Some Statistical Methods Useful in Circulation Research

SYLVAN WALLENSTEIN, CHRISTINE L. ZUCKER, AND JOSEPH L. FLEISS

SUMMARY Some statistical techniques for analyzing the kinds of studies typically reported in Circulation Research are described. Particular emphasis is given to the comparison of means from more than two populations, the joint effect of several experimentally controlled variables, and the analysis of studies with repeated measurements on the same experimental units.

GLANTZ (1980), in an examination of papers published in Circulation Research and Circulation, found that approximately half the studies that used statistical methodology did so incorrectly. He describes the causes and consequences of such misuse and presents methods to minimize the frequency of such errors, including an approach adopted by the Editors of Circulation Research (Rosen and Hoffman, 1978). Glantz's assertion of incorrect statistical methodology in Circulation Research was documented in unpublished correspondence to the Editors.

At the request of the Editors of this Journal, we attempted to verify these findings by reviewing articles published in volume 40 from 1977, as well as articles in volumes published 5 and 10 years before. In general, our review supports Glantz's contention.

By far, the most persistent defect found by Glantz and confirmed by our review was that one pair of techniques, the independent sample and paired sample t-tests, was used to the virtual exclusion of the more appropriate techniques of the analysis of variance. The statistical methods employed by most authors of the articles published in this Journal appear to have remained fairly static over the past 10 years, with an increase in the proportion of authors using statistical methods of any kind but with only a slight increase in the proportion of authors using newer (post-1940) techniques. This finding probably is not too different from what would characterize comparable biomedical Journals.

The purpose of this paper is to describe, in a didactic manner, statistical techniques that are more appropriate than simple t-tests for analyzing data from the kinds of studies typically reported in this Journal (as determined by the review cited above). All the techniques to be described can be performed with simple pocket or desk-top calculators. References to packaged computer programs are included for the benefit of investigators who have access to computers.

The paper will focus on the three kinds of studies most frequently reported in this Journal. In the first type of study, several different groups are compared. For the comparison of two groups, use of the simple t-test usually is correct, but for the comparison of three or more groups, more appropriate procedures will be suggested. In the second type of study, the so-called factorial study, the possible joint effect of several experimentally controlled variables (e.g., treatment, strain, sex) on outcome is examined. In the third type of study, the so-called repeat measurements study, the effect of treatment over time, is of interest. Techniques of analysis appropriate to each kind of study will be considered and illustrated.

The use of these techniques often will result in more sensitive analyses of experimental effects; that is, it may be possible to detect as significant a given
difference using appropriate statistical procedures, but one may fail to do so with currently popular procedures. The use of appropriate analyses will provide greater depth to data exploration, since many hypotheses of interest can be tested within the framework of a single overall analysis. For example, when two or more treatments are given at several times, it is possible to test for overall treatment differences, and time differences, as well as the consistency of treatment differences over time. Repeatedly performing significance tests in the same study inevitably increases the chance of mistakenly declaring significance. The methods presented here also safeguard against this source of error.

Overall Comparison of Several Independent Groups

An example of an experiment comparing several independent groups is a study reported by Coulson et al. (1977) in which five experimental groups were compared. These included a control group, a group with induced congestive heart failure (CHF), a group with induced right ventricular hypertrophy (RVH), a group with induced congestive heart failure followed by 30 days of recovery (CHFR), and a group with induced right ventricular hypertrophy followed by 30 days of recovery (RVHR). Table 1 presents the group sample sizes and the means and standard deviations for heart rate.

Coulson et al. analyzed the data by performing t-tests for each pair of treatments, using as a critical value that for the t-statistic with degrees of freedom $n_i + n_j - 2$, where $n_i$ and $n_j$ are the sample sizes for groups i and j. Thus, to compare groups 1 and 2 at the $P = 0.05$ level, the critical value would be $t_{0.05} (df = 8) = 2.306$. The t-statistics for all 10 pairwise comparisons are given in column 1 of Table 2.

This procedure has two defects. First, full use is not made of all information concerning variability within the groups. Second, if no overall differences exist between any of the groups, the investigator will have about a 30% chance of declaring at least one difference significant, instead of the 5% chance which use of the critical value 2.306 implies.

The lack of use of all information on variability can result in certain anomalies. For example, the observed mean difference between RVH and control was 33 beats/min, and the difference between CHF and CHFR was 49. Nevertheless, the t-ratio

### Table 1  Experimentally Induced Congestive Heart Failure or Right Ventricular Hypertrophy in Cats

<table>
<thead>
<tr>
<th>Group</th>
<th>$n_i$</th>
<th>($X_i - X\bar{.}$)^2</th>
<th>$(n_i - 1)s_i^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>5</td>
<td>239 ± 29.07</td>
<td>80</td>
</tr>
<tr>
<td>CHF</td>
<td>5</td>
<td>182 ± 44.72</td>
<td>14,945</td>
</tr>
<tr>
<td>CHFR</td>
<td>5</td>
<td>231 ± 31.30</td>
<td>80</td>
</tr>
<tr>
<td>RVH</td>
<td>6</td>
<td>272 ± 10.60</td>
<td>8,214</td>
</tr>
<tr>
<td>RVHR</td>
<td>4</td>
<td>248 ± 36.00</td>
<td>676</td>
</tr>
</tbody>
</table>

Abbreviations: $n_i =$ number of observations for group i; $X_i =$ mean for group i; $X\bar{.}$ = overall mean = 135; $s_i =$ standard deviation for group i. The original results reported by Coulson (1977) were mean ± SE, where SE = $s_i/\sqrt{n_i}$.

### Table 2  Results of Pairwise Comparison

<table>
<thead>
<tr>
<th>Group No. 1</th>
<th>Group No. 2</th>
<th>Conventional</th>
<th>Modified</th>
<th>m = 4</th>
<th>m = 10</th>
<th>Tukey</th>
<th>Dunnett</th>
<th>Scheffe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>CHF</td>
<td>2.39</td>
<td>2.77</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>CHFR</td>
<td>0.42</td>
<td>0.39</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>RVH</td>
<td>2.25</td>
<td>1.68</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>RVHR</td>
<td>0.42</td>
<td>0.41</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHF</td>
<td>CHFR</td>
<td>2.01</td>
<td>2.38</td>
<td>NA</td>
<td></td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHF</td>
<td>RVH</td>
<td>4.48</td>
<td>4.57</td>
<td>NA</td>
<td>*</td>
<td>NA</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>CHF</td>
<td>RVHR</td>
<td>2.38</td>
<td>3.49</td>
<td>NA</td>
<td>*</td>
<td>NA</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>CHFR</td>
<td>RVH</td>
<td>2.66</td>
<td>2.08</td>
<td>NA</td>
<td></td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHFR</td>
<td>RVHR</td>
<td>0.84</td>
<td>0.78</td>
<td>NA</td>
<td></td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RVH</td>
<td>RVHR</td>
<td>1.38</td>
<td>1.14</td>
<td>NA</td>
<td></td>
<td>NA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NA = test is not applicable.

* = statistically significant at $P = 0.05$ level; — = not statistically significant at $P = 0.05$ level.

Critical values for above tests (df = 20)

<table>
<thead>
<tr>
<th>Procedure</th>
<th>No. of comparisons tested</th>
<th>Critical 0.05 value</th>
<th>Derived from tables of</th>
</tr>
</thead>
<tbody>
<tr>
<td>NonSimultaneous</td>
<td>1</td>
<td>2.086</td>
<td>t-Distribution</td>
</tr>
<tr>
<td>Bonferroni</td>
<td>4</td>
<td>2.76</td>
<td>t-Distribution</td>
</tr>
<tr>
<td>Bonferroni</td>
<td>10</td>
<td>3.12</td>
<td>t-Distribution</td>
</tr>
<tr>
<td>Tukey</td>
<td>10</td>
<td>3.00</td>
<td>Studentized range</td>
</tr>
<tr>
<td>Dunnett</td>
<td>4</td>
<td>2.65</td>
<td>Special table</td>
</tr>
<tr>
<td>Scheffe</td>
<td>10</td>
<td>3.39</td>
<td>F-distribution</td>
</tr>
</tbody>
</table>
was 2.25 for the former comparison and 2.01 for the latter—a difference that cannot be explained by the small difference in sample sizes.

The second defect is the more serious, because the investigator has nearly a 1 in 3 chance of declaring some differences between five groups to be significant, even if no differences really exist.

To remedy these defects, an overall test based on the analysis of variance should first be performed to give a single test statistic for differences between all five groups. If, and only if, this test indicates that some differences exist, should modified t-tests be performed to identify the sources of these differences.

### The Analysis of Variance

The analysis of variance partitions the total variability in the experiment (the total sum of squares) into components (sums of squares) due to between-treatment and within-treatment variability. These sums of squares are then both divided by the appropriate degrees of freedom to yield a mean square within groups and a mean square between groups. The latter mean square is an average of the variances within each of the groups and is a measure of biological variability. (A variance is simply the square of a standard deviation.) The ratio of the mean square between groups to the mean square within groups, the F-statistic or F-ratio, is a measure of differences among groups. If there are no real differences among the groups, the value of the F-statistic should be close to 1.0. If the F-value is larger than the appropriate critical value (which depends on the degrees of freedom), we conclude that it is unlikely that the observed differences are due to chance alone and state that the differences are statistically significant. All introductory and intermediate level statistics texts contain tables of critical values for the F-statistic.

Table 3 presents an analysis of variance (ANOVA) table and formulas that can be used to assess the group means and standard deviations that have been calculated. The ANOVA table for the data in Table 1 is given at the bottom of Table 3. As indicated on the table, the F-ratio obtained, 5.47, is greater than the tabulated critical value of 4.43, and therefore, we conclude that the group means are significantly different at P < 0.01.

Suppose the analysis of variance indicates the significance of differences among groups. If a comparison between groups i and j is desired, it should be performed using the modified t-statistic $t = (\bar{X}_i - \bar{X}_j) / s \sqrt{1/n_i + 1/n_j}$, where $s^2$, the mean square within groups, is taken from the analysis of variance table, and $\bar{X}_i$ and $n_i$ are the mean and sample size for group i. These modified t-statistics are listed in column 2 of Table 2.

Appropriate critical values for this statistic are discussed in the section on simultaneous multiple comparison procedures.

The overall F-test and the modified t-statistics, together with an indication of significance based on the procedures to be described below, can be obtained using program ONEWAY of SPSS (Nie et al., 1975).

### Assumptions and Alternative Procedures

The analysis of variance assumes that the measurements are obtained under independent conditions, that the data are distributed normally, and that each group has the same underlying standard deviation. (Of course, the sample standard deviations may vary because of sampling variation.)

The assumption of independence is crucial, whereas the assumption of normality is less crucial, especially for large sample sizes.

Statisticians describe the importance of the assumptions in terms of robustness. For a robust procedure, small or moderate departures from the assumptions (such as a ratio of 3:2 in the standard deviations) will have a small effect on the validity of the procedure (e.g., the actual P value may be 0.045 instead of 0.05). When the sample sizes are

### Table 3: One-Way Analysis of Variance Table

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Degrees of freedom</th>
<th>Sum of squares (SS)</th>
<th>Mean squares (MS)</th>
<th>F-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups</td>
<td>k - 1</td>
<td>( \sum_{i=1}^{k} n_i (\bar{X}_i - \bar{X})^2 )</td>
<td>SS (between)/(k - 1)</td>
<td>MS between</td>
</tr>
<tr>
<td>Within groups</td>
<td>N - k</td>
<td>( \sum_{i=1}^{k} (n_i - 1) S_i^2 )</td>
<td>SS (within)/(N - k)</td>
<td>MS within</td>
</tr>
</tbody>
</table>

Analysis of Variance for Data in Table 1

<table>
<thead>
<tr>
<th></th>
<th>Between groups</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups</td>
<td>4</td>
<td>23065</td>
<td>5773.75</td>
<td>5.47</td>
</tr>
<tr>
<td>Within groups</td>
<td>20</td>
<td>21108</td>
<td>1055.40</td>
<td></td>
</tr>
</tbody>
</table>

Critical value $F_{0.05}(4, 20) = 4.43$. Abbreviations: $X_i$ = measurement on the $i^{th}$ unit in the $i^{th}$ group; $k$ = number of groups; $n_i$ = number of units in the $i^{th}$ group; $\bar{X}_i$ = mean of $n_i$ measurements in $i^{th}$ group $= \sum_{j=1}^{n_i} X_{ij}/n_i$; $S_i^2$ = variance of $i^{th}$ group $= \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2/(n_i - 1)$; $N$ = number of units in entire experiment $= \sum_{i=1}^{k} n_i$. $\bar{X}$ = overall mean $= \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}/N = \sum_{i=1}^{k} n_i \bar{X}_i/N$. 

Downloaded from http://circres.ahajournals.org/ by guest on July 10, 2017
nearly equal, the analysis of variance is a robust procedure with respect to the assumption of equal variability. However, in a study with markedly different sample sizes (the largest being more than twice the smallest), the analysis of variance and the subsequent pairwise comparisons are not robust with respect to the assumption of equal variability. In many studies, a transformation of the data will result in the assumptions being more nearly satisfied, but in other cases, an entirely different method (a nonparametric procedure) based on analyzing the ranks of the observations may be in order.

In biological applications, the need for a transformation frequently is indicated by noticing that the standard deviation increases as the mean does, or that, for data that are non-negative, the mean exceeds the standard deviation. (The latter finding is indicative of skewness.) For these types of data, the two most useful transformations are the square root transformation, valid if the data contain no negative quantities, and the logarithmic transformation, which is suitable if the data contain no zeroes and no negative numbers. The square root transformation usually is preferred when the measurements are in the nature of frequencies or counts, and the logarithmic transformation usually is preferred when the measurements are of enzyme activity or other biological characteristics having a time component for which the standard deviation is approximately proportional to the mean. For purposes of interpretation, it is useful to transform the mean of the transformed data back to the original units.

An alternative class of techniques which can be used to analyze the type of data just described, as well as a variety of other trials with minimal assumptions on the nature of the data, is the so-called nonparametric procedures. All of these procedures are based on ranking the \( N \) observations from 1 for the lowest value to \( N \) for the greatest value. For the kind of study considered so far, the mean rank in each treatment group then is calculated, and a summary statistic, the Kruskal-Wallis test statistic, is computed based on the differences between these observed mean ranks. If the differences among groups are significant, multiple comparisons can be performed as described by Miller (1966) or by Hollander and Wolfe (1973). The technique makes no assumptions regarding normality, is almost as sensitive as the \( t \)-test when the data are, in fact, normally distributed, and can be much more powerful otherwise. A defect in the method is that it is not based on descriptive statistics in common use, and thus description of results is more difficult.

**Simultaneous Multiple Comparisons**

Simultaneous multiple comparison procedures control the error rate associated with an entire set of comparisons rather than the error rate for each comparison. The set could consist of all comparisons to control, an arbitrary preplanned number of comparisons, all pairwise comparisons, or more complex comparisons. For a given set, the probability of falsely declaring one or more differences to be significant, when, in fact, all means are equal, is set at a small value, usually 0.05.

It should be noted that large sets (i.e., open-ended investigations) lead to larger critical values than do smaller sets. Thus, an experiment intended to analyze only the differences between control and each of the four other groups in Table 1 would require a smaller critical value (for those comparisons) than a trial intended a priori to investigate each of the 10 possible pairs of differences between treatments. Of the four simultaneous test procedures discussed below, the first is the simplest and is best suited for a small number of preplanned comparisons. The second (Tukey’s test) is best suited to the case for which all pairwise comparisons are of interest, and the third (Dunnett’s test) is to be used only in comparing the control group to each of the other groups. The Scheffé test, the last we shall discuss, is intended to evaluate arbitrary combinations of groups against each other. The reader interested in more detail on these or other methods should consult Miller (1966).

**The Bonferroni Method**

Based on an elementary inequality called Bonferroni’s inequality, a conservative critical value for the modified \( t \)-statistics is obtained from the tables of the \( t \)-distribution using a significance level of \( P/m \), where \( m \) is the number of comparisons between groups to be performed. (The degrees of freedom are, as above, those for the mean square for within group variation from the ANOVA table.) For example, if for the data in Table 1 only four comparisons were of interest, the \( P \) value of 0.05 would be replaced by 0.05/4 = 0.0125, which, with 20 degrees of freedom, yields a critical value of 2.76. Column 3 of Table 2 indicates which of the \( m = 4 \) comparisons with control would be significant if this procedure were used. On the other hand, if all 10 pairwise comparisons were of interest, the \( P \) level becomes 0.05/10 = 0.005, and the critical value becomes 3.12. Note that the determination as to which set of comparisons is to be tested (four or 10 comparisons) must be made beforehand, and not by inspection of the results. Inspection of column 4 indicates that, for \( m = 10 \), only two differences can be declared significant, in comparison to the four differences that would be declared significant in column 2 if no control over simultaneous testing was used.

[In most cases, the critical Bonferroni value cannot be obtained from conventional tables of the \( t \)-distribution but may be approximated from widely available tables of the normal curve by \( t^{*} = z + (z + z^2) / 4n \), where \( n \) is the degrees of freedom and \( z \) is the critical normal curve value for \( P/m \). For the above example, the critical \( z \)-value at \( P = 0.0125 \) is...
2.50 and, thus, $t^* = 2.50 + (2.5 + 2.5^2) / 80 = 2.73$, in close agreement with the exact value of 2.76 given above.]

**Tukey’s Method**

For the procedure suggested by Tukey (1949), the critical value for the modified \(t\)-statistic is obtained by referring to a value in a table of the distribution of the “studentized range.” These tables are available in many intermediate level statistical texts, such as Snedecor and Cochran (1967). The value obtained from the table then is divided by $\sqrt{2}$. For the current example, the value taken from the table is 4.24, and division by $\sqrt{2}$ yields a critical value of 3.00. The differences between treatments found to be significant by means of the Tukey procedure are given in column 5 of Table 2. In this example, use of either Bonferroni’s procedure (with \(m = 10\)) or Tukey’s procedure gives the same result with respect to significance of differences. Although this test theoretically requires equal sample sizes, or at least equal sample sizes in the treatment groups, with a possibly greater number in the control group. If, in the current example, each treatment group were compared to the control group, the critical value would be 2.65. Column 6 of Table 2 gives the results of the tests of significance if each treatment were to be compared to control using this procedure. Note that Dunnett’s method gives, in this example, the same results as the Bonferroni method for all four comparisons to control.

**Dunnett’s Method**

This procedure is applicable when there is a control group and the investigator’s only interest is in comparing each of the other groups to the single control group. The method is that of Dunnett (1964), who also gives special tables for the critical values. Again, as with Tukey’s method, this procedure is in theory limited to the case of equal sample sizes, or at least equal sample sizes in the treatment groups, with a possibly greater number in the control group. If, in the current example, each treatment group were compared to the control group, the critical value would be 2.65. Column 6 of Table 2 gives the results of the tests of significance if each treatment were to be compared to control using this procedure. Note that Dunnett’s method gives, in this example, the same results as the Bonferroni method for all four comparisons to control.

**Scheffe’s Method**

This procedure, suggested by Scheffé (1959), is intended for complicated comparisons such as the mean of groups 1 and 3 vs. the mean of groups 2, 4, and 5. The critical $P = 0.05$ value for this test is $\sqrt{(k - 1)F_{0.05}(k - 1, N - k)}$, where $k$ is the number of groups, $N$ is the number of experimental units, and $F_{0.05}(k - 1, N - k)$ is the critical value for the overall $F$-test. In the present case, $F_{0.05}(4, 20)$ is 2.87, and the Scheffe critical value is 3.39. The last column of Table 2 gives the results of the tests on pairs of treatments if this procedure were used. In this particular example, the results agree with those of Tukey’s test and the Boneferroni procedure, although it should be noted that the critical value is larger.

**Recommendation**

The Bonferroni procedure is recommended for general use since it is easiest to apply, has the widest range of applications, and gives critical values that will be lower than those of other procedures if the investigator is able to limit the number of comparisons—and that will be only slightly larger than those of other procedures if many comparisons are made.

**Factorial Designs**

**The 2 × 2 Design**

In the simplest type of factorial study, usually denoted as the $2 \times 2$ factorial, two treatments are compared in two different populations; i.e., responses are cross-classified by treatment and population. The populations could represent sex, strain of the animal, or even a second pair of experimental conditions.

As an example of a 2 × 2 design, consider data from a study by Cutilletta et al. (1977) in which the effect of nerve growth factor serum (NGFAS) was compared with control serum (SHAM) in spontaneous hypertensive (SH) and normotensive (WKY) rats. Descriptive statistics (mean ± standard deviation) for kidney renin concentration are shown in Table 4.

$k$-Tests performed to compare the two treatments first in SH rats and then in WKY rats lead to the conclusion of a difference due to NGFAS in SH but not in WKY rats. However, in general, this procedure is not the best strategy for at least two reasons: First, mere presence of a significant difference in one strain and its absence in the other strain does not prove conclusively that the groups (strains) differ in the nature of their response. For example, differences between treatments could be the same for both strains, but differences may be statistically significant only in the strain with the larger sample size. Even if the sample sizes are equal, a small difference between the values of the $k$-statistic (e.g., between 1.90 for one strain and 2.10 for the other) is clearly not indicative of a real difference between treatment effects in the two strains. Second, it is possible that both strains show the same nonstatistically significant effect, but, when the results are combined properly, the effect is significant.

The factorial design allows the following questions to be addressed for the above experiment.

Table 4: Descriptive Statistics (Mean ± Standard Deviation) for Kidney Renin Concentration Data

<table>
<thead>
<tr>
<th>TRT</th>
<th>SH</th>
<th>WKY</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(n = 10)</td>
<td>(n = 8)</td>
</tr>
<tr>
<td></td>
<td>2.41 ± 0.17</td>
<td>2.95 ± 0.23</td>
</tr>
<tr>
<td>SHAM</td>
<td>(n = 10)</td>
<td>(n = 8)</td>
</tr>
<tr>
<td></td>
<td>4.24 ± 0.31</td>
<td>2.89 ± 0.27</td>
</tr>
</tbody>
</table>

Source: Cutilletta et al. (1977).

Downloaded from http://circres.ahajournals.org/ by guest on July 10, 2017
the difference between NGFAS and SHAM the same for SH and WKY rats; i.e., is there an interaction between strain and treatment? Averaged over both strains, is there an effect due to NGFAS?

To answer these questions, an analysis of variance table is constructed that partitions the total variability into components (sums of squares) due to treatments, strains, a treatment-strain interaction, and a within-groups term indicative of animal-to-animal variation. These sums of squares then are divided by their appropriate degrees of freedom to yield mean squares. The significance of each term then is evaluated, as in the one-way design (Table 3), by dividing each mean square by the mean square within groups and then comparing this resulting $F$-statistic with the critical value of $F$ with the appropriate degrees of freedom.

A general procedure for calculating sums of squares for this design will be given below. The analysis of variance table for these data is presented in Table 5 and indicates that treatment-strain interaction is significant ($P < 0.01$). Thus, we have shown that the treatment difference is not the same in the two strains, and further analysis now would be directed toward tests for treatment differences within each strain. To perform this test when there are two treatments per strain, one should form the following $t$-statistic for each strain. The numerator of $t$ is equal to the difference between the treatment means for that strain. The denominator of $t$ is the square root of the product of the mean square within groups and the sum of the reciprocals of the sample sizes.

Thus, for SH rats, 
\[
t = \frac{4.24 - 2.41}{\sqrt{0.0597(1/2 + 1/2)}} = 15.25
\]

whereas, for WKY rats, 
\[
t = \frac{2.89 - 2.95}{\sqrt{0.0597(1/2 + 1/2)}} = 0.45
\]

The appropriate critical values for the test should take into account the fact that two separate comparisons were made. Using the Bonferroni method of multiple comparisons, we use the critical value 2.36. [The value is obtained from a table of $t$ with 28 degrees of freedom (corresponding to the degrees of freedom within groups in Table 5) evaluated at a $P$ value of 0.05/2 = 0.025.] In this trial, we reach the same conclusion as Cutilletta et al., i.e., that differences were significant in SH but not in WKY rats.

If the interaction term were not significant, the overall test of significance for treatment differences

\[
t = \frac{4.24 - 2.41}{\sqrt{0.0597(1/2 + 1/2)}} = 15.25
\]

and strain differences would be based on the $F$-statistics given in the first two rows of Table 5.

### The General Factorial Design

The design described above can be extended easily to the case in which $r$ treatments are compared in $c$ strains or populations. Such a design is called an $r \times c$ factorial ($r$ standing for row classification, $c$ for column classification). A typical layout of the resulting means for such a design is given in Table 6.

The computational procedure for an $r \times c$ design when all the sample sizes are equal is straightforward and will be a special case of the procedure to be described below. For the case in which $r$ or $c$ is equal to 2, Snedecor and Cochran (1967, pages 483–489) give an analysis that can be performed by hand calculations. To perform an exact analysis for the case in which the sample sizes vary and both $r > 2$ and $c > 2$ requires, in general, the use of one of several computer packages—ANOVA of SPSS (Nie et al., 1975) or GLM of SAS (1979). However, several approximate procedures can be performed using hand calculators. Below we discuss one such procedure, the unweighted analysis of variance (see Snedecor and Cochran, 1967, pages 475–477) appropriate for the case in which the sample sizes are not too unequal. (A ratio of 2:1 between the largest and smallest sample size is usually considered adequate.)

One problem caused by unequal sample sizes is that of calculating a descriptive measure of the overall treatment mean taken over all strains (columns). Simply dividing the overall sum by the number of animals receiving that treatment can lead to misleading values, as treatments that were received predominantly by a certain strain would reflect unduly the high or low values for that strain. A way of avoiding this problem is to compute the unweighted mean of the mean values for a treatment across all strains. For example, for the data in

### Table 6 Data Layout for Factorial Design

<table>
<thead>
<tr>
<th>Strain</th>
<th>Mean values per cell</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$X_{11}$ $X_{12}$ $X_{1c}$ $X_{1} = \sum_{j=1}^{c} \bar{X}_{ij}/c$</td>
</tr>
<tr>
<td>2</td>
<td>$X_{21}$ $X_{22}$ $X_{2c}$ $X_{2} = \sum_{j=1}^{c} \bar{X}_{ij}/c$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>r</td>
<td>$X_{r1}$ $X_{r2}$ $X_{rc}$ $X_{r} = \sum_{j=1}^{c} \bar{X}_{ij}/c$</td>
</tr>
</tbody>
</table>

Unweighted column means: $\bar{X}_{.j} = \sum_{i=1}^{r} \bar{X}_{ij}/r$ and sample sizes.

Unweighted overall mean: $\bar{X} = \sum_{j=1}^{c} \bar{X}_{j}/c = \sum_{i=1}^{r} \bar{X}_{i}/r$

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Strain 1</th>
<th>Strain 2</th>
<th>Strain c</th>
<th>Unweighted row means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>5.99</td>
<td>5.99</td>
<td>5.99</td>
<td>100.3**</td>
</tr>
<tr>
<td>Strain</td>
<td>0.619</td>
<td>0.619</td>
<td>0.619</td>
<td>10.4**</td>
</tr>
<tr>
<td>Interaction</td>
<td>12.68</td>
<td>12.68</td>
<td>12.68</td>
<td>212.4**</td>
</tr>
<tr>
<td>Within groups</td>
<td>1.671</td>
<td>28</td>
<td>0.0597</td>
<td>—</td>
</tr>
</tbody>
</table>

Source: Cutilletta et al. (1977).

** Significant at $P < 0.01$ level (critical $F$-value = 7.64).
Table 4, the unweighted mean for the SHAM treatment is \((2.41 + 2.95)/2 = 2.68\).

The analysis of unweighted means is based on squaring the difference between each unweighted mean and the overall unweighted mean, and then multiplying this difference by \(n_h\), a measure of "typical" sample size, where

\[
n_h = \frac{rc}{\sum_{i=1}^{r} \sum_{j=1}^{c} (1/n_{ij})},
\]

and \(n_{ij}\) is the sample size for row (treatment) \(i\) and column (strain) \(j\). (The reader may recognize \(n_h\) as the harmonic mean of the sample sizes.)

As above, the analysis of variance partitions the total variability into a row effect, a column effect, an interaction term that determines whether differences between rows are the same for each column, and a within-group term measuring animal-to-animal variation. Each term is tested for significance by comparing its mean square to the mean square for within-group variation. If the interaction term is significant, the conclusion would be that treatment differences varied from one strain to another. Comparisons between treatments then could be performed separately for each strain using the analysis of variance procedures described above. If the interaction term is not significant, a test for the effect of treatments (rows) averaged over strains (columns) would be performed by testing for significance the ratio of the mean square for treatments to the mean square for within groups.

The computations required for the analysis of variance of unweighted means are given in Table 7. As noted previously, Table 5 illustrates these computations for the kidney renin data.

The theoretical assumptions needed for the factorial analysis are that the measurements be obtained under independent conditions, that the data be distributed normally, and that the variability within each group be the same. If these latter assumptions do not appear to be realized, transformations may be employed as described earlier. The methods described here can be extended to the analysis of the joint effect of more than two factors. However, in these more complex designs, there is more chance of marked imbalance of sample sizes, and the introduction of other complexities not discussed here. The investigator is urged to consult a biostatistician before embarking on such trials.

### Repeat Measurements Studies

The paired t-test is used extensively in the literature in comparing pre- and posttreatment scores on the same group of \(n\) animals. Techniques are available in the more general study in which the same \(n\) animals or patients are observed under \(t\), \(t > 2\), different conditions, or at \(t\) different times. These techniques can be generalized to studies in which different groups each are observed at the \(t\) time points.

As an example of such a design, we consider a trial by Yellin et al. (1979) in which mitral regurgitantar orifice areas were compared in five dogs (\(n = 5\)) at peak flow and at three time points following peak flow (\(t = 4\) time points all told). The data are listed in Table 8. In this example, the main interests were in changes from peak flow and in the evaluation of time trends. Before considering the question of actual trends, we will analyze the data for any differences between the time points—an analysis that would be most appropriate for the case in which, instead of time, different treatments or experimental conditions were being compared in the same animals.

### Test for Overall Differences over Time

In the repeated measurements analysis of variance, the total variability is partitioned into differences between experimental units, variation over time, and residual variability. Table 9 indicates the computational formulas used in the analysis. Under assumptions to be described below, the hypothesis of no variation over time is rejected if the \(F\)-statistic formed by the ratio of the mean square for time to the mean square for residual exceeds the critical \(F\) value with degrees of freedom \((t - 1)\) and \((t - 1)(n - 1)\). (The significance of differences between

### Table 7: Analysis of Variance Table for Two-Way Factorial Design

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row (treatment)</td>
<td>(n_c \sum_{i=1}^{r} (\bar{X}_i - \bar{X})^2)</td>
<td>(r - 1)</td>
</tr>
<tr>
<td>Column (strain)</td>
<td>(n_r \sum_{j=1}^{c} (\bar{X}_j - \bar{X})^2)</td>
<td>(c - 1)</td>
</tr>
<tr>
<td>Interaction</td>
<td>(n_0 \sum_{i=1}^{r} \sum_{j=1}^{c} (\bar{X}_{ij} - \bar{X}_i - \bar{X}_j + \bar{X})^2)</td>
<td>((r - 1)(c - 1))</td>
</tr>
<tr>
<td>Within groups</td>
<td>(\sum_{i=1}^{r} \sum_{j=1}^{c} (n_{ij} - 1)s_i^2)</td>
<td>(N - rc)</td>
</tr>
</tbody>
</table>

**Abbreviations:**
- \(s_i\) = standard deviation for row (treatment) \(i\), column (strain) \(j\);
- \(n_{ij}\) = sample size for row \(i\), column \(j\);
- \(n_h\) = harmonic mean;
- \(N\) = total sample size.
experimental units rarely is tested, because it usually is taken for granted.) For our example $t = 4$ and $n = 5$, giving a critical value of $F_{0.05}(3, 12) = 3.49$. The analysis for the data in Table 8 is presented in Table 10 and indicates significant variation across time.

The test described above theoretically requires that the correlations between all the time points be the same, an assumption that rarely is met in practice. Greenhouse and Geisser (1954) give a conservative procedure which uses the same test statistic but requires a much larger critical value, $F_{0.05}(1, n - 1)$. In this example, the critical value of 3.49 would be replaced by 7.71. Wallenstein and Fleiss (1979) give a procedure that is less conservative but requires a much larger critical value, $F_{0.05}(5, 12) = 8.37$. For the data at hand, linearity is hypothesized for the times, $t_1 = 2, t_2 = 3,$ and $t_3 = 4$, following peak flow. The values of the individual regression coefficients are given in the final column of Table 8. The mean slope is $\bar{b} = -8.6$, the standard deviation of the slopes is $s_b = 2.6$, and the $t$ ratio is $t = -8.6 \sqrt{5/2.6} = -7.2$, which indicates a highly significant trend toward decreasing values after peak flow.

### Table 8 Mitral Regurgitant Orifice Areas (from Yellin et al. 1979)

<table>
<thead>
<tr>
<th>Dog</th>
<th>Peak flow</th>
<th>Mean = $\bar{X}_i$</th>
<th>$\bar{X}_i - \bar{X}$</th>
<th>$b_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>33</td>
<td>42</td>
<td>30</td>
<td>1.25</td>
</tr>
<tr>
<td>2</td>
<td>36</td>
<td>43</td>
<td>32</td>
<td>1.75</td>
</tr>
<tr>
<td>3</td>
<td>36</td>
<td>39</td>
<td>27</td>
<td>-1.50</td>
</tr>
<tr>
<td>4</td>
<td>22</td>
<td>25</td>
<td>9</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>51</td>
<td>60</td>
<td>39</td>
<td>46</td>
</tr>
</tbody>
</table>

$\bar{X}_i - \bar{X}_j$: slope computed from times 2, 3, and 4.

For example, for hours 2-4 of animal 2 in Table 8, $t = (2 + 3 + 4)/3 = 3, \bar{x} = (43 + 32 + 27)/3 = 34$, and $b_2 = [(43 - 34)(2 - 3) + (32 - 34)(3 - 3) + (27 - 34)(4 - 3)] / [(2 - 3)^2 + (3 - 3)^2 + (4 - 3)^2] = [(9)(1) + (2)(0) + (7)(1)] / 2 = -8$.

### Test for Trend

Yellin et al. (1979) note that the data following peak flow for each animal in Table 8 can be described by a straight line and suggest that one compute the regression coefficient and perform a test of significance for each animal. Although the results for this study are clear-cut, it is possible to envision cases in which only three of five animals showed a "significant" trend, preventing the statement of firm conclusions. An alternative procedure calls first for fitting a straight line to the data for each experimental unit. The mean, $\bar{b}$, and the standard deviation, $s_b$, of the $n$ regression coefficients are computed. If there is no time trend, $\bar{b}$ should be close to zero. If the $t$-statistic, $t = \bar{b} \sqrt{n} / s_b$, exceeds the critical value of $t$ with $n - 1$ degrees of freedom, then there is evidence of a significant time trend.

The fitting of a straight line to data is a standard statistical procedure covered in most texts. Assuming that animal $i$ has $m$ pairs of observations, $(t_1, x_1), (t_2, x_2), \ldots, (t_m, x_m)$; where $t$ represents time and $x$, response, first calculate the mean trend, $t$, and the mean response $\bar{x}$. Then compute $b_i$, the slope for animal $i$, using the formula

$$b_i = \frac{\sum_{j=1}^{m} (x_j - \bar{x})(t_j - \bar{t})}{\sum_{j=1}^{m} (t_j - \bar{t})^2}.$$
TABLE 10  Analysis of Orifice Areas in Table 8

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Sum of squares</th>
<th>df</th>
<th>Mean square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dog (row)</td>
<td>1600</td>
<td>4</td>
<td>400</td>
<td></td>
</tr>
<tr>
<td>Time (column)</td>
<td>852.6</td>
<td>3</td>
<td>284.2</td>
<td>31.8*</td>
</tr>
<tr>
<td>Residual</td>
<td>107.2</td>
<td>12</td>
<td>8.93</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2559.8</td>
<td>19</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

** Significant at P < 0.01 level.

slopes is informative of differences in the trends. Tests for differences between the groups with respect to the means and slopes can be computed by the one-way analysis of variance described above.

The investigator interested in any possible trend over time (or analyzing a study in which some replication is taken on a factor other than time) can perform a repeat measurements analysis of variance as described by Armitage (1971, page 253). However, as noted above, the validity of the assumptions required for the analysis are questionable, and the researcher is advised to consult a biostatistician for advice in analyzing such trials. This analysis, as well as the analysis for differences between groups with respect to the means and slopes, also can be performed using program BMDP2V of the BMDP series (Dixon and Brown, 1977).

References


Wallenstein, S, Fleiss JL (1979) Repeated measurements analysis of variance when the correlations have a certain pattern. Psychometrika 44: 229-233

Some statistical methods useful in circulation research.
S Wallenstein, C L Zucker and J L Fleiss

Circ Res. 1980;47:1-9
doi: 10.1161/01.RES.47.1.1

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circres.ahajournals.org/content/47/1/1