Statistics for the Molecular Biologist: Group Comparisons

In this appendix, we will consider some of the statistical tests most commonly used (and misused) in biological research. The tests discussed here are those used for comparisons among groups (e.g., *t* test and ANOVA). A number of other important areas (e.g., linear regression, correlation, and goodness-of-fit testing) are not covered. The purpose of the Appendix is to enable you to determine rapidly the most appropriate way to analyze your data, and to point out some of the most common errors to avoid. Toward this end, we have included a flow chart to provide a quick guide to choosing the right statistical test (Fig. A.3I.1). Instead of including the voluminous statistical tables necessary to perform these tests, we assume you will have access to a spreadsheet software program such as Microsoft Excel or to a statistical software package to do the actual calculations involved and to supply critical values. Clearly, this Appendix is a very superficial treatment of statistics! There are of course many statistics texts available; a few that the authors have found particularly useful are listed at the end of this Appendix (see Literature Cited and Key References).

BASIC STATISTICAL BACKGROUND

Samples and Independence

It is important to think about how you will collect your data and what statistical tests you will perform before you begin your experiment. Suppose you want to compare the immune response of normal mice with that of mice carrying a mutation of interest. You examine a sample of mice from each of these populations. For any statistical test to be valid, all subjects in a sample must come from the same population, and subjects must be selected independently of each other. For example, selecting the same mouse twice would violate the independence assumption; on the other hand, selecting mutant mice from different genetic backgrounds to form one sample would violate the assumption that all mice in a sample are coming from the same population. For some experiments, subjects may be *measured* twice (or more) by design, and this information is incorporated into the statistical test. For example, mouse immune response could be measured before and after a treatment (see discussion of Paired *t* Test), or several measurements may be made from each mouse (see discussion of Multiple Comparison Testing). However, each mouse in the sample is still chosen independently of other mice.

Hypothesis Testing

Statistical tests always test a null hypothesis, H_0. The null hypothesis usually states that there is no difference between two (or more) populations. If the null hypothesis is true, any observed differences between samples from these populations arose by chance alone. The alternative hypothesis, H_a, states that differences observed in the samples reflect a real difference between the populations from which they were drawn. We usually hope to reject H_0, in order to show something interesting. In our immune response experiment, for example, our H_0 would be that mutant mice have the same response as normal mice. We hope to be able to reject H_0, and show that mutant mice have a different immune response from normal mice.

Although statistical tests can never tell us with absolute certainty whether our samples arose from two different populations, they can give us an idea about how probable our H_0 is. What we would like to know is the probability that any difference we find between samples arose by chance alone. Unfortunately, statistical tests cannot tell us this probability. What they do tell us is a related probability—i.e., the probability that we would see at least as large a difference between samples as we do, given that the null hypothesis is true. This probability is the famous *p* value.

For example, if $p = 0.50$, it means that there was a 50% chance of getting a difference between samples at least as large as we did, if the populations we sampled were really the same. This is a large probability, so we would accept the null hypothesis H_0, and conclude that there is no difference between the two populations we sampled. The accepted standard is that we should reject H_0 if $p \leq 0.05$—i.e., there was only a 5% chance (or less) of getting at least as large a difference between samples as we did, if the populations sampled were really the same. The value of *p* below which we will reject the H_0 is called the α value; α is usually chosen as 0.05. A difference between samples that results in a

Contributed by Elizabeth F. Ryder and Phil Robakiewicz
Current Protocols in Molecular Biology (1998) A.3I.1-A.3I.22

p value of $\leq\alpha$ is called statistically significant. Note that statistical tests are not black and white; if $p = 0.05$, there is a 5% chance that we will reject the null hypothesis when it is actually true! This is called a Type I error. You might think we could avoid this error by only rejecting H_0 if the p value is even smaller; say, 0.01 or 0.001. However, if we do this, we are likely to accept H_0 when it is false! This is called a Type II error. Setting $\alpha = 0.05$ is a standard that is generally accepted as a good compromise between Type I and Type II errors.

In practice, once you have chosen a statistical test, you calculate (or the computer calculates) the value of a statistic from your sample data using a formula. You then compare this value to a critical value in a table (or provided by the computer) to decide whether to accept or reject H_0. The critical value (e.g., F_c for the F test, or t_c for the t test; see discussion of The

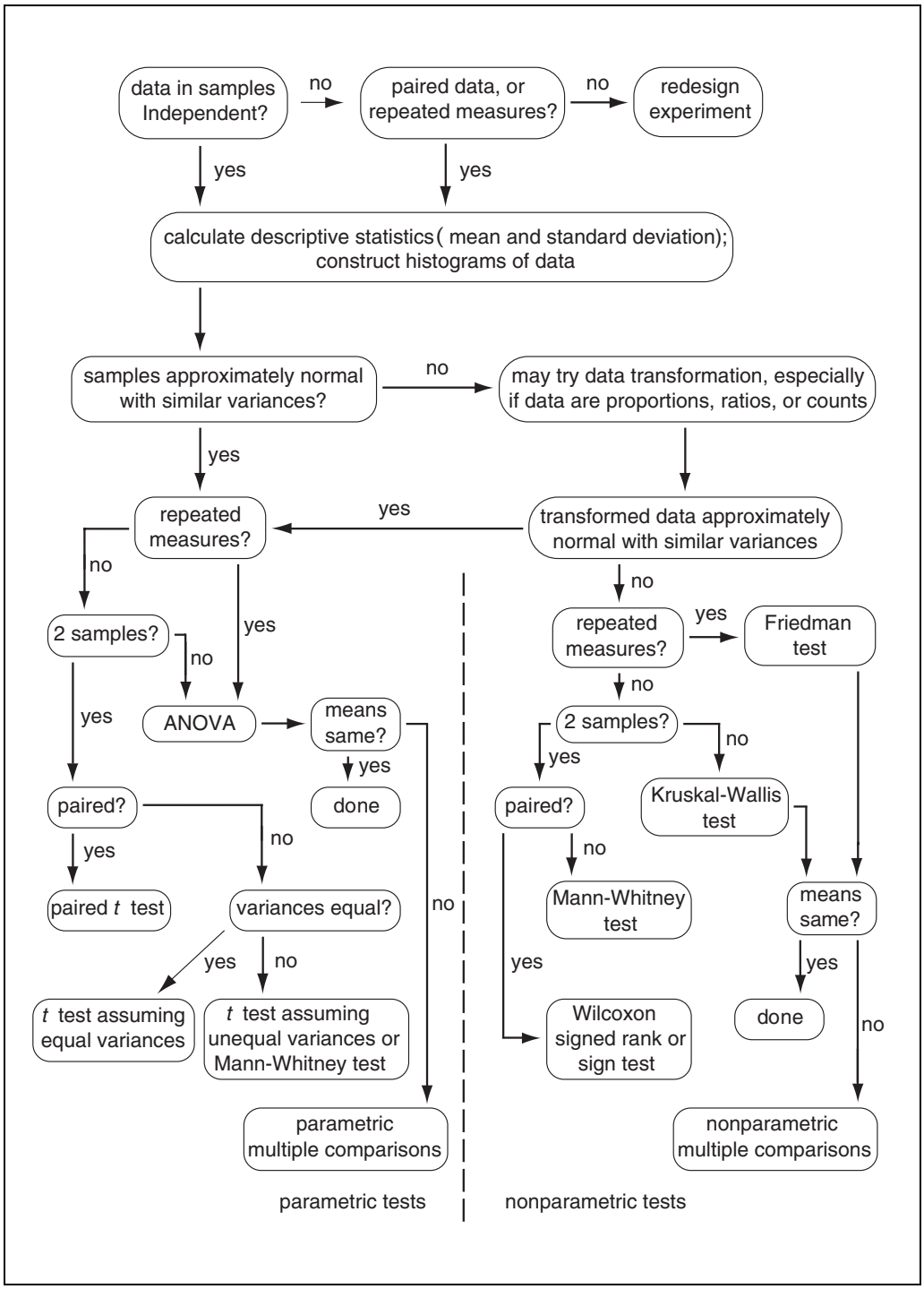

Figure A.3I.1 Flow chart for comparisons among groups.

t Test) is determined by the *p* value you have decided upon (usually, $p = 0.05$) and the degrees of freedom (df) of the data, which vary depending on the test being performed. Usually, you reject H_0 if your statistic is greater than the critical value.

Most tables (and statistical packages) give critical values with one-tailed and two-tailed levels of significance. You are doing a two-tailed test when your H_a does not specify the direction of difference between populations. You perform a one-tailed test when your H_a predicts, based, for example, on a theory that you have, that populations will differ in a particular way. For example, if you think that there may be a difference between the immune response of mutant and normal mice, but you can't predict whether the mutants will do better or worse, you should perform a two-tailed test. If you predict that the response of the mutant mice will be reduced (due to the loss of your gene of interest), you could perform a one-tailed test. There is some controversy in the literature over the use of one-tailed tests. When in doubt, perform a two-tailed test; it is always more conservative (i.e., less likely to reject the null hypothesis).

CHOOSING A TEST—EXPLORATORY DATA ANALYSIS

Parametric and Nonparametric Tests

It can be difficult to sort out what test is appropriate to apply to your data when confronted with a bewildering array of possible tests and formulas. One important distinction you need to make is between parametric and nonparametric tests. Parametric tests (such as the *t* test) assume your data follow a statistical distribution with known mathematical properties (the normal distribution, for the parametric tests we will discuss). Often, parametric tests require that the variability of all the samples you are comparing be the same. Nonparametric tests make fewer assumptions about the distribution of your data. Nonparametric tests are thus less likely than parametric tests to commit a Type I error (rejecting H_0 when it is true), but are usually less powerful (that is, more likely to commit a Type II error, and fail to detect a difference between groups). However, with data sets of reasonable size (say, 25 or more data points per group), nonparametric tests are often nearly as powerful as parametric ones. In order to determine which type of test is appropriate to your data, the first thing you need to do is examine the characteristics of that data.

What Are Parameters?

In order to perform parametric tests, we need to calculate parameters—i.e., numbers that summarize the data in some descriptive way. The statistical parameters we want to calculate are the mean (average) and amount of variability of the population (variance and standard deviation). Your data are a sample from a large population. We estimate the population parameters by sampling a subset of the population. The larger the sample size (n), the closer your sample statistics are to the true population parameters.

Any spreadsheet program or statistical package will allow you to determine the mean, variance, and standard deviation of your data; we show the formulas for these parameters below so that you can better interpret what such programs are telling you about the data.

The following is the equation for the sample mean $\bar{x}$:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Here, x_i is the value of the *i*th data point, and n is the number of data points in the sample.

The sample variance (s^2) is expressed as follows:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{n-1}$$

The sample standard deviation (s), is just the square root of the variance:

$$s = \sqrt{s^2}$$

You can think of the standard deviation as the average deviation from the mean; the magnitude of s tells you how much variability there is in the data. If your data are normally distributed, then ~68% of the data should fall within 1 standard deviation of the mean and ~95% of your data should fall within 2 standard deviations of the mean.

The final parameter of interest is the standard error of the mean (s.e.m.; also called the standard error):

$$\text{s.e.m.} = \frac{s}{\sqrt{n}}$$

The standard error of the mean tells you how much the mean of a sample of size n taken from your population varies. That is, if you were to

take a number of different samples of size n from your population, the mean of each sample would be slightly different. If you were to plot the means of a number of such samples, they would be distributed across a range of values, but there would be far less variability in those values than in the data of the original samples. While the standard deviation gives you a description of the variability of your data, the s.e.m. is the parameter used to compare two sample means and to determine whether or not they are statistically different. For this reason, the s.e.m. is what is normally plotted as the "error bars" around the mean on a graph showing samples that are being compared. Note that the s.e.m. does not tell you much about the variability of your original data; you can see from the equation for s.e.m. that even highly variable data can give a small s.e.m. if the sample size is large enough.

Suppose we are interested in the ability of neurons to sprout neurites in response to growth factors. We want to compare a population of neurons exposed to nerve growth factor (NGF), a factor of known importance, with a population exposed to factor X. Our H_0 would be that neurite outgrowth in response to NGF is the same as neurite outgrowth in response to factor X. We hope to be able to reject H_0, and show that neurons respond differently to factor X. Tables A.3I.1 and A.3I.2 show a (rather small) data set, with calculated sample parameters.

Table A.3I.1 Neurite Outgrowth Data

Neurite outgrowth (μm)	
NGF	Factor X
22.5	43.2
35.4	56.9
38.2	60.2
40.4	62.4
41.3	65.7
46.7	75.8
55.4	85.4
62.5	91.3
81.2	95.3
99.2	105.2

Table A.3I.2 Sample Parameters from Neurite Outgrowth Data

	NGF	Factor X
Mean ($\bar{x}$)	52.28	74.14
Variance (s^2)	535.00	388.26
Standard deviation (s)	23.13	19.70
Standard error (s.e.m.)	7.31	6.23

If these data are normally distributed, then ~68% of the data for NGF should fall within one standard deviation of the mean—i.e., between 29.15 and 75.41 (52.28 ± 23.13). In fact, $^7/_{10}$ or 70% of the data points fall within these limits. About 95% of the data should fall within two standard deviations of the mean—between 6.02 and 98.54 ($52.28 \pm 2 \times 23.13$). In fact, 90% of the data fall within these limits. The distribution of factor X data is similar. This is reasonable agreement with the normal distribution, given the small sample sizes.

Determining Whether Data Are Normally Distributed with Equal Variances

In order to determine whether to use a parametric or nonparametric test on your data, it is important to determine whether the data are normally distributed, with equal variances. Statistical tests for normality exist (e.g., the Kolmogorov-Smirnoff test; Zar, 1984), and we will discuss tests for homogeneity of variance later. However, you can often find out most of what you need to know about your data by simply plotting it and looking at it.

For reasonably large data sets, the parametric tests that we will discuss are not much affected by small departures from normality (this property is called being "robust" to small departures from normality). As a rule of thumb, if you have ≥25 samples in each group that you are comparing, you can simply plot your data using a frequency histogram and examine it visually to see if it falls roughly into a bell-shaped curve. Our neurite outgrowth data is plotted in Figure A.3I.2. In this example, the data points are continuous numbers, so in order to make a histogram, we had to divide the data up into arbitrary "bins" (e.g., 0 to 20, 20 to 40, and so forth). The samples in our data set look approximately normally distributed, although the sample sizes are a bit small to be really sure. The variability in the two data sets also looks similar by eye. Looking back at our sample parameters, the calculated variance and standard deviation of the two data sets look reasonably similar.

If you have a small data set, it may be difficult to determine visually whether your data are normally distributed, and statistical tests also do not have much power to make the determination. If you use a parametric test in this situation, and the data are really not drawn

from a normal distribution, then your may reject your null hypothesis when it is actually true. If you use a nonparametric test, you may accept your null hypothesis when it is false. It is more conservative to use nonparametric tests in this situation; but you can see that the best solution (when possible) is to collect reasonably large data sets!

If visual inspection suggests your data are not normally distributed, or if the variance differs widely among groups, it is sometimes useful to perform a transformation (see discussion of Data Transformations). The transformed data points may then form a normal distribution with similar variances among samples, and parametric statistical tests can be performed on the transformed data points. If neither the original nor the transformed data produce approximately normally distributed data with similar variances, then nonparametric statistical tests generally should be used.

Data Transformations

The parametric tests that we will describe usually assume that the data are normally distributed, and that the variances are the same for all groups being compared (this latter property is known as "homoscedasticity"). There are a number of common situations—e.g., in which the data are expressed as proportions or counts of randomly occurring objects or events—where these assumptions are violated, but you can still use a parametric test if you first perform a data transformation.

Data expressed as proportions of 0 to 1.00 (i.e., 0% to 100%) are not normally distributed. However, you can make them nearly normally distributed, with reasonably homogeneous variances among groups, by performing the following transformation for each data point, and then using the p' values for all of your calculations (e.g., mean, and standard deviation):

$$p' = \arcsin \sqrt{p}$$

When the data are counts of randomly occurring objects (such as the number of butterflies per square foot of a field) or events (such as the number of cars passing by a location), the data will follow a Poisson rather than a normal distribution. Group variances will then be proportional to their means. In this case, the square root transformation, shown in the following equation, should be performed. As with proportions, statistical tests are then performed on the transformed values.

$$x' = \sqrt{x + 0.5}$$

If variances among groups are very different and if the standard deviations are proportional

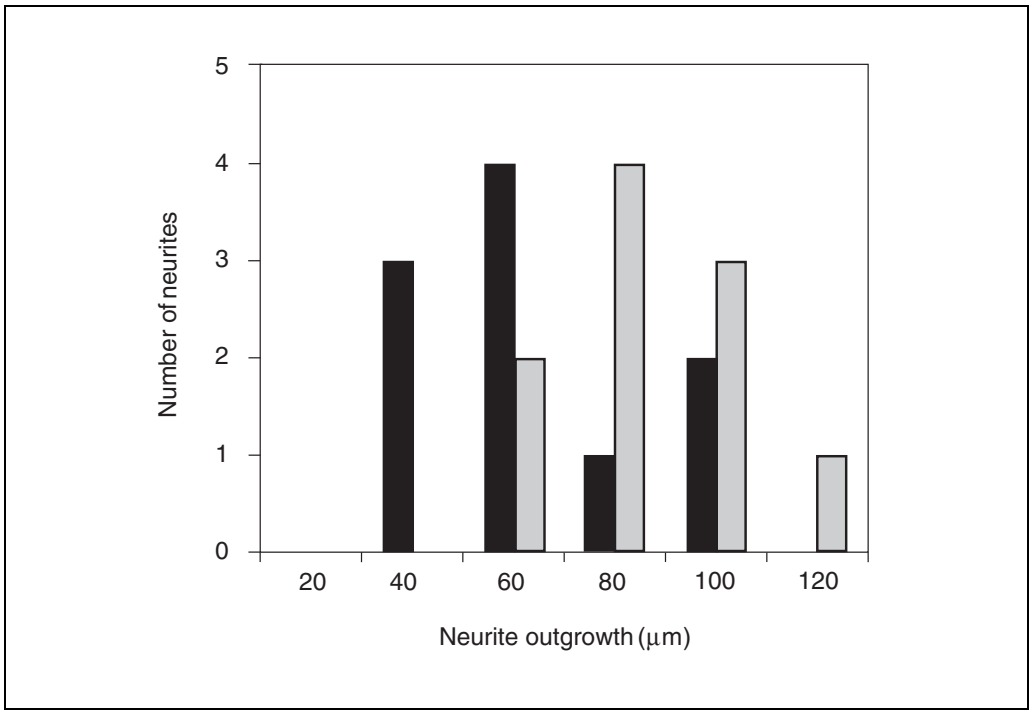

Figure A.3I.2 Neurite outgrowth histogram. Black bars correspond to results for nerve growth factor (NGF) gray bars correspond to results for "factor X."

to the means of your groups, the logarithmic transformation can be used:

$$x' = \log(x + 1)$$

If your data are expressed as ratios, a logarithmic transformation is likely to make them more nearly normally distributed (for an example, see discussion of Ratio Paired *t* Tests).

Other types of transformations are possible (Zar, 1984). If your data still don't fit the assumptions of a parametric test after being transformed, then a nonparametric test should be used.

STATISTICAL TESTS FOR COMPARISONS BETWEEN TWO UNPAIRED GROUPS

The *t* Test

The *t* test is a parametric test that allows us to determine whether the means of two samples are statistically different. The nonparametric equivalent is the Mann-Whitney test (see discussion below; also called the Wilcoxon Rank Sum test).

Assumptions of the t test

1. Only two samples are being compared. If more than two samples are compared, you *must* use a test specifically designed for multiple comparisons, such as ANOVA (see discussion under Multiple Comparison Testing). Using multiple *t* tests to compare among three or more samples is invalid, and the computed *p* values will be wrong. This is one of the most common errors in the use of the *t* test.

2. Subjects in samples were chosen independently. If repeated measurements were made from individual subjects, ANOVA (or the nonparametric equivalent) should be used.

3. The two samples were not paired in any way (such as before-and-after measurements or matched cases and controls). If samples were paired, the paired *t* test (or the nonparametric Wilcoxon Signed Rank test; see discussion below) should be used.

4. The samples are approximately normally distributed. If this assumption is not met, either transform your data (see discussion of Data Transformations) or use the nonparametric Mann-Whitney test instead.

The simplest form of the *t* test assumes that both samples have the same variance. A somewhat more complicated form is necessary if the sample variances are different. Therefore, in order to determine which form of the *t* test to use, we must first determine whether the sample variances are statistically different, by using the variance ratio test, or *F* test. Figure A.3I.3 shows a flow chart for the *t* test.

Performing the F test

The *F* test is summarized as follows:

$H_0: s_1^2 = s_2^2$

$F = \frac{s_1^2}{s_2^2}$ *or* $F = \frac{s_2^2}{s_1^2}$, whichever is larger

(F will always be ≥ 1)

reject H_0 if $F > F_c$; accept H_0 if $F < F_c$

In the above equations, *F* is referred to as the "*F* statistic," s_1^2 and s_2^2 are the variances of sample 1 and sample 2, respectively, and F_c is the critical value.

Statistical computer packages will report the *F* statistic, along with the appropriate critical value, F_c. Note that if the variances of the two samples are very different, *F* will be large, so it makes sense to reject H_0 if $F > F_c$. If we reject H_0, we would use the *t* test assuming unequal variances. If we accept H_0, we would use the *t* test assuming equal variances. If the differences in variance are extreme, a data transformation may be indicated or the (nonparametric) Mann-Whitney test could be performed instead of the *t* test.

F_c is chosen based on the significance level (usually $\alpha = 0.05$), and the degrees of freedom (df) in the data. The value of df is equal to one less than the number of data points in each sample. In the case of the experiment in Table A.3I.1, which will be used as an example here, there are 10 data points in each sample, so df = 9 for each sample. F_c is usually written "$F_{\text{numerator df, denominator df, }\alpha}$." For our example, F_c would be written $F_{9,9,0.05(2)}$. The 2 in parentheses denotes that this is a two-tailed critical value. If the computer package reports both a one-tailed and two-tailed critical value, use the two-tailed value. You would only use a one-tailed value if you had reason to think *before* computing the variances that one variance would be larger than the other. When in doubt, use a two-tailed test.

Performing the t test

The *t* test with equal variances and sample sizes is summarized as follows:

$$H_0: \bar{x}_1 = \bar{x}_2$$

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2 + s_2^2}{n_1}}}$$

reject H_0 if $t > t_c$; accept H_0 if $t < t_c$

In the above equations, t is referred to as the "t statistic" and t_c is the critical value.

Statistical packages will report t and t_c. To choose t_c from a table, the number of degrees of freedom is calculated as the sum of the number of data points in samples 1 and 2, minus 2 ($df = n_1 + n_2 - 2$). The critical value, t_c, is often written $t_{df,0.05(2)}$ for a two-tailed test. As with the F test, use a two-tailed test if both one- and two-tailed critical values are given.

Examining the formula for the t statistic, you can see that t will be large if the difference between the sample means is large. However, as the variances of the samples get larger, the t statistic gets smaller. Thus, the equation makes intuitive sense; you will be able to reject the null hypothesis if the difference between means is large and the scatter in the data is small, so that there is little overlap between the two samples.

There are several different formulas for the t test with unequal variances, which give slightly different results; they are not presented here (see Zar, 1984). The t test with unequal variances tests the same H_0 and follows the same intuitive reasoning given above for the t test with equal variance. However, the calculation of df for this test is quite different than for the test with equal variance—they are calculated using a more complicated formula—a statistical package will calculate them, and print them for you.

Example of the t test

Let us assume that our neurite outgrowth data (see Table A.3I.1 and Table A.3I.2) fit all the assumptions for the t test, and let us perform the t test on this data set.

H_0 for the F test: variances of the two samples are equal.

$F = 535.00/388.26 = 1.378.$

$F_c = F_{9,9,0.05(2)} = 4.03.$

Thus, $F < F_c$, so we accept H_0 and conclude that the variances of the two samples are not statistically different. We proceed to the t test assuming equal variances.

H_0 for the t test: means of the two samples are equal.

We use a spreadsheet to calculate the t statistic. The computer reports:

$t = 2.275$

$t_c = t_{18,0.05(2)} = 2.10.$

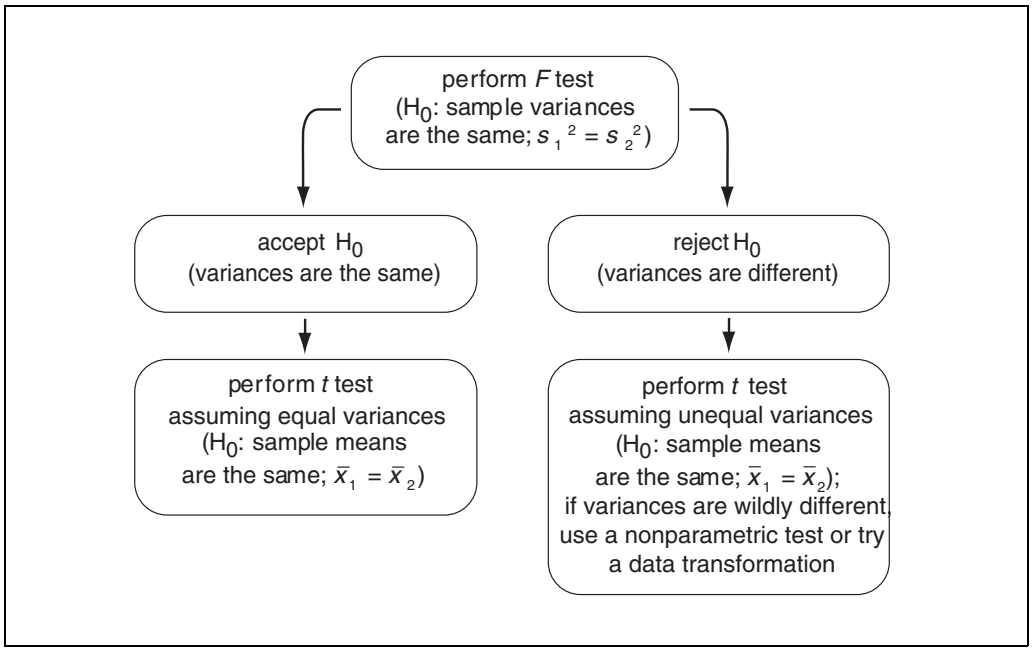

Figure A.3I.3 Flow chart for the t test. Here, we want to compare sample 1 and sample 2, with means $\bar{x}_1$ and $\bar{x}_2$, variances s_1^2 and s_2^2, and sample sizes n_1 and n_2. Whichever t test you use, accepting H_0 for the t test says that the sample means are the same; rejecting H_0 for the t test says that the sample means are different.

Thus, $t > t_c$, so we reject H_0 and conclude that the means of the two samples are different, with statistical significance at the $\alpha = 0.05$ level. The calculated p value for our data is $p = 0.035$. Note that some spreadsheets print out negative values for the t statistic; in this case, simply use the absolute value.

The Mann-Whitney Test

The Mann-Whitney test (also called the Wilcoxon Rank Sum test) is the nonparametric equivalent to the t test, and is used to determine if two samples are statistically different, with no assumptions about whether the data are normally distributed.

This test is nearly as powerful as the t test. When you perform the Mann-Whitney test, you are comparing the order, or rank, of the data rather than its numerical value. To perform the Mann-Whitney test, you rank your entire data set (treating all data from both samples as a single list). For our neurite outgrowth data, the largest data point is 105.2; it has a rank of 1. The next largest is 99.2, with a rank of 2, and so on. You then sum the ranks for each individual data set.

If your two samples have similar values, the ranks would tend to go back and forth between samples. For example, the highest-ranking value might be in sample 1, the next two highest in sample 2, the next in sample 1, and so on. If you then summed the ranks from each sample, the two sums would be about equal. On the other hand, if sample 1 had mostly larger numbers than sample 2, the ranks of sample 1 would all be small numbers, and the sum of the ranks of sample 1 would be much smaller than that of sample 2. The Mann-Whitney test tells us how different the rank sums must be to indicate a significant difference between the two samples.

Assumptions of the Mann-Whitney test

Assumptions 1 to 3 of the t test (see discussion The t Test) apply also to the Mann-Whitney test. The Mann-Whitney test, however, makes no assumption about normality of data

Performing the Mann-Whitney test

The Mann-Whitney test is summarized as follows.

H_0: There is no significant difference between the two samples (of sizes n_1 and n_2).

1. Order all the data from both samples into a single list, from greatest to smallest value, keeping track of which sample the data came from. Tied values get the average of the two ranks they would have had if they had not tied (e.g., if two values tie for 8th largest, they would have been 8th and 9th largest, they each get a rank of 8.5).

2. Sum the ranks for each sample. Call the sums R_1 and R_2.

3. Calculate the U and U' statistics according to the following equations:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U' = n_1 n_2 - U$$

4. If either U or U' is greater than the critical value, U_c ($U_{0.05(2),n1,n2}$, in a table of U values; see, e.g., Zar, 1984), then reject H_0.

5. With more than 20 measurements in one or both samples, the Z statistic is used, as calculated by the following equation. Only U or U' needs to be calculated in this case, not both.

$$Z = \frac{\left|U - \overline{U}\right| - 0.5}{s}$$

where

$$\overline{U} = \frac{n_1 n_2}{2} \text{ and}$$

$$s = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

The Z statistic is approximately normally distributed, even though the original data may not be at all normally distributed. For $\alpha = 0.05$, if $Z > 1.96$, H_0 is rejected.

Example of the Mann-Whitney test

Let us say that we were a bit uncomfortable assuming that our neurite outgrowth data (see Table A.3I.1 and Table A.3I.2) are really normally distributed, and we decide to perform the Mann-Whitney test instead of the t test. These data are shown ranked for the Mann-Whitney test in Table A.3I.3.

From the data in Table A.3I.3 we can make the following calculations, based on the above equations for the Mann-Whitney test:

$U = 100 + 55 - 134 = 21$
$U' = 100 - 21 = 79$
$U_c = U_{0.05,\,10,\,10} = 77$

Since $U' > U_c$, H_0 is rejected and it is concluded that the two samples are significantly different.

Table A.3I.3 Neurite Outgrowth Data Ranked for the Mann-Whitney Test

Outgrowth	Rank (NGF)	Rank (factor X)
NGF		
22.5	20	—
35.4	19	—
38.2	18	—
40.4	17	—
41.3	16	—
46.7	14	—
55.4	13	—
62.5	9	—
81.2	6	—
99.2	2	—
Factor X		
43.2	—	15
56.9	—	12
60.2	—	11
62.4	—	10
65.7	—	8
75.8	—	7
85.4	—	5
91.3	—	4
95.3	—	3
105.2	—	1
Rank sum	134	76

Note that both the *t* and Mann-Whitney statistics were very close to the critical values for these data. The power of the Mann-Whitney test to reject H_0 is almost as great as that of the *t* test.

COMPARING TWO PAIRED GROUPS: PAIRED *t* TEST, WILCOXON SIGNED RANK TEST, AND SIGN TEST

When to Use a Paired Analysis

A paired analysis is indicated under the following circumstances.

1. You measure a variable before and after some treatment is performed upon a subject.
2. You match subjects in pairs (e.g., for age, sex, and exposure) and then treat one of the subjects and not the other (or the other receives alternative treatment).
3. You compare relatives (e.g., siblings or parent/child).
4. You perform an experiment several times, each time with experimental and control preparations treated in parallel.

Paired *t* Test

The paired *t* test is a parametric test used to compare two groups whose members are paired. When the assumptions of a paired *t* test are met, another approach is to perform a repeated-measures ANOVA (see discussion under Multiple Comparison Testing) with only two groups. However, most nonstatisticians find the paired *t* test easier to understand and perform.

Assumptions of the paired t test

1. Pairs must be representative of a larger population. Each pair must be selected independently of other pairs.
2. Pairs must be made *before* data is collected.
3. The calculated *differences* between the two members of each pair must be approximately normally distributed.

Table A.3I.4 Paired Neurite Outgrowth Data

Neurite outgrowth (μm)		
NGF	Factor X	Difference
22.5	43.2	20.7
35.4	56.9	21.5
38.2	65.7	27.5
40.4	62.4	22.0
41.3	75.8	34.5
46.7	85.4	38.7
55.4	95.3	39.9
62.5	60.2	–2.3
81.2	105.2	24
99.2	91.3	–7.9

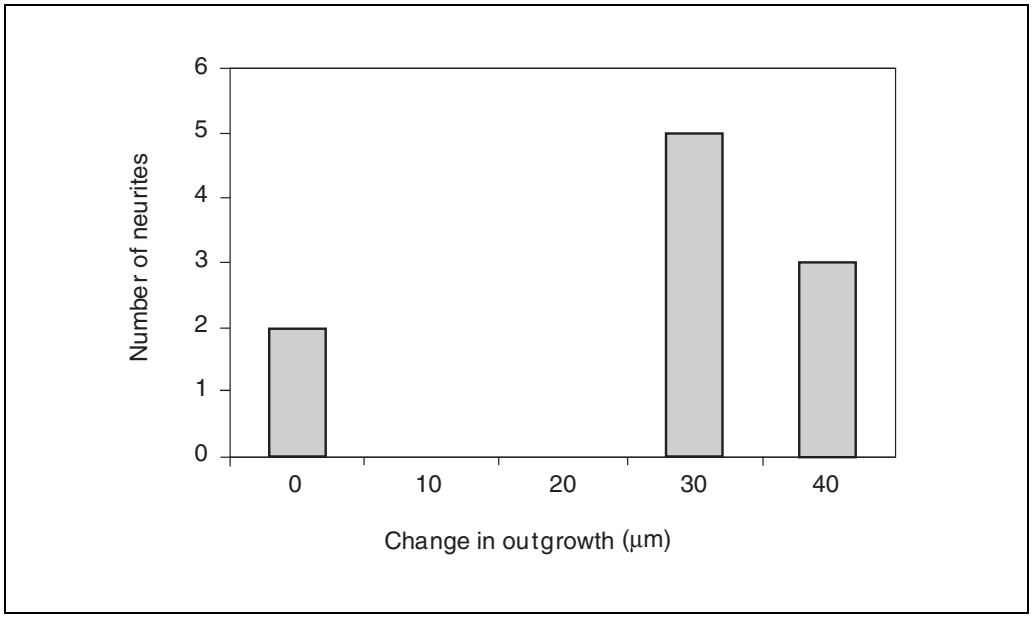

Figure A.3I.4 Change in neurite outgrowth after addition of factor X, in cells treated first with NGF.

Performing the paired t test

The paired *t* test is summarized as follows.

Assume you have a set of *n* pairs: x_{11}, x_{21}, x_{12}, x_{22},....through x_{1n}, x_{2n}.

H_0: There is no difference between members of pairs. (i.e., the mean difference between the pairs, $\bar{d}$, is equal to 0).

1. For each pair, calculate the difference between the pairs ($d_i = x_{1i} - x_{2i}$).

2. Calculate the mean, $\bar{d}$, and the s.e.m. of these differences.

3. Calculate the *t* statistic according to the following equation:

$$t = \frac{\bar{d}}{\text{s.e.m.}}$$

4. Determine t_c from a table using df = n – 1. Note that each pair counts as 1 in computing the df; for the unpaired *t* test, each data point counted as 1.

5. If $t > t_c$, reject H_0.

Example of the paired t test

For our neurite outgrowth data, suppose that instead of looking at two different sets of neurons treated separately with NGF and factor X, we were looking at neurons treated sequentially, first with NGF and then with factor X. We were able to keep track of each individual neurite, and want to know if neurites continued to grow out in the presence of factor X after NGF was removed. We will analyze the same

Table A.3I.5 Original and Transformed Densitometry Data from Northern Blotting of Mutant Versus Wild-Type Gene Expression[a]

Mutant	Wild-type	Ratio: mutant/ wild-type	log(mutant)	log(wild-type)	log(mutant) – log(wild-type)
500	300	1.67	2.6990	2.4771	0.2218
6000	4200	1.43	3.7781	3.6232	0.1549
750	500	1.5	2.8751	2.6990	0.1761

[a]Values expressed in arbitrary densitometry units.

data as before, but now it is paired, as shown in Table A.3I.4 and Figure A.3I.4.

H_0: There is no change in neurite outgrowth when factor X is added to the medium

$\bar{d} = 21.86$
s.e.m. = 5.03
$t = 21.86/5.03 = 4.35$
$t_c = t_{9,0.05(2)} = 2.26$
$t > t_c$; therefore reject H_0.

The p value for the paired t test is $p = 0.0018$ (the p value for the unpaired test on the same data was $p = 0.035$). Clearly, having data that are paired can give us much more power to reject H_0.

If we are uncertain whether our differences are normally distributed, we can use a nonparametric alternative (either the Wilcoxon Signed Rank test, or the Sign test; see discussion below).

Ratio Paired *t* Tests

Sometimes it makes sense to compare values as ratios rather than as differences. For example, you might want to look at densitometric data from a northern blot of your favorite gene's expression in a mutant animal, compared to the signal from a wild-type animal. For each experiment, the densitometry units can be adjusted arbitrarily, and may vary widely depending on how good your probe is and how long the exposure was. However, the ratio of the mutant to wild-type readings is what is important. In this case, what you really want to know is whether the ratio of mutant to wild-type is significantly different from 1 (implying that gene X is expressed either more or less in the mutant than in the wild type).

Ratios are not normally distributed, but logarithms of ratios are quite likely to be normally distributed. The logarithm of a ratio can be written as a difference:

$$\log\left(\frac{\text{mutant}}{\text{wild-type}}\right)$$

$$= \log(\text{mutant}) - \log(\text{wild-type})$$

Thus, to analyze data of this type, simply take the logarithm of each data point and perform a paired t test on this transformed data (Table A.3I.5). If the transformed data do not appear to be approximately normal, the Wilcoxon Signed Rank test can be used instead. The null hypothesis for the transformed data is that the mean difference is 0, as before; this means that for the original (untransformed) data, the null hypothesis is that the ratio of mutant to wild-type expression is 1. Rejecting the null hypothesis says that the amount of expression in mutant and wild-type animals is significantly different.

Let us perform the calculation as follows on the data in Table A.3I.5.

H_0: Ratio of mutant to wild-type expression = 1 (there is no difference between mutant and wild-type animals).

Transform data, and perform paired t test of differences of logarithms of data as described for the paired t test.

$\bar{d} = 0.1843$
s.e.m. = 0.019755
$t = 9.33$
$t_c = t_{2,0.05(2)} = 4.30$
$t > t_c$

Thus, reject H_0 and conclude that there is a significant difference in expression between mutant and wild-type animals. (the calculated p value was 0.01). Note that this is a very significant result, even with only three pairs of data.

Wilcoxon Signed Rank Test

The Wilcoxon Signed Rank test is a nonparametric equivalent to the paired t test. It is used for data that may not be normally distributed.

Table A.3I.6 Differences Computed from Neurite Outgrowth Data, and Ranked for Use in the Wilcoxon Signed Rank Test

Difference	Rank (–)	Rank (+)
20.7	—	3
21.5	—	4
27.5	—	7
22.0	—	5
34.5	—	8
38.7	—	9
39.9	—	10
–2.3	1	—
24	—	6
–7.9	2	—
Rank Sum	3	52

Assumptions of the Signed Rank test

Assumptions 1 and 2 of the paired *t* test apply also to the Wilcoxon Signed Rank test. An additional assumption is made that the sampled population is symmetrical about the median. If the data do not fit this assumption very well, the Sign test (see discussion below) can be performed.

Performing the Signed Rank test

The Wilcoxon Signed Rank test is summarized as follows.

H_0: There are no differences between members of pairs.

1. For each pair, calculate the difference between the members of the pair. Keep track of the sign.
2. Rank the absolute value of all differences from low to high.
3. Sum the ranks of the negative differences (call the sum T^-) and the ranks of the positive differences (call the sum T^+).
4. Reject H_0 if *either* T^- or T^+ is *less than or equal to* the critical value, T_c, as determined from a Wilcoxon Signed Rank table.

Example of the Signed Rank test

If we were concerned about the normality of our neurite outgrowth differences (Table A.3I.1 and Table A.3I.2), we could perform the Wilcoxon Signed Rank test on these data rather than the paired *t* test we did earlier. Consider the data in Table A.3I.6.

H_0: There is no change in neurite outgrowth when factor X is added to the medium.

$T^- = 3$; $T^+ = 52$

$T_c = 8$

$T^- < T_c$; therefore reject H_0.

The Sign Test

The Sign test is another nonparametric paired test that comes in especially handy when you are sure of the *direction* of a difference between the members of your pair (positive or negative), but you may not know exactly how much difference there is.

Assumptions of the Sign test

The assumptions of the Sign test are the same as those of the Wilcoxon Signed Rank test, but no symmetry assumption is made. This is a less powerful test.

Performing the Sign test

The Sign test is summarized as follows.

H_0: There are no differences between members of pairs.

1. For each pair, record whether the first member is greater than the second (+) or less than the second (–). Ignore any pairs with equal members. C^+ is the total number of (+) pairs; C^- is the total number of (–) pairs.
2. Reject H_0 if either C^+ or C^- is less than or equal to the critical value of C for the Sign test (C_c).

Example of the Sign test

H_0: There is no change in neurite outgrowth when factor X is added to the medium.

For our neurite outgrowth data (Table A.3I.6), $C^+ = 8$ and $C^- = 2$. For $n = 10$ and $p =$

0.05, the critical value, C_c is equal to 1. Therefore, H_0 cannot be rejected.

MULTIPLE COMPARISON TESTING

Why Not Lots of *t* Tests?

If we find that we want to compare multiple groups in an experiment, then the *t* test just will not do. Using pairwise comparisons ignores the experimental effects in our design in all but the two groups being compared. Even if it were appropriate to use multiple pairwise comparisons (and it is not), we would still run afoul of the problem of hypothesis rejection. If we used the typical $\alpha = 0.05$, that would mean that for every 20 pairwise tests we ran, we would expect to get the wrong answer for one of them. This is not a desirable situation. There are a number of methods, however, that avoid these problems by allowing us to consider our entire design in our model for analysis. The most common in current literature is the analysis of variance (ANOVA). There are classical and nonparametric methods to analyze variance differences, and we will first consider the classical, or parametric case.

Assumptions of ANOVA

As with any statistical test, there are several assumptions of the classical analysis of variance that need to be examined. Some of these are more critical than others, and we will try to make that clear. In this chapter, we are dealing only with one-way analysis of variance; additional assumptions hold for more complex ANOVA procedures. If in doubt about the ability of your own data to meet these criteria, use nonparametric alternatives, for which the assumptions are less stringent.

Independent observations

Observations must be independent even if you end up doing nonparametric analysis. Data points need to have been collected such that the magnitude of one observation does not affect another. Certain tests, such as the repeated-measures analysis of variance or Friedman's nonparametric analysis of variance, however, allow for multiple measurements on a single subject. Care should be taken to avoid interdependence of data during your experimental design because no amount of adjustment or transformation can rectify the problems inherent in nonindependent data. An example of nonindependent data would be taking three measurements of pH from a single cell culture dish and entering them as $n = 3$ in your ANOVA. In this case, $n = 1$, and you are just as well off taking a single measurement as you are averaging the three readings.

Continuous variable

This assumption is not an extremely strong one as long as you bear several things in mind. The data need to be numerical and ordinal. Continuous data, that is, data that can take on any possible numerical value, are the kind of data for which ANOVA was designed to be used. Deviations from this will affect the sensitivity of the test. It is permissible for data to be discrete (having only a certain set of possible values, like counting numbers) if the numerical values have a fairly wide range. For ANOVA, this would mean something on the order of a 1 to 10 scale or greater. The numbers applied to observations must have some tie to a "greater than or less than" scale; they cannot simply be a classification system of your data. If you have discrete data, you should pay closer attention to the adherence of your data to the additional assumptions detailed below, or just choose a nonparametric test from the outset.

Homogeneity of variance

This is probably the strongest assumption of ANOVA, since, as we will see later, the whole test hinges on the variability within each of our experimental groups being similar. As in the case of the *t* test (see discussion of The *t* Test), the variances for each group need to be tested for similarity. The big difference here, however, is that for the greater-than-two-groups case, there is no way to continue with the test if the variances of the groups are significantly different. Instead, we need to transform the data mathematically to make the variances more similar, or move on to a nonparametric equivalent.

Checking for homogeneity of the variances should start with drawing a frequency histogram of each group and seeing how spread out the data are. It is also a good idea to calculate variances for each group and see if the numbers look close. In many cases, this is all the checking that needs be done. If you remain unsure, there is a simple test called the F_{max} test, which is a simple matter of setting up the following ratio:

$$\frac{\text{largest group variance in the study}}{\text{smallest group variance in the study}} = F_{max}$$

This is equivalent to using an *F* ratio of two groups in the *t* test (see discussion of The *t* Test),

but the tabled critical value to which it is compared is from a cumulative *F* distribution. Degrees of freedom are calculated as $n - 1$, where n is the number of observations in a single group. If n varies among groups, then the lowest value should be chosen, to be most conservative in the test. As in the case of the *t* test, if you fail to reject the null hypothesis, then the variances can be assumed to be roughly equivalent.

Normality

Although this is touted as an important assumption in most classical statistics, it is not as strong an assumption in practice for ANOVA as the assumption of homogeneous variances. Again, using a frequency histogram to give an idea of the shape of your distribution is often a good enough test of normality (see discussion of Determining Whether Data Are Normally Distributed with Equal Variances). If the data are distributed with a marked hump toward the average value and trailing off to either side, then ANOVA will probably work fine. Computer programs provide several different tests of normality that can also be used for this purpose.

Partitioning the Variance

It may seem odd that we wish to analyze variance when, in the long run, our question is to find the differences in means among groups in our experiment. Analysis of variance, however, takes advantage of the underlying assumption that the variation in a measurable quantity is similar within all groups in a population. One or more of these groups, however, may have a very different mean relative to the other groups. If this were the case, the total variance of the whole population would increase relative to the amount seen in each group. It is this observation that gives us the null hypothesis for ANOVA, which is that the means of all the groups are the same:

$$H_0: \mu_1 = \mu_2 = \mu_3 = ... = \mu_n$$

In ANOVA, we break down the total variance into two basic components: within-group and among-group variance. Within-group or error variance is the naturally occurring variance among individuals due to inherent genetic differences and the differential effects of environment that each individual experiences. This contrasts with among-group or treatment variance, which is the variance observed among groups due to perturbations of the experiment. The treatment variance is the mathematical difference between the total variation observed in our experimental model and the natural variance measured across individuals (the error variance). Partitioning the variance into components allows us to establish the ratio of the two components of variance, defined as $F_s$as shown in the following equation:

$$\frac{\text{among-group (treatment) variance}}{\text{within-group (error) variance}} = F_s$$

Testing the Null Hypothesis

After partitioning the variance into its two components (see discussion of Partitioning the Variance), we can then compare our calculated F_s to the *F* distribution, which is tabled in numerous statistics texts. The tables are usually organized with the critical values listed for many combinations of the numerator (treatment) versus denominator (error) degrees of freedom. Treatment degrees of freedom are the number of groups minus 1, and the error degrees of freedom are the total number of individual observations in all groups minus the number of groups. While it is unlikely that you will ever have to look up tabled *F* values, since computers do it for us, it is important to understand the convention used for *F* ratios. The *F* value given for a particular tested proportion is written as:

$$F_{(\text{treatment df, error df, } \alpha)} = \text{critical value from table.}$$

So, the critical value of *F* for an hypothetical experiment with a rejection level (α) of 0.05, four treatment groups, and a total experimental population of 48 organisms would be written as:

$$F_{(3,44,0.05)} = 2.82.$$

If α is chosen as 0.05, it is often omitted from the statistical reporting for brevity's sake. This critical value is important in our analysis because, if our calculated *F* ratio gives a value higher then the critical value from the table, we reject our null hypothesis. For ANOVA, the null hypothesis is that all the means for all the groups are equal. Rejecting the null hypothesis tells us only one thing: that at least one of the groups differs from the others. It does not tell us how many differ or from how many of the other groups any one might differ. For this answer we need to proceed to a method of multiple-means comparisons.

The ANOVA Table

Assuming a hypothetical experiment such as the one summarized in Table A.3I.7 and Table A.3I.8, the output from any ANOVA computer package is likely to produce an

Table A.3I.7 Hypothetical Optical Density Data for Bacterial Cultures

Control	Treatment A	Treatment B	Treatment C
0.53	0.37	0.52	0.7
0.57	0.44	0.55	0.66
0.6	0.44	0.59	0.72
0.49	0.39	0.59	0.69
0.42	0.62	0.7	0.55
0.4	0.58	0.72	0.53
0.41	0.6	0.73	0.5

Table A.3I.8 Descriptive Summary Table for Hypothetical Optical Density Data Shown in Table A.3I.7

Groups	Count	Sum	Mean	Variance
Control	7	3.42	0.488571	0.006581
Treatment A	7	3.44	0.491429	0.011081
Treatment B	7	4.4	0.628571	0.007448
Treatment C	7	4.35	0.621429	0.008381

Table A.3I.9 Single-Factor ANOVA Table for Hypothetical Optical Density Data Shown in Tables A.3I.7 and Table A.3I.8[a]

Source of Variation	SS	df	MS	F	p value	F_c
Among groups	0.12778214	3	0.042594	5.09	0.01	3.01
Within groups	0.20094286	24	0.008373	—	—	—
Total	0.328725	27	—	—	—	—

[a]Abbreviations: F_c, critical value of F statistic; SS, sum of squares; MS, mean square.

ANOVA table similar to Table A.3I.9, originally generated by Microsoft Excel. The data are hypothetical optical density readings for plasmid-transformed bacteria that have been transformed by three different experimental protocols, along with a control that has not been transformed.

The first part of the data analysis is a simple descriptive summary table (Table A.3I.8) that gives the sample size, mean, and variance for each group. Based on this table, it is fairly clear that the variances within groups are similar, but that the means for some of the groups seem to be different from others. The ANOVA table (Table A.3I.9) then gives the variance components and degrees of freedom accounted for among and within groups. The "SS" column indicates the sum of squares for each component, and the "MS" indicates mean square. Sum of squares is the summation of the square of the amount by which each observation differs from the total mean. The mean square is the same as the variance, and is calculated by dividing the SS by the appropriate degrees of freedom. The mean square within-groups value is often called the "mean square error," and figures prominently into tests of differences among the means.

The table also reports the calculated F ratio for the among versus within variance comparison. The computer then provides the exact p value associated with this ratio and sample size, and further indicates the critical F value that is necessary for rejection at the 0.05 level.

Which Means Are Different?

We have finally reached the point where we are ready to answer the question we asked, which was whether the average value of one or more of our groups differs from the others. While statisticians differ on this issue, it is the authors' opinion that if you fail to reject the null

hypothesis of your ANOVA you should not test for significant means differences unless you planned certain comparisons *before* you ran the experiment. That having been said, it is possible that you will find significant means differences even if you fail to reject the ANOVA null hypothesis. If that is the case, however, it is worth asking how robust your initial test was—did you have large-enough sample sizes, were your variances "equal enough," and were the distributions of your data groups normal or at least all similarly shaped? If not, you would probably be better off reworking your analysis with a nonparametric equivalent.

Assuming that you have rejected H_0, how can you tell which means differ? Again, a first approximation is to look at the average values for each group and see which appear most different from the others. Next, you will want to proceed to a test that will provide you with a *p* value for those differences. For the special the case of testing a control versus treatments, the Dunnett test is appropriate. Other test names you might see are Bonferroni or Tukey; computer packages have many options and you should refer to the statistical documentation of your package to determine which option best suits your data.

Since means comparisons are cumbersome to do by hand and are part of most statistical packages, we will not go into the specific calculations here. Make sure, however, that you understand how the computer is reporting the significance of your data to you. The computer will usually give you a *p* value. As you perform more mean-versus-mean comparisons, your acceptable α level will change, and the computer rarely keeps track of this, so it is important to know how to calculate these adjustments in α. One of these techniques is discussed below.

Planned or unplanned comparisons

The use of certain techniques in your final means comparison will depend, in part, on whether you planned before the experiment among which groups you were going to look for differences (a priori tests) or whether you waited to peruse your data first (a posteriori). This turns out to be a theoretically very complex subject and one that is not worth delving into here. A detailed treatment can be found in Sokal and Rohlf (1981).

It is worth pointing out here a common mistake in choosing significance levels for multiple means comparisons. In general, planned comparisons allow the researcher a greater rejection level (bigger α or *p* value) than unplanned comparisons. The reasoning is simple: if you first look at your data, then choose your comparisons, you are introducing a bias to your analysis. To account for this, the level of rejection needs to be tightened. Even if you have planned your comparisons, however, it is easy to abuse the power of multiple means tests.

For example, declaring a priori that you are going to conduct all possible pairwise comparisons in your experiment is tantamount to using multiple *t* tests (see discussion of Why Not Lots of *t* Tests?). If you have *k* groups in your analysis, the number of all possible pairwise comparisons would be $k(k-1)/2$ Since each comparison is not independent of all others, there is a need to alter the significance level of the tests. In order to use the experiment-wise error (i.e., $\alpha = p = 0.05$), the degrees of freedom for all the tests you perform should not exceed $k-1$. So, if you have four groups (*a,b,c,d*), you might declare to test *a* versus *b,c,d* (1 df); *a* versus *d* (1 df); and *b,c* versus *d* (1 df). That would burn up 3 degrees of freedom at $\alpha = 0.05$. Note that in the very common case of comparing each treatment to a control, we stay within this limit as well (assume *a* is the control and compare it pairwise with each of the three treatments *b,c,d*).

If we were to go beyond this level of comparison, however, we would need to adjust the rejection level for *all* the tests we perform to account for a loss of independence. A simple and conservative technique that does this is the Bonferroni method (Sokal and Rohlf, 1987). To make an adjustment to the experiment-wise error term, α, calculate a test significance, α', by dividing the experiment-wise error by the number of degrees of freedom used in your selected tests (df_T):

$$\alpha' = \frac{\alpha}{df_T}$$

Applying this technique to the example above, if we choose to compare the control *a* to each of the three treatment groups (3 df), and perform two additional internal treatment tests of *b* versus *c,d* (1 *df*) and *c* versus *d* (1 *df*), we have a total of 5 degrees of freedom, which exceeds our "allowable" number by 2. The Bonferroni method would then require that we adjust α' as follows:

$$\alpha' = \frac{0.05}{5} = 0.01$$

Therefore, for all the tests we perform, including all three of the control-versus-treatment tests, we have decreased our rejection

Table A.3I.10 Turbidity Data (see Table A.3I.7) Scored on a Scale of 1 (least turbid) to 10 (most turbid)

Control	Treatment A	Treatment B	Treatment C
5	1	4	8
5	3	4	9
5	2	6	9
5	3	5	8
6	2	6	8
5	5	7	9
5	1	6	10

Table A.3I.11 Turbidity Data (see Table A.3I.10) Ranked from 1 to *N*

	Control	Treatment A	Treatment B	Treatment C
	11.5	1.5	7.5	23
	11.5	5.5	7.5	26
	11.5	3.5	18.5	26
	11.5	5.5	11.5	23
	18.5	3.5	18.5	23
	11.5	11.5	21	26
	11.5	1.5	18.5	28
Rank sums	87.5	32.5	103	175
Rank averages	12.50	4.64	14.71	25.00

level to $p = 0.01$, a level that could make a big difference in the significance of our results. This should point out the importance of good experimental design and of choosing only those tests that are important to the experimental goals.

The Nonparametric ANOVA Equivalent

If the data you end up with do not meet the assumptions of ANOVA, you can still perform multiple comparisons with a nonparametric test. For a one-way ANOVA, the nonparametric equivalent most popularly used is the Kruskal-Wallis one-way analysis of variance. The assumptions of the Kruskal-Wallis test are less stringent than those of the parametric ANOVA. The samples are drawn from a population with a continuous distribution and the sampling must be random. The samples must be independent. Their values must be rankable, that is, on at least an ordinal scale. Many computer programs provide the option of calculating the Kruskal-Wallis statistic. It is simple to do by hand, and is worth outlining here.

In the hypothetical example of the optical densities for transformed bacteria strains (see Table A.3I.7, Table A.3I.8, and Table A.3I.9), assume that the researcher, instead of using a spectrophotometer, simply used an ordinal scale of 1 to 10 to approximate the turbidity of the sample (see Table A.3I.10). (Okay, so it's not very precise science, but, who knows, maybe the spec was broken that day and the samples needed to be measured!) In this case, there is an expectation that the scores would not be distributed in such a way as to meet the criteria for a parametric ANOVA.

To proceed through the analysis, it is then necessary to find the average value for the ranks of the data in each of the groups. To do this, we will give the data across all groups a rank score. The smallest number will have the lowest rank, and so on until we have ranked all the data. Equivalent values are termed "ties" and are given the average rank of the tied observations (see Table A.3I.11).

From these ranks, the test statistic (KW) is calculated according to the following formula:

$$KW = \left[\frac{12}{N(N-1)} \sum_{j=1}^{k} n_j \bar{R}_j^2 \right] - 3(N+1)$$

In the above equation, N = total population size, n_j = number of observations in a group (note that groups need not be equal in size), k = number of groups, and $\overline{R}_j$= rank average for each group.

For small samples (usually $k \leq 3$ with $n_j \leq 5$) there are special tabled values for the KW statistic. For larger samples, the resulting value is compared to a χ^2 distribution with $(k - 1)$ degrees of freedom. If the KW value is larger than the tabled valued at a chosen α, then H_0 is rejected. As in the case of the parametric ANOVA, a significant result implies that at least one of the groups is different from the rest. A means comparison test needs to be done to show which group it is. For the data in Table A.3I.11, the calculation is as follows:

$$KW = \left[\frac{12}{(28)(29)}\right]\Big[7(12.5)^2 + 7(4.64)^2 + 7(14.71)^2 + 7(25)^2\Big] - [3(28+1)]$$

$$= 18.45$$

The critical value for $\alpha = 0.05$, with df = $(k - 1) = 3$, is 7.82. Since our KW test statistic is larger than that, our test is significant and we reject H_0. A problem that we might have encountered with our data, however, is that we ended up with many tied ranks. There are mathematical corrections for ties that will yield a slightly higher KW statistic than the calculation without the correction. This means that if you reject H_0 without the correction, you will also reject with the correction, so it is unnecessary if your results are significant. If, on the other hand, if your significance is marginal and you are concerned about the ties in your data, you may choose to perform a ties correction. Siegel and Castellan (1988) gives one such correction. For our example, the ties correction would yield KW = 19.0.

Nonparametric Means Comparison

Having demonstrated a significant difference somewhere among the groups with the Kruskal-Wallis test (see discussion of The Nonparametric ANOVA Equivalent), it is then appropriate to proceed to find out where those differences are. The technique that we will apply will allow us to perform only pairwise comparisons, but it is adjusted so that all possible pairwise comparisons can be made with the same level of rejection.

Treatment versus treatment

Multiple comparisons among all treatments, as in the case of ANOVA, require attention to the error inherent in the entire experiment relative to each pairwise test we choose to do. To account for this, we adjust our test statistic by the total possible pairwise comparisons in the experiment. Whether we plan to do only some of them or all of them is irrelevant to the adjustment; however, there is a special case, outlined below, if all we wish to do is compare a single control to several treatments.

The many possible differences between pairs of observations in a data set will become approximately normal as the sample size increases, regardless of the underlying distribution of the data. This allows us to use the normal distribution probabilities to test for significant differences between groups. In effect, we can use the normal distribution to establish a confidence interval for the difference between two groups. If they fall beyond this value, we can say that their rank average values differ significantly.

To be able to use the normal approximation, we first need to adjust the value we look up in a normal probability table (z) by the number of pairwise group comparisons possible in the experiment:

$$z_{\alpha'} = z_{\alpha/k(k-1)}$$

In this equation, k is the number of groups in the experiment. For the data in Table A.3I.11, if we chose an experiment-wise $\alpha = 0.05$, this would yield $z_{\alpha'} = z_{0.05/(4)(3)} = z_{0.004}$. The numerical value of $z_{\alpha'}$ can be looked up on a normal table or derived from a computer program. In our case, $z_{0.004} = 2.65$.

Now that we have determined the correct normal approximation for our test, we apply it to determine the critical value for rejection at $\alpha = 0.05$ (not $\alpha' = 0.004$) between any two groups in the population. If the mathematical difference between the two rank averages of the groups is greater than the critical value, then we reject at $p < 0.05$. The calculation of the critical value (CV) is dependent on sample sizes:

$$CV = z_{\alpha'}\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_u} + \frac{1}{n_v}\right)}$$

In the above equation, N is the total population size and n_u and n_v are the sample sizes of the two groups that you are comparing (they may be different sizes).

Table A.3I.12 Pairwise Comparisons of Hypothetical Turbidity Ranked Data in Table A.3I.11

Comparison	Calculation	Significance[a]
Control vs. A	$\|12.5 - 4.64\| = 7.86$	NS
Control vs. B	$\|12.5 - 14.71\| = 2.21$	NS
Control vs. C	$\|12.5 - 25\| = 12.5$	*
A vs. B	$\|4.64 - 14.71\| = 10.07$	NS
A vs. C	$\|4.64 - 25\| = 20.36$	*
B vs. C	$\|14.71 - 25\| = 10.29$	NS

[a]NS, not significant; asterisk refers to differences significant at $p < 0.05$.

For our example (see Table A.3I.11), since all group sizes are the same, the CV will be identical for any and all of the six possible pairwise comparisons. If sample sizes are different, you need to calculate a new CV for each comparison:

$$\mathrm{CV} = 2.65\sqrt{\frac{28(29)}{12}\left(\frac{1}{7}+\frac{1}{7}\right)}$$

$$= 11.65$$

To determine which differences are significant, we first calculate the absolute value of the difference between all pairs of rank averages (Table A.3I.12). If the difference exceeds the critical value, then the difference can be determined to be significant at $p < 0.05$.

From this analysis it is clear that while the entire model was quite significant the differences in the data set are due mostly to the large value of Treatment C.

Control versus treatment

The test above provides a powerful nonparametric tool for comparing multiple samples. A researcher often is interested, however, in a less broad model—one that looks only at the differences between a control and two or more treatments. Following the same procedure as above, it is possible to approximate these differences with a normal distribution. A different correction is needed, however, to account for fewer, more defined comparisons. In this case the following equation applies:

$$z_{\alpha'} = z_{\alpha/2(k-1)}$$

For our example, this equation would yield $z_{\alpha'} = z_{0.005/(2)(3)} = z_{0.008} = 2.41$. Because fewer comparisons are being done, this smaller z value will lead to a smaller critical value and provide greater probability of finding a difference between the control and a treatment. Furthermore, if we expected a priori that our treatments would yield a directional result (either smaller or greater, but not both) relative to the control, we could reduce this number even more by adjusting by $\alpha/(k-1)$ rather than $\alpha/2(k-1)$.

Since we did not have this expectation in our data, we will use the two-tailed option and calculate as follows:

$$\mathrm{CV} = z_{\alpha'}\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_c}+\frac{1}{n_u}\right)}$$

In the equation above, N is the total population size, n_c is the sample size of the control, and n_u is the sample size of the treatment being compared.

Again, because all of our treatments have the same number of observations, we need calculate only one value for CV.

$$\mathrm{CV} = 2.41\sqrt{\frac{28(29)}{12}\left(\frac{1}{7}+\frac{1}{7}\right)}$$

$$= 10.60$$

This value is then compared as above to the differences between the control and treatment rank averages. We would reject any difference greater than 10.60. Despite the tighter rejection interval (compared with 11.65 from the all comparisons model) the only treatment that differs significantly from control at $p < 0.05$ is Treatment C. Note that the previously significant comparison (A versus C) is not valid in this model.

Table A.3I.13 Hypothetical Immunoprecipitation Data for Repeated-Measures ANOVA

	IgG precipitate (mg/ml)				
Mouse ID	Day 1	Day 2	Day 3	Day 4	Day 5
A	0	0	0.02	0.7	1.3
B	0.7	2.5	3.2	4.5	3
C	0.09	1.8	4	7.1	2.2
D	0.01	0.01	0.6	1.8	0.1
E	0.6	5.5	6.8	9.6	8.5
F	0.2	1.1	2.4	2.5	1.5
G	0	0.05	0.9	3	2.7
H	0.2	0.5	0.7	2.5	6

Table A.3I.14 Repeated-Measures ANOVA Table for Hypothetical Immunoprecipitation Data in Table A.3I.13[a]

Source of variation	SS	df	MS	F	p value	F_c
Rows	121.8885	7	17.41265	8.745258	1.12E-05	2.359258
Columns	68.29064	4	17.07266	8.574505	0.00012	2.714074
Error	55.75068	28	1.991096	—	—	—
Total	245.9298	39	—	—	—	—

[a]Abbreviations: F_c, critical value of F statistic; SS, sum of squares; MS, mean square.

Repeated Measures—A Special Case

Sometimes we wish to collect more than one sample from an individual organism in our study. Given that we would expect there to be less variability in samples taken from one individual than in samples taken among several individuals, it would seem that we need to adjust our analysis of variance somewhat. And indeed we do. Instead of a simple one-way analysis of variance, we perform a repeated-measures analysis of variance. This technique allows us to statistically control for the variance attributable to the inherent differences in the test subjects relative to the variance in response to the treatments. Once you determine that a difference exists somewhere in the data set, you can test for means differences in the same way as described above for one-way ANOVA.

Parametric

The same assumptions hold true for the repeated measures ANOVA as for the one-way ANOVA. For our parametric ANOVA example, let us consider the following case (see Table A.3I.13). You are producing an antibody in a strain of mice and want to know when antibody production reaches a peak. To collect data, you perform immunoprecipitation on serum collected for 5 days from eight study animals.

As you can see, certain animals have higher values overall than others, some respond with IgG production very quickly, others lag a bit and so on. If we were to do a one-way ANOVA on these data, the variability among the animals (rather than across days) might compromise the significance of our test. To avoid this problem, the repeated-measures ANOVA calculates variances for the rows and columns separately. The final output for such an analysis would resemble Table A.3I.14, originally produced by Microsoft Excel.

In this table there are significance values reported for the rows (mice) and columns (days). For our example, both factors are clearly significant. This implies that there is a difference in the precipitation of IgG over the 5 days of measurement and a difference in IgG expression in different mice.

Nonparametric

If you have repeated measures that violate the assumptions of the parametric ANOVA,

Table A.3I.15 Hypothetical Immunoprecipitation Data in Table A.3I.13 Ranked for Friedman's Analysis of Variance

Mouse ID	IgG precipitate (ranks)				
	Day 1	Day 2	Day 3	Day 4	Day 5
A	1.5	1.5	3	4	5
B	1	2	4	5	3
C	1	2	4	5	3
D	1.5	1.5	4	5	3
E	1	2	3	5	4
F	1	2	4	5	3
G	1	2	3	5	4
H	1	2	3	4	5
Rank sums	9	15	28	38	30

then there is a nonparametric equivalent for the repeated-measures ANOVA as well. As in the turbidity example above, a common case would be when a response was scored on an ordinal scale rather than measured on a continuous one.

The nonparametric equivalent in this case is the Friedman's analysis of variance. The data are handled in a fashion similar to the Kruskal-Wallis test (see discussion of The Nonparametric ANOVA Equivalent), except that ranks are assigned only within an individual rather than across the entire data set. In this way, an individual that overall got higher responses than the others would not contribute unduly to the overall variance of the population. To illustrate, let us simply use the immunoprecipitation data in Table A.3I.13. We begin by ranking the values within each individual from 1 to 5 (see Table A.3I.15).

We then sum the ranks for each column. Under the null hypothesis, the ranks should be distributed in random order for any individual, which would result in the column rank sums being approximately equal. Clearly the rank sums are not equal, but are the differences significant? We can determine this by using the Friedman test and comparing the resulting value to a χ^2 distribution with $k - 1$ degrees of freedom for our selected α level of rejection. The calculation of the statistic (Fr) is similar to the calculation of the Kruskal-Wallis statistic:

$$\mathrm{Fr} = \left[\frac{12}{Nk(k+1)} \sum_{j=1}^{k} R_j^2 \right] - 3N(k+1)$$

In the above equation, N is the number of rows (subjects), k is the number of columns (treatments), and R_j is equal to the rank sum of each column.

For our example, this would yield the following:

$$\mathrm{Fr} = \left[\frac{12}{(8)(5)(6)} (81 + 225 + 784 + 1444 + 900) \right] - (3)(8)(6)$$

$$\mathrm{Fr} = 27.7$$

The χ^2 critical value for $\alpha = 0.05$ with 4 degrees of freedom is 9.48. Since our Fr statistic is greater than that, we reject the null hypothesis and conclude that there are differences in IgG production over the course of 5 days (columns), irrespective of differences inherent in the test subjects (rows).

CONCLUSION

While the methods for dealing with comparisons of data outlined above are just a small sampling of those available, we hope they add to the arsenal of tools available to biologists. By performing simple descriptive analyses of your data and following the guidelines for test selection outlined in this Appendix, you should be able to avoid many of the statistical pitfalls frequently encountered in groups comparisons.

LITERATURE CITED

Siegel, S. and Castellan, N.J., Jr. 1988. Nonparametric Statistics for the Behavioral Sciences, 2nd ed. McGraw-Hill, New York.

Sokal, R.R. and Rohlf, F.J. 1981. Biometry, 2nd ed. W.H. Freeman, New York.

Sokal, R.R. and Rohlf, F.J. 1987. Introduction to Biostatistics, 2nd ed. W.H. Freeman, New York.

Zar, J.H. 1984. Biostatistical Analysis, 2nd ed. Prentice-Hall, Englewood Cliffs, New Jersey.

KEY REFERENCES

Hampton, R.E. 1994. Introductory Biological Statistics. William. C. Brown Publishers, Dubuque, Iowa.

Provides excellent outlines for many statistical tests and uses relevant, comprehensible biological examples.

Motulsky, H. 1995. Intuitive Biostatistics. Oxford University Press, New York.

A nice, nonmathematical introduction to biostatistics, covering standard topics as well as many areas important to biology that are often omitted from introductory texts.

Contributed by Elizabeth F. Ryder and Phil Robakiewicz
Worcester Polytechnic Institute
Worcester, Massachusetts