwise comparisons and determines that there is a significant difference between the control rats and rats receiving placebo with cecal ligation and puncture ($P = 0.049$). Is this statistical analysis valid?

Just as methodological flaws in research design can influence the interpretation of trial results, failure to use appropriate statistical tests may result in inaccurate conclusions. Readers must be knowledgeable enough to recognize data analytic errors and to interpret the reported statistical findings. However, in a survey of 91 fifth year surgery residents in 1987, 92% reported less than 5 hours of instruction in statistics during their residency [1]. In a more recent survey reported in 2000 of 62 surgical residency programs, only 33% included education in statistics as a formal component of their curricula [2].

[1] To whom correspondence and reprint requests should be addressed at Department of Surgery, University of Texas Health Science Center at Houston, Kelley St., Suite 30S 62008, Houston, TX. E-mail: Lillian.S.Kao@uth.tmc.edu.

Given the growing impetus to practice evidence-based medicine, surgeons must be able to understand basic statistics to interpret the literature. Although descriptive statistics and *t*-tests are the most widely used statistical methods [3–5], researchers are employing increasingly sophisticated techniques for analyzing data. A review of trends in statistical techniques in surgical journals in 2003 compared to 1985 reported that statistical analyses have become more complicated with time [5]. In particular, the most significant changes were increases in the use of analysis of variance (ANOVA), **nonparametric tests**, and **contingency table** analyses. While the use of more advanced statistical methods may reflect increasing contributions of statisticians and epidemiologists to study design and interpretation, researchers must still be able to understand basic statistical concepts so as to choose the appropriate test. Additionally, surgeons must be able to judge the validity of the statistical methods and results reported in the literature both for research purposes and for clinical application.

Over the past several decades, not only have statistical analyses become more sophisticated, but the appropriate application of tests has improved as well. For example, in 2003, out of 187 randomly selected articles from surgical journals, 14 (7%) study authors incorrectly used *t*-tests instead of ANOVA for comparison of means for three or more groups [5]. In comparison, in 1985, 50 journal articles from the New England Journal of Medicine were analyzed, of which 27 (54%) used inappropriate statistical methods for comparison of multiple means [6]. Although advancements have been made in the statistics included in medical journals, errors still occur. Inappropriate statistical analyses were identified in 27% of studies examined from 2003 surgical journals [5]. Therefore, readers must be able to recognize common errors and the appropriate methods for addressing them. The primary purpose of this paper is to address the problem with multiple comparisons and to discuss why, when, and how to use ANOVA. The intended audience for the main text is surgical researchers and clinicians and, therefore, the concepts and applications of ANOVA are highlighted. For interested readers, the calculations for the main test statistic for a simple, one-way ANOVA are included (Appendix 1). A simulated example is also provided with calculations and basic computer programming (Appendix 2). The appendices' purposes are to provide concrete examples for the readers to reinforce the concepts presented in the paper and to increase the readers' confidence with using ANOVA. Lastly, definitions of the statistical terms used but not explained in the paper are included in a Glossary section.

Student's t-Test Versus ANOVA

ANOVA expands on the basic concepts used in performing a *t*-test. In a previous article in the Journal of Surgical Research, Livingston discussed the use of Student's *t*-test to detect a statistical difference in means between two **normally distributed populations** [7]. The **F-ratio** or **F-statistic**, which is used in ANOVA, can also be used to compare the means of two groups, and yields equivalent results to the ***t*-statistic** in this situation. In fact, mathematically, when comparing only two groups, the F-ratio is equal to the square of the *t*-statistic. However, there are several key differences between the two tests. First, ANOVA can be used for comparing the means of more than two groups and is in fact more statistically powerful in this situation. Moreover, variants of ANOVA can include **covariates**, which allow one to control statistically for **confounders** and to detect **interactions** whereby one variable moderates the effects of another variable.

t-Tests and F-tests vary essentially in the method of quantifying the variability around the group means. The *t*-statistic is calculated using the actual difference between means, while the F-statistic is calculated from the squared sums of the differences between means. This difference has implications for the probability distributions and the interpretation of the two test statistics. To better understand these differences, a discussion of the *t*- and f-families of probability distributions and **degrees of freedom** is necessary. Degrees of freedom is a parameter that is dependent upon sample size, which is used to calculate the probability distributions for certain statistical models. Degrees of freedom may be considered a measure of parsimony, as it is a measure of the number of observations available to vary, to estimate additional parameters. In other words, as the precision increases in estimating model parameters, fewer degrees of freedom are available.

The *t*-test is based upon the ***t*-distribution**, which is similar to a normal distribution (e.g., resembles a bell-shaped curve whereby 95% of data points lie within two standard deviations and 99.7% lie within three standard deviations of the mean) except for the use of the sample rather than the true population standard deviation [7]. The *t*-distribution approaches a normal distribution as the sample size, *n*, increases. A smaller sample size and fewer degrees of freedom (n − 1) result in the tails of the *t*-distribution being denser, containing a greater percentage of the data points. Thus, there is a family of *t*-distributions that are dependent upon the degrees of freedom. All members of the family of *t*-distributions are symmetric around zero as depicted in Fig. 1A.

The probability density function or equation for generating the family of f-distributions is also dependent upon the sample size, *n*. The total degrees of freedom for the f-distribution, like for the *t*-distribution, is $n - 1$. However, the total degrees of freedom is divided up into the between and within groups degrees of freedom, both of which contribute to the probability distribution. Because the f-distribution is based on squared sums, the f-distribution is always positive (Fig. 1B). The flatness and skewness of the distribution depend upon the between and within groups degrees of freedom. For more about the calculations of degrees of freedom for the F-ratio, refer to Appendix 1.

These differences in probability distributions result in two main distinctions between the *t*- and the F-tests. First, directionality of hypothesized statistical relations can be evaluated using a one-tailed *t*-test, which answers the question of whether the mean of one group is larger than the other. In contrast, the F-test cannot determine the direction of a difference, only that one exists. The reason is that for a *t*-test, the **critical value**, or the value at which the *t*-statistic is significant, can be either positive or negative (since the distribution is centered about zero). Therefore, the *t*-test can evaluate hypotheses at either tail. In contrast, the F-ratio is always a positive number. Second, *t*-tests are not additive; that is, multiple *t*-tests cannot be summed together to identify a difference between multiple groups. For example, if the *t*-statistic for a comparison between A and B is −3 and the *t*-statistic for a comparison between B and C is −3, then the *t*-statistic for a comparison between A and C is not 0; that is, one cannot conclude that there is no difference between A and C. On the other hand, the F-test can identify an overall difference between three or more means using a single test that compares all of the groups simultaneously; thus, the F-test is referred to as an **omnibus test**.

Problem of Multiple Comparisons

One important advantage of the F-test is that as an omnibus test, it maintains an appropriate **familywise error rate** in hypothesis testing. In contrast, multiple *t*-tests result in an increased probability of making at least one **Type 1 error**. The problem of multiple comparisons is important to recognize in the literature, especially since the increase in the error rate may be substantial. As an example, in an analysis of 40 studies in orthopedic surgery journals, 182 significant results were reported. However, after adjustment for multiple comparisons, only 59.3% of these remained statistically significant [8]. Therefore, the Type 1 error or false positive rate was much greater than the standard, predetermined rate of 5%.

The probability of at least one Type 1 error increases exponentially with the number of comparisons. The mathematical explanation for this increase is derived as follows: assuming an α equal to 0.05, the probability that an observed difference between two groups is not due to chance variability is 1 − α or 0.95. However, if two comparisons are made, the probability that an observed difference is true is no longer 0.95. Rather, the probability is $(1-\alpha)^2$ or 0.90, and the likelihood of a Type 1 error is 1 − 0.90 or 0.10. Therefore, the probability that a Type 1 error occurs if *k* comparisons are made is $1-(1-\alpha)^k$; if 10 comparisons are made, the Type 1 error rate increases to 40%.

When all pairwise comparisons are made for *n* groups, the total number of possible combinations is *n**(*n* − 1)/2. However, some pairwise comparisons may not be biologically plausible and other pairwise comparisons may be related to each other. Therefore, the true overall Type 1 error rate is unknown. Nonetheless, the take-home message is that the false-positive error rate can far exceed the accepted rate of 0.05 when multiple comparisons are performed.

Different statistical methods may be used to correct for inflated Type 1 error rates associated with multiple comparisons. One such method is the Bonferroni correction, which resets the *P*-value to α/*k* where *k* represents the number of comparisons made. For example, if 10 hypotheses are tested, then only results with a *P*-value of less than 0.05/10 or 0.005 would be considered statistically significant. The Bonferroni correction therefore results in fewer statistically significant results. However, the resultant trade-off for minimizing the likelihood of a Type 1 error is a potential inflation of the **Type 2 error** rate. Another statistical method to minimize the number of comparisons performed is to use an omnibus test, such as the F-ratio in ANOVA, thereby diminishing the Type 1 error rate.

In the initial example, the total number of pairwise comparisons that can be made between four groups of rats is 4*(4 − 1)/2 or six. Therefore, the probability of at least one Type 1 error is $1-(1-0.05)^6$ or 0.26, which is significantly higher than the predetermined level for rejecting the null hypothesis of 0.05. Using a Bonferroni correction, the adjusted *P*-value would be 0.05/6 or 0.008 for each comparison. Therefore, a *P* value of 0.049 would not be considered statistically significant. Rather than having to perform six separate pairwise comparisons, ANOVA would have identified whether any significant difference in means existed using a single test. An F-ratio less than the critical value would have precluded further unnecessary testing.

Basic Concepts and Terminology

ANOVA was developed by Sir Ronald A. Fisher and introduced in 1925. Although termed analysis of **variance**, ANOVA aims to identify whether a significant difference exists between the means of two or more groups. The question that ANOVA answers is: are all of the group means the same? Or is the variance between the group means greater than would be expected by chance? For example, consider the data in Table 1 representing 23 observations distributed among four groups. Expressed in words, the null hypothesis in ANOVA is that the means of all four groups are equivalent; that is, the means for each column are equal. Expressed as an equation, the null hypothesis is:

$$\mu_1=\mu_2=\mu_3=\cdots=\quad\mu_4$$

where μ_j represents the mean of the jth group. The alternative hypothesis is then that the means of all four groups are not equivalent. Expressed as an equation, the alternative hypothesis is:

$$\mu_1\neq\mu_2\neq\mu_3\neq\cdots\quad\mu_4$$

Although not intuitive, testing of the null hypothesis is accomplished by examining the total variance as an aggregated measure of all mean differences; the total variance is then partitioned into the variance due to the factors of interest (the independent variables) and the variance due to random error. In other words, the variation among observations within each column is compared to the variation among observations between columns.

Figure 2 illustrates pictorially the comparison of four group means. In one scenario, the group means are different (Fig. 2A), which would result in a statistically significant F-statistic. In the second scenario, the group means are equivalent (Fig. 2B); a non-significant F-statistic would then preclude further statistical testing.

As with the *t*-test, the F-test is used when the outcome of interest is a **continuous variable**; the outcome is designated the **dependent variable** in an ANOVA. The variable postulated to explain or predict the outcome in ANOVA is referred to as the **independent variable** or factor; that is, the variable responsible for the group classification is the independent variable. Other explanatory or predictor variables (covariates) can be included in the analysis, which is referred to as ANCOVA or analysis of covariance. MANOVA, or multivariate analysis of variance, allows analysis of multiple dependent variables. MANCOVA, or multivariate analysis of covariance, is similar to ANCOVA but includes more than one dependent variable. All of these variants belong to the same family of statistical models called **general linear models**, which also includes linear regression models and Student's *t*-test. Ultimately, researchers should become familiar with all of these techniques so as to be able to choose the model with the best fit.

Take the example at the start of this paper where IL-6 levels are being compared in rats receiving placebo or antibiotic and sham laparotomy or cecal ligation and puncture. The hypothesis for the experiment is that the mean IL-6 levels of the groups are the same. The alternate hypothesis is that there is a difference in mean levels between at least two of the groups, presumably due to the antibiotic and/or the operation. The independent variables are the antibiotic and the operation, and the dependent variable is the IL-6 level. The null hypothesis is tested by apportioning the total variance into systematic variance and error variance, or more specifically, variance due to differences resulting from the interventions being tested (variance between groups or systematic variance) and random variation within groups, which are due to chance (variance within groups or error variance). If the null hypothesis is rejected and the alternate hypothesis supported, then the researcher concludes that there is a difference in the levels of IL-6 between at least two of the groups that is due to either the antibiotic or the operation or the combination of the two.

Comparison of systematic and error variance is accomplished in ANOVA with the F-test. The F-ratio or F-statistic is the value obtained from the ratio of the variance between groups and the variance within groups. The F-test represents the determination of significance of the F-ratio by comparing it to a critical value derived from the probability distribution (e.g., the value along the f-distribution above, which 5% of the area under the curve lies, $P < 0.05$). If the F-ratio is greater than the critical value, then the F-test supports rejection of the null hypothesis. The critical value is never less than 1 because if the F-ratio is 1, the variance between groups is the same as that within groups, (which is assumed to be due to chance.) Therefore, an F-ratio of 1 or less represents no significant difference between groups. As the F-ratio increases, the more the variation in the outcome is explained by differences in the independent variable. Because the F-test is an omnibus test, if the F-test is statistically significant, then there is at least one significant difference in means. (See Appendix 1 for more detailed calculations of the F-ratio). Post-hoc tests can then be used to perform specific comparisons for the purpose of discovering the origin(s) of the difference.

In describing ANOVA, there are several important conventions based on the number of factors and levels being analyzed. The term *factor* describes the independent variable by which the groups are determined. The number of subgroups defined by each factor is referred to as the number of *levels* of the factor. A *one-way* ANOVA refers to a single factor analysis; that is, a one-way ANOVA tests for a difference in outcome between two or more levels of a single independent variable or factor. For example, a researcher studying the effects of three different dosages (levels) of an experimental drug would use a one-way ANOVA. A *factorial* ANOVA is used for two or more factors or independent variables; thus, a 2-way ANOVA compares two independent variables as in the initial example, e.g., the effects of different medications and different operations on IL-6 levels. A 2 × 2 ANOVA is a two-way factorial ANOVA, which is used to compare two levels of one independent variable and two levels of a second independent variable. The IL-6 example is a 2 × 2 ANOVA comparing rats receiving one of two levels of medication (placebo *versus* antibiotic) and one of two levels of operation (sham laparotomy *versus* cecal ligation and puncture).

A *fixed effects* ANOVA is used when inferences are being made only about the specific levels of the factor included in the study whereas *random effects* ANOVA is used when inferences are being made about the levels of the factor not included in the study. A fixed effects model assumes random allocation of the level of a factor, but not random sampling. Therefore, the results of the trial would only be applicable to the specific levels studied and not to all levels possible. On the other hand, a random effects model assumes random sampling of the levels assigned to the factor of interest and the results can be generalized to the population. As an example, in an experiment evaluating the effect of a drug on enzyme levels, a fixed effects model might specify three different dosages to be tested, 1 mg, 5 mg, and 10 mg. Results would then only be applicable to those drug dosages. No conclusions could be made about enzyme levels at a dosage of 20 mg. A random effects model would randomly select the dosages to be evaluated and, therefore, the results would be generalizable to the drug at all dosages, even dosages not specifically studied.

Assumptions

There are three assumptions that must be satisfied to apply ANOVA. The first assumption is that of normality; the outcome variable should be normally distributed within each group. This assumption can be evaluated by examining a histogram of the observations, which should resemble a bell-shaped curve, or using formalized tests such as the Kolmogorov- Smirnov test or the Shapiro-Wilks test (see Appendix 2 for an example of how to test the assumption of normality using a computer program). However, the F-test is relatively resistant or **robust** to violations of this assumption; that is, the Type 1 error rate does not appear to be greatly affected by skewed populations, particularly if the group distribution is **balanced**. The statistical **power** of the test also does not appear to be substantially affected by violation of this assumption, although it may be diminished with smaller group sizes. Alternative tests are available when the data are skewed, such as the Kruskal-Wallis non-parametric procedure.

The second assumption is that the variance in each group is the same (homogeneity of variance), which can be assessed using the Levene test (see Appendix 2 for an example). The F-test is also fairly robust to violations of the assumption of homogeneity of variance. Balanced designs where sample sizes are equal across groups guarantee homogeneity of variance. In unbalanced designs, however, error rates are more likely to be inflated. For example, when the smallest group has the largest variance or the largest group has the smallest variance, then error rates will be increased. Welch's or O'Brien's ANOVA are alternative approaches for analyzing data that violate the homogeneity assumption.

The third assumption is that the observations are independent; that is, the observations are not correlated or related to each other. This requirement is often addressed during study design. For serial observations within subjects, repeated measures ANOVA can be used as long as the subjects are independent from each other. For example, authors assessing the same outcome measure at different time points should be analyzed using repeated measures ANOVA.

Post-Hoc Analyses

If the F-ratio is significant, indicating that a difference between means exists, then post-hoc analyses can be performed to uncover the source of the significance or, in other words, to determine which specific means are different. A significant F-test may occur unexpectedly, in which case specific comparisons between factor levels, or contrasts, may be conducted. These contrasts are referred to as post-hoc comparisons. The appropriate post-hoc analysis is dependent upon the number and type of comparisons planned. If specific comparisons are planned or hypothesized up front, then these contrasts are referred to as *a priori* comparisons.

In the example from the beginning of the paper, suppose that preliminary experiments were conducted with the same antibiotic, but at a lower dosage—rats received cecal ligation and puncture and either no treatment or the antibiotic at the lower dosage. Suppose that there was no difference in IL-6 levels between the two groups. Now suppose that the dosage used in this experiment is significantly higher than previously tested. Additionally, suppose that a criticism of the prior experiment was the lack of a control group (sham laparotomy). Therefore, the experiment as originally described above, with four groups, was conducted. If the F-test were significant, then the researcher would wish to explore whether the source of the difference(s) detected was due to the inclusion of the control group or due to the higher dosage of antibiotics or both. Different strategies for performing these contrasts are described below.

Tukey’s HSD (Honestly Significant Difference) procedure allows the comparison of all pairs of means. When used with equal sample sizes, the familywise error rate is exactly equal to α, which is usually set at 0.05. However, when used with unequal sample sizes (also referred to here as the Tukey-Kramer procedure), the procedure yields a conservative estimate of the chance of a Type 1 error. The Tukey procedure also allows for the derivation of **confidence intervals** about the mean difference.

Scheffe’s procedure differs from Tukey’s in that it allows for comparisons of all types, not just pairwise. Scheffe’s procedure is the most conservative of all of the post-hoc analyses, meaning that the critical F-test value for significance is the largest and that the familywise error rate is minimized in the setting of the largest number of possible comparisons. Therefore, if only pairwise comparisons are planned, Tukey’s procedure should be used because it will result in narrower confidence limits. Nonetheless, if the F-test using Scheffe’s procedure is statistically significant, then at least one contrast out of all possible contrasts is statistically significant. The likelihood of a Type 1 error for Scheffe’s test is exactly α regardless of whether the sample sizes are equal.

For a limited number of planned comparisons, Bonferroni’s procedure can be used. This procedure is superior to Tukey’s if the number of contrasts of interest is equal to or less than the number of factor levels. However, if all pairwise comparisons are of interest, then Tukey’s is superior and will result in smaller confidence intervals.

Another post-hoc analysis is the Newman-Keuls procedure, which ranks groups according to their means and then takes into account the number of steps between the groups in calculating the critical value for significance. Duncan’s procedure is similar to the Newman-Keuls test but is less conservative. Therefore, Duncan’s test is more likely to result in a difference when larger groups are used.

There are other post-hoc analyses that can be performed. Dunnett’s test is used for comparison of groups with a control such that for n groups, there are $n - 1$ comparisons. Hsu’s multiple comparisons with the best (MCB) test is used for comparison of the group with the highest mean *versus* each of the other groups. The appropriate post-hoc analysis therefore depends upon whether the comparisons were planned or unplanned and the number and type of comparisons. However, they all address the problem of multiple comparisons and thus minimize the Type 1 error rate.

Power and Sample Size Calculations

In a previous paper in the Journal of Surgical Research, Livingston discussed sample size calculations for analyses using the Student's *t*-test [9]. For a *t*-test, the determinants of sample size include the magnitude of the hypothesized effect or effect size, standard deviation, and probabilities of Type 1 and 2 errors. The calculations for sample size for ANOVA are more complicated and beyond the scope of this paper. However, the basic principles are similar. First, the researcher must determine the hypothesized effect size based on a number of factors including the proposed difference between means, the within group standard deviation, and the number of groups being compared. The sample size is then based on the proposed distribution of means if the null hypothesis is rejected and the alternate hypothesis is supported. Based on this distribution and the desired α and β, the sample size can be calculated using either a statistical program or standardized table. Similarly, the power can be calculated based on the probability of obtaining the critical F-value given the adjusted F-distribution if the null hypothesis were to be rejected.

CONCLUSIONS

In summary, the appropriate use and interpretation of statistical tests is necessary to evaluate scientific data. While there is significant overlap between different statistical analyses, depending upon the research question and design, there are advantages and disadvantages to each. ANOVA is an appropriate test for evaluating the effect of categorical independent variables on a continuous outcome variable. ANOVA minimizes the inflation of a Type 1 error due to multiple comparisons, reduces the number of tests required to identify a significant difference in means when comparing more than two groups, prevents further unnecessary analysis if the omnibus test (F-test) is not statistically significant, and is relatively robust to violations of assumptions in balanced study designs.

ACKNOWLEDGMENTS

This article was supported by an NIH grant (5K23RR20020-2) (PI: L.S.K.).

The authors acknowledge Dr. Virginia Moyer and Dr. Robert Lasky for their critical reading of the manuscript.

APPENDIX

Appendix 1: Calculation of the F-Test

Although ANOVA can be performed using basic statistical packages, the F-ratio can be calculated manually for a one-way ANOVA. Understanding of the derivation of the F-ratio allows readers to be able to interpret tables with ANOVA results and allows researchers to understand the components required by software programs to perform ANOVA. Thus, calculation of the F-ratio for a one-way ANOVA is described here. Calculations for two-way or factorial ANOVA are more complicated and will not be discussed.

As stated previously, the total variance in the population is the sum of the variance between groups and the variance within groups. The variability of the model is based on how different the individual observations are from the overall population in general. Thus, the total variance is derived from the sums of the squared differences between individual observations and the grand mean of the population. The variance of a sample population with *n* observations is derived from the squares of the differences between each observation (x_i) and the sample population mean $(x)\tilde{}$ divided by $n - 1$.

$$\sigma^2 = \sum (x_i - \bar{x})^2 / (n - 1)$$

In performing ANOVA, the sum of the squared deviations of observations from the mean is known as the *Sum of Squares* or SS. If x_{ij} represents the i^{th} observation in the j^{th} group, then the total sum of squares ($\mathrm{SS_T}$) can be expressed as:

$$SS_T = \sum (x_{ij} - \bar{x})^2$$

For calculation purposes, the formula can be rewritten as follows:

$$SS_T = \sum (x_{ij} - \bar{x})^2 = \sum \left(x_{ij}{}^2\right) - \left(\sum x_{ij}\right)^2 / n$$

The total sum of squares can then be divided into the sum of squares between groups ($\mathrm{SS_B}$) and the sum of squares within groups ($\mathrm{SS_W}$). $\mathrm{SS_B}$ represents the contribution to the total variance in the model by differences due to the independent variables while $\mathrm{SS_W}$ represents the contribution to the total variance by random error.

The between groups sum of squares ($\mathrm{SS_B}$) quantifies how different each individual group is different from the population. Therefore, $\mathrm{SS_B}$ is calculated based on the squared deviations between the group means $(\overline{x}_j)$ and the grand mean of the sample population. $\mathrm{SS_B}$ can be expressed as:

$$SS_B = \sum (\overline{x}_j - \bar{x})^2$$

For calculation purposes, the formula can be alternatively expressed as:

$$SS_B = \sum (\overline{x}_j - \bar{x})^2 = \sum_j n_j \left(x_j{}^2\right) = \left(\sum x_{ij}\right)^2 / n$$

To calculate the within group sum of squares $\mathrm{SS_W}$, the squared deviations between each individual in a group and that individual's group mean are calculated for each group and summed for all of the groups:

$$SS_W = \sum_j \sum_i (x_{ij} - \overline{x}_j)^2$$

Since the total variability in the model is the sum of the between groups and within groups variability, an easier method of computing $\mathrm{SS_W}$ is to obtain the difference between $\mathrm{SS_T}$ and $\mathrm{SS_B}$.

$$SS_W = SS_T - SS_B$$

The next step is to calculate the mean squares. The *Mean Square* (MS) refers to the average squared deviation of observations from the grand mean, or the mean of the entire sample population. This is derived by dividing the sum of squares ($\mathrm{SS_T}$) by the total number of *degrees of freedom* (df), or $n - 1$. In statistics, degrees of freedom represent the number of observations in a population of size n required to obtain a given grand mean; the value of the final observation can be determined if $n - 1$ observations have assigned values. Thus, only $n - 1$ observations can be assigned values without restriction while the n^{th} observation is fixed.

The mean square between groups ($\mathrm{MS_B}$) is calculated as the sum of squares between groups ($\mathrm{SS_B}$) divided by the degrees of freedom between groups ($\mathrm{df_B}$). The degrees of freedom ($\mathrm{df_B}$) for the mean square between groups are one less than the number of groups or $j - 1$.

$$MS_B = SS_B/df_B$$

Similarly, the mean square within groups is calculated as the sum of squares within groups divided by the degrees of freedom within groups (df_W). The degrees of freedom for the mean square within groups (MS_W) is the difference between the total degrees of freedom ($n-1$) and the between groups degrees of freedom ($j - 1$).

$$Df_{total} = df_B + df_W$$

$$df_W = Df_{total} - df_B = (n-1) - (j - 1)$$

The F-test is then the ratio of the mean square between groups and the mean square within groups:

$$\mathbf{F = MS_B/MS_W}$$

The critical F-value upon which statistical significance is determined for different levels of α. df_B, and df_W can be obtained from tables derived using the F distribution. See Appendix 2 for an example of how to calculate the F-test using a clinical example.

A representative table of results is shown in Table 2.

Appendix 2: Example with Calculations and Computer Programming

The following example uses SAS (SAS 9.1, SAS Institute, Cary, NC), but ANOVA can be performed using other statistical programs as well.

SAS was used to generate a data set with random allocation of 24 subjects to four groups. The means of the groups were set at 30, 40, 50, and 50, and all of the standard deviations were set at 10 (refer to Table 1).

One-way ANOVA is an appropriate statistical test to compare the groups given that the outcome variable is continuous and that there are more than two groups. These data could represent the results of either an observational or an experimental trial, and ANOVA could be used to analyze the data in both situations. This data set is realistic in that group sizes may not necessarily be equal. Although balanced designs are preferred, ANOVA can still be used with unequal group designs, provided the assumptions are met.

Assuming a fixed-effects ANOVA, the null hypothesis of the study is that there is no difference in the means between the four groups. The alternate hypothesis is therefore that at least two of the means are different from each other.

The first step after selection of ANOVA as the appropriate test is to determine whether the assumptions have been met. The data observations are independent, in that they are not correlated with each other. The other assumptions of homogeneity of variance and normality must then be assessed.

To evaluate for homogeneity of variance, the following code is entered into SAS:

```
proc glm data = sample_analysis;

class group;

model y = group;

means group/hovtest;
```

```
run;
```

The first line asks the program to perform PROC GLM or a general linear model procedure on the dataset entitled “sample-analysis”. PROC ANOVA can also be used, but PROC GLM is more flexible. (Remember that ANOVA is one member of the family of general linear models.) The second line informs the program of the independent variable(s) in the model; “class” refers to classification level or factor of interest. In this example, the independent variable is the group. The third line informs the program that the desired general linear model type is ANOVA; in SAS, the model for one-way ANOVA is coded as the name of the dependent variable is equal to the name of the independent variable of interest (e.g., y = group). The term “means” in the fourth line requests the program to perform multiple comparisons of the means. The command “hovtest” asks the program to assess for homogeneity of variance among the groups whose means are reported.The default procedure for assessing homogeneity is the Levene test. The results of running the above code are illustrated in Table 3. Since the *P*-value is 0.59, the null hypothesis is not rejected; that is, homogeneity of variance is not rejected and the assumption is met.

Normality of a dataset is determined using the residuals, which are the differences between the predicted and actual outcome variables in an analysis. The residuals should be normally distributed if the assumption is met.

```
proc glm data = anova;
class group;
model y = group;
output out = new p = yhat r = resid stdr = eresid;
run;
proc univariate data = new normal normaltest;
var resid;
histogram resid/normal
cframe = ligr;
inset mean std normal(ksd ksdpval);
qqplot resid/normal(mu = est sigma = est color = yellow l = 2 w = 2 noprint)
square cframe = ligr;
probplot resid/normal
(mu = est sigma = est)
color = yellow
l = 2 w = 2 noprint)
square cframe = ligr;
run;
```

The first three lines of the code are as previously explained. The fourth line asks the program to output into a new file the following: the predicted values of y (yhat), the residual values of y (resid), and the standard errors of the residuals. The next set of instructions utilizes the univariate procedure or “PROC UNIVARIATE”. The “var” and “histogram” commands ask the program to graph the residuals from the ANOVA model as a histogram to assess normality. On the inset for the graph (Fig. 3), the mean, standard deviation, and Kolmogorov-Smirnov test results are printed. Since the probability is greater than 0.05, then the assumption of normality is accepted.

The qqplot command describes a quantile-quantile plot which is used to assess whether the residuals from a data set follow a **normal distribution**. This command graphs quantiles of the data set on the y axis against quantiles of a standard normal distribution on the x axis. If the data are normally distributed, then the plot should be a straight line. The probplot command is similar to the qqplot command except that instead of quantiles, percentiles are plotted on the x-axis. Figure 4 illustrates that the residuals from the example follow a normal distribution.

Since the assumptions of normality, homogeneity of variance, and independence are met, one-way ANOVA is an appropriate statistical test.

After determining the appropriate statistical test and testing the assumptions, the F-ratio should be determined. Performing the calculations first manually, the grand mean can be obtained by adding together all of the observations and dividing by the total number in the dataset. The grand mean is 1097.01/24 or 45.71. Summing the squared deviations of each observation from the grand mean results in the total sum of squares (SS_T), which can alternatively be calculated using the following formula:

$$SS_T = \sum (x_{ij} - \bar{x})^2 = \sum \left(x_{ij}^2\right) - \left(\sum x_{ij}\right)^2 / n$$

$SS_T = 53{,}366 - (1097.01)^2/24 = 53{,}366 - 50{,}143 = 3223.$

The next step is to calculate the between group sum of squares (SS_B), which is calculated as follows:

$$SS_B = \sum (\bar{x}_j - \bar{x})^2 = \sum_j n_j \left(x_j^2\right) - \left(\sum x_{ii}\right)^2 / n$$

$$\begin{aligned} SS_B = \; & 3^*(x_1)^2 + 5^*(x_2)^2 + 10^*(x_3)^2 + 6^*(x_4)^2 \\ & -\left(\sum x\right)^2 / 24 \end{aligned}$$

$SS_B = 3 * (28.11)^2 + 5 * (40.32)^2 + 10 * (48.85)^2 + 6 * (53.77)^2 - (1097.01)^2/24 = 2371 + 8127 + 23{,}863 + 17{,}344 - 50{,}143 = 1562.$

The within group sum of squares (SS_W) is the difference between the total and between-group sums of squares or:

$$SS_W = SS_T - SS_B$$

$SS_W = 3223 - 1562 = 1661$

The total degrees of freedom are equal to $n - 1$ or $24 - 1$ or 23. The degrees of freedom between groups are equal to one less than the number of groups or 3. Therefore, the degrees of freedom within groups are 20, or the difference between the total and between groups degrees of freedom. The mean square between and within groups can then be calculated from the above numbers.

$MS_B = SS_B/df_B = 1562/3 = 521$

$MS_W = SS_W/df_W = 1661/(23 - 3) = 83$

The F-test is then the ratio of the mean square between groups and the mean square within groups:

$F=MS_B/MS_W$ or 521/83 = 6.3

Looking at a table (available in statistical textbooks) where α is equal to 0.05, $df_B = 2$ and $df_W = 12$, the $F_{critical}$ is 3.10. Since the F-test is greater than the $F_{critical}$ value, the null hypothesis should be rejected; that is, there is at least one significant difference among the means of the four groups. Further post-hoc testing can be performed to identify the source of the difference.

Now, repeating the analysis using SAS, the following code is entered:

```
proc glm data = sample_analysis;

class group;

model y = group;

run;
```

As stated previously, this code asks the program to perform the GLM procedure, ANOVA, as specified by the command "model y = group". The independent variable is defined by the "class" command as group. The dataset is titled "sample_analysis." The results (Table 4) are the same as those calculated by hand. The right hand column reports that the probability of the results occurring by chance is 0.003. Since this probability is less than an α of 0.05, the omnibus test is significant and at least one difference in means exists.

Posrt-hoc analyses can be performed to determine the origin(s) of this difference. The following code is used:

```
proc glm data = sample_analysis;

class group;

model y = group;

means group/bon Duncan Scheffe Tukey;

run;
```

The commands following "means group" ask the program to perform the following post-hoc analyses: Bonferroni's, Duncan's, Scheffe's, and Tukey's Honestly Significant Difference procedures. The results of Tukey's procedure are listed in Table 5. Note that the critical value for the F-test is 3.96, which is higher than the critical value required for the omnibus test. This adjusts for the multiple comparisons being performed so as to keep the familywise error rate at 0.05 (the first row labeled α). The results illustrate that groups 1 and 2 have similar means and groups 2, 3, and 4 have similar means. However, the mean of group 1 is significantly different from the means of groups 3 and 4. Based on the parameters used to generate the data set (mean of group 1 of 30 and means of groups 3 and 4 of 50), these results are consistent with the data.

REFERENCES

1. Reznick RK, Dawson-Saunders E, Folse JR. A rationale for the teaching of statistics to surgical residents. Surgery 1987;101:611. [PubMed: 3576452]
2. Cheatham ML. A structured curriculum for improved resident education in statistics. Am Surg 2000;66:585. [PubMed: 10888136]
3. Emerson JD, Colditz GA. Use of statistical analysis in the New England Journal of Medicine. N Engl J Med 1983;309:709. [PubMed: 6888443]
4. Feinstein AR. Clinical biostatistics. XXV. A survey of the statistical procedures in general medical journals. Clin Pharmacol Ther 1974;15:97. [PubMed: 4808744]

5. Kurichi JE, Sonnad SS. Statistical methods in the surgical literature. J Am Coll Surg 2006;202:476. [PubMed: 16500253]

6. Godfrey K. Statistics in practice. Comparing the means of several groups. N Engl J Med 1985;313:1450. [PubMed: 4058548]

7. Livingston EH. Who was student and why do we care so much about his *t*-test? J Surg Res 2004;118:58. [PubMed: 15093718]

8. Bhandari M, Whang W, Kuo JC, et al. The risk of false-positive results in orthopaedic surgical trials. Clin Orthop Relat Res 2003;413:63. [PubMed: 12897597]

9. Livingston EH, Cassidy L. Statistical power and estimation of the number of required subjects for a study based on the *t*-test: A surgeon's primer. J Surg Res 2005;126:149. [PubMed: 15919413]

Glossary

GLOSSARY

Assumption

a criterion that must be met by the data for a statistical test to be valid.

Balanced design

study design whereby all groups are equally sized.

Comparisonwise error rate

the probability of making a Type 1 error for a single comparison, in contrast to the Familywise error rate.

Confounder

also known as confounding variable; an additional variable that is related to both the predictor variable and the outcome of interest, causing the predictor and outcome falsely to appear related; for example, playing pool and lung cancer may appear to be related, but the relationship is confounded by smoking—people who play pool tend to smoke more and therefore develop lung cancer more often than those who don't.

Confidence intervals

the range as defined by an upper and lower limit which contains the population parameter of interest with a probability *P*; e.g., a 95% confidence interval for estimating an unknown population mean can be interpreted as 95 out of 100 times, the mean will be contained within that interval.

Contingency table

a table that is used to record the frequency of two or more categorical variables and to describe their relationship.

Continuous variable

a variable with a numerical value that is measured on a continuum.

Covariate

a predictor or explanatory variable that is related to the outcome of interest.

Critical value

the minimal cutoff value of a test statistic necessary to reject the null hypothesis.

Dependent variable

the outcome variable in a study.

Familywise error rate

the probability of making at least one Type 1 error among all of the hypotheses tested or comparisons performed from the results of a single experiment.

F-distribution

the probability distribution used in analysis of variance.

F-ratio or F-statistic

the test statistic used in analysis of variance which is calculated as the ratio of the systematic variance (mean square between groups) and the error variance (mean square within groups); see Appendix 1 for calculation of F-ratio for one-way ANOVA.

General linear model

a group of statistical models whereby a continuous outcome variable is predicted by a weighted sum of independent variables; includes linear regression models, ANOVA, *t*-test among other tests.

Independent variable

an explanatory or predictor variable in a study.

Interaction

a relationship between two variables whereby the level of one variable moderates the effect of the other; e.g., gender and a treatment may have an interaction whereby males have a good response and females have a poor response.

Non-parametric tests

statistical tests that do not make assumptions regarding the probability distribution of the data; for example, the Kruskal-Wallis non-parametric procedure is used instead of ANOVA when the independent variables are not normally distributed.

Normal distribution

also known as the Gaussian distribution; the probability distribution described by the symmetrical, bell-shaped curve with a mean μ and a standard deviation σ whereby 95% of the observations are contained within the interval $(\mu - 2\sigma, \mu + 2\sigma)$.

Omnibus test

a test used to evaluate a global or overall hypothesis; if significant, further post-hoc analyses may be performed to evaluate sub-hypotheses.

Power

the probability of appropriately rejecting the null hypothesis or the probability of detecting an effect or difference when one is present; power is equivalent to 1-β or 1-probability of a Type 2 error.

***P* value**

the likelihood that the observed difference or effect was a result of chance assuming the null hypothesis to be true; the *P* value is typically assigned a significance level of α.

Robustness

lack of susceptibility of a statistical test to violations of the assumptions of that test.

Skewed distribution

the probability distribution described by an asymmetrical curve; a distribution skewed to the left has a tail toward the right or fewer large observations and vice versa.

t-Distribution

the probability distribution described by a symmetrical, bell-shaped curve with a mean of 0 and a standard deviation estimated by the sample (because the population standard deviation is unknown); as the sample size increases, the t-distribution approaches the normal distribution.

t-Statistic

the test statistic used in comparing the means of two groups.

Type 1 error

the error that results when the null hypothesis is incorrectly rejected - a false-positive error; the probability of a Type 1 error is represented by α. When performing a *t* test on two independent groups with an α equal to 0.05, the chance of concluding that there is a difference in means between the two groups when the difference is really due to chance is 5%. Restated, if an experiment is conducted 100 times, a difference will be found when none really exists 5 out of 100 times.

Type 2 error

represented by β, which occurs when the investigator incorrectly fails to reject the null hypothesis. Thus, a β of 0.20 indicates that the likelihood of concluding that there is no difference in the means between the two groups when one really exists is 20%.

Variance

the measure of dispersion or spread about the mean; calculated as the sum of the squared differences between the observations and the mean.

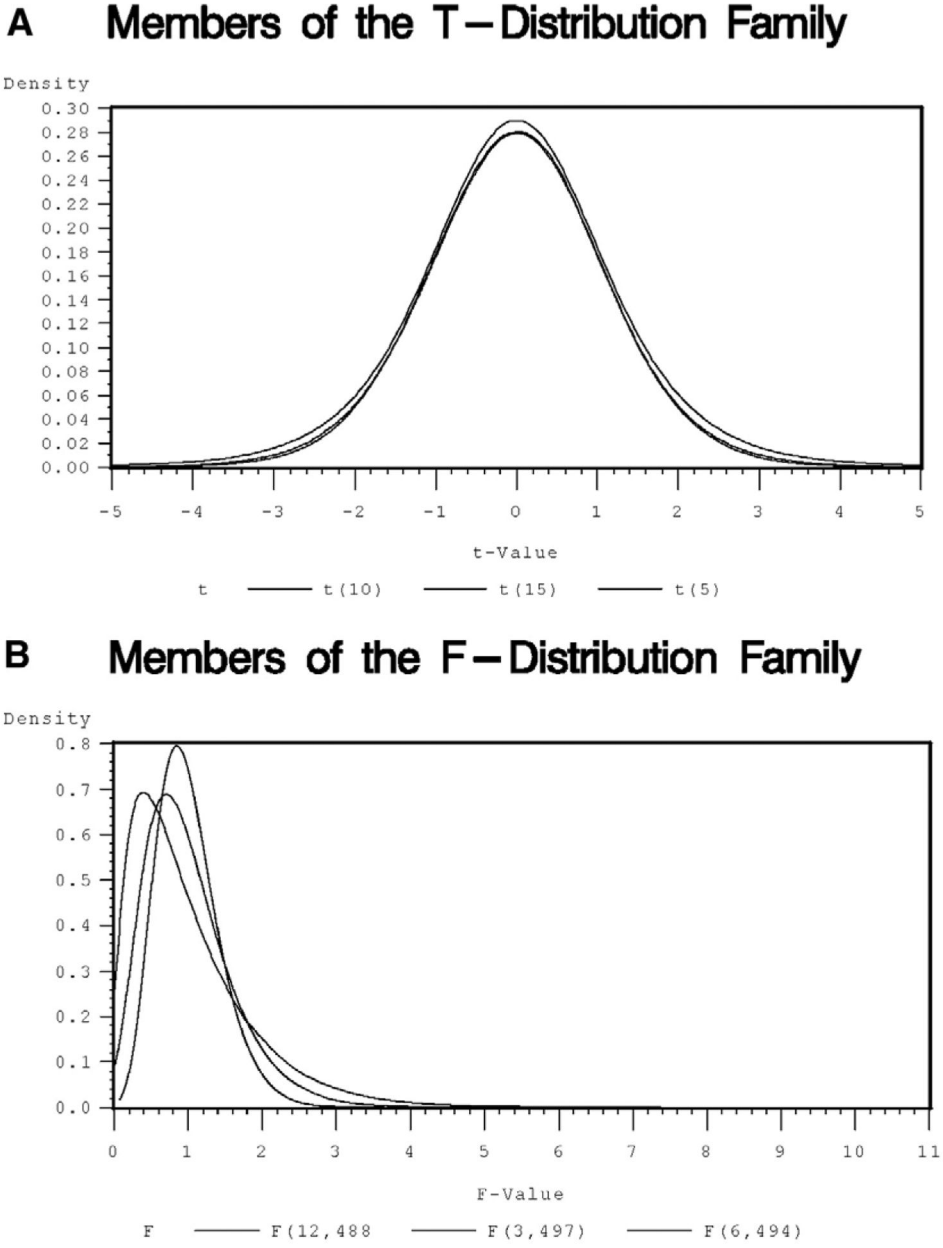

FIG. 1.

(A) The student’s *t*-distribution is similar to a normal or Gaussian distribution except that the sample standard deviation is used instead of the population standard deviation. The critical value indicating statistical significance can be either positive or negative. (B) The F-distribution is a binomial distribution. The critical value is always positive.

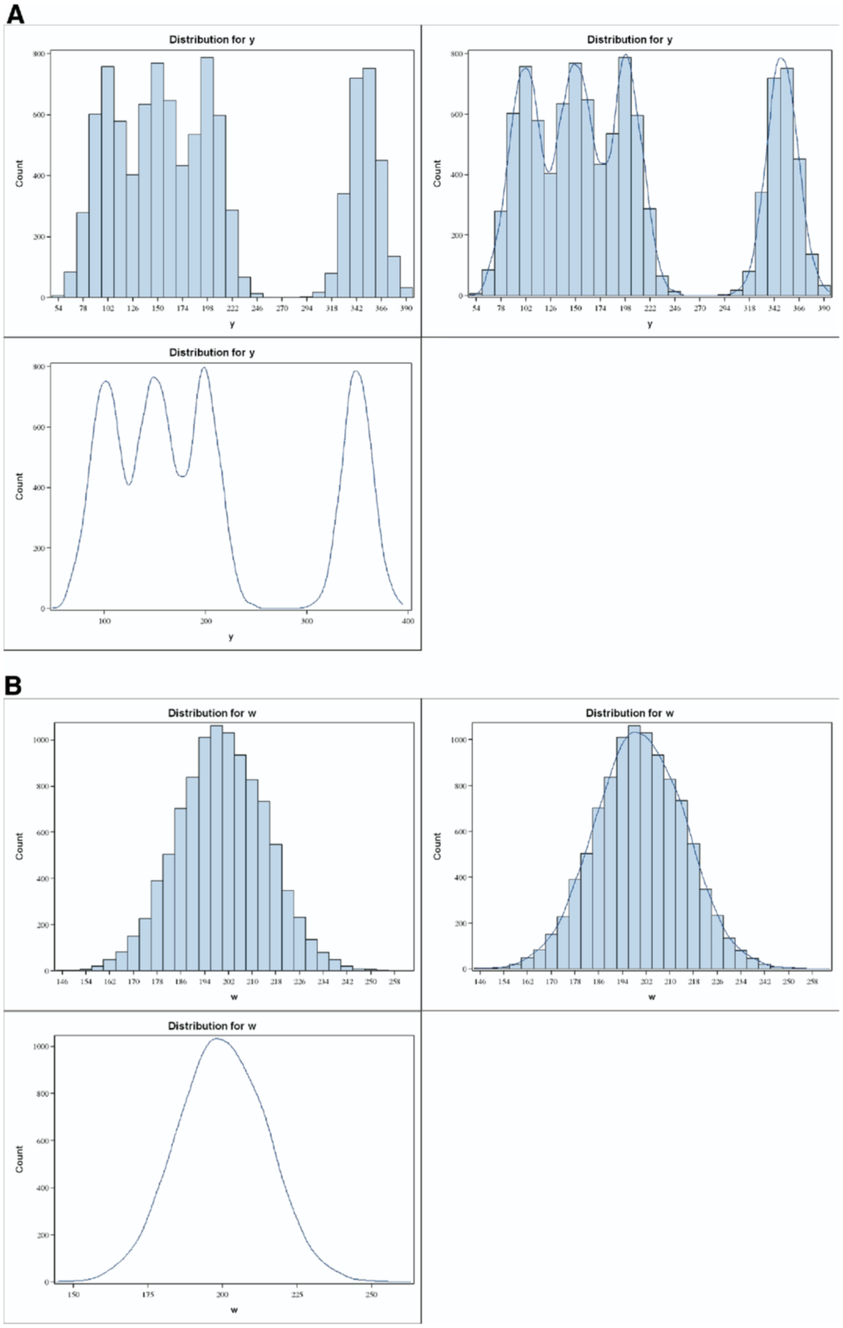

FIG. 2.

A graphical representation of ANOVA: suppose that there are four groups of 2500 data points each. Two different possible scenarios are represented. (A) The means of the four groups are 100, 150, 200, and 350. The histograms and overlying curves, representing the distributions of each group, demonstrate clearly that the mean of one group is different from that of the other three. In this case, the F-test would be statistically significant, prompting further exploration of the differences. (B) The mean of all four groups is 200. The overlying curve demonstrates that a single mean characterizes the entire population. In this case, the F-test would not be statistically significant, precluding further testing for differences. (Color version of figure is available online.)

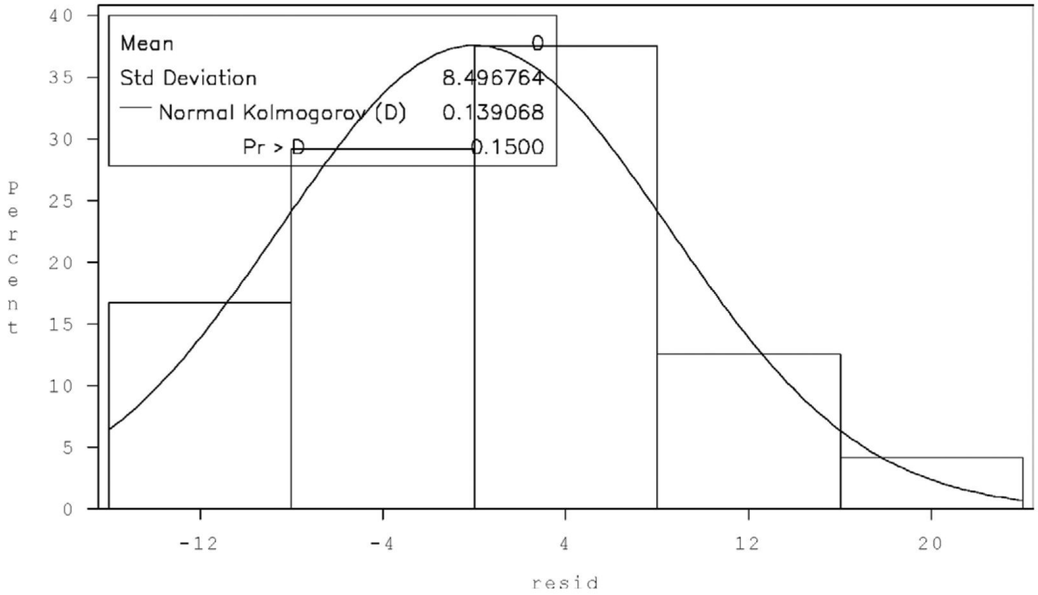

FIG. 3.
The histogram of the residuals follows a normal distribution. The Kolmogorov-Smirnov test result also confirms the assumption of normality.

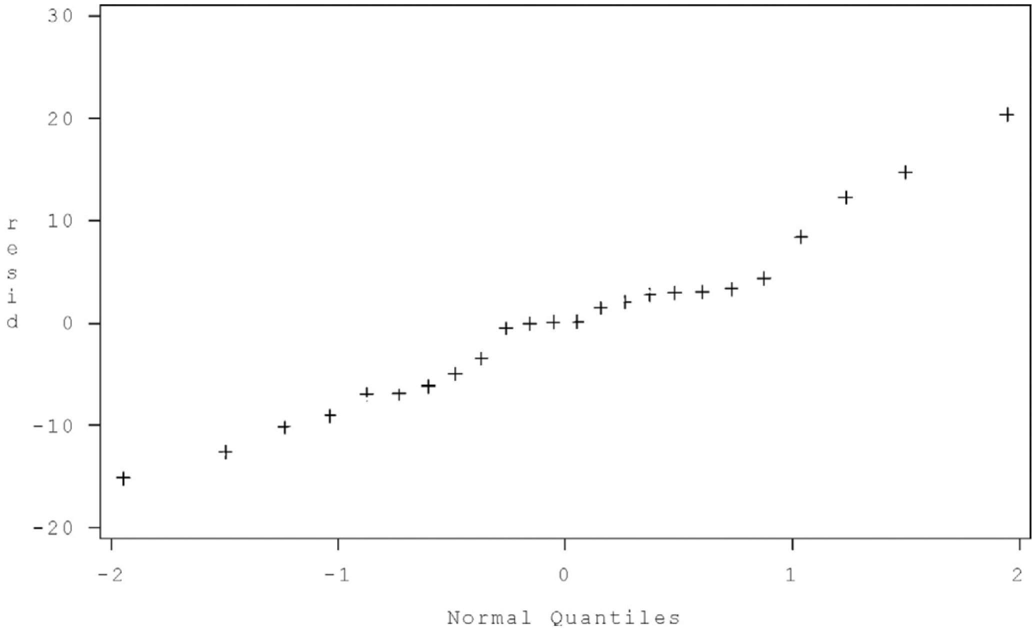

FIG. 4.

The QQ plot or quantile-quantile plot, which graphs quantiles from a normal distribution on the x-axis and quantiles of the residuals, on the y-axis appears to be a straight line. The QQ plot provides additional confirmation of the assumption of normality.

TABLE 1

Sample Data

Group 1	Group 2	Group 3	Group 4
36.55	52.59	42.66	74.15
24.63	25.15	49.01	46.94
23.16	39.81	51.62	56.81
	40.48	38.64	56.69
	43.65	53.23	41.17
		63.55	
		50.39	
		39.79	
		50.91	
		48.79	

TABLE 2

ANOVA Table of Results

Source	SS	df	MS	F
Factor of interest (between groups)	$SS_B = \sum(\bar{x}_j - \bar{x})^2 = \sum_j n_j(x_j^2) - (\sum x_{ij})^2 / n$	$df_B = j - 1$	$SS_{B/}df_B$	MS_B/MS_W
Error (within groups)	$SS_W = \sum_j \sum_i (x_{ij} - \bar{x}_j)^2$	$df_w = (n - 1) - (j - 1)$	SS_W/df_W	
Total	$SS_T = \sum(x_{ij} - \bar{x})^2 = \sum(x_{ij}^2) - (\sum x_{ij})^2 / n$	$df = n - 1$		

TABLE 3

Test for Homogeneity of Variance

Levene’s test for homogeneity of y variance ANOVA of squared deviations from group geans					
Source	**DF**	**Sum of squares**	**Mean square**	**F Value**	**Pr > F**
Group	3	20764.1	6921.4	0.65	0.5950
Error	20	214537	10726.8		

TABLE 4

Results of Analysis of Variance

Source	DF	Sum of squares	Mean square	F Value	Pr > F
Model	3	1562.476950	520.825650	6.27	0.0036
Error	20	1660.485046	83.024252		
Corrected total	23	3222.961996			

TABLE 5

Results of Post-Hoc Analysis Using Tukey's Procedure

Alpha	0.05
Error Degrees of Freedom	20
Error mean square	83.02425
Critical value of studentized range	3.95829
Minimum significant difference	16.13
Harmonic mean of cell sizes	5

Means with the same letter are not significantly different

Tukey Grouping		Mean	*N*	group
	A	53.765	6	4
	A			
	A	48.850	10	3
	A			
B	A	40.316	5	2
B				
B		28.111	3	1