

Myers, A. (1987). Experimental Psychology.
Monterey, CA: Brooks/Cole
Chapters 12 & 13

In Chapter 11 we discussed the principles of statistical inference. We focused on the logic behind statistical tests. By now you may be thinking, "Okay, now I have all these principles. But I have some data and I still do not know what to do with them." By the end of this chapter, you will know how to begin your data analysis and how to carry it through when you have a two-group experiment.

We will begin by looking at the results of an experiment that has two independent groups. We will go through the basic stages of data analysis. We will organize and summarize the data and trace the process of selecting a statistical test for them. Then we will actually carry out that test. Finally, we will apply some of the same principles to handle a two-condition experiment that was run within subjects.

In Chapter 11 we considered a hypothetical experiment to compare the time estimates of two groups of subjects, those having fun and those not having fun. One group saw cartoons with captions for 10 minutes. The other group saw the same cartoons with the captions missing. Our assumption was that the incomplete cartoons would not be nearly as much fun for subjects. Compared to subjects who saw the incomplete cartoons, we predicted that subjects who saw the complete cartoons would estimate that less time had passed.

This experiment had a two independent groups design. Subjects were assigned at random to one of the two treatments. The exact hypothesis of the experiment was disguised. Subjects were merely told that they should examine the cartoons carefully and that they would be asked to rate them for "funniness" later. Because there were no strong arguments against it, we chose a $p < .05$ significance level for this experiment.

Suppose we have actually run the experiment and we have the data. What do we do with them? Where do we begin? There are three basic steps in analyzing any set of data: First, we organize the data. Second, we summarize them. Third, we apply a statistical test to interpret our results. We will look at each of these steps in detail.

If you have access to computers, you may wish to obtain packaged software to carry out the statistical procedures covered in this and

the next chapter. Several comprehensive software catalogs list what is available—for example, *Omni Complete Catalog of Computer Software*, 1984 and *CP/M Software Finder*, 1983. Minitab, SPSS™, SPSSX, Psychostat, and SAS are some software programs that are currently used by psychologists. However, you will develop a greater understanding of what the procedures accomplish if you follow the examples in the chapters and work out the practice exercises yourself first.

Organizing Data

We start by organizing the data. The hypothetical data from our time estimation experiment have been organized in Table 12-1. You can see that these data have been laid out in column form. Subjects' responses are divided into two columns, one for each of the two treatment groups. Each subject in each group is listed by number next to his or her datum. Statistical work will go more quickly and be more accurate if you begin by organizing the data and labeling them in a clear and orderly way. Especially with more complex designs, you will avoid a great deal of confusion if you take the time to prepare data tables. Many students find it easiest to use columnar paper, such as bookkeeping paper. At the very least, do your work on lined paper so you will be sure which datum belongs to which subject. You can simplify the task by planning an orderly data sheet that you can use to record subjects' responses throughout the experiment.

Summarizing Data: Using Descriptive Statistics

raw data

Published articles rarely contain the data obtained from every single subject in the experiment. The data we record as we run an experiment are called **raw data**. Raw data are like raw potatoes—they are unprocessed and sometimes hard to swallow.

summary data

Whenever we report the results of an experiment, we report **summary data** rather than raw data. Usually readers are not as interested in the scores of individual subjects, and neither are we.

Table 12-1 Laying out organized data

Group 1 (Incomplete Cartoons)		Group 2 (Complete Cartoons)	
Subject	Time Estimate (min)	Subject	Time Estimate (min)
1	11.2	6	13.6
2	16.2	7	10.9
3	13.3	8	5.5
4	12.1	9	8.8
5	18.2	10	9.2

descriptive statistics

We want to compare treatment effects, and we do that by comparing group data. (The only exception is a small *N* experiment.) When we have group data, we summarize them with **descriptive statistics**, shorthand ways of describing data. They represent standard procedures for summarizing results. When we want to order a shade, we do not carry the window frame to the hardware store. Instead, we summarize its characteristics using the standard dimensions of length and width. Similarly, we can summarize and describe data by using some of the standard descriptive statistics: measures of central tendency (mean, median, and mode) and measures of variability (range, variance, and standard deviation).

Measures of Central Tendency

As you know, statistics are quantitative indexes of the characteristics of our samples. Some of the most commonly computed and reported statistics are **measures of central tendency**, summary statistics that describe what is typical of a distribution. The **mode** is the score that occurs most often. The **median** is the score that divides the distribution in half so that half the distribution lies above the median, half below. The **mean** is the arithmetic average: Add all the scores together and divide by the total number of scores and you have the mean.

Together the mean, median, and mode are useful indicators of the *shape* of the distribution of values we have. If the distribution is symmetrical and has only one mode (no scores tied for most frequent), the mean, median, and mode will coincide. When the distribution is asymmetrical, the mean, median, and mode will be different, and each may lead to different impressions about the data.

The mean is particularly sensitive to asymmetry and is pulled in the direction of extreme scores. Consider this set of scores:

3 4 5 6 6 6 7 8 9

The mean is 6, the mode is 6, and the median is also 6. Suppose we substitute a higher score into the set:

3 4 5 6 6 6 7 8 18

The mode and median are still 6. However, the mean is now 8. The mean has increased because of one exceptionally large score.

Even when the means of two distributions are the same, the distributions may be quite different. Reichmann (1961) cited these examples:

a. 5 5 6 6 6 6 7 7

b. 1 2 4 9 10 10

In (a) the mean is 6 and truly is typical of the data. In (b) the mean is also 6, but clearly 6 is not a usual score. Note that the mode is useful information. The mode of (a) is 6, and the mode of (b) is 10.

Clearly, the particular statistic we choose to report can make a difference in the impression we create through our data. Wages are often cited as one category of data that is subject to distortion. The distribution of wages in the United States is not symmetrical; it is skewed to the right: Lots of people earn relatively small amounts of money, whereas a few earn a great deal. The mean income is thus always higher than the median or mode. Millionaires pull the mean up. Corporate management would prefer to use means in salary negotiations; labor would prefer to talk in terms of medians. When evaluating any descriptive data, it is important to ask who is reporting what statistic and for what purpose. (Huff's *How to Lie with Statistics* is a delightful treatise on this topic.) Figure 12-1 shows the relationship between the mean, median, and mode for various distributions.

Measuring Variability

We also use descriptive statistics to measure the amount of variability in data. So far, we have talked about variability in a commonsense way: We say that anything that fluctuates has variability. When we do statistical tests, variability has more specific meanings. It is defined numerically by one of several descriptive statistics: the range, the variance, and the standard deviation. By using these statistics, we can compare the variability of one sample with that of another.

range

The simplest measure of variability is the **range**, the difference between the largest and smallest scores in a set of data. If the scores on an exam varied from a high of 100 points to a low of 74, we would say that the range is 26. If the price of a 6-ounce bag of potato chips varies from 39 cents in one store to 53 cents in another, we would say that the price range is 14 cents. The range is often a useful measure. It can be computed quickly, and it gives a straightforward

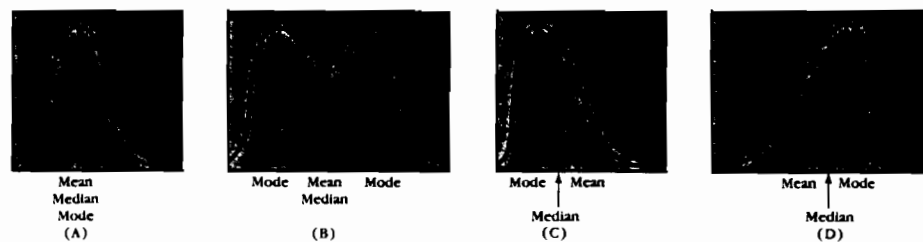


Figure 12-1 Mean, median, and mode in different distributions. From *Fundamental Statistics for Psychology*, 3rd Ed., by Robert B. McCall. Copyright © 1980 by Harcourt Brace Jovanovich, Inc. Reproduced by permission of the publisher.

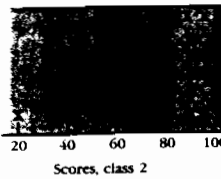
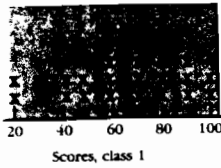


Figure 12-2 Two distributions of test scores with the same range (80 points).

indication of the spread between the high and low scores in a distribution.

The problem with using the range is that it does not reflect the amount of variability in all the scores. Figure 12-2 shows you two distributions of test scores that have the same range. However, you can see that the distributions are really very different from each other. Distribution for class 1 indicates that the test scores varied a great deal from student to student. In class 2, however, most students got similar scores: One extreme score accounts for the relatively large size of the range in this case. Knowing that these distributions have the same range tells us very little about them.

Computing the Variance

When we measure variability, we would like to be able to compare different samples in a more meaningful way. Computing the statistic we call the variance enables us to do that. Computing the variance is a way of transforming variability into a standard form that provides a good but simple description of how much individual scores differ from one another. By using the variance, we can talk about the variability of all our scores without having to present an entire set of data each time, just as we can order a shade without carrying a window frame to the shop.

The **variance** is the average squared deviation of scores from their mean. The easiest way to explain this is to show you how we get the variance. Table 12-2 shows the steps we follow to compute the

Table 12-2 Computing the variance: Time estimates of subjects who saw incomplete cartoons (in minutes)

Subject	X_i	$(X_i - \bar{X})^2$
1	11.2	$(-3.0)(-3.0) = 9.00$
2	16.2	$(2.0)(2.0) = 4.00$
3	13.3	$(-0.9)(-0.9) = 0.81$
4	12.1	$(-2.1)(-2.1) = 4.41$
5	18.2	$(4.0)(4.0) = 16.00$

Step 1. List each subject's score (X_i):

Step 2. Add the scores together:
 $\Sigma X_i = 71.0$

Step 3. Compute the mean:
 $\bar{X} = \frac{\Sigma X_i}{N}$
 $\bar{X} = \frac{71.0}{5}$
 $\bar{X} = 14.2 \text{ min}$

Step 4. Compute the deviation from the mean for each subject ($X_i - \bar{X}$):
 $11.2 - 14.2 = -3.0$
 $16.2 - 14.2 = 2.0$
 $13.3 - 14.2 = -0.9$
 $12.1 - 14.2 = -2.1$
 $18.2 - 14.2 = 4.0$

Step 5. Square the deviation from the mean for each subject ($X_i - \bar{X}$):
 $(-3.0)(-3.0) = 9.00$
 $(2.0)(2.0) = 4.00$
 $(-0.9)(-0.9) = 0.81$
 $(-2.1)(-2.1) = 4.41$
 $(4.0)(4.0) = 16.00$

Step 6. Add the squared deviations together:
 $\Sigma(X_i - \bar{X})^2 = 34.22$

Step 7. Compute the variance:
 $s^2 = \frac{\Sigma(X_i - \bar{X})^2}{N - 1}$
 $s^2 = \frac{34.22}{5 - 1}$ or 8.6 min

variance of the scores of one group of subjects. The scores are those of the hypothetical group of subjects who estimated the time that passed while they were looking at a series of cartoons with missing captions. The steps are numbered so that you can follow them more easily. One way or another, you will be computing the variance in just about every statistical test that you do, so it is important to master this concept.

Step 1. We have already listed each subject's score, so step 1 has been completed. Note that each subject's score is represented by X_i .

Step 2. If we have not already done so, we must compute the group mean. The scores of all the subjects in the group are added together. The total is represented in mathematical notation by the Greek letter *sigma* (Σ).

Step 3. Once you have the total of all the scores in the group, compute the mean. Divide the total of all the scores in the sample (ΣX_i) by the number of scores you have (N). The mean is represented by \bar{X} . The mean, an average, gives us an idea of what the overall sample is like. The mean time estimate of this sample is 14.2 minutes. You can see that some subjects gave estimates that are above the mean and others gave estimates that are below the mean. In fact, in this sample there are no subjects whose estimates are exactly equal to the mean. The mean is an overall description, but how representative it is of the sample depends on other factors, such as the variance.

Step 4. In step 4 we compute the deviations from the mean for each subject. Deviations are differences between what we see and what is typical. If someone behaves in an unusual way, we say that his or her behavior deviates, or is deviant from the norm.

A sample of data has a mean—that is, an average that gives us an idea of what the sample is like. The mean of this sample is 14.2 minutes. This tells us that when we added up the estimates of all the subjects in the sample and divided by the number of subjects, we got 14.2 minutes. It does not say that *each* subject estimated 14.2 minutes. Actually, we know that no one in this sample said 14.2 minutes had passed: All the subjects deviated from the group mean. But by how much? We can find out simply by calculating the difference between each subject's estimate and the group mean. We subtract 14.2 (\bar{X}) from the actual estimate each subject gave.

Step 5. Next, we square the deviations from the mean. Remember that when we multiply a negative number (for example, -30) by itself, the result is a positive number. That means that all squared deviations from the mean will be positive numbers even though some of the subjects fell below the mean and so had negative deviations from the mean.

Step 6. In step 6 we add all the deviations together to get a total of squared deviations from the mean.

Step 7. Finally, we compute the variance. You can see that we need the results of steps 1 through 6 before we can do step 7. The formula is just a shorthand way of writing all the operations we perform to get the variance. The variance for a sample is represented by s^2 . The s comes from the Greek letter sigma used to represent the variance of a population. The formula tells us that to get the variance of a sample, we have to divide the sum of the squared deviations from the mean by $N - 1$. N is the number of scores.

The variance formula actually tells us to compute an average—the average of the squared deviations from the mean. That is exactly what we mean by variance. The variance is an indicator of how much variability there is in our data. The more variability, the larger the variance. You may be wondering why the variance formula tells us to divide by $N - 1$, not N : When we compute the mean, we divide by N . When we compute the variance for a sample, we are actually trying to estimate how much variability there is in the population we are sampling. Since we cannot measure the whole population, we draw inferences from samples. We can get a more accurate idea or estimate of how much variability there is in the population if we compute the variance of a sample by dividing by $N - 1$ instead of N .

Our estimate of the variance would be too small if we divided by N . Because deviations from the mean are squared when we compute the variance, extreme members of the population will enlarge the variance a great deal. A deviation of 4 will add 16 to the total of deviations from the mean; a deviation of 8 will add 64. When we sample, we are likely to miss some of those extreme individuals because there are relatively few of them in the population. Dividing by $N - 1$ gives us a larger variance estimate and corrects for the fact that we miss some of the extremes.

The variance for the subjects who were shown the incomplete cartoons is 8.6 minutes.¹ If we take the square root of the variance (s^2), we have another useful measure of variability, the **standard deviation**, or s . It reflects the average deviation of scores about the mean. We cannot compute that average directly by adding up the deviations; the total of the deviations from the mean is always zero. So we use the square root of the variance to return to the original unsquared units of measurement. The standard deviation of our “no fun” group means that, on average, we can expect each individual subject to deviate from the group mean by 2.93 minutes. We use the same procedures for each treatment group. To save time, let me tell you that those computations yield a group mean of 9.6 minutes and a

¹ We have rounded this off to the nearest tenth. Remember the conventions used for rounding: We round down if the last digit is less than 5. We round up if the last digit is more than 5. If the last digit is 5, we follow this rule: Round up if the digit to be rounded off is an odd number. Simply drop the 5 if the digit to be rounded is an even number. Thus, 8.35 becomes 8.4, and 4.65 becomes 4.6.

Which Test Do I Use?

variance of 8.8 minutes for the “fun group.” subjects who saw cartoons complete with caption. You may want to verify those figures by working through the formulas on your own.

When we report the results of our experiments, we usually report these summary statistics, in place of raw data. We give the mean and the variance of each treatment group. We may also report the range, although the range is often not given in published reports because of its limited use. We have now completed the first two stages of analyzing our results: We have organized and summarized the data. We will use the summary data again when we do the statistical tests.

How to choose a statistical test was postponed until now so that the various aspects of data analysis could be presented in one section of the text. However, in practice, it is best to select a test (and a significance level) as you plan the design of the experiment.

When we looked at experimental designs, we developed a set of questions to help us choose the best design for an experiment. We can make decisions about what statistical tests to use in much the same way. The number of independent variables is still important. How many do you have? How many treatment groups? Is the experiment within or between subjects? Did you use matching? As you become more familiar with choosing and using statistics, you will not need to go through all these steps. But you will find it much easier to choose the right test if you begin with these questions. In addition to the number of independent variables and treatment conditions, we need to consider the type of data we are analyzing. The way we measured the dependent variable makes a big difference in the way we handle the results. There are different statistical tests for different kinds of data.

Levels of Measurement

level of measurement

ratio scale

Recall from Chapter 4 that the **level of measurement** is the kind of scale used to measure a variable. There are four levels of measurement: ratio, interval, ordinal, and nominal. To review quickly, a **ratio scale** has equal intervals between all its values and an absolute zero point. These attributes enable us to express relationships between values on these scales as ratios: We can say 2 pounds are twice as heavy as 1 pound.

interval scale

An **interval scale** also measures magnitude or quantitative size and has equal intervals between values. However, it has no absolute zero point.

ordinal scale

An **ordinal scale** reflects differences only in magnitude, where magnitude is measured in the form of ranks. We cannot be sure that the intervals between values are equal, and the scale has no absolute zero.

standard deviation

nominal scale

A **nominal scale** classifies items into distinct categories that have no quantitative relationship to one another. Nominal scaling provides the least information. It tells nothing about magnitude, nor does it have equal intervals between values.

Variables may be measured by using one of these four different types of scales. We have looked mainly at ratio and interval data in our examples because these scales yield the most information and researchers tend to prefer them. But remember that different techniques are needed for different types of data.

Selecting a Test for a Two-Group Experiment

To select the right statistical test, first decide which level of measurement is used to measure the dependent variable, and answer the other questions summarized in Table 12-3. You now have all the information you need to select a statistical test.

Table 12-3 The parameters of data analysis

1. How many independent variables are there?
2. How many treatment conditions are there?
3. Is the experiment run between or within subjects?
4. Are the subjects matched?
5. What is the level of measurement of the dependent variable?

Let us return to our example. How will we know what statistical test to use? First, we answer the key questions from Table 12-3.

1. There is one independent variable (*fun*).
2. There are two treatment conditions (*fun versus no fun*).
3. The experiment is run between subjects. (*There are different subjects in each treatment condition.*)
4. The subjects are not matched.
5. The dependent variable is measured by a ratio scale. (*Time estimates have magnitude, equal intervals, and an absolute zero—there is no way to score below zero.*)²

With this information, we can select a possible test to use for the data. Table 12-4 shows the most common statistical tests, organized by the number of independent variables they can handle, the level of measurement of the dependent variable, and whether the experiment is within or between subjects. We will not discuss all the tests in detail; the table note supplies sources to consult for further information.

² It can be argued that time estimates do not constitute a *true* ratio scale because we have no way of knowing whether subjects use equal intervals between values. The 1st minute of waiting in line, for example, might seem longer or shorter than the 31st. The same argument can be made regarding other dependent measures that rely on subjects' perceptions, such as ratings of attractiveness. This difficulty is often ignored by researchers because the statistical tests commonly used are reasonably accurate in spite of it.

Table 12-4 Selecting a possible statistical test by number of independent variables and level of measurement

Level of Measurement of Dependent Variable	One Independent Variable				Two Independent Variables	
	Two Treatments		More Than Two		Factorial Designs	
	Two Independent Groups	Two Matched Groups (or Within Subjects)	Multiple Independent Groups	Multiple Matched Groups (or Within Subjects)	Independent Groups	Matched Groups (or Within Subjects)
Interval or ratio	<i>t</i> test for independent groups	<i>t</i> test for matched groups	One-way ANOVA (randomized)	One-way ANOVA (repeated-measures)	Two-way ANOVA	Two-way ANOVA (repeated measures)
Ordinal	Mann-Whitney U test	Wilcoxon test	Kruskal-Wallis test	Friedman test	—	—
Nominal	Chi square test	—	Chi square test	—	—	—

Note: Shaded boxes list tests not discussed in this text. You can find explanations of them in most standard texts on statistics. A good source is McCall, R. B. (1980), *Fundamental statistics for psychology* (3rd ed.), New York: Harcourt Brace Jovanovich. See Winer, B. J. (1971), *Statistical principles in experimental design*, New York: McGraw-Hill, for repeated measures procedures.

There are other tests that we do not even list, but they are used less often and you may not need them until you take more advanced courses. Here and in the next chapter we will focus on the tests you are most likely to need for your first experiments.

The table indicates "possible" tests; it does not tell you what you will definitely need in all cases. The reason is that it may be possible to use more than one test. Also, before using any test, we must be sure we have the kind of data the test was designed to handle. This goes beyond asking whether our level of measurement is appropriate. As you learn about the tests, you will also learn that each test has its own additional requirements.

The *t* Test

For our hypothetical experiment, the test suggested in Table 12-4 is the *t* test for independent groups. When we want to evaluate interval or ratio data from a two-group experiment, we compute the test statistic *t*. The *t* statistic is a computational way of relating differences between treatment means to the amount of variability we would expect to see between any two sets of data drawn from the same population. When we evaluate the likelihood of obtaining a particular value of *t*, we are performing a ***t* test**.

The exact probabilities of each value of *t* have been calculated for us. However, the distribution of these values changes depending on the number of subjects in the samples. Before actually computing *t*

for the time estimation example, let us examine the family of t distributions and the effects of sample size.

Effects of Sample Size

Size of sample is very important. If we take both small and large samples from the same populations, we will generally find that small samples vary more from the mean of the population than large samples do. You know that test statistics represent a relationship between treatment effects and variability. If sample size affects variability, it also affects the size of the test statistics.

For a test statistic such as t , sample size is critical because the exact shape of the distribution of t changes depending on the size of the samples. The t statistic has a whole family of distributions, some of which are shown in Figure 12-3. The t distributions resemble the normal curve we looked at in Chapter 11. They are symmetrical, with the greatest concentration of values around the mean. The shape of the t distribution becomes more and more like the normal curve as the sample size increases. With small samples, the t distribution has a flatter and wider shape.

Sample size is also important because of the assumptions we make whenever we apply t . One of the requirements of a t test is that the data to be analyzed (interval or ratio) come from populations that are normally distributed. We must be able to assume that if we could somehow measure all the members of the population, their scores on the dependent variable would form a normal curve. If the data come from populations that are not normally distributed, we have a problem. The odds of getting each individual t value were worked out for populations that are normally distributed. If the data do not come from such populations, the odds that have been worked out for t will be wrong for those data. Of course, we can hardly ever measure all the members of a population. We get around this problem by using large samples so that the correct odds of each t value are very close to what they would be if the population were normally distributed. This is rarely a problem because the t test is relatively

robust **robust**—its assumptions can be violated without creating serious errors. If there are at least 10 to 20 subjects in each treatment group, a t test is probably safe when the other conditions are met.

Degrees of Freedom

degrees of freedom We select the appropriate t distribution based on **degrees of freedom**. [Figure 12-3 refers to degrees of freedom (df) rather than number of subjects.] The degrees of freedom tell us how many members of a set of data could vary or change value without changing the value of a statistic we already know for those data. Samples that are the same size can have different degrees of freedom depending on the way the experiment is designed and on the statistic being computed. Let us say we know the mean of the data. Then the de-

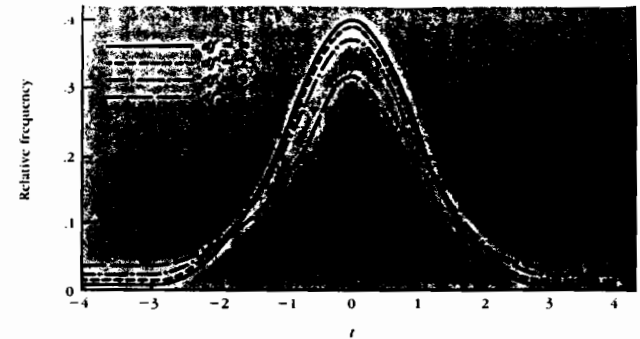


Figure 12-3 Some members of the family of t distributions. Adapted from *Quantitative Methods in Psychology*, by D. Lewis. Copyright © 1960 by McGraw-Hill, Inc. Reprinted by permission.

grees of freedom tell us how many members of that set of data could change without altering the value of the mean.

Imagine that my phone number is a set of data. It has seven digits. Suppose I tell you that the total of the seven digits in my number is 37 and that the first six digits of the number are 8, 9, 4, 3, 9, and 2. Can you find the last digit? Of course you can. Since you know the total and six of the digits, you can easily compute the value of the last digit, which is *not free to vary*. Different combinations of the first six digits are possible. But once their values have been set, the value of the last digit is also set if the total must equal 37. If we tried to substitute any other value for the seventh digit, the total, or the known statistic, would no longer be correct. The degrees of freedom for my phone number therefore equal 6. If we include the area code in the data, we would say that my telephone number has 10 digits. It now totals 42. Its degrees of freedom now equal $10 - 1$, or 9. Clearly, the degrees of freedom are related to the number of digits, or data, in a sample.

Similarly, the degrees of freedom in the distribution of a statistic vary in a way related to the number of subjects sampled. However, we compute degrees of freedom differently for different test statistics. Sometimes all but one value of a set of data can change, sometimes fewer. The way we compute the degrees of freedom can also be different for different applications of the same statistic. The way we compute the degrees of freedom for t , for instance, changes depending on what we are testing with t . If we are using different statistics or the same statistic applied in different ways, we may have different degrees of freedom even though sample sizes are identical. That is why the critical values of test statistics are always presented and organized by degrees of freedom rather than by number of subjects.

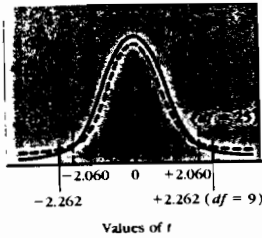


Figure 12-4 The t distribution for 25 and 9 degrees of freedom and critical values at $p < .05$.

The Critical Value of t

Let us look more closely at two distributions of t to get a clearer idea of how degrees of freedom will affect the critical value of t . Figure 12-4 shows distributions of t for 25 and 9 degrees of freedom. It also shows the cutoffs for the $p < .05$ significance level using a two-tailed test. What is the relationship between these levels?

As the t distribution changes shape, the value of t needed to reject the null hypothesis also changes. Remember that the significance level, or critical level, refers to probabilities. We are looking to see whether the value of t that we compute is more or less likely than our chosen critical value. If the experimental manipulation was effective, the computed value of t should be more extreme than the chosen critical value. In terms of probabilities, this may mean, for instance, that the observed value of t is so unlikely it could have occurred by chance less than 5 percent of the time.

It is easy to see that the distribution for 9 degrees of freedom is flatter and wider than the other curve shown. In terms of probabilities, you can see that the most extreme 5 percent of this distribution falls relatively far out on the curve. Smaller degrees of freedom mean more variability between samples. That means that more and more cases will be far from the mean of the population; large differences between samples will occur relatively often. Thus, the tails of the t distribution will get fatter as sample size (and degrees of freedom) get smaller. With 25 degrees of freedom, the tails of the t distribution are thinner: The most extreme 5 percent of the distribution falls closer to the mean.

Can you see what this means for our decision about the null hypothesis? We will reject the null hypothesis if the computed value of t is more extreme than the most extreme 5 percent of the distribution. Because the distribution changes shape as the degrees of freedom change, the critical value of t also changes. In fact, the critical value of t (the value needed to reject the null hypothesis) gets larger as the degrees of freedom get smaller. The fewer the subjects, the less likely it is that we will be able to reject the null hypothesis. And with fewer subjects, we have a greater chance of making a Type 2 error.

Using the t Test

Before we can use the t test, we have to know how to find the critical values of t . Fortunately, they have been worked out for us and organized into tables in which we can find them quickly and easily. To use the statistical tables, we need three pieces of information.

1. Will we use a directional or nondirectional test (that is, a one- or two-tailed test)?
2. What is our significance level?
3. How many degrees of freedom do we have?

In our experiment on fun, we have a directional hypothesis. Our directional hypothesis predicts that the average time estimate of the "no fun" group will be greater than the average time estimate of the "fun" group. That means that we may use a one-tailed test. We chose a $p < .05$ significance level. The degrees of freedom (df) for this experiment equal the total number of subjects in both groups minus the number of groups. Here, $df = 5 + 5 - 2$, or 8. Now, turn to Appendix B in the back of the book and find Table B2. Table B2 shows the critical values of t for both one- and two-tailed tests and several degrees of freedom. Find the critical value of t for 8 df , a one-tailed test, and $p < .05$.

The critical value of t for this experiment is 1.86. If our observed value of t (the value we compute) is less than 1.86, we must accept the null hypothesis. Computed values of t that are less extreme than the critical value indicate that differences between our treatment groups are not large enough to be significant. They were probably caused by chance variations between the samples. If the computed value of t is equal to or greater than 1.86, we reject the null hypothesis. If the computed value of t is more extreme than the table value, it is unlikely that the differences between treatment groups can be explained simply by chance.

The t Test for Independent Groups

We cannot just compare raw data or absolute differences between treatment groups; we must evaluate the results by taking variability into account. Treatment groups are likely to differ even if the independent variable had no effect. Computing a test statistic gives a numerical index of this relationship; t is just one of many test statistics we can compute. We use the **t test for independent groups** when we have two different randomly selected samples of subjects, randomly assigned to two treatments, and interval or ratio data. Let us use t to analyze the results of the time estimation experiment. The hypothetical data from that experiment are summarized in Table 12-5.

Table 12-5 Summary data for a hypothetical experiment on fun and time estimation

No Fun Group (Group 1)	Fun Group (Group 2)
$\bar{X}_1 = 14.2$ min	$\bar{X}_2 = 9.6$ min
$s_1^2 = 8.6$ min	$s_2^2 = 8.8$ min
$N_1 = 5$	$N_2 = 5$

Note. Since we predicted that the "no fun" group will make larger time estimates, we labeled that group "group 1." Given our prediction, $\bar{X}_1 - \bar{X}_2$ should be a positive number and our computed value of t should be positive. It does not really matter which way the groups are labeled as long as we set up the critical value of t (in a positive or negative direction) consistent with our predictions and our computations of t . We are using hypothetical data here to keep things simple. If you were actually running this experiment, you would want to have more than five subjects in each treatment group.

t test for independent groups

The table tells us at a glance that the performance of the groups is different. The mean time estimate of subjects in the "no fun" group is 14.2 minutes; that of subjects in the "fun" group is 9.6 minutes. Of course, we know that this absolute difference may not be significant. We have to evaluate the difference in terms of the amount of variability we find between any samples drawn from the population. We have to decide whether to accept or reject the null hypothesis. To do that we must compute the observed value of t for these data using this formula:

$$t_{\text{obs}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{(N_1 + N_2 - 2)}\right) \cdot \left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

The computation of t for these data is shown in Table 12-6. The formula is just a shorthand way of writing the steps to get t . If you take it slowly and step by step, you should have no trouble.

When we first talked about test statistics in a general way, we said that they represent a relationship between treatment effects and variability. If you think about what is shown in the formula for t , you will see very clearly how that principle is applied. The numerator (the top) tells us to find the difference between the means of the treatment groups. If the independent variable had a large effect, we would expect this difference to be relatively large. Notice that the denominator of the formula (the bottom) is a collection of terms that represent the variances of the treatment groups and the number of subjects in each. The denominator is an estimate of variability. If the ratio between the two components is relatively large, we may be able to reject the null hypothesis. We will reject it if the computed value of t is more extreme than 1.86. If the observed value of t is more extreme than that, the odds are good that it did not occur by chance.

The computed value of t turns out to be 2.47. Since this is more extreme than the critical value, we reject the null hypothesis: There is a significant difference between the time estimates of subjects who had fun and subjects who did not. How much importance we wish to attach to these findings depends partly on our assessment of the quality of this experiment. Were control procedures adequate? Were variables defined appropriately? These issues will influence our final judgment. The possibility of a Type 1 error must be considered. Before we can draw any sweeping conclusions from the findings, they must be replicated. We will return to these issues when we discuss the interpretation of results.

The t Test for Matched Groups

The procedures we have discussed so far assume that the two samples of subjects are independent, randomly selected groups. We need different procedures when we look at the data for matched

Table 12-6 Computation of t for the data presented in Table 12-5, fun and time estimation example

Step 1. Lay out the formula.	$t_{\text{obs}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{(N_1 + N_2 - 2)}\right) \cdot \left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$
Step 2. Put in all the quantities needed.	$t_{\text{obs}} = \frac{14.2 - 9.6}{\sqrt{\left(\frac{(5 - 1)8.6 + (5 - 1)8.8}{(5 + 5 - 2)}\right) \cdot \left(\frac{1}{5} + \frac{1}{5}\right)}}$
Step 3. Calculate the difference between treatment means; begin simplifying the denominator.	$t_{\text{obs}} = \frac{4.6}{\sqrt{\left(\frac{(4)8.6 + (4)8.8}{(8)}\right) \cdot \left(\frac{1}{5} + \frac{1}{5}\right)}}$
Step 4. Continue simplifying the denominator.	$t_{\text{obs}} = \frac{4.6}{\sqrt{\left(\frac{34.4 + 35.2}{(8)}\right) \cdot \left(\frac{1}{5} + \frac{1}{5}\right)}}$
Step 5. Remember to complete all operations inside the parentheses first.	$t_{\text{obs}} = \frac{4.6}{\sqrt{\left(\frac{69.6}{8}\right) \cdot \left(\frac{2}{5}\right)}}$
Step 6. Convert all fractions in the denominator to decimals.	$t_{\text{obs}} = \frac{4.6}{\sqrt{(8.7) \cdot (.40)}}$
Step 7. Complete the multiplication.	$t_{\text{obs}} = \frac{4.6}{\sqrt{(3.48)}}$
Step 8. Remember to take the square root of the denominator.	$t_{\text{obs}} = \frac{4.6}{1.86}$
Step 9. Divide the numerator by the denominator and you have the computed value of t . Compare it with the critical value.	$t_{\text{obs}} = 2.47$

Note. $df = N_1 + N_2 - 2$; $df = 5 + 5 - 2$, or 8.

groups of subjects. If we did statistical tests for these experiments in the same way as for an independent groups experiment, we would overestimate the amount of variability in the population sampled.

You know that subjects are apt to differ on a dependent variable simply because subjects are not all the same. Even if we are testing rats, we find that some run faster than others. One source of variability is individual differences. Subjects' scores vary because subjects differ from one another. Even the scores of the *same* subjects measured at different times vary, but they usually do not vary quite as much as the responses of different subjects. Neither do the scores of

subjects who are matched on a relevant variable. For these reasons, the way that we compute variability changes when we use matched groups or a within-subjects design. You will get a better sense of how these procedures compare by looking at an example of a within-subjects experiment done with two treatment conditions. The research problem is summarized in Box 12-1.

What statistical test would you use for the Yin experiment? (See Table 12-4.) This experiment is similar to the time estimation example. We are looking at one independent variable, two treatment conditions, and a ratio scale of measurement (that is, number of errors). However, this experiment is also quite different because it was run with only one group of subjects: It was a within-subjects experiment. The appropriate statistical test is a *t* test for matched groups. Obviously, the treatment groups were not matched per se in this experiment; they are simply the same subjects, and there may be no better match than that.

t test for matched groups

The *t* test for matched groups uses the same family of *t* distributions you have already seen. It also applies to interval and ratio data and requires the assumption that the population sampled is normally

Box 12-1 A two-group within-subjects example

Robert Yin had already done research that showed that people had trouble recognizing photographs of faces they had seen upside down (Yin, 1969). (Figure 12-5 shows an upside-down example.) Now he wanted to test whether inversion would affect recognition of line drawings as well. He showed the same subjects two sets of drawings of faces.* After each set, he showed subjects pairs of drawings and asked them to pick out the drawing they had seen before in each pair. Subjects saw one set of drawings and test pairs in the usual upright position; the other set was shown to them upside down. The order of presentation of the blocks was counterbalanced across subjects to control for order effects.



Figure 12-5 Do you recognize this person?
Reprinted by permission.

* Yin tested for effects on the memory of costumed figures as well as faces. Since he used the same procedures to handle both kinds of pictures, we will focus only on faces. Yin's results for the costumed figures were similar but less dramatic than the effect of inversion on faces.

distributed on the dependent variable. But because it is used to evaluate data from an experiment in which the treatment groups are not independent, the computations are handled differently.

Table 12-7 shows how data and computations from a within-subjects (or matched groups) experiment would look. The data are hypothetical and are presented just to illustrate the procedures simply; Yin used many more subjects. The scores for each subject represent the number of errors each made under each treatment condition. The table also illustrates the computation of *t* for matched groups: We use the same procedures for within- and matched-groups experiments with two treatment conditions.

From the table, you can see that we compute *t* for these data by looking at differences between each subject's performance under the two treatment conditions. This reflects the logic behind the design we are using. We are evaluating the effect of the independent variable *within* each subject. Similarly, when we have matched pairs

Table 12-7 Evaluating data from a within-subjects experiment on memory for inverted faces (scores represent number recalled correctly)

Subject (or Pair)	Upright Faces (X_1)	Inverted Faces (X_2)	Difference Scores ($X_1 - X_2$) = D_i	D_i^2
1	5	3	(5 - 3) = 2	4
2	3	2	(3 - 2) = 1	1
3	4	3	(4 - 3) = 1	1
4	5	3	(5 - 3) = 2	4
5	3	0	(3 - 0) = 3	9
			$\Sigma D_i = 9$	$\Sigma D_i^2 = 19$

Computing *t*

Step 1. This formula for *t* requires difference scores (D_i). The computation of *t* is based on differences between pairs of scores rather than group means. (Note how difference scores were computed above.)

$$t_{\text{obs}} = \frac{\Sigma D_i}{\sqrt{\frac{N\Sigma D_i^2 - (\Sigma D_i)^2}{N - 1}}}$$

Step 2. Put in all the required values.

Note that *N* stands for the number of pairs of data.

$$t_{\text{obs}} = \frac{9}{\sqrt{\frac{5(19) - (9)^2}{5 - 1}}}$$

Step 3. Simplify the denominator. Remember to take the square root.

$$t_{\text{obs}} = \frac{9}{\sqrt{\frac{(95) - (81)}{4}}} \text{ or } \frac{9}{\sqrt{4}}$$

Step 4. Our computed *t*. We are now ready to make a decision on the null hypothesis.

$$t_{\text{obs}} = 4.81$$

$df = N - 1$, where *N* is the number of pairs of scores.

Note: A *t* test for matched groups, to be used for two matched groups or a two-condition within-subjects design, ratio or interval data only. These hypothetical scores represent the number of drawings correctly recognized out of five presented in each condition.

of subjects, we want to look at the effects of our independent variable *within* each pair. The observed value of t for these data is $t_{\text{obs}} = 4.81$. How does that compare to the critical value of t ? Can we reject the null hypothesis? Who knows? Unfortunately, we never bothered to figure out what the critical value of t would be. Of course, we need to look at Table B2 (see Appendix B). Assume that the researcher had decided to use a $p < .05$ significance level. Should we use a one- or a two-tailed test? Although Yin might have made a directional prediction based on prior evidence, he was simply testing the notion that inversion would affect line drawings of faces. He did not specify the direction of the effect, so a two-tailed test is appropriate.

Since we computed t for this experiment with different procedures, we also have to compute df differently. Because we are looking at differences between pairs of scores, our df are based on the number of pairs. The df for two matched groups is $N - 1$, where N is the number of pairs. The df for this experiment is $5 - 1$ or 4 . If you look at Appendix B, Table B2, you will see that the critical value of t for 4 df and a two-tailed test ($p < .05$) is ± 2.776 . Since the computed value of t (4.81) is more extreme than the critical value, we reject the null hypothesis, which says that these data were sampled from the same population. (The actual data from Yin's experiment yielded a significant difference: Subjects had significantly more difficulty recognizing drawings presented upside down.)

Notice that using the within-subjects procedures affects the critical value of t . In the last example, we had five subjects in each group and 8 df ($5 + 5 - 2$). The critical value of t was 1.86 . Even though we have the same number of actual scores in both examples, we have about half as many degrees of freedom in the within-subjects (or matched groups) experiment. The critical value of t for 4 df ($p < .05$) is ± 2.776 . It takes a more extreme t to reach significance in the matched groups or within-subjects experiment.

If we need a larger value of t in the matched groups experiment, why bother matching groups at all? It would be easier to find a significant difference with the independent groups design. Actually, it would not. We have not yet discussed all the reasons for variability in data, but one thing should be clear: If we measure the responses of *different* subjects, we are likely to get more variability than we will if we measure the same subjects, or matched subjects. Using the matched groups design lowers the amount of variability in the data. Look at the formulas for t : The denominator (the bottom) of each formula reflects variability. When we reduce variability among individual subjects, we make the denominator of the t formula smaller. That in turn makes the computed value of t larger. To put it more simply, when we use a matched groups or a within-subjects design, we have a tradeoff: We lower the degrees of freedom for the experiment, but we also lower the amount of variability produced by factors other than the independent variable. And, as you already know, that can give us a more precise measure of the effect of the experimental manipulation.

Summary

There are four basic steps in analyzing results: (1) organize the data; (2) summarize them; (3) apply the appropriate statistical test; (4) interpret the outcome of the test. We organize data by making sure that all subjects' responses are labeled clearly and separated by treatment condition. We summarize data by computing *descriptive statistics*, shorthand representations of data. Some commonly used descriptive statistics are the measures of central tendency (mean, median, and mode) and measures of variability (range, variance, and standard deviation).

The *mean* is the arithmetic average of all scores in a group. The *mode* is the most frequent score. The *median* is the score that divides the distribution in half.

We also want to know how much variability there is among subjects' scores: how much they differ from one another. The *range* is the difference between the largest and the smallest scores in a set of data. Two distributions with the same range can look quite different; the range shows only how much the highest and lowest scores differ. The *variance* is a more precise indication of the amount of variability. It reflects the amount of variability among all the scores in a distribution, and so it is a more useful indicator than the range. The larger the variance, the more subjects' scores differ from one another. The *standard deviation* is the square root of the variance. It reflects the average deviation of scores about the mean. Finding the standard deviation converts "squared deviations" back to the original unsquared units of measurement. The larger the standard deviation, the more each individual subject is apt to differ from the group mean.

Five basic questions help us choose an appropriate statistical test: (1) How many independent variables are there? (2) How many treatment conditions are there? (3) Is the experiment run between or within subjects? (4) Was matching used? (5) What is the level of measurement of the independent variable?

Two common statistical tests are the *t test for independent groups* and the *t test for matched groups*. We use t tests to evaluate interval or ratio data from two-group experiments. The t statistic is a computational way of relating differences between treatment means to the amount of variability expected between any two samples of data from the same population. One of the assumptions of the test is that the data come from populations that are normally distributed on the dependent variable. A t test is done by computing a t statistic for the data. The computed value of t is compared to the table, or critical, value of t based on the chosen significance level. If the computed value of t is more extreme than the table value, the null hypothesis is rejected; the difference between treatment means is statistically significant.

The t statistic has a whole family of distributions. The appropriate distribution is selected based on the *degrees of freedom* for the

experiment. The degrees of freedom indicate how many members of a set of data could vary or change value without changing the value of a statistic already known for that data. With two independent groups, the degrees of freedom for t are equal to the total number of subjects minus 2. As the degrees of freedom of t get larger, the critical value of t gets less extreme. In addition to the degrees of freedom, the critical value of t also depends on whether the hypothesis is directional or nondirectional.

A within-subjects design or matching in a two-group experiment requires a different statistical procedure, the t test for matched groups. The same family of t distributions that apply to the independent groups procedures are used for this test. However, t for matched or within-subjects data is computed by looking at the differences between each pair of responses. Using the matched groups procedures to compute t reduces the estimate of variability in the data and results in a more precise measure of the effect of the independent variable. However, given the same total number of subjects, the experiment with a matching procedure has roughly half the degrees of freedom as one carried out with independent groups. If the matching was on a variable not highly correlated with the dependent variable, the chances of detecting the effect of the independent variable are reduced.

dy Questions

1. What are the four basic stages of analyzing the results of an experiment?
2. What are descriptive statistics? Why do we need them in an experiment?
3. Define each of the following and explain what each tells us about a set of data:
 - a. The mean
 - b. The range
 - c. The variance
 - d. The standard deviation
4. Two families have the same average income per person. Does this mean that each person in both families earns the same amount of money? Why or why not?
5. Here are two distributions of scores on a memory test. Find the mean, range, variance, and standard deviation of each group.

Group 1	Group 2
5	3
6	1
8	3
3	2
1	5

6. What five basic questions do we have to answer before we can select the appropriate statistical test for an experiment?
7. What is a t test? When is it used?
8. Briefly outline the difference between the t test for independent groups and the t test for matched groups.
9. Our computed value of t is more extreme than the table value, or critical value. What does that mean? Do we accept or reject the null hypothesis?
10. Our computed value of t is less extreme than the table value. What does that mean? Do we accept or reject the null hypothesis?
11. Our computed value of t is 3.28. Our critical value of t is ± 2.048 . We have 28 degrees of freedom and we are using a two-tailed (nondirectional) test. Draw a simple figure to illustrate the relationship between the critical and the computed values of t for this result.
12. Poor Jack is getting more and more confused. He says, "Anyone can see that my group means are different. Why do I have to go through all the trouble of making all these computations of t ?" Can you explain to him why these procedures are necessary? What advantage do they have over simply doing the "eye test"?
13. Our computed value of t is -1.07 . We have made a directional prediction and our critical value is -1.734 . Make a rough illustration of the relationship between the computed and table values of t in this case. Is there a significant difference between the treatment means?
14. A researcher has studied subjects' ability to learn to translate words into Morse code. He has experimented with two treatment conditions: In one condition, the subjects are given massed practice. They spend 8 full hours working on the task. In the other condition, subjects are given distributed practice. They also spend 8 hours practicing, but their practice time is spread out over 4 days; they practice 2 hours each day. After the subjects have completed their practice, they are given a test message to encode. The dependent variable is the number of errors made in encoding the test message. Since intelligence may affect the learning of this new skill, the researcher has matched the subjects on that variable. The results are given below. Decide what statistical test would be appropriate for these data, carry out the test, and evaluate the outcome. Assume that the researcher has chosen a .01 level of significance and that the direction of the outcome has not been predicted.

Massed Practice		Distributed Practice	
S_1	5	S_1	6
S_2	3	S_2	4
S_3	2	S_3	3
S_4	2	S_4	5
S_5	3	S_5	2

15. Assume that the Morse code researcher did not match the subjects.
 - a. What statistical test would be appropriate? Carry out that test and evaluate the outcome for $p < .01$ and a nondirectional prediction.
 - b. Follow the same procedure as in (a), but assume that the researcher has now predicted that the massed practice group will make *more* errors.
16. Alice has decided that the procedures for finding t for a matched groups design are a little easier to do, so she will just make sure she can match her subjects on some variable. That way she can save a little time on the computations. What is wrong with her approach? What is she forgetting?

-
- CP/M software finder* (1983). Indianapolis, IN: QUE.
- Davies, O. (Ed.). (1984). *Omni complete catalog of computer software*. New York: Collier Books, Macmillan.
- Huff, D. (1954). *How to lie with statistics*. New York: Norton.
- Lewis, D. (1960). *Quantitative methods in psychology*. New York: McGraw-Hill.
- McCall, R. B. (1980). *Fundamental statistics for psychology* (3rd ed.). New York: Harcourt Brace Jovanovich.
- Reichmann, W. J. (1961). *Use and abuse of statistics*. London: Methuen.
- Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.
- Yin, R. (1969). Perception of inverted faces. *Journal of Experimental Psychology*, *81*, 141–145.

CHAPTER 13

Analyzing Results: Multiple-Group and Factorial Experiments

Analysis of variance
 Sources of variability
 A one-way analysis of variance
 Within-groups variance
 Between-groups variance
 Computing and evaluating the F ratio
 Analyzing data from a factorial experiment
 A two-way analysis of variance
 Assumptions behind the two-way analysis of variance
 Evaluating the F ratios
 Summary
 Review and study questions
 References

So far we have covered some of the techniques for data from experiments with only two groups. The *t* tests for matched and independent groups are used when we have interval or ratio data in two-group experiments. However, very often we need to test more than two levels of an independent variable. We may need three or more groups to give an adequate idea of the way that variable operates. We may even want to study more than one independent variable at a time. For those experiments, we need other kinds of statistical procedures.

In this chapter we will look at procedures that can be used for interval or ratio data from multiple-group and factorial experiments. These procedures fall under the general heading of analysis of variance. By the end of this chapter you will know how these procedures work and why they are needed; you will also be ready to carry them out. We will begin by taking a general look at analysis of variance.

ance

f variance

The **analysis of variance (ANOVA)** is a statistical procedure used to evaluate differences among two or more treatment means. It is used with interval or ratio data.¹ The name reflects the basic nature of the test—the variance in the data is analyzed into component parts, which are then compared and evaluated for statistical significance. Treatment means are not compared directly. In Chapter 12,

¹ Other forms of the basic analysis of variance can be used for noninterval data. However, in this chapter we will talk only about ANOVA in its most common form. We will also assume that we have different subjects in each of our treatment groups.

the *t* test was used to evaluate the data from a two-group experiment. You may be wondering why we need another procedure at all. After all, a multiple-group experiment is just a continuation of a two-group design. We could use *t* tests to compare all the treatment means. We could analyze one pair, and then the next, and just keep doing that until we did all the pairs. But computing several *t* tests for each experiment is very bothersome. With five treatment levels, you would need ten different *t* tests to account for all possible pairs of means.

There is also a more serious problem. The more *t* tests in one experiment, the more apt you are to make a Type 1 error. Remember that when you do a single test, the odds of rejecting the null hypothesis by mistake are equal to your significance level (for example, 5 percent). Doing many *t* tests in the same experiment distorts those odds and increases the possibility that you will reject a null hypothesis that is true. Although it would be inappropriate to use several *t* tests in a multiple-group experiment, many of the principles of statistical analysis in the last two chapters still apply. We are still testing a null hypothesis. From the samples we have tested, we draw inferences about the population. We also use distribution curves to evaluate the results according to the significance levels we have chosen.

Still, there are many differences. The analysis of variance does not work like a *t* test. When we computed *t*, we calculated differences between treatment groups—differences between treatment means for the independent groups design and differences between pairs of scores in the matched groups design. We looked at those differences in relation to our estimates of the amount of variability in the populations sampled. The analysis of variance enables us to test the null hypothesis in a slightly different way. It breaks up the variability in the data into component parts. Each part represents variability produced by a different combination of factors in the experiment.

In the simplest analysis of variance, all the variability in the data can be divided into two parts: within-groups variability and between-groups variability. **Within-groups variability** is the degree to which the scores of subjects in the same treatment group differ from one another (that is, how much subjects vary from others in the group). **Between-groups variability** is the degree to which different treatment groups differ from one another (that is, how much difference there is between the scores of subjects under different levels of the independent variable). The proportions of the within-groups and between-groups variability differ from one experiment to another. Sometimes between-groups variability is relatively large; sometimes the two parts are about the same. Their relative proportions vary depending on the impact of the independent variable. When we carry out the analysis of variance, we are actually evaluating the likelihood that the proportions we observe could occur by chance. To understand the way this process works, we need to look more closely at the sources of variability that produce these components.

within-groups variability

between-groups variability

Ideally, when we run an experiment we would like to be able to show that the pattern of data obtained was caused by the experimental manipulation. However, you already know that if we observe changes in the dependent variable across treatment conditions, those changes may not be entirely due to the effects of the independent variable. But what else accounts for changes in the dependent variable? What else might produce variability in the scores of subjects across treatment conditions?

One common source of variability is individual differences. Whether we test children or chimps, we find that some do better than others. Within each treatment group, subjects' scores will differ from one another because subjects are different from one another. We use random assignment or matching in each experiment so that these differences do not confound the results of the experiment. We do not want differences between groups to be produced solely by extraneous subject variables. However, no two groups will be identical in every respect, so individual differences may lead to variability between groups as well as within the same group.

There are other sources of variability in data. Some differences between scores will be due to things we did not handle well in the experiment. For instance, we may have made small mistakes in measuring lines that subjects drew or in timing their answers. Extraneous variables of all kinds can produce more variability. They may cause changes in subjects' behavior that we may not detect. One subject is tested when the room is cool and so does a little better than the others. Like individual differences, these factors can lead to variability within the same group of subjects, as well as between different treatment groups. We can lump all these factors together in a single category called **error**: Individual differences, undetected mistakes in recording data, and a host of extraneous variables are all aspects of error that produce variability in subjects' data both within and between treatment groups.

error

Another major source of variability in the data is the experimental manipulation. We test subjects under different treatment conditions (that is, various levels of an independent variable). We predict that these conditions will alter subjects' behavior; we expect subjects under different treatment conditions to behave differently from one another. In other words, we expect our treatment conditions to create variability among the responses of subjects who are tested under different levels of an independent variable.

But the experimental manipulation does not operate like other sources of variability in the experiment. Error leads to variability between different treatment groups; it also produces variability within the same group. Unlike those sources of variability, treatment conditions produce variability only between the responses of different treatment groups. Subjects within the same treatment group are all treated in the same manner. Their scores may differ because of

individual differences or error but not because they were exposed to different levels of the independent variable: Subjects in the same treatment group all receive the *same* level of the independent variable.

When we do an analysis of variance, we break the variability in our data into parts that reflect the sources of variability in the experiment: within-groups variability and between-groups variability. *Within-groups variability is the extent to which subjects' scores differ from one another under the same treatment conditions.* The factors that we call error explain the variability that we see within groups. *Between-groups variability is the extent to which group performance differs from one treatment condition to another.* Between-groups variability is made up of error and the effects of the independent variable. These components are summarized in Table 13-1.

We can evaluate the effect of the independent variable by comparing the relative size of these components of variability. The logic behind this is straightforward. The variability within groups comes from error and nothing else. The variability between groups comes from both error and treatment effects. If the independent variable had an effect, the between-groups variability should be larger than the within-groups variability. We compare the relative sizes of these components by computing a ratio between them called the **F ratio**. Conceptually, it looks like this:

F ratio

$$F = \frac{\text{Variability from treatment effects} + \text{error}}{\text{Variability from error}}$$

or

$$F = \frac{\text{Variability between groups}}{\text{Variability within groups}}$$

Theoretically, if the independent variable had no effect, the *F* ratio should equal 1. There should be just as much variability within groups as there is between them: The same sources of variability would be operating both within and between treatments. However, the larger the effect of the independent variable, the larger the *F* ratio should be: The independent variable will lead to greater differences between the scores of subjects who receive different levels of

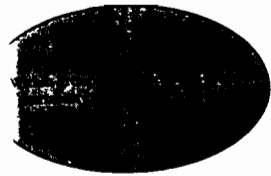
Table 13-1 Sources of variability in an experiment with one independent variable

Variability within Groups	Variability between Groups
Error	Error
Individual differences	Individual differences
Extraneous variables	Extraneous variables
	Treatment Effects

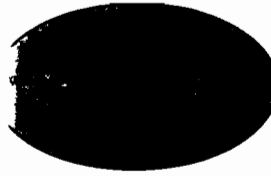
the independent variable. Figure 13-1 represents both possibilities graphically.

We use the distribution of F to evaluate the significance of the F ratio that we compute. F , like t , is actually a whole family of distributions. The shape of the distribution changes as the size of the samples changes. We will use the degrees of freedom to choose the right distribution and critical value for each experiment. If the F ratio is statistically significant, the amount of between-groups variability is large compared to the amount of within-groups variability. It is so large, it is unlikely that all the group means belong to the same population. If F is significant, we reject the null hypothesis that all the treatment means were drawn from the same population: We confirm the existence of differences across the groups that were probably produced by the independent variable.

Now that you have a general idea of how ANOVA works, let us turn to an example of a multiple-group experiment. We will look at some hypothetical data obtained from three groups of subjects who learned high-, medium-, and low-frequency words. We will proceed step by step through the computation of the F ratio. We will test F with a $p < .05$ significance level. Since this experiment has only one independent variable, the statistical test we do is called a one-way analysis of variance.



No treatment effects. Within- and between-groups variability are about equal.



Large treatment effects. Within-groups variability is small relative to between-groups variability.

Figure 13-1 The components of variability in an experiment with two possible outcomes.

4 One-Way Analysis of Variance

Assume that we have done a very simple experiment on learning. We have made up three lists of words that have varying frequencies in English. We know that frequency is likely to affect subjects' ability to learn lists of words: Words with greater frequencies are apt to be easier to learn because they are more familiar and also more meaningful to us. We have made up lists of high-, medium-, and low-frequency two-syllable nouns. We defined these categories on the basis of Howes's 1966 count of the frequency of words in spoken English. Our high-frequency words occurred at least 50 times in Howes's sample of 250,000 words. Our medium-frequency words occurred more than once but less than 50 times. The low-frequency words occurred only once in Howes's sample. Words such as "hundred," "mother," and "question" are high-frequency words by these criteria. "Coffee" and "paper" are medium-frequency words. "Turtle" and "textbook" are low-frequency words. Our hypothesis is that a list of words that has higher frequency in the spoken language will be recalled with greater accuracy than a list that has lower frequency. We ran the experiment using a between-subjects design: Our subjects were randomly assigned to three different treatment groups, one for each of the three word frequencies. The procedures for presenting the lists and testing recall were identical for all the groups.

Table 13-2 Hypothetical data from an experiment on word frequency and learning

	Group 1 (Low Frequency)	Group 2 (Medium Frequency)	Group 3 (High Frequency)
S_1	2	1	3
S_2	2	3	4
S_3	1	3	2
S_4	0	3	3
S_5	1	3	4
	$\bar{X}_1 = 1.2$	$\bar{X}_2 = 2.6$	$\bar{X}_3 = 3.2$
	$s_1^2 = .7$	$s_2^2 = .8$	$s_3^2 = .7$

Note. Scores represent the number of words recalled from a list. All subjects saw the same number of items.

This is a three-group experiment. It has one independent variable—the frequency of the words on each list. The dependent variable is the number of words subjects are able to recall, a ratio measure. As in any experiment, we will test the null hypothesis: The means of the three groups were sampled from the same population. Let us use a significance level of $p < .05$ for the data. Table 12-4 indicates that the data from this experiment may be analyzed by using a one-way analysis of variance. Table 13-2 shows some hypothetical data. As you can see, there are three groups. Their scores represent the number of words they recalled from the list they were shown. We have already begun our analysis of the data by organizing and summarizing it in table form.

Certain assumptions about our data must be met if we are to use analysis of variance procedures appropriately. First, the procedures we will use here require that treatment groups are independent from each other and that observations are sampled at random. They also require that the populations from which the groups are sampled are normally distributed on the dependent variable and that the variances of those populations are roughly equal, or homogeneous. However, the F test is relatively robust. If we have fairly large groups of subjects, the assumptions can be violated without serious errors. The computations shown here are for illustration only. We are looking at very few subjects so that the procedures will be clear. In practice, it would be better to have larger treatment groups because of the assumptions of the ANOVA procedures.

Within-Groups Variance

In order to compute an F ratio for these data, we need two pieces of information: the within-groups variability and the between-groups variability. We begin with the procedures for finding within-groups variability because they are a little simpler.

Think about the definition of within-groups variability—the extent to which subjects' scores vary from the scores of other subjects

within the same group. If we had only one group, we could measure its variability by computing the variance. We would use the formula

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{N - 1}$$

We would begin by finding the sum of squared deviations from the group mean. For each score in the group, we would calculate the difference between that score and the group mean ($X - \bar{X}$). We would square each of those differences ($(X - \bar{X})^2$) then add them together (Σ). To get a good estimate of the amount of variability in the population sampled, we would find the variance by dividing the sum of the squared deviations by $N - 1$, the degrees of freedom.

The variance we compute for one group is the average squared deviation from the mean of that group. Of course, when we do an analysis of variance, we are working with several groups at the same time. We need to get an estimate of within-groups variability. That estimate must pool, or combine, the variability in all treatment groups. We do that in two stages that are exactly parallel to the way we find the variance of a single group. First, we compute the **sum of squares (SS)**. The sum of squares is just a shorthand way of talking about the sum of squared deviations from the group mean. To get the sum of squares for within-groups variability, we compute the squared deviation of *each* score from *its* group mean.² Then, we simply add them all together.

We can summarize all those steps with this formula:

$$SS_w = \Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_2 - \bar{X}_2)^2 + \dots + \Sigma(X_p - \bar{X}_p)^2$$

The letter p stands for the number of groups in our experiment. In our example we have only three groups, but the analysis of variance procedures can be used with any number of groups. We simply keep adding up the squared deviations of scores from their group means until we have accounted for all the scores in all the groups.

Once we have the sum of squares within groups, we are ready to find the within-groups variance. With only one group, you know that we get a good estimate of the variance in the population if we divide our sum of squared deviations by $N - 1$. $N - 1$ is actually the degrees of freedom for one group. When we compute within-groups variance for several groups, we also need to divide by the degrees of freedom. For one group, the degrees of freedom are $N - 1$. For more than one group, the degrees of freedom are $N - p$. N is the total number of scores; p is the number of groups. We divide

² The formulas throughout this chapter are definitional formulas—direct statements of the operations they define. They are presented to clarify the logic behind the ANOVA procedures. Computational formulas are shown in Appendix A. They are derived from the definitional formulas and, although computational formulas are sometimes easier to use, the rationale behind them is less clear.

through by that number (df_w) to get within-groups variance. The W stands for within groups.

Note that although we are calculating the variance within groups, analysis of variance has its own peculiar terminology. You have learned sum of squares, which is an understandable abbreviation. However, when we finally obtain what we would otherwise call a variance estimate, we change terms rather abruptly. Dividing the SS_w by df_w gives us the mean square. This is actually an abbreviation too, although its origin is not as clear. The mean is an average. The variance is the average squared deviation from the mean. So the **mean square (MS)** is an average squared deviation. The existence of a plot to confuse students on this point has been suggested many times, but it has never been confirmed. The important point is this: We are still talking about variance in the data. The mean square within groups (MS_w) is one estimate of the amount of variability in the population sampled. It represents one portion of the variability in the data, the portion produced by the combination of sources that we call error. Table 13-3 (p. 274) shows how to compute SS_w and MS_w for our three-group example.

mean square (MS)

Between-Groups Variance

Once we have MS_w , we are ready to calculate the second component of the F ratio, a measure of the variability between groups of subjects. It is a measure that reflects both error and treatment effects: Between-groups variability is the extent to which group performance differs from one treatment condition to another. Let us look a little more closely at the implications of that definition.

If the independent variable had no effect in this experiment, the subjects in all the groups would have done about equally well. We would not expect to see any dramatic difference in recall from one group to another; the only differences we would see would be due to error. If that were the case, the means of the individual treatment groups would all be about the same. We could compute one overall mean, or **grand mean**, an average of all the treatment means. If our independent variable had no effect, the grand mean would describe the data about as well as three separate means, one for each of the three separate groups. But imagine what would happen if the independent variable really did have an effect on subjects' recall. We could still compute an overall grand mean that would represent the average of all the subjects' scores. However, the means of the individual groups would be quite different from the grand mean. They would also be quite different from one another.

grand mean

We can measure the amount of variability within groups by finding the total variance of scores from the individual group means. Similarly, we can measure the variability between groups by finding the variance of the group means from their mean, the grand mean of the experiment. Now that you are familiar with the logic behind the procedures, let us compute the between-groups variance. The pro-

Table 13-3 Computing within-groups variance for a three-group example

<i>Step 1.</i> Compute the deviation of each score from its group mean.	Group 1 (low frequency)	$(X_1 - \bar{X}_1)$	$(X_1 - \bar{X}_1)^2$
	S_1 2	.8	.64
	S_2 2	.8	.64
	S_3 1	-.2	.04
	S_4 0	-1.2	1.44
	S_5 1	-.2	.04
	$\bar{X}_1 = 1.2$		$\Sigma(X_1 - \bar{X}_1)^2 = 2.80$
<i>Step 2.</i> Square the deviation of each score from its group mean.	Group 2 (medium frequency)	$(X_2 - \bar{X}_2)$	$(X_2 - \bar{X}_2)^2$
	S_1 1	-1.6	2.56
	S_2 3	.4	.16
	S_3 3	.4	.16
	S_4 3	.4	.16
	S_5 3	.4	.16
	$\bar{X}_2 = 2.6$		$\Sigma(X_2 - \bar{X}_2)^2 = 3.20$
<i>Step 3.</i> Total the squared deviation scores for each group.	Group 3 (high frequency)	$(X_3 - \bar{X}_3)$	$(X_3 - \bar{X}_3)^2$
	S_1 3	-.2	.04
	S_2 4	.8	.64
	S_3 2	-1.2	1.44
	S_4 3	-.2	.04
	S_5 4	.8	.64
	$\bar{X}_3 = 3.2$		$\Sigma(X_3 - \bar{X}_3)^2 = 2.80$
<i>Step 4.</i> Add all the group totals together to find SS_w .	$SS_w = \Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_2 - \bar{X}_2)^2 + \Sigma(X_3 - \bar{X}_3)^2 = 8.80$		
<i>Step 5.</i> Find df_w .	$df_w = N - p$	$N =$ Number of scores	
	$df_w = 15 - 3$	$p =$ Number of groups	
	$df_w = 12$		
<i>Step 6.</i> Find MS_w .	$MS_w = \frac{SS_w}{df_w}$		
	$MS_w = \frac{8.80}{12}$		
	$MS_w = .73$		

Note. The same procedures apply when the groups are unequal in size.

cess is carried out in Table 13-4 (p. 276–277) for our three-group example. We begin by computing the grand mean, the average of all the treatment means. We then compute deviations of the group means from the grand mean, and next we obtain the squared deviations. Notice, however, that the formula for SS_B (B

stands for between groups) is a little different from the one for SS_w . Instead of simply adding together all the squared deviations, each one is first multiplied by n_j , the number of subjects in each respective treatment group.

Next, we find the variance about the grand mean, the mean square between groups (MS_B). To get MS_B , we divide SS_B by its degrees of freedom. We are now working with group means rather than individual subjects' scores. Hence, our degrees of freedom for SS_B (df_B) are equal to $p - 1$, where p is the number of groups. The MS_B gives us a second estimate of the amount of variability in the population. MS_B reflects the amount of variability produced by error *and* treatment effects in the experiment.

Computing and Evaluating the F Ratio

We now have both components of variability that we need to compute our F ratio. The F ratio represents this relationship:

$$F = \frac{\text{Variability from treatment effects} + \text{error}}{\text{Variability from error}}$$

We have transformed the components of this formula into the numerical terms MS_B and MS_w . Thus, the statistical form of the F ratio is as follows:

$$F = \frac{MS_B}{MS_w}$$

If we substitute our computed values into this formula, we find that for our three-group example,

$$F = \frac{5.28}{.73} \text{ or } 7.23$$

To test our F ratio for significance, we need to find the critical value. As you know, F is a whole family of distributions. We use our degrees of freedom to locate the appropriate distribution. But is there a problem? As we computed F , we actually calculated two different degrees of freedom, one to get MS_B and another to get MS_w . Which do we use? Since the F test can be used with any number of groups as well as any number of subjects, we need both. The F distribution changes as the size of treatment groups changes; it also changes as the number of treatment conditions changes.

If you look in Appendix B, you will find that Table B3 lists critical values of F . The table is organized by the degrees of freedom. The values listed across the top refer to the degrees of freedom of the numerator, or top, of the F ratio—here, df_B . Values listed vertically down the side of the table indicate the degrees of freedom of the

Table 13-4 Finding the between-groups variance for our three-groups example

	Group 1 (low frequency)	Group 2 (medium frequency)	Group 3 (high frequency)
	$\bar{X}_1 = 1.2$	$\bar{X}_2 = 2.6$	$\bar{X}_3 = 3.2$
<i>Step 1.</i> Compute the grand mean, the mean of all the group means.			
			<div style="border: 1px solid black; padding: 5px; width: fit-content; margin: auto;"> Grand mean (\bar{X}_G) $\bar{X}_G = \frac{\sum \bar{X}}{p}$ $\bar{X}_G = \frac{1.2 + 2.6 + 3.2}{3}$ $\bar{X}_G = \frac{7}{3}$ $\bar{X}_G = 2.3$ </div>
<i>Step 2.</i> Compute the differences between each group and the grand mean.	$\bar{X}_1 - \bar{X}_G =$ $1.2 - 2.3 = -1.1$	$\bar{X}_2 - \bar{X}_G =$ $2.6 - 2.3 = .3$	$\bar{X}_3 - \bar{X}_G =$ $3.2 - 2.3 = .9$
<i>Step 3.</i> Put these differences in the SS_B formula; n is the number of subjects in each group; p is the number of groups—this general formula can handle any number of groups.	<div style="border: 1px solid black; padding: 5px; width: fit-content; margin: auto;"> $SS_B = n_1(\bar{X}_1 - \bar{X}_G)^2 + n_2(\bar{X}_2 - \bar{X}_G)^2 + n_3(\bar{X}_3 - \bar{X}_G)^2 \dots n_p(\bar{X}_p - \bar{X}_G)^2$ $SS_B = 5(-1.1)^2 + 5(.3)^2 + 5(.9)^2$ </div>		

Step 4. Square all deviations from the grand mean.

$$SS_B = 5(1.21) + 5(.09) + 5(.81)$$

Step 5. Carry out all multiplications.

$$SS_B = 6.05 + .45 + 4.05$$

Step 6. Obtain the total SS_B .

$$SS_B = 10.55$$

Step 7. Calculate the degrees of freedom; p is the number of groups.

$$df_B = p - 1$$

$$df_B = 3 - 1 \text{ or } 2$$

Step 8. Divide SS_B by df_B to find the mean square between groups, the second estimate of population variance.

$$MS_B = \frac{SS_B}{df_B}$$

$$MS_B = \frac{10.55}{2}$$

$$MS_B = 5.28$$

Note: At this point you can check your work by computing SS_T , which represents the total sum of squares for the data. Since we are simply dividing the variability into two components, $SS_B + SS_W$ should equal SS_T . You can compute SS_T with this formula: $SS_T = \sum(X^2) - \frac{(\sum X)^2}{N}$. N is the number of scores. For this example,

$$SS_T = 101 - \frac{(35)^2}{15}$$

$$SS_T = 101 - \left(\frac{1225}{15}\right)$$

$$SS_T = 101 - 81.67$$

$$SS_T = 19.33$$

Check:

$$SS_T = SS_B + SS_W$$

$$19.33 = 10.55 + 8.80$$

(The small discrepancy is due to rounding error.)

denominator of the F ratio—here, df_w . To find the appropriate critical value, first, locate df_b along the top of the table, then, locate df_w along the side. Now, find the place in the table where those two lines meet. We are looking for $df_b = 2$ and $df_w = 12$. (These are simply the df values we computed to get mean squares.) If we look at a portion of the table, we see this:

	df numerator						
	1	2	3	4	5	6	...
df denominator	1						
	2						
	3						
	...						
	12	3.88	6.93				

The value in light type, 3.88, is the critical value of F at the .05 level; **6.93**, shown in bold face, is the critical value of F at the .01 level. Remember, these values apply only to an F test with 2 and 12 degrees of freedom. We have to look up the critical values for each experiment. Figure 13-2 illustrates the distribution of F with 2 and 12 degrees of freedom. It also shows the distribution of F with 2 and 6 degrees of freedom. As you can see, the critical values change dramatically as the degrees of freedom change.

We chose a significance level of $p < .05$ for our three-group experiment. Our computed value of F was 7.23. The table value of F is 3.88 at the .05 level: Therefore, our computed F is significant. We reject the null hypothesis that the treatment means came from the same population. Our computed F is actually also significant at the .01 level: It is greater than 6.93, the critical value of F at the .01 level.

Preparing a summary table. By now it may seem that we have gone through a thousand steps to evaluate the data from this study. The count is actually slightly less than that, but you can understand why we need to prepare a simple, comprehensive summary of the findings. We would not present all the steps and calculations in an actual report. Instead, we summarize our computations in a summary table. The summary table for our example is shown in Table

Table 13-5 Analysis of variance summary table

Source	df	SS	MS	F
Between groups	2	10.55	5.28	$\frac{MS_b}{MS_w} = 7.23^*$
Within groups	12	8.80	.73	
Total	14	19.35		

* $p < .01$

13-5. The table includes all the basic information needed to compute F , along with the actual computed value. However, we do not include the table values of F . Since we have given the degrees of freedom, readers can always consult their own tables to get the critical value if they need it. The format of the table is used by convention, and you should follow it exactly; list between-groups variance first, and so on.

Graphing the results. Another useful way of summarizing the results of an experiment is graphing. We can transform our findings into a picture that shows the reader the overall results at a glance. Look closely at Figure 13-3, which presents the results of our experiment as a graph.

The figure illustrates several general points you should keep in mind. Notice that the figure is well proportioned; the vertical axis is roughly three-fourths the size of the horizontal axis. Other proportions are not as pleasing to look at. Notice also that the independent variable is plotted on the horizontal axis; the dependent variable is plotted on the vertical axis. Finally, note that the data points represent group means. We usually do not graph the data of individual subjects unless we have a small N design. Of course, the axes are labeled clearly so that readers will know exactly what the figure represents.

Interpreting the results. We know that the computed F was significant for these data. We will have more to say about interpreting the meaning of that outcome in the next chapter. However, there are some points about the F test that should be made before we go further.

From the graph of our results (Figure 13-3), you can see quite clearly that subjects in different groups performed differently from one another in this experiment. Subjects in the high-frequency group recalled the most items; subjects in the low-frequency group recalled the least. You can also see clearly from the figure that the variation between different groups was not uniform. The high- and low-frequency groups differed more from each other than from the medium group. When we compute F , we test only the overall pattern of treatment means. Our significant F in this example tells us that across all the group means, there is a significant difference. Since the F test does not test the differences between each pair of means, we

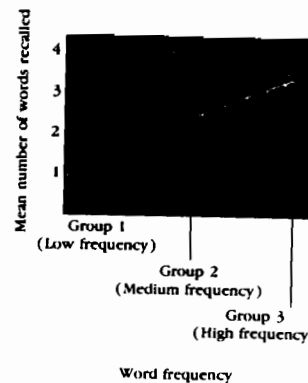


Figure 13-3 Mean number of words recalled as a function of word frequency.

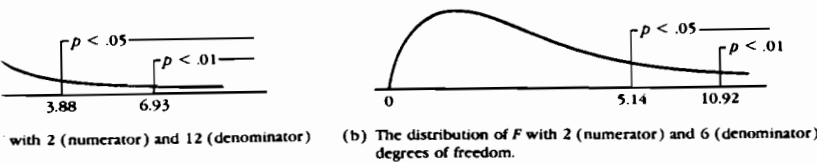


Figure 13-2 The distribution of F with varying degrees of freedom.

do not know exactly where the difference is. From the graph, it seems likely that the high and the low groups are significantly different from each other. After all, they differ the most. However, the difference between, say, the low- and the medium-frequency groups may or may not be significant.

When we need to pinpoint the exact source of the differences across several treatment groups, we need to go beyond the *F* test. There are several additional tests that are called **post hoc tests**, tests done after the overall analysis indicates a significant difference. We will not go into the details of these tests here, but some of the names you will see are the Tukey and the Scheffé tests. These tests have essentially the same function: They can be used to make pair-by-pair comparisons to pinpoint the source of a significant difference across several treatments.

Let us briefly summarize what we have accomplished in this chapter so far. We took a multiple-group experiment and selected a suitable statistical test on the basis of three dimensions: We considered the number of independent variables, the number of treatment groups, and the level of measurement in making our choice. The experiment had one independent variable (word frequency) and three treatment groups (low, medium, and high frequency). The dependent variable (number of words recalled) was measured by a ratio scale. The data therefore required a one-way analysis of variance, or *F* test, which we carried out and evaluated. We also prepared a summary table of our analysis and graphed the group means.

The basic principles of the analysis of variance apply in many multiple-group experiments. However, those principles can also be extended to handle more complex research designs. We will carry them further in our next example, an experiment with a factorial design.

post hoc tests

ata from a Factorial Experiment

Factorial experiments are designed to look at the effects of more than one independent variable at a time. They also enable us to look at the interaction between variables. The impact of one independent variable may differ depending on the values of the other independent variables in the experiment. When we analyze the data from a factorial experiment, we evaluate both kinds of effects. We look at the impact of each independent variable; we assess whether there is a main effect of each independent variable. We also measure the size of any interaction between the variables. Let us look at an example of a simple factorial experiment and see what statistical procedures are used to accomplish these goals.

Assume we have set up and run another experiment to explore the relationship between word frequency and recall. Half the subjects saw high-frequency words, and half saw low-frequency words. This time, though, we have run a factorial experiment. In addition to evaluating the effect of frequency, we have manipulated our testing

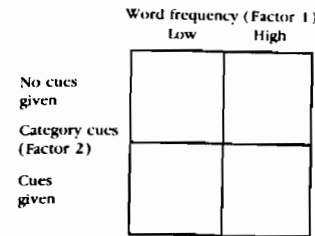


Figure 13-4 Diagram of 2 × 2 factorial experiment on the effects of word frequency and cuing on recall.

procedures in a 2 × 2 design. The design is diagrammed in Figure 13-4. Half the subjects have been asked simply to recall the words they saw on the original list. The other half have been given cues to aid them in remembering the words they saw. For instance, suppose subjects saw the word “camel” on the original list. If they were in the “no cue” condition, they were simply asked to recall the word. If they were in the “cued” condition, we provided the name of the category the word belongs to—animal. Category cues were given for all words on the list.

Our hypothesis, which is based on prior research (such as Tulving and Pearlstone, 1966), is that cuing will enhance recall. Frequency will also affect recall, with the more frequent words being easier to recall. We have two independent variables in this experiment—word frequency and category cues. Our dependent variable is the number of words correctly recalled from each list, a ratio measure. We will use $p < .05$ as our significance level. If you consult Table 12-4, you will find that the statistical test indicated for these data is a two-way analysis of variance.

You already know how to do the basic, or one-way, ANOVA. The same principles apply to all ANOVA procedures. But when we have a factorial design, additional complexities arise. The procedures for the one-way ANOVA are not designed to give us as much information as we want to get from a factorial experiment. We want to be able to evaluate the effect of each independent variable: We need to know whether the word frequency, as well as the category cues, affects ability to recall words. Of course, we also want to know whether there was any interaction between the two variables. We want to assess whether the effects of using different frequencies might differ depending on whether or not cues are given—or perhaps the effect of cues varies depending on whether the word to be recalled is relatively frequent or infrequent.

To answer all these questions with an analysis of variance, we need to break down the variance in the data into more components than we had before. In the one-way ANOVA, we had one independent variable. We divided all the variability in the data into just two parts: within-groups and between-groups variability. Within-groups variability is created by all those sources of error in the experiment: individual differences, the experimenter’s mistakes, and other extraneous variables. Between-groups variability is created by all those sources of error plus the effect of the independent variable.

The same is true in a factorial experiment. We can separate variability into within-groups and between-groups variance. However, the picture is more complex. Between-groups variability comes from error and treatment effects, but there are several sources of treatment effects in the factorial experiment. Every independent variable may produce its own unique treatment effects; each can produce a portion of the between-groups variability or a main effect. The interaction of the independent variables can produce another portion. This is represented graphically in Figure 13-5, which compares the

Table 13-6 Summarizing the analysis of variance for a two-factor experiment

Source	df	SS	MS	$F = \frac{MS}{MS_e}$
Between groups				
Factor 1				
Factor 2				
Interaction 1 × 2				
Within groups				
Total				

components of variability for the one-way analysis of variance against a two-way analysis of variance for a two-factor experiment.

We can begin our analysis of variance for the factorial experiment by finding within- and between-groups variability. However, we will also need to break between-groups variability into its component parts: the variability associated with *each* of the independent variables and the variability associated with the *interaction* between them.

When we did the one-way ANOVA, we used a summary table to organize our computations. When we do a two-way ANOVA, it is helpful to plan the summary table in advance. We can use it to keep track of computations as we do them. The outline of the summary table for a two-factor experiment is shown in Table 13-6. You can see that all the sources of variability in the experiment are represented. We must find all these components in order to compute the *F* ratios needed to judge significance. We will calculate a different *F* ratio for every source of between-groups variability in the experiment and use those ratios to decide whether the effects of each independent variable are significant. We will also decide whether there is a significant interaction.

Table 13-7 Data from a two-factor experiment: The effects of word frequency and category cues on recall in a list-learning task

		Word Frequency (Factor 1)			
		Low	High		
Category Cues (Factor 2)	No cues given	2	4	$\bar{X}_1 = 3$	$\bar{X}_2 = 5$
		3	5		
		1	4		
		4	6		
		5	6		
	Cues given	4	7	$\bar{X}_3 = 6$	$\bar{X}_4 = 8$
		6	6		
		5	9		
		6	8		
		9	10		

Note: Scores represent number of words correctly recalled from a list.

A Two-Way Analysis of Variance

Table 13-7 presents some hypothetical data from our experiment to test the effect of word frequency and category cues on recall. Since this is a 2×2 factorial design, the data are divided into four treatment groups. We have already begun the data analysis by computing the mean number of words correctly recalled by each treatment group. Let us take an overall look at the steps required to complete the analysis of variance. The actual computations are shown in Tables 13-8 through 13-11, which follow our written description.

Assumptions behind the Two-Way Analysis of Variance

The procedures and formulas for a two-way ANOVA require the same basic assumptions as the one-way ANOVA procedures we examined earlier. They assume that the treatment groups are independent from each other and that the observations were randomly sampled. They also assume that the population from which each treatment group is sampled is normally distributed on the dependent variable. Finally, they assume that the variances of the populations are all about equal (homogeneous). The computations here are done just for the sake of illustration, so we have only five subjects per cell. However, the assumptions behind the ANOVA procedures are more likely to be met with larger groups of subjects. In addition, the procedures shown here assume *equal* number of subjects (*n*) in each group and more than one subject per group. If you have unequal *n*'s, you will need more complicated procedures. The same is true if you have used within-subjects procedures. In either case, consult your instructor or see Winer's *Statistical Principles in Experimental Design* (1971).

Finally, these procedures are set for fixed models, experiments in which the values of the independent variables are fixed by the experimenter. In other words, the experimenter chooses to run subjects at certain levels of each independent variable. In our example, the experimenter has chosen to use high and low word frequencies and two levels of the category cue variable—cues versus no cues. However, in experiments with random models (randomly selected values of the independent variables) or mixed models, different statistical procedures are required. Most experiments follow the fixed model, as we do here.

Step 1: Computing within-groups variance. We begin ANOVA by filling in the within-groups section of the summary table. We need the degrees of freedom (*df*), the sum of squares within groups (SS_w), and the mean square within groups (MS_w). To get them, we follow the same basic procedures that we used for the one-way ANOVA. The calculations for our 2×2 experiment are shown in Table 13-8 (p. 286). Remember that the mean square within groups

represents variability produced by individual differences, extraneous variables, and other sources of error in the experiment. We will use MS_w to evaluate the impact of the independent variables and their interaction in the experiment.

Step 2: Computing between-groups variability. We continue our ANOVA by finding the total sum of squares between groups, SS_B . We need the SS_B because it represents all the variability we have among treatment groups. To complete our ANOVA, we will have to divide the SS_B into its main components: the parts associated with each of the independent variables and the part associated with the interaction between them. Table 13-9 (p. 287) illustrates the procedures for finding SS_B for our factorial example.

Step 3: Computing main effects. As you know, the ANOVA procedures have some special terms associated with them. "Sum of squares" and "mean square" refer to variability in the data. When we want to discuss the variability associated with a single independent variable in a factorial design, we call it a **main effect**, the change in the dependent variable produced by the various levels of one independent variable. In our 2×2 example, we are looking for a main effect of word frequency. We are also looking for a main effect of category cues. The number of main effects to be tested in a factorial experiment is determined by the number of independent variables in the experiment.

When we carry out our ANOVA, we evaluate whether or not each main effect in the experiment is significant: We compute an F ratio to test the impact of each independent variable. When we test for a significant main effect, we are simply asking again whether subjects' scores on the dependent variable differ depending on the levels of one independent variable that we have manipulated. To measure the total between-groups variability, we calculated the deviation of group means around their grand mean. In effect, we asked how much individual treatment groups differed from the average of all the groups. When we measure a main effect, we want to look only at a particular portion of the total variability. We want to measure how much variability occurs between groups because of the impact of one independent variable.

We can ask a straightforward question: How much do the means of groups under different levels of one variable, say, word frequency, differ from the grand mean of all the groups? This is like the logic we followed in doing the one-way ANOVA: The larger the effect of the independent variable, the larger the differences from the grand mean. In our example, a large main effect of word frequency would mean that subjects' recall varied depending on whether the words to be learned were relatively common or uncommon. When we evaluate the main effect of one independent variable, we treat the data as if that variable is the only one in the experiment. We simply ignore all

the other experimental manipulations that were done: We say we collapse the data across the other conditions of the experiment. In effect, we pretend that those conditions did not exist. Table 13-10 (p. 288) shows how this is done as we compute SS_1 for our first independent variable, word frequency.

We also need to test for a main effect of the second variable. We know that this second variable may have contributed to the total variability between treatment groups. We can evaluate the main effect of the second variable by using the same basic procedures we followed to get the main effect of word frequency. In effect, we ask whether subjects' recall differed depending on whether they were given category cues or not: Did it differ regardless of whether they were shown high- or low-frequency words? We look at the effects of our second independent variable by simply disregarding the word-frequency manipulation. We collapse across the word-frequency conditions. Table 13-11 (p. 289) shows the procedures.

Step 4: Computing the interaction. The variability associated with the interaction of the two independent variables is simply what remains after the main effects of the independent variables have been taken into account. The variability between groups that is not explained by either independent variable may be explained by their interaction.

Since we have two independent variables, the SS_B must be divided into three parts: the variability associated with the first independent variable (SS_1); the variability associated with the second independent variable (SS_2); and the variability associated with the interaction of the two ($SS_{1 \times 2}$). Once we have computed the total SS_B , SS_1 , and SS_2 , the simplest way to find $SS_{1 \times 2}$ is by subtracting:

$$SS_{1 \times 2} = SS_B - SS_1 - SS_2$$

The sum of squares for the interaction is entered in the summary table, Table 13-12 (p. 289).

Step 5: Computing the F ratios. We have now completed nearly all the computations that we need to evaluate the results of our experiment. We summarize our calculations in a summary table, Table 13-12. The table is similar to the one we prepared for the simple ANOVA. The only difference is in the way we represent the sources of variability. Because we have two independent variables in this experiment, we have three sources of variability: Factor 1, Factor 2, and their interaction. The within-groups variability (MS_w) is used as the denominator of all three F ratios required to evaluate the significance of these sources. The three F ratios have been computed and are also shown in the summary table.

Table 13-8 Step 1: Computing within-groups variability (MS_w) for a 2×2 factorial experiment

		Word Frequency (Factor 1)					
		Low			High		
No cues given	X_1	$(X_1 - \bar{X}_1)$	$(X_1 - \bar{X}_1)^2$	X_2	$(X_2 - \bar{X}_2)$	$(X_2 - \bar{X}_2)^2$	
	2	-1	1	4	-1	1	
	3	0	0	5	0	0	
	4	-2	4	4	-1	1	
	5	1	1	6	1	1	
	$\bar{X}_1 = 3$		$\Sigma(X_1 - \bar{X}_1)^2 = 10$	$\bar{X}_2 = 5$		$\Sigma(X_2 - \bar{X}_2)^2 = 4$	
Cues given	X_3	$(X_3 - \bar{X}_3)$	$(X_3 - \bar{X}_3)^2$	X_4	$(X_4 - \bar{X}_4)$	$(X_4 - \bar{X}_4)^2$	
	4	-2	4	7	-1	1	
	6	0	0	6	-2	4	
	5	-1	1	9	1	1	
	6	0	0	8	0	0	
	$\bar{X}_3 = 6$		$\Sigma(X_3 - \bar{X}_3)^2 = 14$	$\bar{X}_4 = 8$		$\Sigma(X_4 - \bar{X}_4)^2 = 10$	

$$SS_w = \Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_2 - \bar{X}_2)^2 + \Sigma(X_3 - \bar{X}_3)^2 + \Sigma(X_4 - \bar{X}_4)^2 + \dots + \Sigma(X_{pq} - \bar{X}_{pq})^2$$

$$SS_w = 10 + 4 + 14 + 10$$

$$SS_w = 38$$

$$df_w = N - pq \quad N = \text{Number of scores}; pq = \text{Number of rows} \times \text{number of columns}$$

$$df_w = 20 - 4 \text{ or } 16$$

$$MS_w = \frac{SS_w}{df_w}$$

$$MS_w = \frac{38}{16}$$

$$MS_w = 2.38$$

Table 13-9 Step 2: Computing the between-groups variability (SS_b) for a 2×2 factorial experiment

	Group 1	Group 2	Group 3	Group 4	Grand Mean (\bar{X}_G)
	$\bar{X}_1 = 3$ $n_1 = 5$	$\bar{X}_2 = 5$ $n_2 = 5$	$\bar{X}_3 = 6$ $n_3 = 5$	$\bar{X}_4 = 8$ $n_4 = 5$	Total of all group means $\bar{X}_G = \frac{\text{Number of groups}}{3 + 5 + 6 + 8}$ $\bar{X}_G = \frac{22}{4}$ $\bar{X}_G = 5.5$
Step 1. Compute the grand mean, the mean of all the group means.					
Step 2. Compute the deviation of each group mean from the grand mean.	$\bar{X}_1 - \bar{X}_G = 3 - 5.5 \text{ or } -2.5$	$\bar{X}_2 - \bar{X}_G = 5 - 5.5 \text{ or } -.5$	$\bar{X}_3 - \bar{X}_G = 6 - 5.5 \text{ or } .5$	$\bar{X}_4 - \bar{X}_G = 8 - 5.5 \text{ or } 2.5$	
Step 3. Put the deviations in the SS_b formula; n is the number of subjects in each group.	$SS_b = n_1(\bar{X}_1 - \bar{X}_G)^2 = 5(-2.5)^2 = 31.25$	$SS_b = n_2(\bar{X}_2 - \bar{X}_G)^2 = 5(-.5)^2 = 1.25$	$SS_b = n_3(\bar{X}_3 - \bar{X}_G)^2 = 5(.5)^2 = 1.25$	$SS_b = n_4(\bar{X}_4 - \bar{X}_G)^2 = 5(2.5)^2 = 31.25$	
Step 4. Square all deviations from the grand mean.					
Step 5. Complete all computations.					

Step 1. Find the mean at each level of Factor 1: Ignore Factor 2 (rows) and find the mean of each column. N is the number of scores.

Step 2. Find the difference between each column mean and the grand mean.

Step 3. Put those differences in the SS_1 formula; n is the number of subjects in each group; p is the number of rows; q is the number of columns. This general formula will handle any number of columns.

Step 4. To get MS_1 , divide SS_1 by df_1 .

		Word Frequency (Factor 1)	
		Low	High
Category Cues (Factor 2)	No cues given	$\bar{X}_1 = 3$ ($N = 5$)	$\bar{X}_2 = 5$ ($N = 5$)
	Cues given	$\bar{X}_3 = 6$ ($N = 5$)	$\bar{X}_4 = 8$ ($N = 5$)
	Column means	$\bar{X}_{col1} = 4.5$	$\bar{X}_{col2} = 6.5$
	Column mean - Grand mean $\bar{X}_G = 5.5$	$\bar{X}_{col1} - \bar{X}_G = 4.5 - 5.5 = -1.0$	$\bar{X}_{col2} - \bar{X}_G = 6.5 - 5.5 = 1.0$

$$SS_1 = np \sum [(\bar{X}_{col1} - \bar{X}_G)^2 + (\bar{X}_{col2} - \bar{X}_G)^2 + \dots + (\bar{X}_{colq} - \bar{X}_G)^2]$$

$$SS_1 = 5(2) \sum [(-1.0)^2 + (1.0)^2]$$

$$SS_1 = 10[(1) + (1)]$$

$$SS_1 = 10(2)$$

$$SS_1 = 20$$

$$MS_1 = \frac{SS_1}{df_1} \quad df_1 = q - 1$$

$$df_1 = 2 - 1 = 1$$

$$MS_1 = \frac{20}{1}$$

$$MS_1 = 20$$

Table 13-11 Step 3: Finding the main effect for Factor 2 (category cue) in a 2 × 2 factorial experiment

Step 1. Find the mean at each level of Factor 2: Ignore Factor 1 (columns) and find the mean of each row.

Step 2. Find the difference between each row mean and the grand mean.

Step 3. Put those differences in the SS_2 formula; n is the number of subjects in each group; q is the number of columns; p is the number of rows. This general formula will handle any number of rows.

Step 4. To get MS_2 , divide SS_2 by df_2 .

		Word Frequency (Factor 1)		Row means	Row mean - grand mean
		Low	High		
Category Cues (Factor 2)	No cues given	$\bar{X}_1 = 3$	$\bar{X}_2 = 5$	$\bar{X}_{row1} = 4$	$\bar{X}_{row1} - \bar{X}_G = 4 - 5.5 = -1.5$
	Cues given	$\bar{X}_3 = 6$	$\bar{X}_4 = 8$	$\bar{X}_{row2} = 7$	$\bar{X}_{row2} - \bar{X}_G = 7 - 5.5 = 1.5$

$(\bar{X}_G = 5.5)$

$$SS_2 = np \sum [(\bar{X}_{row1} - \bar{X}_G)^2 + (\bar{X}_{row2} - \bar{X}_G)^2 + \dots + (\bar{X}_{rowp} - \bar{X}_G)^2]$$

$$SS_2 = 5(2)[(-1.5)^2 + (1.5)^2]$$

$$SS_2 = 10[2.25 + 2.25]$$

$$SS_2 = 10(4.50) \text{ or } 45$$

$$MS_2 = \frac{SS_2}{df_2} \quad df_2 = p - 1$$

$$df_2 = 2 - 1 = 1$$

$$MS_2 = \frac{45}{1}$$

$$MS_2 = 45$$

Table 13-12 Summary table: Analysis of variance for a 2 × 2 factorial experiment and computed F ratios (includes step 4)

Source	df	SS	MS	F
Between groups		65 ^a		
Factor 1 (word frequency)	$q - 1 = 1$	20	20	$F_1 = \frac{20}{2.38}$ or 8.40*
Factor 2 (category cues)	$p - 1 = 1$	45	45	$F_2 = \frac{45}{2.38}$ or 18.91**
Interaction 1 × 2	$(p - 1)(q - 1) = 1$	0 ^b	0	$F_{1 \times 2} = \frac{0}{2.38}$ or 0
Within groups	$N - pq = 16$	38	2.38	
Total	$N - 1 = 19$			

* $p < .05$

** $p < .01$

^a SS_B is usually not shown in published articles.

^bWe find the sum of squares for the interaction of our two variables by subtracting:

$$SS_{1 \times 2} = SS_B - SS_1 - SS_2$$

The $SS_{1 \times 2}$ represents all the between-groups variability that is not explained by the main effect of either independent variable. Its degrees of freedom depend on the degrees of freedom for the main effects. Here, there is no interaction.

To judge whether the computed F ratios are significant, we compare them to the table values of F . We get those values from our table of F values in Appendix B, Table B3. The procedures are the same as those used for the simple ANOVA. We locate the proper value of F by using the degrees of freedom of the F ratio. We look across the top of Table B3 to find the degrees of freedom that belong to the top, or numerator, of the F ratio. We look along the side of the table to find the degrees of freedom of the denominator of our F ratio. Each F ratio we compute has its own degrees of freedom. That means that each ratio has its own critical value or table value of F . When we evaluate each F ratio, we must be sure we are using the correct degrees of freedom and the correct critical value.

Practice finding the correct critical value by looking up the table values of F for the F ratios we have computed. The F ratio for Factor 1 (word frequency) has 1 degree of freedom for the numerator (MS_1); it has 16 degrees of freedom for the denominator (MS_w). The table value of $F(1, 16)$ is 4.49 at $p < .05$ and 8.53 at $p < .01$. Our computed value of F for Factor 1 is 8.40. Therefore, the effect of Factor 1 is significant at $p < .05$. Our computed value of F is more extreme than the table value. This means the main effect of word frequency is so large that it is probably not due to chance. Whether the lists contained high- or low-frequency words made a significant difference in subjects' recall. We reject the null hypothesis that the means of the high- and low-frequency groups were sampled from the same population.

The F ratio for Factor 2 (category cues) also has 1 degree of freedom for the numerator (MS_2); it has 16 degrees of freedom for the denominator (MS_w). We know that the table value of $F(1, 16)$ is 4.49 at $p < .05$ and 8.53 at $p < .01$. (They turn out to be the same as they were for Factor 1 because we had the same number of treatment levels and subjects for both word frequency and category cues.) Our computed value of F for Factor 2 is 18.91. The main effect of Factor 2 is significant at $p < .01$: Our computed value of F for Factor 2 is more extreme (that is, in this case larger) than the table value at $p < .01$. Subjects who received category cues recalled significantly more items than subjects who did not receive cues. We reject the null hypothesis that the means of the groups under the two levels of Factor 2 were sampled from the same population.

The computed F for the interaction is 0. This is clearly not significant. In effect, this tells us that the variability between treatment groups can be explained by the effect of either word frequency or category cues acting separately on subjects' scores. Also, the impact of each independent variable was unrelated to the value of the other independent variable: The effect of word frequency was the same regardless of whether or not subjects received category cues. Similarly, the effect of giving category cues was the same whether subjects saw high- or low-frequency words.

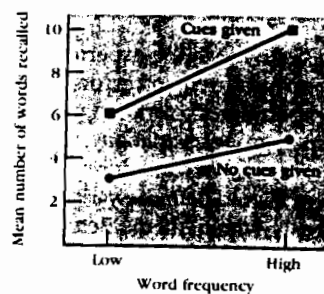


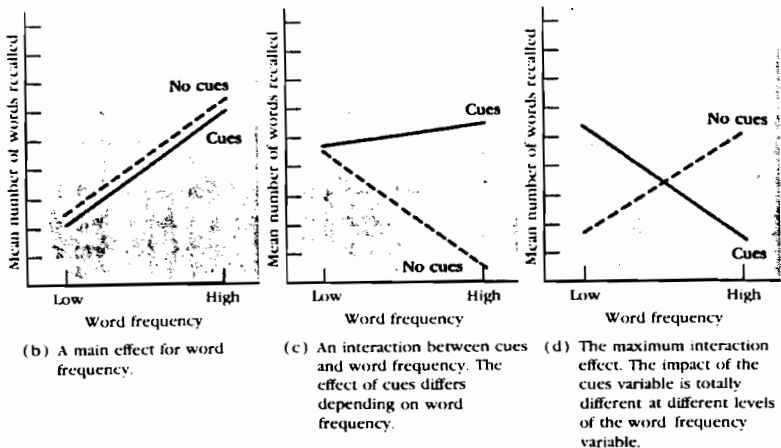
Figure 13-6 Graphing the results of a two-factor experiment: Recall of word lists as a function of word frequency and category cues.

If the interaction had been significant, we would be limited in what we could conclude about the main effects in this experiment. Generally, the existence of a significant interaction makes a discussion of simple main effects unnecessary. If there is a significant interaction, it is usually more useful to discuss the impact of the independent variables in combination with each other. A significant interaction means that the impact of one independent variable differs depending on the value of the other. We can make accurate predictions about subjects' performance only when we know the subjects' position with respect to both variables. For instance, in this example, a significant interaction would mean that we could accurately predict about how many items the average subject would recall—but only if we knew both that the subject saw high-frequency words and that the subject received cues. Without the interaction, we can make a reasonably good prediction if we know the subject's position on only one variable. If we know that Carl was given category cues, we automatically also know that he probably did better than the subjects who did not get cues, regardless of whether he saw high- or low-frequency words.

Graphing the results. When we had only one independent variable, we had only one line to graph. However, in a factorial experiment we need to do more. The results of our experiment are presented graphically in Figure 13-6. Notice that the vertical axis still represents the dependent variable. The horizontal axis represents the different levels of one independent variable. Each line that is graphed presents the data from a different level of the other independent variable. One line represents the recall of subjects who were given category cues; the other stands for recall under the "no cues" condition. You can see from the location of the lines that there are differences between the scores of subjects under the various conditions of this experiment.

Our example yielded two significant main effects. The distance between the two lines (cues versus no cues) reflects the impact of the category cues variable. If giving category cues had no effect on the number of items subjects recalled, the two lines would fall in the same place on the graph. Similarly, the impact of the word-frequency variable is indicated by the relative position of the data points along the Y axis. If word frequency had no effect on recall, subjects would recall about the same number of items in both the high- and low-frequency conditions. These and other possible outcomes are illustrated in Figure 13-7.

Notice that interactions appear on the graphs as lines that are not parallel. If the lines converge, diverge, or intersect, we may have a significant interaction effect. Such graphs are useful ways of summarizing the results of an experiment to give an overall view of the findings. They are especially useful in constructing summaries of findings for experimental reports. But they are not substitutes for statistical analysis. Even though the results look impressive, we need



(b) A main effect for word frequency. (c) An interaction between cues and word frequency. The effect of cues differs depending on word frequency. (d) The maximum interaction effect. The impact of the cues variable is totally different at different levels of the word frequency variable.

Figure 13-7 Illustrating other main effects and interactions: Some hypothetical outcomes of an experiment on the effects of word frequency and category cues on list recall.

to carry out all the statistical procedures before we can make precise statements of whether we will accept them as significant findings.

The *analysis of variance* (ANOVA) procedures are used in experiments having more than two treatment conditions and interval or ratio data. An analysis of variance evaluates the effect of treatment conditions by looking at the variability in data. In the one-way ANOVA, all the variability in the data can be divided into two parts: within-groups variability and between-groups variability. *Within-groups variability* is the degree to which the scores of subjects in the same treatment group differ from one another; *between-groups variability* is the degree to which different treatment groups differ from one another.

Variability is caused by all the sources of *error* in the experiment: differences between subjects as well as measurement errors and other extraneous variables. Error also contributes to within-groups and between-groups variability. Between-groups variability, however, reflects variability due to error and treatment conditions. If the independent variable had an effect, there should be more variability between treatment groups than there is within them: Between-groups variability should be large relative to the amount of variability within each group.

The relationship of within- and between-groups variability is evaluated by computing the statistic called *F*. *F* represents the ratio between the variability observed *between* treatment groups and the variability *within* the groups. The larger the *F* ratio, the more likely it

is that the variability between groups was caused by the independent variable. *F* is found by computing these quantities: sum of squares between groups (SS_B) and sum of squares within groups (SS_W). Each of these quantities is divided by its respective degrees of freedom (*df*) to obtain the mean square between groups (MS_B) and mean square within groups (MS_W). The degrees of freedom of the experiment are used to locate the critical value of *F* in standardized tables. If the computed value of *F* is more extreme than the table value, the null hypothesis that treatment means were sampled from the same population is rejected.

The basic ANOVA procedures may be extended to handle the data from experiments with more than one independent variable. In factorial experiments, the variability in data may be caused by several sources: It can be produced by error; it may also be produced by each independent variable in the experiment. The effects of each independent variable are called *main effects*. In addition, there may be variability due to the *interaction* or combination of variables in the experiment. There is an interaction when the effect of one independent variable changes depending on the value of another independent variable in the experiment. In analyzing data from a factorial experiment, an *F* ratio is computed to evaluate each main effect and each possible interaction. The basic computational procedures for a two-factor experiment (two-way ANOVA) are similar to those for a one-way analysis of variance.

Review and Study Questions

- When do we use a one-way analysis of variance?
- What is within-groups variability?
- What are the sources of within-groups variance?
- What is between-groups variance?
- What are the sources of between-groups variance in an experiment with one independent variable?
- Explain how the one-way analysis of variance works: How do we use within- and between-groups variance?
- Briefly explain each of these terms:
 - Sum of squares within groups (SS_W)
 - Sum of squares between groups (SS_B)
 - Mean square within groups (MS_W)
 - Mean square between groups (MS_B)
 - The *F* ratio
- A researcher computed the *F* ratio for a four-group experiment. The computed *F* is 4.86. The degrees of freedom are 3 for the numerator and 16 for the denominator.
 - Is the computed value of *F* significant at $p < .05$? Why or why not?
 - Is it significant at $p < .01$? Why or why not?