APPENDIX

Online Introduction to Statistics

CHOOSING THE CORRECT ANALYSIS

To analyze statistical data correctly, you must choose the correct statistical test. The test you should use when you have interval data is not the same test you should use when you have nominal data. The test you should use when you are comparing each participant with himself or herself is not the same test you should use when you are comparing one group to another group. The test that would work when you only had two conditions may not work when you are comparing more than two conditions. In other words, there are at least three factors you should take into consideration when choosing a statistical test: (a) the scale of measurement—the type of numbers—that your measure provides (to learn more about scales of measurement, see the table below or see Chapter 6); (b) the type of comparison you are making (one group of participants compared to one or more other groups [between-subjects] or each participant compared to himself or herself [within-subjects]); and (c) the number of conditions you have. In the next three sections, we will show you how to take each of these three factors into account so that you can choose the right analysis for your study.

Scales of Measurement

Often, the type of statistical test depends on what type of data you have. For example, you will do one test if your scores do not represent amounts of a quality but instead represent what *kind* or *type* of response a participant made (e.g., responses are *categorized* as helped or did not help, cheated or did not cheat, or preferred one product over another product), and you will do a different test if your scores represent *amounts* of a quality (e.g., how loud a person yelled, how much they agreed with a statement). To get more specific information about how the type of data you have affects how you

should summarize and analyze those data, see the following table (if you want more information on scale of measurement, see Chapter 6).

Scale of measurement	Example	Average	Measure of correlation	Typical statistical analysis
Nominal	When numbers represent categories that are not ordered, such as 1 = yelled, 2 = frowned, 3 = cried	Mode (most common score) or simply describe the percentage of participants in each category	Phi coefficient	Chi-square
Ordinal	Ranks (e.g., 1st, 2nd, 3rd)	Median (middle score)	Spearman's rho	Mann-Whitney (if testing two groups), Kruskal-Wallis(if testing more than two groups), Friedman test (if using within-subjects design)
Interval	Rating scales	Mean	Pearson *r*	*t* test, ANOVA
Ratio	Height, magnitude estimation	Mean	Pearson *r*	*t* test, ANOVA

Within-Subjects Versus Between-Subjects Designs

Another factor that determines which statistics you should use is whether you are using a within-subjects design (comparing each participant with himself or herself) or a between-subjects design (comparing one group of participants with a different group of participants). For example, if you were using a two-condition within-subjects design, rather than using a between-subjects ANOVA or an independent groups *t* test, you should use either a dependent groups *t* test or a within-subjects ANOVA.

Number of Conditions

Finally, you must also consider the number of conditions you are comparing. For example, if you have interval data and are comparing only two conditions, you can use a *t* test. If, however, you are comparing more than two conditions, you must use ANOVA instead. To get more specific information about how the number of conditions should affect how you analyze your data, consult the following table.

	Number of conditions	
Type of data	Two	More than two
Nominal, between-subjects	Chi-square	Chi-square
Nominal, within-subjects or matched pairs	McNemar test	Cochran Q test
Ordinal, between-subjects	Mann-Whitney test	Kruskal-Wallis test

Type of data	Number of conditions	
	Two	More than two
Ordinal, within-subjects or matched pairs	Wilcoxon matched-pairs	Friedman test
Interval/ratio, between-subjects	independent groups *t* test or between-subjects ANOVA	between-subjects ANOVA
Interval/ratio, within-subjects or matched subjects	dependent *t* test or within-subjects ANOVA	within-subjects ANOVA

Performing the Correct Analysis: An Overview of the Rest of This Appendix

If you refer to the information we just discussed or follow our flowchart (see Figure 1), you will choose the right statistical test. But should you conduct a statistical significance test on your data? Not everyone agrees that you should (to understand both sides of this issue, read Box 1).

Despite the controversy surrounding significance testing, most experts agree that statistical significance provides good evidence that a finding is reliable. Largely because statistically significant tests are helpful in preventing us from mistaking a coincidence for a genuine relationship, almost all articles you read will report the result of a significance test. Therefore, the rest of this appendix will be devoted to discussing the logic and computations behind the most commonly used statistical tests.

We will begin by discussing the independent groups *t* test. Learning about the *t* test will not only teach you about one of the most commonly used statistical techniques, but it will also give you the foundation for understanding other statistical techniques. We will then discuss the most common technique for analyzing the results of an experiment that has more than two groups: ANOVA. We will finish our discussion of techniques that students typically use to analyze data from experiments with a description of the dependent *t* test.

After talking about techniques commonly used to analyze the results of experiments, we will discuss techniques commonly used to analyze data from surveys and other correlational research. We will begin by talking about how to compute the Pearson *r*. Then, we will show you how to calculate and interpret the coefficient of determination. Next, we will show you how to find out if a Pearson *r* in your sample indicates that the two variables are really related in the population. Following this discussion of techniques that are commonly used when you have interval data, we show you how to do comparable analyses when you have nominal data. Finally, we will discuss more sophisticated ways of analyzing correlational data, including multiple regression, mediational analyses, and factor analysis.

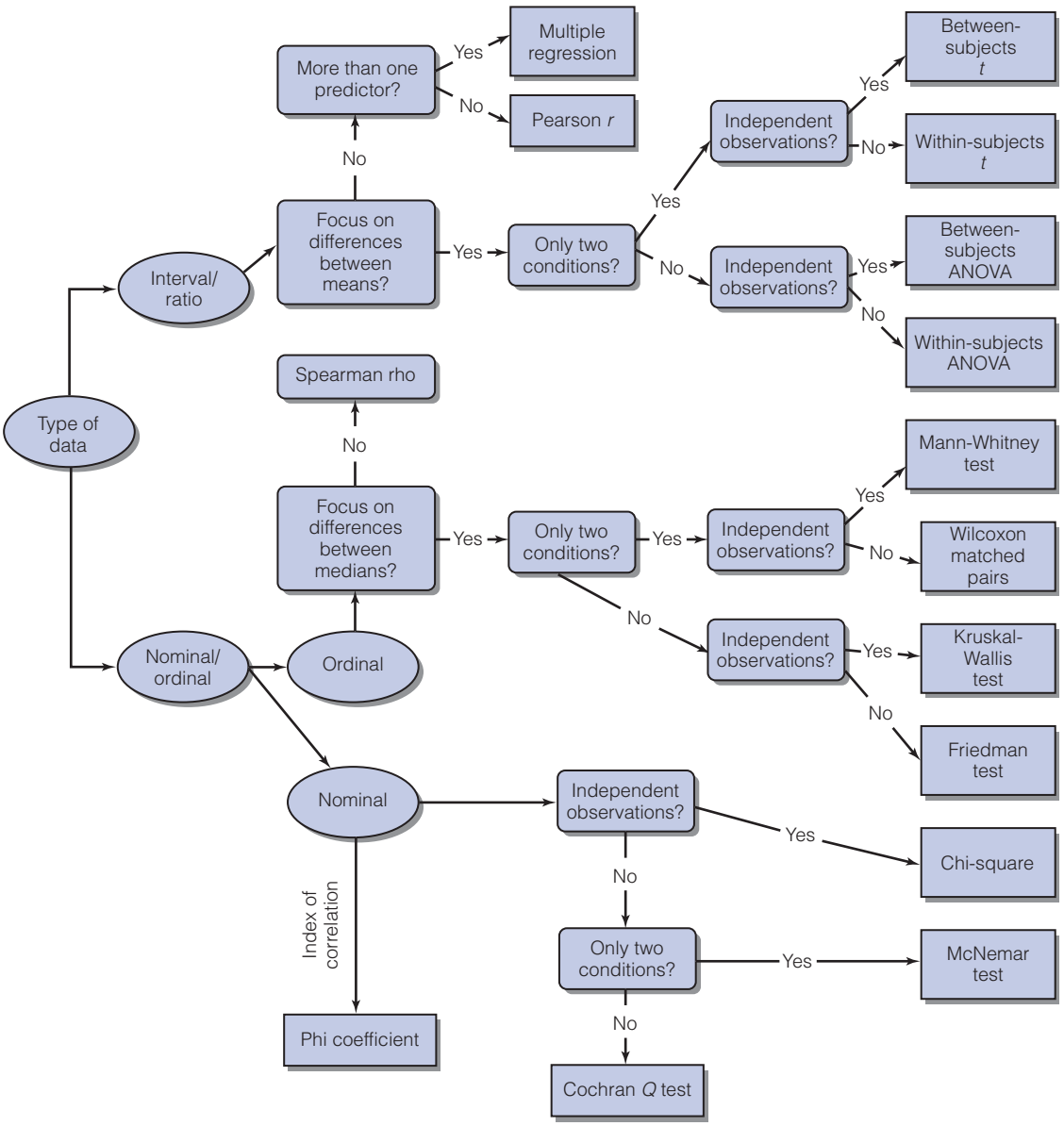


FIGURE 1 Choosing the Right Statistical Test

BOX 1 Ban Statistical Significance Testing?

Although many have criticized the use of statistical significance tests, psychologists—even most of the critics of such tests—still use them (Greenwald, Gonzalez, Harris, & Guthrie, 1996). To understand why, we will consider the major objections to statistical significance and how defenders of the statistical significance tests would respond to those objections. As you can see, the responses to the attacks on statistical significance fall into three general categories: (a) the attack reflects a problem not with statistical significance tests themselves but with how people use or think about statistical significance tests, (b) the attack would still be made if people used alternatives to significance tests, or (c) the attack is misguided.

Objection to statistical significance testing	Reply to objection	Defender's general comment about the attack
The significance test doesn't tell you anything because the null is always false. Everything is connected to everything else. That is, two variables will always be related.	1. The evidence for the view that the null is always false is less than overwhelming. Scientists who have really wanted to believe that a treatment had an effect have used strong treatments and large sample sizes, and still failed to find significant effects. Indeed, most psychologists know that their work will not be published unless their results are statistically significant and yet they often fail to obtain significant effects for the variables of interest. 2. As Hagan (1998) points out, if opponents mean that every treatment has some effect on the measured dependent variable, then that means that (a) there are no Type 1 errors, (b) all null results are Type 2 errors, and(c) quack therapies really work. 3. Even if the null was always false, we would still want to know the direction of the effect (does it help or hurt?). Significance testing is good at detecting the direction of an effect (Harris, 1997).	The attack is misguided.
The $p < .05$ level is arbitrary. Why would a p of .051 fail to be significant, whereas a p of .049 would be significant?	1. Before significance testing, people could just decide that their results "felt real and reliable." Currently, with the .05 criterion, they have to meet some objective standard. 2. In the situation described, any reasonable investigator would follow up on any interesting hypothesis that had a p value of .051. Usually, the investigator would replicate the study using more participants so that the study would have more power. 3. Generally, if we are going to err, we should error on the side of caution. By sticking to the $p < .05$ level, we are unlikely to report that a treatment has one kind of effect when the treatment actually has the opposite effect.	The problem is not as serious as critics allege—and alternative methods have similar problems.

(Continued)

BOX 1 Continued

Objection to statistical significance testing	Reply to objection	Defender's general comment about the attack
	4. As Greenwald, Gonzalez, Harris, and Guthrie (1996) point out, people need "yes versus no" answers to many questions, such as "Is this treatment superior to a placebo?" (p. 178). When we must decide whether to act or not to act, we must use an arbitrary cutoff. For example, how sure do you have to be that going to a doctor would be the best thing to do before you actually go? If we used the same tactics as those who argue that the .05 significance level is arbitrary, we could make any cutoff seem arbitrary. For example, if you said "60%," we could reply, "so you would not go if you were 59.99% sure?" Note, however, that you are not being irrational: You have to have some cutoff or you would never act. 5. An alternative approach, using confidence intervals instead of significance tests, has the same problem. (To learn more about confidence intervals, see Box 10.2.)	
The logic behind statistical significance is not—according to the rules of formal deductive logic—valid (Cohen, 1994). Statistical significance does not tell us how likely it is that the null is false. Instead, it gives us the probability of getting a set of results given that the null is true (Cohen, 1994).	1. The fact that significance testing is not valid according to the rules of formal logic does not mean it is illogical (Hagan, 1998). Most of what physical scientists do is not valid in terms of formal, deductive logic (T. A. Lavin, personal communication, July 18, 2002). 2. Philosophers would say that the arguments behind significance testing are logically valid abductive arguments (J. Phillips, personal communication, September 4, 2005). 3. Hagan (1998) argues that DNA testing uses a similar logic. 4. With simulated data, Type 1 error rates are what significance tests would predict (Estes, 1997). 5. In practice, significance tests (a) are very good at telling us the direction of an effect and (b) provide information about the probability that a replication of the study would obtain a significant effect (Greenwald et al., 1996).	The problem is not as serious as critics allege.
Statistical significance is misunderstood. 1. Many think null results mean accepting the null (Shrout, 1997).	Would physics researchers change their methodology because the average person did not understand their methods? If there is a concern about significant results being misunderstood, there are alternatives to eliminating significance testing. For example, the public or the media could be educated about what	The problem is due to people misunderstanding the term "statistical significance,"

BOX 1 **Continued**

Objection to statistical significance testing	Reply to objection	Defender's general comment about the attack
2. Many think statistical significance means the same as important (Shrout, 1997).	"statistical significance" means or the term could be changed so that people did not confuse its meaning with the meaning of the word "significant." Similarly, Scarr (1997) suggests that the term "reliable" replace "significance."	not with statistical significance itself.
Statistical significance is not a measure of effect size. It is a measure of sample size and effect size. Therefore, we should measure effect size instead.	1. "Effect size" does not give you a pure measure of a variable's effect because "effect size" depends on (a) the power of the original manipulation, (b) how homogenous the sample is, and (c) the reliability and sensitivity of the dependent measure. Thus, with a different treatment manipulation, measure, or sample, you will obtain a different effect size. In addition, an effect that would be large in a well-controlled laboratory study might be tiny in a real-world setting. Even when we (a) study variables in real world settings and (b) use measures that give us estimates of effect size, we still are unable to establish the relative strength of variables (e.g., the nature/nurture debate on IQ). 2. Estimates of effect size can be obtained from significance tests (see Box 10-2). 3. As Prentice and Miller (1992) point out, the fact that a weak manipulation of a variable has any effect at all can be convincing evidence for the importance of that variable. 4. A large effect size on a trivial variable (ratings on a scale) may be unimportant, whereas a small effect size on an important variable (health) may be important. 5. A large effect that does not persist may be less important than a small, lasting effect that accumulates over time.	The problem is not as serious as critics claim.
Significant effects may be small.	1. Small effects may be very important (a) when evaluating the relative validity of two theories, (b) when looking at an important variable (e.g., the effect size for aspirin on preventing heart attacks is tiny, but has enormous practical implications), and (c) when looking at variables that have effects that accumulate over time (e.g., if we produce even a tiny effect for a single commercial we present in the lab, the implications may be enormous because the person will	The problem is with people misunderstanding the meaning of statistical significance rather than with statistical significance tests

BOX 1 Continued

Objection to statistical significance testing	Reply to objection	Defender's general comment about the attack
	probably see more than a million ads in the course of a lifetime). 2. Any informed person could determine whether the effect was small.	
Significance testing made our science unreliable, unlike physical sciences.	1. Our findings are as replicable as those in physics (Hedges, 1987). 2. Health research's abandonment of statistical significance seems to have made their research more conflicting rather than less. 3. Significance reduces our risk of mistaking a chance difference for a genuine effect. It also prevents us from believing that our data support whatever pattern we desire. Thus, significance testing has made our findings more—not less—reliable (Scarr, 1997). 4. Impressions that studies conflict often reflect a misunderstanding of significance tests. If one study is significant and the other is not, then one study refutes the null and the other fails to refute the null. The two studies are not in conflict because the second study does not support the null. 5. Other social scientists tend to want to model our approach because it has been so successful in producing reliable, objective findings.	The attack is misguided.
Significance tests are misused.	Everything can and will be misused (Abelson, 1997).	The attack is misguided.
The p value doesn't tell you the probability that you would get similar results if you repeated the study. For example, if your results are significant at the $p = .05$ level, there is only about a 50% chance that if you repeated the study, you would again get significant results in the predicted direction. Therefore, at the very least, researchers should use p_{rep} rather than p. (For more information on p_{rep}, see the text's website).	There is a relationship between the p value and the chance of obtaining the same results in an exact replication (Greenwald et al., 1996). p_{rep} is based on p and is controversial.	This is a problem with people misunderstanding statistical significance rather than with significance testing.

(Continued)

BOX 1 Continued

Objection to statistical significance testing	Reply to objection	Defender's general comment about the attack
Significance tests don't tell us anything because observed differences that are not significant would have been significant with a larger sample size.	As Hagen (1997) points out, with larger sample sizes, the observed differences will tend to get smaller. That is, with a small sample, the standard error of the difference is large. However, with many participants, the standard error of the difference shrinks, and large differences will be less likely to occur by chance.	The attack is misguided.
People doing statistical significance tests ignore power and thus make many Type 2 errors (Shrout, 1997).	Researchers should use studies that have more power. If they fail to do so, the problem is not with statistical significance testing, but with the researcher.	This is a problem with researchers rather than with significance testing.
A confidence interval (CI), in which the researcher would be 95% confident that the effect was more than ___ but less than ___, would be more informative than significance tests.	1. CI has many of the same problems as significance testing (Abelson, 1997). 2. An informed reader could construct confidence intervals from the reports of a significance test.	CIs should supplement, rather than replace, statistical significance testing.

ANALYZING DATA FROM THE SIMPLE, TWO-GROUP EXPERIMENT: THE INDEPENDENT GROUPS *t* TEST

To use the independent groups *t* test, you *must* meet the following three criteria:

1. You must have two groups.
2. Your observations must be independent.
3. You must be able to assume that your data are either interval or ratio.

In addition, each of your groups *should* have approximately the same variance, and your scores *should* be normally distributed.

As long as your data meet these assumptions, then you can use the *t* test to analyze your data. Thus, the *t* test can be used to look at differences on any measure, such as between men and women, computer users vs. nonusers, or any two independent groups. However, the most common use of the *t* test is to analyze the results of a simple (two-group, between-subjects) experiment.

To understand why you can use the *t* test to analyze the results of a simple experiment, remember why you did the simple experiment. You did it to find out whether the treatment would have an effect on a unique population—all the individuals who participated in your experiment. More specifically, you wanted to know the answer to the hypothetical question, "If I had put all my participants in the experimental condition, would they have scored differently

than if I had put all of them in the control condition?" To answer this question, you need to know the averages of two populations:

> Average of Population 1: what the average score on the dependent measure would have been if all your participants had been in the control group
>
> Average of Population 2: what the average score on the dependent measure would have been if all your participants had been in the experimental group

Unfortunately, you cannot measure both of these populations. If you put all your participants in the control condition, you won't know how they would have scored in the experimental condition. If, on the other hand, you put all your participants in the experimental condition, you won't know how they would have scored in the control condition.

Estimating What You Want to Know

Because you cannot directly get the population averages you want, you do the next best thing—you estimate them. You can estimate them because—thanks to independent random assignment—you started your experiment by dividing all your participants (your population of participants) into two random samples: one of these random samples from your original population of participants was the experimental group; the other random sample was the control group.

The average score of the random sample of your participants who received the treatment (the experimental group) is an estimate of what the average score would have been if all your participants received the treatment. The average score of the random sample of participants who received no treatment (the control group) is an estimate of what the average score would have been if all of your participants had been in the control condition.

Calculating Sample Means

Even though only half your participants were in the experimental group, you can assume that the experimental group is a fair sample of your entire population of participants. Thus, the experimental group's average score should be a reasonably good estimate of what the average score would have been if all your participants had been in the experimental group. Similarly, you can assume that the control group's average score is a fairly good estimate of what the average score would have been if all your participants had been in the control group. Therefore, the first step in analyzing your data will be to calculate the average score for each group. Usually, the average you will calculate is the *mean:* the result of adding up all the scores and then dividing by the number of scores (e.g., the mean of 0, 2, and 4 would be $[0 + 2 + 4]/3 = 6/3 = 2$).

Comparing Sample Means

Once you have your two sample means, you can compare them. We can compare them because we know that, before the treatment was administered, both groups represented a random sample of the population consisting of

every participant who was in the study. Thus, if the treatment had no effect, at the end of the experiment, the control and experimental groups would both still be random samples from that population.

As you know, two random samples from the same population will be similar to each other. For example, two random samples of the entire population of New York City should be similar to each other, two random samples from the entire population of students at your school should be similar to each other, and two random samples from the entire group of participants who were in your study should be similar to each other. Thus, if the treatment has no effect, at the end of the experiment, the experimental and control groups should be similar to each other.

Because random samples of the same population should be similar, you might think all we need to do is subtract the control group mean from the experimental group mean to find the effect. But such is not the case: Even if the treatment has no effect, the means for the control group and experimental group will rarely be identical. To illustrate, suppose that Dr. N. Ept made a serious mistake while trying to do a double-blind study. Specifically, although he succeeded in not letting his assistants know whether the participants were getting the real treatment or a placebo, he messed up and had all the participants get the placebo. In other words, both groups ended up being random samples of the same population—participants who did not get the treatment. Even in such a case, the average scores (the means) of the two groups may be very different.

Dr. N. Ept's study illustrates an important point: Even when groups are random samples of the same population, they may still differ because of random sampling error. You are aware of random sampling error from reading about public opinion polls that admit to a certain degree of sampling error or from reading about two polls of the same population that produced slightly different results.

Because of random sampling error, some random samples will not be representative of their parent population. Because of the possibility that a sample may be strongly affected by random sampling error, your sample means may differ even if the real, parent population means do not.

Inferential Statistics: Judging the Accuracy of Your Estimates

We have told you that random error can throw off your estimates of population means. Because of random error, the treatment group mean is an imperfect estimate of what would have happened if all the participants had received the treatment and the control group mean is an imperfect estimate of what would have happened if none of the participants had received the treatment. Thus, the difference between your experimental group mean and control group mean could be due to random error. Consequently, finding a difference between the treatment group mean and the no-treatment group mean doesn't prove that you have a treatment effect.

If the difference between your group means could be due to random error, how can you determine whether a difference between the sample means is due to the treatment? The key is to know how much of a difference random error could make. If the actual difference between your group means was much bigger than the difference that chance could make, you could conclude that the treatment had an effect.

Estimating the Accuracy of Individual Sample Means

How can you determine whether the difference between your sample means is too large to be due to random error? Knowing the accuracy of each of your individual sample means should help. For example, suppose you knew the control group mean was within one point of its true population mean. Furthermore, suppose you knew that the experimental group mean was also within one point of its real population mean. In other words, you knew that (a) the estimate for what the mean would be if everybody had been in the control group was not off by more than one point, and that (b) the estimate for what the mean would be if everyone had been in the experimental group was also not off by more than one point.

If you knew all that, and if your control group mean differed from your experimental group mean by 20 points, then you would know that your two sample means represent different population means. In other words, you could assume that if all your participants had been given the treatment, they would have scored differently than if they had all been deprived of the treatment.

If, on the other hand, the two group means had differed by less than one point but each of your estimates could be off by a point, a one-point difference between the groups could easily be due to random error. In that case, you would not be able to conclude that the treatment had an effect.

Consider Population Variability: The Value of the Standard Deviation. You have seen that a key to determining whether your treatment had an effect is to determine how well your two sample means reflect their population means. But how can you do that?

One factor that affects how well a mean based on a random sample of the population reflects the population mean is the amount of variability in the population. If there is no variability in the population, all scores in the population will be the same as the mean. Consequently, there would be no sampling error. For example, if everyone in the population scored a 5, the population mean would be 5, and the mean of every random sample would also be 5. Thus, because all Roman Catholic cardinals hold very similar positions on the morality of abortion, almost any sample of Roman Catholic cardinals you took would accurately reflect the views of Roman Catholic cardinals on that issue.

If, on the other hand, scores in a population vary considerably (~~for example~~, ranging anywhere from 0 to 1,000), then independent random samples from that population could be extremely inaccurate. In that case, you might get sample means ranging from 0 to 1,000. Thus, two sample means from such a heterogeneous population could be very different.

To recap, you have seen that the variability of scores in a population affects how accurately individual samples will reflect that population. Because the extent of the variability of scores in the population influences the extent to which we have random sampling error, we need an index of the variability of scores within a population.

The ideal index of the population's variability is the population's **standard deviation:** a measure of the extent to which individual scores deviate

BOX 2 How to Compute a Standard Deviation

Assume we have, as a result of random sampling, obtained four scores (108, 104, 104, 104) from a population. We could estimate the population's standard deviation by going through the following steps.

STEP 1:	**STEP 2:**	**STEP 3:**
Calculate the mean (*M*).	Subtract scores from mean (105) to get differences.	Square differences.
108 —	105 = +3	$(+3)^2 = +9$
104 —	105 = –1	$(-1)^2 = +1$
104 —	105 = –1	$(-1)^2 = +1$
104 —	105 = –1	$(-1)^2 = +1$
420 = Total		SS = 12
Mean = 420/4 = **105**		

STEP 4: Add (sum) the squared differences obtained in step 3 to get sum of squared differences, otherwise known as sum of squares. Sum of squares is often abbreviated as (SS). Sum of squares (SS) = 12.

STEP 5: Get variance by dividing SS (which was 12) by one less than the number of scores (4–1 = 3). This division yields a variance of 4 (because 12/3 = 4).

STEP 6: Get the standard deviation by taking the square root of variance. Because the variance is 4, the standard deviation is 2 (because the square root of 4 is 2).

For those preferring formulas,

$$S = \sqrt{(\Sigma X - M)^2/N - 1}$$

where *X* stands for the individual scores, *M* is the sample mean, *S* is the estimate of the population's standard deviation, and ~~*n*~~ is the number of scores (so, ~~*n*~~–1 is one less than the number of scores).

from the population mean. Unfortunately, to get that index, you have to know the population mean (for the control condition, the average of the scores if all the participants had been in the control condition; for the experimental condition, the average of the scores if all the participants had been in the experimental condition). Obviously, you don't know the population mean for either the control or experimental condition—that's what you are trying to find out!

Although you cannot calculate the population standard deviation, you can estimate it by looking at the variability of scores within your samples. In fact, by following the steps in Box 2, you can estimate what the standard deviation would have been if everyone had been in the control group (by looking at variability within the control group) and what the standard deviation would have been if all your participants had been in the experimental group (by looking at variability within the experimental group).

One reason the standard deviation is a particularly valuable index of variability is that many populations can be completely described simply by knowing the standard deviation and the mean. You probably already know that the mean is valuable for describing many populations. You know that for many populations, most scores will be near the mean and that as many scores will be above the mean as will be below the mean.

What you may not know is that for many populations, you can specify precisely what percentage of scores will be within a certain number of

standard deviations of the mean. For instance, you can say that 68% of the scores will be within one standard deviation of the mean, 95% will be within two standard deviations of the mean, and 99% of the scores will be within three standard deviations of the mean. If a population's scores are spread out (distributed) in this manner, the population is said to be *normally distributed*.

As the term "normally distributed" suggests, many populations are normally distributed—from test scores to the heights of American women. Because normally distributed populations are common, graphing the distribution of scores in a population will often produce a **normal curve:** a bell-shaped, symmetrical curve that has its center at the mean (see Figure 2).

It's convenient to summarize an entire distribution of scores with just two numbers: the mean, which gives you the center of a normal distribution; and the standard deviation, which gives you an index of the width of the distribution. It's comforting to know that 68% of the scores will be within one standard deviation of the mean, that 95% of the scores will be within two standard deviations of the mean, and that virtually all the scores will be within three standard deviations of the mean.

But the standard deviation has more uses than merely describing a population. You could use the standard deviation to make inferences about the population mean. For example, suppose you don't know the population's mean, but you know that the distribution is normally distributed and that its standard deviation is 3. Then, you don't need much data to make certain inferences about that population. Specifically, you know that if you randomly selected a single score from that population, there would be a 68% chance that the population mean would be within 3 points (one standard deviation) of that score and a 95% chance that the population mean would be within 6 points (two standard deviations) of that score.

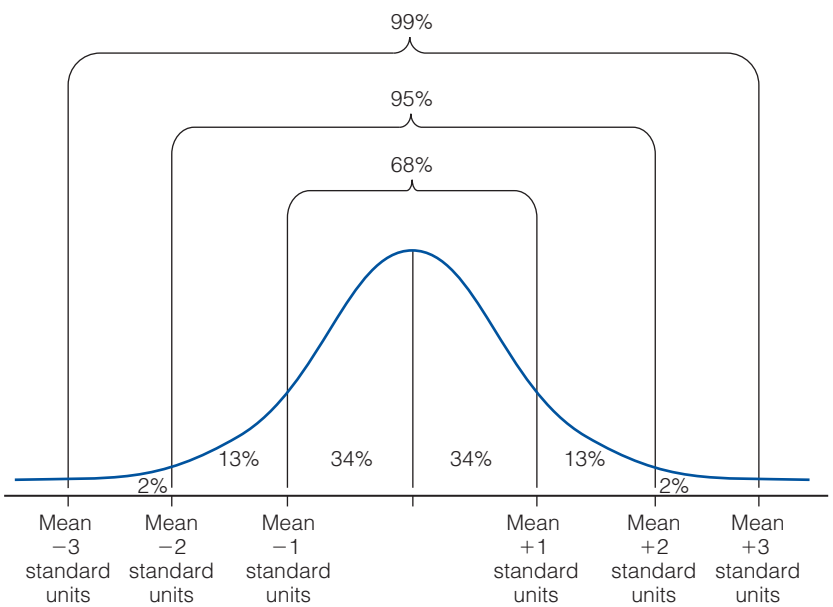


FIGURE **2** The Normal Curve

Consider Sample Size: The Role of the Standard Error. Of course, to estimate your control group's population mean, you would not use just one score. Instead, you would use the mean you calculated by averaging all the scores from your control group. Intuitively, you realize that using a sample mean based on several scores will give you a better estimate of the population mean than using a single score.

You also intuitively realize that your sample mean will be a better estimate of the population mean if your sample mean is based on many scores than if it is based on only a few scores. In other words, the bigger your independent random sample, the better your random sample will tend to reflect the population—and the closer its mean should be to the population mean.

As you have seen, the accuracy of your sample mean depends on (a) how much the scores vary and (b) how many scores you use to calculate that mean. Thus, a good index of the degree to which a sample mean may differ from its population mean must include both factors that influence the accuracy of a sample mean, namely:

1. population variability (the less population variability, the more accurate the sample mean will tend to be)
2. sample size (the larger the sample, the more accurate the sample mean will tend to be)

Although the standard deviation tells you how much the scores vary, it does not take into account how many scores the sample mean is based on. The standard deviation will be the same whether the sample mean is based on 2 scores or 2,000. Because the standard deviation does not take into account sample size, the standard deviation is not a good index of your sample mean's accuracy. However, both of these (population variability and sample size) are included in the formula for the **standard error of the estimate of the mean (also called the standard error):** an index of the degree to which random error may cause a sample mean to be an inaccurate estimate of its population mean.

The standard error (of the estimate of the population mean) equals the standard deviation (a measure of population variability) divided by the square root of the number of participants (an index of sample size). Thus, if the standard deviation were 40 and you had 4 people in your sample, the standard error would be

$$\frac{40}{\sqrt{4}} = \frac{40}{2} - 20$$

Note that dividing by the square root of the sample size means that the bigger the sample size, the smaller the standard error. Thus, the formula reflects the fact that you have less random sampling error with larger samples. Consequently, in the example above, if you had used 100 participants instead of 4, your standard error would have shrunk from $20(40/\sqrt{4})$ to $4(40/\sqrt{100})$.

What does the standard error tell you? Clearly, the larger the standard error, the more likely a sample mean will misrepresent the population mean. But does this random error contaminate all samples equally or does it heavily infest some samples while leaving others untouched? Ideally, you would like to know precisely how random error is distributed across samples. You want

to know what percentage of samples will be substantially tainted by random error so that you know what chance your sample mean has of being accurate.

Using the Standard Error. Fortunately, you can know how sample means are distributed. By drawing numerous independent random samples from a normally distributed population and plotting the means of each sample, statisticians have shown that the distribution of sample means is normally distributed. Specifically, most (68%) of the sample means will be within one standard error of the population mean, 95% will be within two standard errors of the population mean, and 99% will be within three standard errors of the population mean. Therefore, if your standard error is 1.0, you know that there's a 68% chance that the true population mean is within 1.0 points of your sample mean, a 95% chance that the population mean is within 2.0 points of your sample mean, and a 99% chance that the population mean is within 3.0 points of your sample mean.

When you can assume that the population is normally distributed, you can estimate how close your sample mean is to the true population mean. You do this by taking advantage of the fact that sample means from normally distributed populations will follow a very well-defined distribution: the normal distribution. But what if the underlying population isn't normally distributed?

Even then, as the *central limit theorem* states, the distribution of sample means will be normally distributed—if your samples are large enough (30 or more participants). To understand why the central limit theorem works, realize that if you take numerous large random samples from the same population, your sample means will differ from one another for only one reason—random error. Because random error is normally distributed, your distribution of sample means will be normally distributed—regardless of the shape of the underlying population. Consequently, if you take a large random sample from any population, you can use the normal curve to estimate how closely your sample mean reflects the population mean.

Estimating Accuracy of Your Estimate of the Difference Between Population Means

Because you know that sample means are normally distributed, you can determine how likely it is that a sample mean is within a certain distance of its population mean. But in the simple experiment, you are not trying to find a certain population mean. Instead, you are trying to find out whether two population means differ. As we mentioned earlier, you want to know whether there was a difference between two hypothetical population means: (a) what the mean score would have been if all your participants had been in the control group, and (b) what the mean score would have been if all your participants had been in the experimental group. Put another way, you are asking the question: "If all the participants had received the treatment, would they have scored differently than if they had all been in the control group?"

Because you want to know whether the treatment made a difference, your focus is not on the individual sample means, but on the difference between the two means. Therefore, you would like to know how differences between sample means (drawn from the same population) are distributed.

How the Differences Between Means Are Distributed: The Large Sample Case. Statisticians know how differences between sample means drawn from the same population are distributed because they have repeated the following steps thousands of times:

1. Take two random samples from the same population.
2. Calculate the means of the two samples (Group 1 and Group 2).
3. Subtract the Group 1 mean from the Group 2 mean to get the difference between Group 1 and Group 2.

From this work, statisticians have established three basic facts about the distribution of differences between sample means drawn from the same population.

First, if you subtracted the Group 1 mean from the Group 2 mean an infinite number of times, the average of all these differences would equal zero. This is because, in the long run, random error averages out to zero. Because random error averages out to zero, the mean of all the Group 1 means would be the true population mean—as would the mean of all the Group 2 means. Because the Group 1 means and the Group 2 means both average out to the same number, the average difference between the Group 1 and Group 2 means would be zero.

Second, the distribution of differences would be normally distributed. This makes sense because (a) the only way random samples from the same population can differ is because of random error, and (b) random error is normally distributed.

Third, the standard unit of variability for the distribution of differences between means is neither the standard deviation nor the standard error. Instead, it is the standard error of the difference between means.

The standard error of the difference *between* means is larger than the standard error *of* the mean. This fact shouldn't surprise you. After all, the difference between sample means is influenced by the random error that affects the control group mean *and* by the random error that affects the experimental group mean. In other words, sample means from the same population could differ because the first sample mean was inaccurate, because the second sample mean was inaccurate, or because both were inaccurate.

The formula for the standard error of the difference between means reflects the fact that this standard error is the result of measuring two unstable estimates. Specifically, the formula is

$$\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}$$

where s_1 is the estimate of the population standard deviation for Group 1, and s_2 is the estimate of the population standard deviation for Group 2, N_1 is the number of participants in Group 1, and N_2 is the number of participants in Group 2.

We know that with large enough samples, the distribution of differences between means would be normally distributed. Thus, if the standard error of the difference was 1.0, we would know that (a) 68% of the time, the true difference would be within one point of the difference we observed; (b) 95% of the time, the true difference would be within two points of the difference we observed; and (c) 99% of the time, the true difference would be within three

points of the difference we observed. In that case, if our two sample means (the control group mean and the experimental group mean) differed by more than three points, we would be confident that the treatment had an effect. In other words, we would be confident that the groups were samples from populations that had different means. Therefore, we would conclude that if all the participants had received the treatment, their mean score would be different than if they had all been in the control condition.

If, however, we observed a difference of 1.0, we realize that such a difference might well reflect random error, rather than the groups coming from different populations. That is, with a difference of 1.0 and a standard error of the difference of 1.0, we could not disprove the null hypothesis. In other words, we would not be able to conclude that the treatment had an effect.

How Differences Are Distributed: The Small Sample Case. Although the distribution of differences would be normally distributed if you used large enough samples, your particular experiment probably will not use enough participants. Therefore, you must rely on a more conservative distribution, especially designed for small samples: the *t* distribution.

Actually, the *t* distribution is a family of distributions. The member of the *t* distribution family that you use depends on the sample size. That is, with a sample size of 10, you will use a different *t* distribution than with a sample size of 11.

The larger your sample size, the more the *t* distribution will be shaped like the normal distribution. The smaller your sample size, the more spread out your *t* distribution will be (see Figure 3). Thus, with small samples, a difference between means of more than two standard errors of the difference

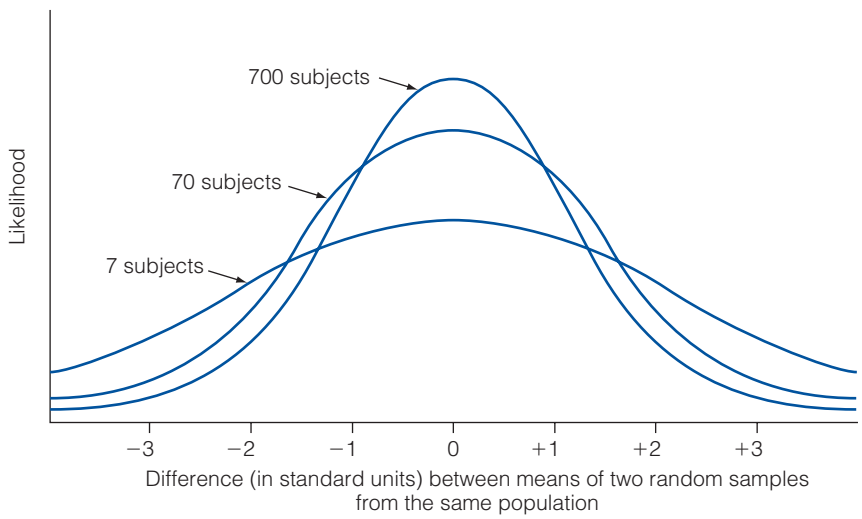


FIGURE **3** With Larger Samples, *t* Distributions Approximate the Normal Curve

might not be statistically significant (whereas such a difference would be significant with a large sample).

Although the particular *t* distribution you use depends on sample size, you do not determine which particular *t* distribution to use by counting how many participants you have. Instead, you determine how many degrees of freedom (*df*) you have.

To calculate your degrees of freedom, simply subtract 2 from the number of participants in your experiment. For example, if you had 32 participants, your *df* would be 30 (because $32-2=30$).

Executing the t Test

You now understand that the difference between your experimental group mean and control group mean could be due to random error. You also realize that to estimate the chances that a difference between means could be due to random error, you need to do two things.

First, you need to compare the difference between the means to the standard error of the difference. In other words, you need to find out how far apart—*in terms of standard errors of the difference*—the two group means are.

Second, you need to use a *t* distribution to figure out how likely it is that two means could differ by that many standard errors of the difference. The particular *t* distribution you will use depends on your degrees of freedom.

Now that you understand the basic logic behind the *t* test, you're ready to do one. Start by subtracting the means of your two groups. Then, divide this difference by the standard error of the difference (see Box 3). The number you will get is called a *t* ratio. Thus, *t* = difference between means/standard error of the difference. Less technically, the *t* ratio is simply the difference between your sample means divided by an index of random error.

Once you have your *t* ratio and your degrees of freedom, refer to a *t* table to see whether your *t* ratio is significant. Specifically, you would look under the row corresponding to your degrees of freedom. As we mentioned

BOX 3 Calculating the Between-Subjects *t* Test for Equal-Sized Groups

$$t = \frac{\text{Group 1 Mean} - \text{Group 2 Mean}}{\text{Standard Error of the Difference}}$$

And where the standard error of the difference can be calculated in either of the following 2 ways:

1. $$\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}$$

   Where S^1 = standard deviation of Group 1 (see Box 1), S^2 = standard deviation of Group 2, N_1 = number of participants in Group 1, and N_2 = number of participants in Group 2.

2. $$\sqrt{\frac{SS\ \text{Group 1} + SS\ \text{Group 2}}{N-2} \times (1/N_1 + 1/N_2)}$$

   Where SS = the sum of squares (see Box 1), N_1 = the number of participants in Group 1, N_2 = the number of participants in Group 2, and N = the total number of participants.

before, the degrees of freedom are two fewer than the number of participants. Thus, if you studied 32 participants, you would look at the *t* table in Appendix under the row labeled 30 *df*.

When comparing the *t* ratio you calculated to the value in the table, act like your *t* ratio is positive. That is, even if you have a negative *t* ratio, treat it as if it is a positive *t* ratio. In other words, take the absolute value of your *t* ratio.

If the absolute value of your *t* ratio is not bigger than the number in the table, then your results are not statistically significant at the $p < .05$ level. If, on the other hand, the absolute value of your *t* ratio is bigger than the number in the table, then your results are statistically significant at the $p < .05$ level.

If your results are statistically significant at the $p < .05$ level, there's less than a 5% chance that the difference between your groups is solely due to chance. Consequently, you can be reasonably sure that your treatment had an effect. You might report your results as follows: "As predicted, the experimental group's mean recall (8.12) was significantly higher than the control group's (4.66), $t(30) = 3.10$, $p < .05$."

ANOVA: ANALYZING DATA FROM A MULTIPLE-GROUP EXPERIMENT

To analyze data from a multiple-group experiment, most researchers use analysis of variance. To use analysis of variance, your observations must be independent, and you must be able to assume that your data are either interval or ratio. Although ANOVA also assumes that your scores are normally distributed and that each of your groups should have approximately the same variance, you can often work around these latter two assumptions. For example, if you have more than 30 participants in each group, you do not have to worry about whether your scores are normally distributed.

In analysis of variance, you set up the ***F* ratio:** a ratio of the between-groups variance (measuring differences between the different group averages, differences that could be due to the treatment as well as to random error) to the within-groups variance (measuring differences between each group's average score and the individual scores making up that average, differences that could only be due to random error). To use more precise terminology, you set up a ratio of mean square between (*MSB*) to mean square within (*MSW*).

To calculate mean square within groups, you must first calculate the sum of squares for each group. You must subtract each score from its group mean, square each of those differences, and then add up all those squared differences. If you had the following three groups, your first calculations would be as follows.

	Group 1	Group 2	Group 3
	5	6	14
	4	5	12
	3	4	10
Group Mean:	4	5	12

Sum of squares within for Group 1:

$$(5-4)^2+(4-4)^2+(3-4)^2=(1)^2+(0)^2+(-1)^2=1+0+1=2$$

Sum of squares within for Group 2:

$$(6-5)^2+(5-5)^2+(4-5)^2=(1)^2+(0)^2+(-1)^2=1+0+1=2$$

Sum of squares within for Group 3:

$$(14-12)^2+(12-12)^2+(10-12)^2=(2)^2+(0)^2+(-2)^2=4+0+4=8$$

To get the sum of squares within groups, you add (sum) all of these sums of squares together ($2+2+8=\mathbf{12}$).

To get the mean square within groups, you divide the sum of squares within groups (*SSW*) by the within-groups' degrees of freedom. In a multiple-group experiment, the within-groups' degrees of freedom equals the number of participants–number of groups. You had 9 participants and 3 groups. Therefore, your within-groups' degrees of freedom is 6 (because $9-3=6$). In this case, because your sum of squares within is 12 and your within-groups degrees of freedom is 6, your *MSW* is 2 (12/6).

To get the mean square between groups, calculate the variance of the group means as follows:

Calculate the mean of the group means ($4+5+12)/3=21/3=7$.

Subtract each group mean from the overall mean and square each difference:

$$4-7=-3;-3 \text{ squared}=9$$
$$5-7=-2;-2 \text{ squared}=4$$
$$12-7=5;5 \text{ squared}=25$$

Add up all these squared differences ($25+9+4=38$).

Divide this term by one less than the number of groups. Since you have three groups, divide by two.

So, your between groups variance is 19 ($38/2=19$).

To transform your variance between groups to a mean square between, multiply it by the number of participants in each group. In this case, you have three participants per group, so you multiply 19×3 and get 57.

Your *F* ratio is the ratio of mean square between (*MSB*) to mean square within (*MSW*). In this case, your *MSB* is 57 and your *MSW* is 2. Therefore, your *F* ratio is 57/2, or 28.5.

Thus, at this point, your ANOVA summary table would look like this:

Source of variance	Sum of squares	Degrees of freedom	Mean square	*F* ratio
Treatment	?	?	57	28.5
Error	12	6	2	

To fill in the rest of the table, you need to know the sum of squares treatment and the degrees of freedom for the treatment. The degrees of freedom (*df*) for the treatment is one less than the number of groups. Because you

have 3 groups, your *df* for the treatment is 2. To get the sum of squares for the treatment, simply multiply the *df* for the treatment by the mean square for the treatment ($2 \times 57 = 114$).

Thus, your completed ANOVA summary table would look like this:

Source of variance	Sum of squares	Degrees of freedom	Mean square	*F* ratio
Treatment	114	2	57	28.5
Error	12	6	2	

To determine whether the *F* of 28.5 is significant at the $p < .05$ level, you would look in the *F* table (Appendix D, Table D.3) for the critical *F* value for 2 degrees of freedom in the numerator and 6 degrees of freedom in the denominator. If 28.5 is larger than that value, the results would be statistically significant.

ANALYZING DATA FROM THE TWO-CONDITION WITHIN-SUBJECTS EXPERIMENT (OR THE MATCHED-PAIRS DESIGN): THE DEPENDENT *t* TEST

If you are comparing two conditions (treatment and no treatment), but your observations are not independent because you are collecting a treatment and no-treatment score from each participant, you cannot use the independent groups *t* test. Similarly, you cannot use an independent *t* test if your scores are not independent because you used a matched-pairs design.

If you are using a two-condition within-subjects design (or a matched-pairs design) and you have interval data, you could analyze your data with a dependent groups (within-subjects) *t* test. The formula for the dependent *t* boils down to dividing the average difference between the conditions by the standard error of the difference. However, to calculate the *t* by hand, you need to execute the following seven steps.

STEP 1: For each matched pair (in the matched-pairs design) or for each participant (in the two-condition, within-subjects design), subtract the Condition 2 score from the Condition 1 score.

Pair or participant	Condition 1 score	Condition 2 score	Difference
1	3	2	1
2	4	3	1
3	5	4	1
4	2	1	1
5	3	2	1
6	5	2	3

Pair or participant	Condition 1 score	Condition 2 score	Difference
7	5	2	3
8	4	3	1
9	3	4	–1
10	5	6	–1

SUM OF DIFFERENCES = 10

AVERAGE DIFFERENCE = 10/10 = 1

STEP 2: Sum up the differences between each pair of scores, then divide by the number of pairs of scores to get the average difference.

STEP 3: Calculate the variance for the differences by subtracting each difference from the average difference. Square each of those differences, sum them up, and divide by one less than the number of pairs of scores.

Pair or participant	Average difference (*AD*)	Observed difference (*D*)	AD–D	AD–*D* squared
1	1	1	0	0
2	1	1	0	0
3	1	1	0	0
4	1	1	0	0
5	1	1	0	0
6	1	3	–2	4
7	1	3	–2	4
8	1	1	0	0
9	1	–1	2	4
10	1	–1	2	4

TOTAL SUM OF SQUARES = 16

VARIANCE OF DIFFERENCES = SUM OF SQUARES/$N-1$ = 16/9 = 1.77

STEP 4: Take the square root of the variance of the differences to get the standard deviation of the differences.

$$\text{Standard deviation of the differences} = \sqrt{\text{variance of the differences}} = \sqrt{1.77} = 1.33$$

STEP 5: Get the standard error of the difference by dividing the standard deviation of the differences by the square root of the number of pairs of scores.

$$\frac{\text{Standard deviation of the differences}}{\sqrt{N}} = \frac{1.33}{\sqrt{10}} = .42$$

STEP 6: Set up the *t* ratio by dividing the average difference (*AD*) by the standard error of the difference (*SED*).

$$t = \frac{AD}{SED} = \frac{1}{.42} = 2.38$$

STEP 7: Calculate the degrees of freedom by subtracting 1 from the number of *pairs* of scores. In this example, because we have 10 pairs of scores, we have 9 degrees of freedom. Then, compare your obtained *t* value to the *t* value needed to be statistically significant. That value is listed in Table D.1 of Appendix D. In this case, the *t* value needed to be significant at the .05 level with 9 degrees of freedom is 2.262. Because our value (2.380) is higher than that, our results are statistically significant at the $p < .05$ level.

CORRELATIONAL ANALYSES

If you are examining the relationship between scores on two or more variables, you may decide to use a correlational analysis. The type of analysis you use will depend on (a) whether you want to describe the data from your sample or whether you want to make inferences about the population that you sampled from and (b) whether your data are at least interval scale (your scores tell you *how much* of a characteristic that participant has).

If you want to describe your data—and you have interval data—you would probably compute a Pearson *r*. If, on the other hand, your data are less than interval (scores do not tell how much more of a quality one participant has than another), you may choose to describe the relationship between your variables using a phi coefficient.

If you want to make inferences about whether the variables that are related in your sample are really related in the population, the type of test you should use depends on your data. If you have interval data (your scores can tell you not only that one participant has more of a quality than another but can also tell you *how much* more of the quality that participant has), you should determine whether the Pearson *r* between the variables is significantly different from zero. If, on the other hand, you have nominal data (higher scores do not reflect more of a variable but instead reflect different kinds of responses), you should do a chi-square test. Soon, we will show you how to perform these tests. However, before we show you how to determine whether the relationship you observed in your sample indicates that the variables are

related in the population, we will start by showing you how to describe the relationship that you observed in your sample.

Computing the Pearson *r*

If two variables are related, you can describe that relationship with a scatterplot. However, if both variables are interval-scale variables, you will probably also want to know what the Pearson *r* correlation coefficient is between the two variables.

The formula for the Pearson *r* is

$$\frac{\Sigma XY - [(\Sigma X \times \Sigma Y)/N]}{N \times sd\,x \times sd\,y}$$

where ΣXY = multiplying each pair of scores together and then adding up all those products, ΣX = the sum of all the scores on the first variable, ΣY = the sum of all the scores on the second variable (so $\Sigma X \times \Sigma Y$ means to add up all the scores on the first variable, add up all the scores on the second variable, and then multiply those two sums), N = number of participants, $sd\ x$ = standard deviation of the x scores (the first set of scores), and $sd\ y$ = standard deviation of the y scores (the second set of scores).

This formula for the Pearson *r* makes sense once you realize three important facts.

1. The formula must produce an index of the degree to which two variables (which we will denote as *X* and *Y*) vary together.
2. The formula must produce positive numbers when the variables are positively correlated, negative numbers when the variables are inversely related, and the number zero when the variables are unrelated.
3. The formula must produce numbers between −1 and +1. That is, the formula can't produce numbers above +1 (or below −1), no matter how many scores there are and no matter how large those scores may be.

Because the Pearson *r* is an index of the degree to which two variables vary together, each pair of scores is multiplied together. Specifically, the *X* member of each pair is multiplied by the *Y* member of the pair. We then add up all these $X \times Y$ products. Note that if *X* and *Y* are positively correlated, we will be multiplying the biggest *X* values by the biggest *Y* values and get some large products. If, on the other hand, *X* and *Y* are negatively correlated, we will be multiplying the biggest *X* values by the smallest *Y* values and the biggest *Y* values by the smallest *X* values, thus giving us relatively small products. Although these products will be relatively small, they won't be negative if *X* and *Y* are always positive (e.g., we are correlating height and salary).

So, how do we get a negative correlation coefficient (which we need when *X* and *Y* are inversely related) if scores on *X* and *Y* are all positive? Given we would never get a negative number if all we did was multiply *X* times *Y* for each pair of scores and then added up those products, there must be more to the Pearson *r* formula than just adding up all the $X \times Y$ products.

To allow ourselves to get negative numbers when the variables are negatively (inversely) related, we subtract a number from the sum of the $X \times Y$ products. That number is an estimate of what the sum of all the $X \times Y$ products would have been if the two sets of scores were completely unrelated. Thus, if the variables are positively related, subtracting this estimate will still

leave us with a positive number. If the variables are not related, subtracting this estimate will leave us with zero. If the variables are inversely related, subtracting this estimate from the actual product of $X \times Y$ will result in a negative number.

To this point, we have a formula that can produce positive and negative numbers. The formula does not, however, meet our final criterion for the correlation coefficient: Coefficients must always be between -1 and $+1$. The numbers produced by our incomplete version of the correlation formula might be far outside of the -1 to $+1$ range, especially if

1. we have many pairs of scores
2. the scores are extremely spread out

The more *XY* pairs there are, the more scores there will be to add up and the larger the total will tend to be. Similarly, the more spread out the scores, the more extreme the products of the scores can be. For example, if scores range from 1 to 5 on both variables, the individual $X \times Y$ products cannot be greater than 25 (because $5 \times 5 = 25$). However, if the scores on both variables can range from 1 to 10, the $X \times Y$ products can be as large as 100 (10×10).

You have seen that our incomplete formula would produce "correlation coefficients" that would be far outside the -1 to $+1$ boundaries of conventional correlation coefficients. More importantly, the correlation coefficients would be influenced by two factors that have nothing to do with the extent to which two variables are related to each other: (a) the number of pairs and (b) the variability (spread) of the distributions. Therefore, we need to add one more step to our formula. Specifically, we need to take the number we have obtained so far and divide it by an index composed of (a) the number of *XY* pairs, (b) a measure of the variability of the *X* scores (the first set of scores), and (c) a measure of the variability of the *Y* scores (the second set of scores).

By adding this final step, you now have a formula that will produce a correlation coefficient that will range between -1 and $+1$, regardless of whether you compute a correlation based on 5 pairs or 5,000 pairs and regardless of whether participants' raw scores range from 1.5 to 1.6 or from 200 to 200,000. Thus, as we stated before, one formula for the Pearson *r* is

$$\frac{\Sigma XY - [(\Sigma X \times \Sigma Y)/N]}{N \times sd\,x \times sd\,y}$$

where ΣXY= multiplying each participant's *x* score (the participant's score on the first variable) by that participant's *y* score (the participant's score on the second variable) and then adding up all those products, ($\Sigma X \times \Sigma Y$) = adding up all the *x* scores, getting a total, adding up all the *y* scores, getting a total, and then multiplying the total of the x scores by the total of the y scores, $N =$ number of participants, $sd\,x =$ standard deviation of the *x* scores (the first set of scores), and $sd\,y =$ standard deviation of the *y* scores (the second set of scores).

To see this formula in action, imagine that you collected data from five students at your school on self-esteem (*X*) and grade-point average (*Y*). Furthermore, assume that self-esteem and grade-point average are interval-scale variables. To see if the variables were related, you would use the following steps to compute a Pearson *r*.

STEP 1: List each pair of scores in the following manner:

	Score for *X*	Score for *Y*	*X* Times *Y*
First pair of scores	1	1	1
Second pair of scores	2	2	4
Third pair of scores	3	2	6
Fourth pair of scores	4	4	16
Fifth pair of scores	5	3	15

STEP 2: Sum the scores in each column (to get ΣX, ΣY, ΣXY).

	Score for *X*	Score for *Y*	*X* times *Y*
First pair of scores	1	1	1
Second pair of scores	2	2	4
Third pair of scores	3	2	6
Fourth pair of scores	4	4	16
Fifth pair of scores	5	3	15
	$\Sigma X = 15$	$\Sigma Y = 12$	$\Sigma XY = \mathbf{42}$

STEP 3: Calculate the means for variables *X* and *Y*.

$15/5 = 3$	$12/5 = 2.4$
(Mean of $X = \bar{X}$)	(Mean of $Y = \bar{Y}$)

STEP 4: Calculate the sum of squares (*SS*) for variables *X* and *Y*.

$(X - \bar{X})^2$	$(Y - \bar{Y})^2$
$(1-3)^2 = 4$	$(1-2.4)^2 = 1.96$
$(2-3)^2 = 1$	$(2-2.4)^2 = 0.16$
$(3-3)^2 = 0$	$(2-2.4)^2 = 0.16$
$(4-3)^2 = 1$	$(4-2.4)^2 = 2.56$
$(5-3)^2 = 4$	$(3-2.4)^2 = 0.36$
10	**5.2**

STEP 5: Calculate the variance for X and Y (Variance = SS/N).

$$\mathbf{10}/5 = 2.0 \qquad \mathbf{5.2}/5 = 1.04$$

STEP 6: Calculate the standard deviations for X and Y (sd = square root of the variance).

$$\sqrt{2.0} = 1.41 \quad \sqrt{1.04} = 1.02$$

STEP 7: Multiply the total of the first set of scores (ΣX) by the total of the second set of scores (ΣY). (We calculated these two sums in Step 2). Then, divide by the number of pairs of scores.

$$(15 \times 12)/5 = 180/5 = \mathbf{36}$$

STEP 8: Subtract the result that we calculated in Step 7 (**36**) from the sum we calculated in Step 1 (42).

$$42 - \mathbf{36} = 6$$

STEP 9: Divide the result (6) by the number of pairs times the standard deviation of X times the standard deviation of Y.

$$6/(5 \times 1.41 \times 1.02) = .83$$

Calculating the Coefficient of Determination

One problem with correlation coefficients is that they give you only a rough idea of the strength of the relationship between two variables. For example, if you compared a relationship described by a correlation of .1 with a relationship described by a correlation of .5, you would probably not immediately realize that the .5 relationship was 25 times stronger than the .1 relationship. Squaring the correlation coefficient gives you a better index of the strength of the relationship: the ***coefficient of determination***.

The coefficient of determination represents the degree to which knowing a participant's score on one variable helps you know (determine) the participant's score on the other variable. The coefficient of determination can range from 0 (knowing participants' scores on one variable is absolutely no help in guessing what their scores will be on the other variable) to +1.00 (knowing participants' scores on one variable allows you to know exactly what their scores will be on the other variable).

If you had a correlation of +1, you would have a coefficient of determination of 1 (because $+1 \times +1 = 1.00$). Similarly, if you had a correlation coefficient of −1, you would have a coefficient of determination of 1 (because $-1 \times -1 = 1.00$). Thus, with either a +1 or −1 correlation coefficient, if you know a participant's score on one variable, you can predict that person's score on the other variable with 100% (1.00) accuracy.

The coefficient of determination tells you the amount of scatter in your scatterplot. If the coefficient of determination is near 1, there is almost no scatter in your scatterplot. If you draw a straight line through your scatterplot, most of the points would be on or near that line. If, on the other hand, the coefficient of determination is near zero, there is a lot of scatter in your

scatterplot. If you draw a straight line through the scatterplot of that data, very few of the points will be close to your straight line.[1]

To get a better idea of what the coefficient of determination indicates, imagine the following scenario. Participants take a test. The average score for those participants is 30. For each participant, the researcher has recorded the participant number ("1" for the first participant, "2" for the second, etc.) and the participant's score on the test. The researcher plots these data. As you can see from Figure 4 (and the researcher confirms by computing a correlation coefficient), there is no relationship between participant number and participant test score.

The researcher then asks you to predict people's scores on the test knowing only the average score (30). For every participant, you should guess "30." The researcher could represent your predictions as a line that went across the graph (see Figure 5).

To see how far off your guesses were, the researcher could look at the distance between each of the data points and your prediction line. To assign you a score that would provide a rough index of how far off your estimates were, the researcher could (a) for each participant, measure the difference between the data point representing the participant's actual score and the point on the prediction line representing the participant's predicted score, (b) square each of those differences, and then (c) add (sum) up all those squared differences. If your guesses had been perfectly accurate, the researcher would have obtained a score of 0 on this crude index. However, your guesses were not perfect: You obtained a score of 1,000 on the researcher's makeshift index of inaccuracy.

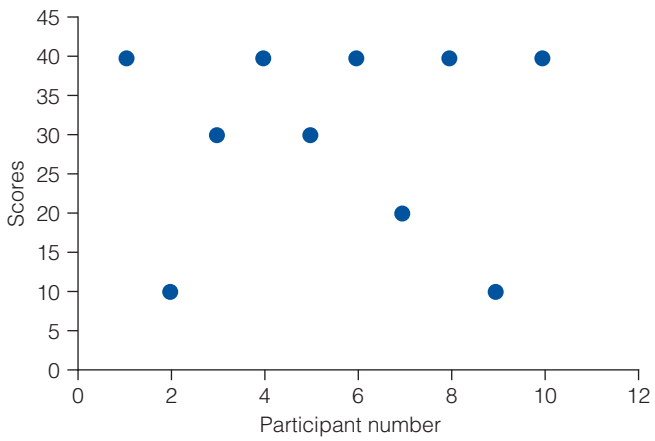


FIGURE 4 Plot Indicating No Relationship Between Participant Number and Scores

[1]There are two cases in which you can have a zero correlation and yet draw a line through all the points: (1) when the line goes straight up and (2) when the line goes straight across the graph. However, you could draw such lines only when there was no variability in scores for one of the variables. In our self-esteem and grade-point average example, you would have a zero correlation if all your participants scored a 5 on the self-esteem measure (producing a vertical line). You would also get a zero correlation if all your participants had a 3.0 grade-point average (producing a horizontal line).

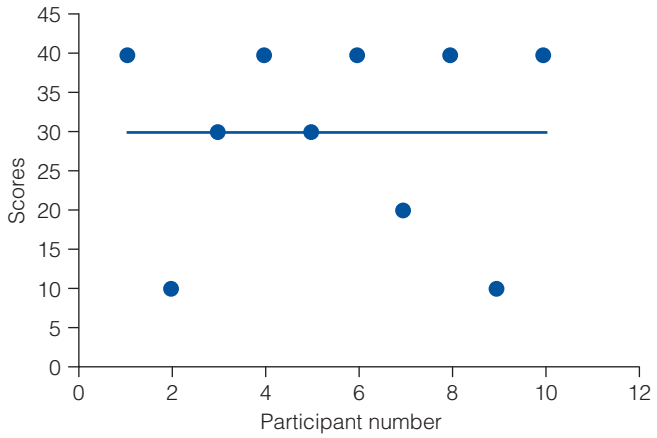


FIGURE **5** As Shown by This Best Fitting Prediction (Regression) Line, Predicting the Mean Is the Best Strategy When the Predictor Is Not Correlated With the Outcome Variable

Next, the researcher asks you to predict the scores again, but this time gives you a piece of worthless information (the participant's number). If you had to guess what a certain person's score was, you would not base your guess on the worthless information. Instead, you should again guess the mean: "30." Because, just as before, for every participant, you are guessing "30," your prediction line would be the same as before and your score on her unsophisticated index of inaccuracy would be the same as before: 1,000.

What you are doing now is **regression:** you are using your knowledge of how two variables are associated to predict one from the other. Your prediction line is a regression line. Your goal in regression is for your predicted scores to match the actual scores. In other words, your predicted scores should match the actual scores on two key characteristics: (1) the average of your predicted scores should be the same as the average of the actual scores, and (2) your predicted scores should differ from each other as much as the actual scores vary from each other (and so your predicted scores should vary around the mean to the same extent that the actual scores vary around the mean). In this case, you accomplished the first goal: the mean of your predicted scores is the same as the mean of the actual scores (both were 30). However, you failed miserably at the second goal: Your predicted scores are all the same as the mean (30) so they do not vary from each other to the same degree that the actual scores vary from each other. Given that your actual scores vary but your predicted scores do not, your predicted scores cannot match the actual scores.

In technical terminology, the coefficient of determination measures the accuracy of your predictions by looking at "*the percent of variance accounted for.*" In other words, the coefficient of determination assesses the accuracy of predictions by looking at the overlap between the predicted scores and the actual scores. Mathematically, this overlap is expressed as a ratio of

$$\frac{\text{how much your predicted scores vary around the mean}}{\text{how much the actual scores vary around the mean}}$$

In this case, your accuracy, as measured by the coefficient of determination, is

$$\frac{0(\text{none of your predicted scores vary from the mean})}{\text{how much the actual scores vary around the mean}} = 0$$

Ideally, you would like perfect accuracy: You would like your predicted scores to perfectly match up with the actual scores in terms of both mean and variation around the mean. If the variability of the predicted scores was the same as the variability of the actual scores, your ratio of predicted variance to actual variance—and your coefficient of determination—would be 1. For example, if the actual variance was 6, and the variance of your predicted scores was also 6,

$$\frac{6}{6} = 1$$

Although it may be unrealistic to expect perfect predictions that account for all the variance in scores, you would like to make predictions that account for some of the variance in the scores and are thus better than just guessing the mean. To make better predictions, you need a predictor that correlates with test scores. The more that predictor is correlated with test scores, the more your estimates will improve. As you can see from Table 1, if the r between your predictor and the test is .1, knowing the person's scores on the predictor reduces the error in your guesses only slightly. Even with an r of .2, your score on her particular index of inaccuracy would still be practically 1,000—what it was when you guessed "30" (the mean) for everybody's score.

Put another way, correlations between −.2 and +.2 do little to improve the accuracy of predictions. As you can see from Table 1, an r of even .2 produces a coefficient of determination (r^2) that is very close to zero.

Determining Whether a Pearson *r* Is Statistically Significant

In addition to determining whether the relationship between your variables in the sample data is substantially above zero, you may want to determine whether the relationship between the variables is different from zero in the population. To illustrate why you might want to determine whether the Pearson r in the sample data indicates that the two variables are related in the population, suppose you collected self-esteem and grade-point average data from a random sample of 5 students at your school and found that $r = +.58$. In that case, you could use your sample data to determine whether there is a relationship between self-esteem and grade-point average for the entire school.

STEP 1: Compute a t value, using the formula

$$t = \frac{r \times \sqrt{(N-2)}}{\sqrt{1-(r \times r)}}$$

where r = the Pearson r and N = number of participants.

TABLE 1
Pearson r, the Coefficient of Determination, and Errors in Prediction

r[a]	r^2 (also called η^2)	Index of inaccuracy[b]
0	.00	1000
.1	.01	990
.2	.04	960
.3	.09	910
.4	.16	840
.5	.25	750
.6	.36	640
.7	.49	510
.8	.64	360
.9	.81	190
1.0	1.00	0

[a]Note two indications that accuracy of prediction increases as r increases: (a) r^2 increases and (b) an index of inaccuracy decreases.

[b]The numbers in this column are the total of the squared errors in prediction you would make if you (a) based all your predictions of participants' scores entirely on a best-fitting prediction line that used one predictor, (b) that one predictor correlated with participants' scores to the degree stated in the leftmost ("r") column, and (c) you were predicting all the participants' scores for the one particular sample we used for this example. Lower scores indicate more accuracy (less inaccuracy), whereas higher scores indicate less accuracy (more inaccuracy). Thus, 0 on the index reflects perfect accuracy (no errors in prediction).

Note that, all other things being equal, the bigger N is, the bigger t will be. Also, note that the bigger r is, the bigger t will tend to be. Not only does a larger r increase the size of the numerator, but it shrinks the size of the denominator. In other words, the larger the relationship and the more participants you have, the greater the chance of finding a statistically significant result.

$$t = \frac{.58 \times \sqrt{(5-2)}}{\sqrt{1-(.58 \times .58)}}$$

$$= \frac{.58 \times 1.73}{\sqrt{1-.34}}$$

$$= \frac{1.00}{.81} = 1.23$$

STEP 2: After computing the t value, look the value up in the t table (Table D.1 in Appendix D) under 3 degrees of freedom ($N-2$) for the .05 level of significance. That value is 3.182. Because 1.23 does not reach that value, you would conclude that the correlation coefficient was not significantly greater than zero. Note that your results are

inconclusive: If you had used a larger sample, you might have found a statistically significant relationship.

Computing a 2 × 2 Chi-Square and the Phi Coefficient

Calculating the Pearson *r* is a good way to describe the relationship between two interval-scale variables in your sample. Testing whether a Pearson *r* is statistically significant is a good way to determine whether there is a relationship between two interval-scale variables in the population.

But what if, instead of having interval scale data, you only have nominal data? In that case, rather than calculating a Pearson *r*, you should compute a phi coefficient—and instead of testing whether the Pearson *r* is statistically significant, you should do a chi-square test.

To see how to do such tests, imagine that you asked men and women whether they believed gay men deserved the same employment opportunities as heterosexual men. If you wanted to know whether there was a gender difference in their responses, you could find out by calculating a chi-square using the following steps.

STEP 1: Set up a table like the following one.

Women	Men	Total
Yes	*A*	*B*
No	*C*	*D*
	(*N*) = Total Number of Participants	

STEP 2: Replace the letter *A* with the number of women who said "yes."

Replace the letter *B* with the number of men who said "yes."

Replace the letter *C* with the number of women who said "no."

Replace the letter *D* with the number of men who said "no."

Replace *N* with the total number of participants.

By the end of this process, your table might look like the following one.

	Women	Men	Total
Yes	20 (*A*)	15 (*B*)	35
No	55 (*C*)	10 (*D*)	65
Totals	75	25	(*N*) 100

STEP 3: Multiply the number in the (*B*) square by the number in the (*C*) square. Then, multiply the number in the (*A*) square by the number in the (*D*) square. For our data, that would be

$$B \times C = 15 \times 55 = 825$$
$$A \times D = 20 \times 10 = 200$$

STEP 4: Plug in the appropriate numbers in the following formula:

$$X^2 = \frac{N(B \times C - A \times D)^2}{(A + B) \times (C + D) \times (A + C) \times (B + D)}$$

$$= \frac{100\,(825 - 200)^2}{35 \times 65 \times 75 \times 25}$$

$$= \frac{100 \times 390{,}625}{4{,}265{,}625} = \frac{39{,}062{,}500}{4{,}265{,}625}$$

STEP 5: Turn to the Chi-Square Table (Table D.2 in Appendix D), and find the row corresponding to 1 degree of freedom. (For a 2 × 2 chi-square, your degrees of freedom will always be 1 because *df* equals the number of rows minus 1 times the number of columns minus 1. Because a 2 × 2 chi-square has 2 rows and 2 columns, its $df = (2-1) \times (2-1) = 1 \times 1 = 1$.)

STEP 6: Determine whether your chi-square is one-tailed or two-tailed. If you predicted only that the groups would differ, then you have a two-tailed test. For example, if you predicted only that there would be a difference between the genders in views toward gay men's employment rights, you have a two-tailed test. If, on the other hand, you predicted which group would score higher than the other, then you have a one-tailed test. Thus, if you predicted that men were less likely to think that gay men should have equal employment opportunities, then you have a one-tailed test.

STEP 7: If you have a two-tailed test with a value of 3.84 or more, your test is significant at the .05 level. Our value of 7.937 exceeds that value, so our test would be significant at the .05 level.

To compute the phi coefficient, use the following formula:

$$\frac{B \times C - A \times D}{\sqrt{(A + B) \times (C + D) \times (A + C) \times (B + D)}}$$

In this case,

$$\frac{825 - 200}{\sqrt{4{,}265{,}625}} = 30$$

Introduction to Multiple Regression

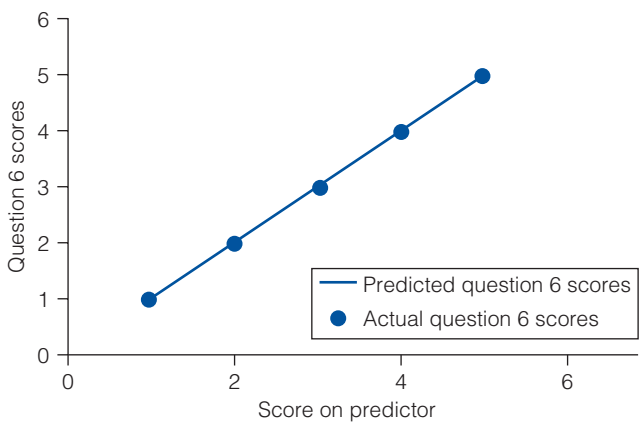
Thus far, we have used correlational analyses to look at the relationship between two variables. However, some correlational analyses, such as multiple regression, can be used to look at the relationships among several variables. With most standard regression analyses, you end up with an equation that uses one or more predictors to predict scores on a question or measure.

For example, suppose that you conducted a survey composed exclusively of 5-point questions and you want to find a set of predictors that will help you predict the answer to question 6. If only one of the predictors is useful, your regression equation might be *the answer to question 5 = the predicted score on question 6*. By substituting the possible values of question 5 into the equation, we could make use of that equation to construct the following table:

Participant's score on question 5	Participant's predicted score on question 6 ($\hat{Y}$)	Participant's actual score on question 6 (Y)	Difference (residuals)
1	1	1	0
2	2	2	0
3	3	3	0
4	4	4	0
5	5	5	0

If we wanted to compare predicted scores (column 3 of our table) to the actual scores (column 4 of our table), we could subtract those two sets of scores. The differences between the predicted and actual scores are called residuals.

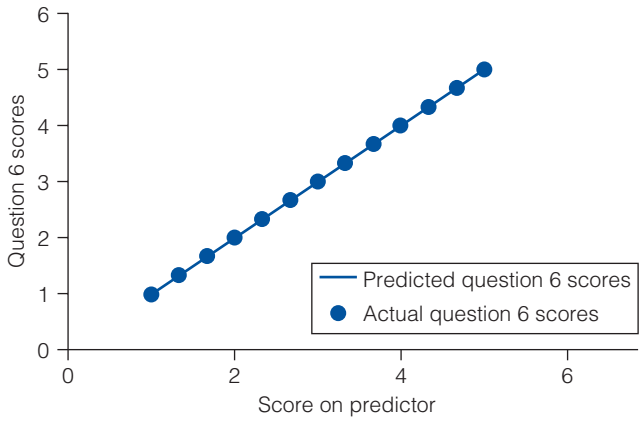
We could also compare the predicted scores to the actual scores with a graph. If we plotted a line based on the scores predicted by the equation (the numbers in the third column of our table), and then plotted the actual scores (the numbers in the fourth column of our table), we would construct the following graph:



If two of your predictors were useful, your equation might be *2/3 × the answer to question 5 + 1/3 × the predicted answer to question 10 = the predicted score on question 6*. By substituting the possible values for questions 5 and 10 into the equation, we could use that equation to create the following table:

Participant's score on question 5	Participant's score on question 10	Participant's predicted score on question 6	Participant's actual score on question 6
1	1	1.00	1.00
1	2	1.33	1.33
1	3	1.67	1.67
1	4	2.00	2.00
1	5	2.33	2.33
2	1	1.67	1.67
2	2	2.00	2.00
2	3	2.33	2.33
2	4	2.67	2.67
2	5	3.00	3.00
3	1	2.33	2.33
3	2	2.67	2.67
3	3	3.00	3.00
3	4	3.33	3.33
3	5	3.67	3.67
4	1	3.00	3.00
4	2	3.33	3.33
4	3	3.67	3.67
4	4	4.00	4.00
4	5	4.33	4.33
5	1	3.67	3.67
5	2	4.00	4.00
5	3	4.33	4.33
5	4	4.67	4.67
5	5	5.00	5.00

Alternatively, we could use the equation and the actual scores (the last two columns of the table) to construct the following graph:



The line that we have drawn through the points is called a regression line. If you have the computer draw a regression line for your data, the line should appear to fit those data: The line's predicted scores should be close to the actual scores. If you could perfectly predict scores, every data point would be on your regression line (as in the two previous examples). If your equation was fairly accurate, then most of the points would be close to your regression line. If your equation was not very accurate, then the line would not fit the points.

You will not need to eyeball your data to determine how accurate your regression equation is. Almost all computerized statistics programs will provide an indicator of how accurate your equation is. This estimate of how well your predictors, as a group, predict your outcome measure is called "multiple *R*-squared." Multiple *R*-squared can range from 0 (using the regression equation to predict each participant's score would be no more accurate than predicting that each participant's score was the mean score) to 1 (your prediction equation can predict scores in your sample with 100% accuracy). (Note that most statistics programs will refer to multiple *R*-squared as either "R^2" or "*R* square.")

Most statistics programs will also provide you with an indication of which predictors are least important for predicting your outcome variable and which are most important. The least important predictors will tend to be left out of the final regression equation. The most important ones tend to be those that, when added to the equation, increase *R*-squared the most.[2]

How to Avoid Being Tricked by Multiple Regression

As we have said, computer programs can provide you with important information. However, that information may be misleading, especially if the

[2]Another way to determine the relative contributions of your predictors is to look at their beta weights in the final, standardized regression formula. The larger the beta weight is (often referred to as standardized coefficients and often abbreviated as β), the bigger the predictor's contribution.

analysis is a stepwise regression and the ratio of predictors to participants was less than 15 to 1 (e.g., there were 30 participants and 3 predictors).

The equation that the computer generates may be greatly affected by an extreme score from a single research participant. Consequently, the regression equation that you get in one sample may be very different from the one that you would get if you were to repeat your study. To determine whether a few extreme scores are dramatically affecting the equation, you should scan the data for extreme scores and re-run your analysis without those extreme scores.[3] If you obtain essentially the same results on this second analysis, you can be relatively confident that your results are not being thrown off by an extreme score.

The multiple *r*-squared can be deceiving because it will tend to give you an inflated impression of how well the predictors correlate with the outcome variable. Keep in mind that the equation did not really predict your outcome variable. Instead, after looking at your outcome variable, an equation was generated to fit the data from your particular sample. Thus, just as you would not be surprised if someone was able to draw a line to fit your plotted data, you should not be surprised if a computer could fit a line to your existing data. Given a large number of predictor variables and a small number of scores, a formula can be made to fit almost any set of scores.

Regression is like "the Texan who shoots holes in the side of the barn and then draws a bull's-eye around the bullet holes" (Carroll, 2003, p. 375). Consequently, you may find that a multiple *r*-squared that seems large (e.g., .50) is not statistically different from chance. Therefore, before deciding whether a regression equation can predict scores on your outcome variable, you should determine whether the multiple *r*-squared is statistically significant. To do this, look for an *F* test (ANOVA) testing either "Model," "Regression," or "R^2." To be statistically significant, the *p* value of the test (often abbreviated as either "Sig." or "Prob > *F*") should be less than .05.

A significant multiple *r*-squared tells you that your equation does more than just capitalize on chance: It produces an equation that fits the data better than an equation that used variables that were uncorrelated with your outcome variable. In other words, if you were to use the same equation on a new sample of data, your multiple r^2 would be greater than 0. However, you probably want to know more than that your equation's *r*-squared, after adjusting for chance, is greater than zero. You want to know how much greater than zero. To find out, look at the *adjusted r-squared*. The adjusted *r* -squared subtracts a value from the multiple *r*-squared to take into account that even an equation full of variables that were uncorrelated with the outcome variable could be made to produce values that would correlate with the outcome variable. In short, if you look at the multiple *r*-squared instead of the adjusted *r*-squared, you can be fooled about how good you are at *predicting* participants' scores.

---

[3]You may be able to spot an extreme score in a graph of your data by just looking for scores that seem to be almost off the graph. Another tactic is to look for scores that are more than 3 standard deviations from the mean. If your computer program lists the *h* values or *D* values of data points, consider extreme scores to be those with *h* values above .5 or *D* values greater than 1.

Not only can you be fooled about how good your equation is at predicting scores, but you may also be fooled about the relative importance of an individual predictor variable. The amount that a predictor increases *r*-squared often depends on (a) when it was entered into the equation and (b) whether related variables were already entered into the equation.

To illustrate that it matters when the predictor is added, suppose we were doing a survey and trying to predict responses to item 11 (whether people strongly disagree, disagree, neither agree nor disagree, agree, or strongly agree with the statement "I like college students"). Suppose agreement with item 6 ("College is stressful for students") significantly correlates with answers to "I like college students." In that case, if item 6 was the *first* variable we entered into the equation, item 6 would be certain to be a statistically significant predictor for two reasons. First, it doesn't have to do much: It only has to make *r*-squared significantly greater than zero. Second, it doesn't have to share credit with any other variables: Any increase in *r*-squared is attributed to item 6.

If, on the other hand, we added item 6 to the equation only after entering all the other items as predictors, adding item 6 might not significantly improve our equation's ability to predict item 11 responses because (a) we already have a large *r*-squared so improving it significantly would be difficult and (b) some of the variability that item 6 could account for has already been eaten up by related, competing variables (especially if we had the following items: "Colleges need to spend more time on students' emotional development" and "College students work hard on their studies" that, like item 6, tap into concerns about college being stressful). Thus, if we looked at the "Model Summary" section of an SPSS printout, we might find that the "*R*-Square Change" for our model with question 6 added was not significantly different from our model without question 6 (e.g., "Sig. *F* Change" was greater than .05). Similarly, if we looked under the "Coefficients" table in the printout, we might find that the variable we labeled "Question 6" was not significant (e.g., the "*t*" associated with question 6 was less than 2 and the "Sig." in the Question 6 column was greater than .05).

To help you understand and remember how a regression equation may mislead you about the relative importance of a predictor, realize that the regression equation is sensitive to the *unique* contribution of each predictor. In a way, the same things that would allow you to make a large and unique contribution to an electronic discussion list are the same things that allow a predictor to make a large and unique contribution to the equation. It is easier to make a large and unique contribution if you are one of the first to enter the discussion, just as it is easier for a predictor to have a large and significant effect if it is the first entered into the equation. It is also easier to make a large and unique contribution if your viewpoint is different from that of the people who have already entered the discussion. Thus, a comment that you make in one list might be unique and contribute much, whereas the same comment might seem redundant in another list. Similarly, whether a predictor appears to be relevant may depend on the other variables in the equation. In more technical terminology, intercorrelations among predictors (sometimes called collinearity or multicollinearity) can cause the regression equation to underestimate the strength of a particular predictor variable.

Therefore, before deciding that a variable is unimportant for predicting your outcome variable, there are two things you should do. First, look at the

Pearson *r* between the potential predictor and (a) the outcome variable and (b) the predictors that did make it into the regression equation. You may find that the potential predictor correlates well with the outcome variable but was left out of the equation because it correlates highly with a predictor that is already in the equation. In such a case, you might see what happens when you enter your potential predictor variable into the regression equation while leaving out predictors that correlate highly with that variable. Second, see whether your computer program provides the variance-inflation factor (VIF) statistic. If the VIF is greater than 5, do not trust the equation's estimates about the relative importance of your predictors.

Using Regression to Test for Moderator Variables

Although the results from multiple regression can be misleading, multiple regression is a flexible technique that has many uses. It can even help you find a moderator variable: a variable that alters the relationship between two other variables; a predictor that, when occurring in combination with another predictor, is related to the outcome measure in a way that could not be predicted from knowing only the individual predictors' relationships with the outcome measure.

To see how multiple regression can help you find a moderator variable, consider the following example. First, suppose that (a) newly married couples who had positive expectations tended to be happier with the marriage than those who entered with negative expectations, and (b) couples who tended to be skilled at interacting with each other in a positive constructive manner were happier with the marriage than those who were not skilled. From these findings, we might create a crude regression-type equation in which we would say that a (expectations) + b (skills) = c (predicted marital happiness). To plug numbers into our equation, we could give couples a +1 for positive expectations but a −1 for negative expectations and a +1 for good skills but a −1 for poor skills. Thus, a couple with positive expectations (+1) and good skills (+1) would have a predicted score a +2 (1 + 1 = 2), whereas a couple with low expectations (−1) and poor skills (−1) would have a predictor score of −2 (−1 + −1 = −2). In this model, our prediction is just a function of adding the values of our individual variables. Thus, a couple with good skills (+1) and low (−1) expectations would get a 0. Let's say that this additive model predicted actual marital happiness with some degree of accuracy.

To see whether we had a moderator variable, we would need to see whether certain combinations of expectations and skills (e.g., positive expectations combined with positive skills) had effects that were beyond what we would get from just adding the values of the individual variables together. For example, suppose couples with positive expectations (+1) and positive skills (+1) did not score a 2 but instead scored a 3 on our marital happiness scale. Or, suppose that couples with negative expectations (−1) and negative skills (−1) did not score a −2 but instead scored a 0 on our marital happiness scale. In both cases, adding the individual, average values of the predictors does not give us the right predicted value. Put another way, both cases suggest that skills *moderate* expectations.

How could we get multiple regression to tell us that the combination of our predictors has a relationship with marital happiness that is more than—and different from—the sum of the predictors' individual relationships with

marital happiness? The basic strategy would be to see whether adding a variable that represents the combination of the two variables can improve the equation. In this case, the goal would be to better predict marital satisfaction by changing the formula "expectation + skills = satisfaction" to "expectation + skills + combination of expectations and skills = satisfaction."

To get the term expressing the combination (interaction) of the two variables, we could multiply the scores of the individual variables together. Multiplying those values gives us a positive number when there is a match between expectations and skills (the combination of positive expectations with positive skills produces +1, as does the combination of negative expectations with negative skills) and a negative number when there is a mismatch between expectations and skills (positive expectations combined with negative skills produces a −1, as does negative expectations combined with positive skills). Thus, our new formula is not

$$a\ (\text{expectations}) + b\ (\text{skills}) = c, \text{ but rather}$$
$$a\ (\text{expectations}) + b\ (\text{skills}) + a \times b\ (\text{combination of expectations and skills}) = c.$$

As you can see from Table 2, the two equations make different predictions. If the formula including a term expressing the combination (interaction) of the two variables does a significantly better job of predicting actual marital happiness, you have solid evidence that skill is a moderator variable. As it turns out, McNulty and Karney (2004) found that an equation including the interaction (combination) term does do a better job of predicting actual marital happiness. Thus, skill does moderate the effect of expectations: Couples with positive skills are better off having high expectations, but couples with poor skills are better off having low expectations.

TABLE **2**

Two Regression Equations Predicting Marital Satisfaction on a –3 to +3 Scale

Couple's characteristics	A	B	A × B	Formula 1 prediction (A + B)	Formula 2 prediction (A + B + A × B)
Positive expectations Positive skills	1	1	+1	2	3
Positive expectations Negative skills	1	–1	–1	0	–1
Negative expectations Positive skills	–1	1	–1	0	–1
Negative expectations Negative skills	–1	–1	+1	–2	–1

Note: Column A refers to expectations (positive = +1, negative = –1) and Column B refers to skills (positive = +1, negative = –1).

To help yourself understand how skill—or any other moderator variable—modifies the relationship between two other variables, you could go back and compute two correlation coefficients between those two other variables: (1) a correlation between the two other variables for those cases that are above the mean on the moderator variable and (2) a correlation between the two other variables for those cases that are below the mean on the moderator variable. For example, you might find that the correlation between expectations and marital happiness is +.30 for couples who have above average skills, but that the correlation between expectations and marital happiness is −.20 for couples who have below average skills.[4]

Using Multiple Regression to Look for Mediator Variables: Answering "How" Questions

Suppose that, instead of showing that you have found a moderator variable, you want to show that you have found a *mediating variable*: a mental or physiological mechanism that causes the relationship between two other variables. That is, you may want to show that your predictor variable (Variable 1) does not have a direct effect on your outcome variable (Variable 3), but instead affects a mediating variable (Variable 2) and that mediating variable, in turn, affects your outcome variable (Variable 3). How can you make the case for this chain of events?

To make the case that, like a chain reaction involving three dominoes, the first affects the second, which in turn, affects the third, you can use multiple regression. For example, take Sargent's (2004) finding that people who most believe in punishing criminals tend to score low on the need for cognition scale: a measure of how much people enjoy thinking. You might suspect that the reason for this relationship is that people who (1) don't like to think (2) may not think of the cultural, environmental, and situational reasons for a person's behavior and therefore would have a (3) greater desire to punish the person for the person's behavior.

To see whether thinking about situational causes for behavior mediates the relationship between need for cognition and punishment, you would measure all three variables. Then, you would go through Baron and Kenney's (1986) four steps (see Figure 6):

1. You would establish that need for cognition was related to punishment by finding a significant correlation between those two variables. (If you were going to argue that knocking over the first domino causes the second domino to fall, which, in turn causes the third domino to fall, you would have to show that knocking over the first domino correlates with the third domino falling. Likewise, if you want to argue that your predictor influences the outcome variable, your predictor better correlate with the outcome variable.)
2. You would establish that need for cognition was related to your measure of thinking about situational, rather than personal, causes for behavior

[4]If you want to see whether the correlation coefficients are significantly different, go to this book's website to do the appropriate statistical test.

by finding a significant correlation between the two variables. (If you were going to argue that knocking over the first domino causes the second domino to fall, which, in turn causes the third domino to fall, you would have to show that knocking over the first domino correlates with the second domino falling. Likewise, if you want to argue that your predictor influences the outcome variable by influencing the mediator variable, your predictor better correlate with the mediating variable.)

3. You would show that your mediating variable has an effect beyond that of your predictor by showing that when you add your mediating variable (thinking about situational causes) to a regression equation that has already used your predictor (need for cognition), your mediating variable improves the equation's ability to predict the amount of punishment a person gives. (If you were going to argue that knocking over the first

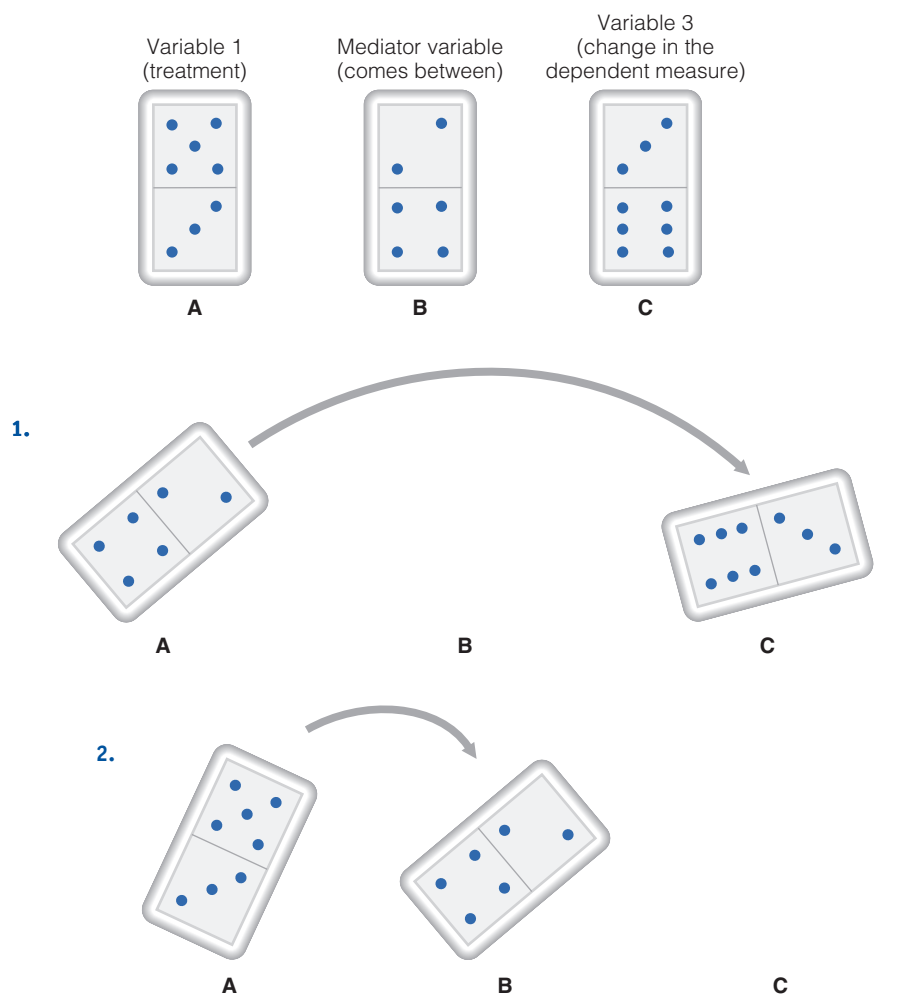


FIGURE 6 Mediating Variables

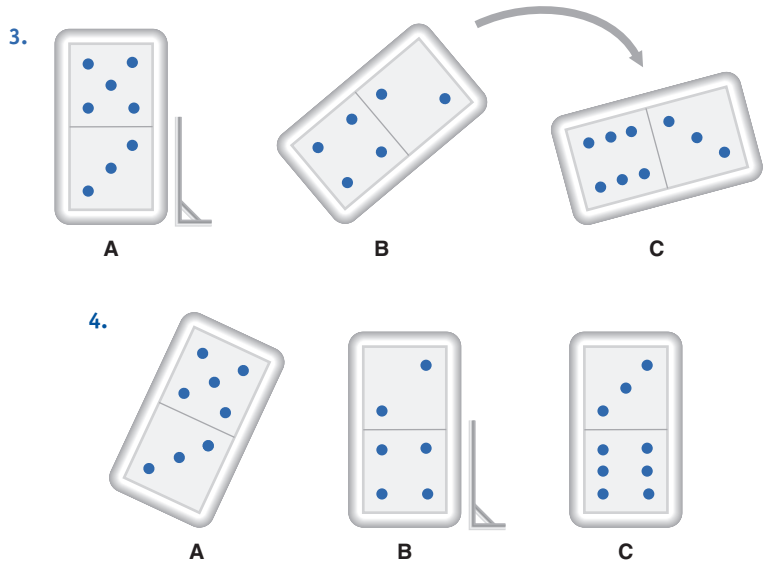


FIGURE 6 (Continued)

domino causes the second domino to fall, which, in turn, causes the third domino to fall, you would have to show that, regardless of what happens to the first domino, knocking over the second domino causes the third domino to fall. Similarly, if your mediating variable causes changes in the outcome variable, it should be able to do so independently of your predictor variable.)

4. You could argue that your predictor's effect is entirely through your mediating variable by showing that when you add your predictor (need for cognition) to an equation that has already used your mediator (thinking about situational causes), the predictor does not improve the equation's ability to predict the amount of punishment a person gives. (If you were going to argue that knocking over the first domino causes the second domino to fall, which, in turn causes the third domino to fall, you would have to show that, when you have already knocked down the second domino, there is no relationship between knocking down the first domino and the third domino falling. Similarly, if your predictor's effect is entirely through the mediating variable, the predictor variable will not have an effect that is independent of your mediating variable.)

Making the Case for Cause–Effect Relationships: Attempts to Answer "Why" Questions With Correlational Data

When we were discussing mediators, we were asking how a predictor variable had an effect. Before finding a mediator, we usually need to first establish that the predictor variable had an effect.

How do we know that the predictor had an effect? Sometimes, we know because an experiment allowed us to establish it. But if we had only correlational data, how can researchers argue that the predictor had an effect on the outcome variable? After all, correlational techniques cannot establish cause–effect relationships because (1) with correlation, you do not know which variable came first so you can mistake causes for effects and (2) because both your variables may be effects of some other (third) variable, this third variable may be responsible for the relationship between your two variables. However, some researchers try to overcome these two problems with correlational data.

Researchers are sometimes able to establish which of their variables came first by using longitudinal and prospective methods—methods in which they measure a variable one time and then measure a second variable later. For example, if you collected individuals' scores on a mental health measure when they were 7 and then, 20 years later, you collected their college grade-point average, you know their college grade-point average could not have caused them to score poorly on a mental health measure when they were 7.

In terms of ruling out third variables, researchers may be able to rule out some third variables by statistically controlling ~~for~~ them. Usually, researchers would measure the suspected third variable and then try to rule out its effects using either a simple technique such as a partial correlation or ANCOVA or using a sophisticated technique such as structural equation modeling.

A partial correlation between two variables attempts to calculate the association between two variables when the effects of a third variable are accounted for. Thus, if the relationship between two variables (e.g., mother's skill at reading her child's mind and child's self-esteem) was due to a third variable (divorce leading to mothers being worse at reading their child's mind and divorce hurting a child's self-esteem), the partial correlation between mother's mind reading and child's self-esteem (controlling for divorce) would be zero.

In analysis of covariance (ANCOVA), a researcher might create two groups (children whose mothers were accurate mind readers and children whose mothers were poor mind readers) and use moms' self-esteem as a variable (a covariate) in the analysis. If, even after statistically controlling for moms' self-esteem, children of accurate mind readers had higher self-esteem than the children of poor mind readers, you could be confident that mom's self-esteem wasn't the third variable causing both poor mind reading and poor self-esteem. The problem is that you don't know whether there is some *other* third variable causing both poor mind reading and poor self-esteem.

Structural equation modeling (SEM) is better than ANCOVA or partial correlations at ruling out the effects of third variables. However, like ANCOVA and partial correlations, SEM can account for only third variables that the researcher knew about and measured. Furthermore, SEM, like ANCOVA and partial correlations, can confuse cause and effect. For example, if the child's high self-esteem causes the mother's accurate mind reading, all three methods might incorrectly conclude that the mother's accurate mind reading causes the child's high self-esteem.

In short, researchers cannot use correlational methods to make cause–effect statements. However, they can often make a better case that one variable has an effect on another by using logic and sophisticated statistical techniques than by using only correlation coefficients.

Introduction to Factor Analysis

We have shown you how researchers can use correlational analyses (a) to describe the relationship between two variables (with the Pearson *r,* the coefficient of determination, and the phi coefficient), (b) to determine whether two variables that were related in a sample are also related in the population (testing to see whether the Pearson *r* is statistically different from 0 or using the chi-square test), (c) to determine which combination of predictors allows you to best estimate scores on a measure, (d) to identify moderator variables, (e) to identify mediating variables, and (f) to make a case that one variable causes changes in another. However, we have not shown you a very common use of correlational analyses: to help assess the validity of a measure.

To see how this works, suppose you want to measure love, and you think that love has two different dimensions (sexual attraction and willingness to sacrifice for the other). Furthermore, you believe that these dimensions are relatively independent. For example, you believe that a person could be high on sexual attraction, but low on willingness to sacrifice—and vice versa.

One approach would be to make up a love scale that had two different subscales. If the subscales are really measuring two different things, then the following should apply:

1. A participant's answers to each question in the first subscale should correlate (correspond, agree) with each other.
2. A participant's answers to each question in the second subscale should correlate with each other.
3. A participant's score on the first subscale should not correlate highly with that participant's scores on the second subscale.

In our case, all the responses to items related to sexual attraction should correlate with one another, and all the responses to items related to sacrifice should correlate with one another. However, the sexual attraction items should not correlate highly with the sacrifice items.

A more sophisticated and extremely common approach to determining whether the items on a test correlate with each other is to do a factor analysis (Reis & Stiller, 1992). We can define *factor analysis* as a statistical technique designed to divide the many questions on a test into as few coherent groups as possible. Put another way, rather than explaining how participants answer the test by talking about how participants answer each individual question, factor analysis tries to explain participants' patterns of answers in terms of a smaller number of underlying hypothetical factors.

The logic behind factor analysis is straightforward: We assume that when participants' answers to one group of questions correlate with each other, then those questions all measure the same factor. For example, imagine that we have a 10-item test. In that test, participants answered the first six questions similarly: If we know how they answered any one of those questions, we can make a reasonable prediction about how they answered the other five. Similarly, their responses to the last four items were highly correlated. However, their responses to the first six questions did not correlate very well with their answers to the last four questions. In such a case, factor analysis would say that because the test seems to be composed of two groups of items, the test measures two factors. In technical terminology, the first six items of the test would load on one factor, the last four items would load on

another factor. Each question's *factor loading* tells us the degree to which it appears to be measuring a given factor.

Factor loadings, like correlation coefficients, can range from −1 to +1. Ideally, questions designed to measure a certain factor would have a factor loading of 1.0. However, because of unreliability and other measurement error, a question's factor loadings will usually be well below 1.0. Indeed, a factor loading of +.7 is considered very high and some researchers are happy when a question has a factor loading above +.3.

You have seen that factor analysis tries to find out how many factors are being measured by a test and how well individual questions measure those factors. But what results would you want to obtain from a factor analysis of your love scale? In this case, you would hope for two outcomes.

First, you would hope that the factor analysis supported the view that there were two different factors being measured by the test. You would be disappointed if the factor analysis reported that, based on participants' responses, your test seemed to be composed of three types of items. If the factor analysis supports the view that there are two factors, you might be able to report something like, "The two-factor solution accounts for a large amount (at least 60%) of the variability in participants' responses."

Second, you would hope that the factor analysis found that the items that you thought made up the sexual attraction subscale all corresponded to one factor and the items that made up the sacrifice subscale all corresponded to another factor. In technical terminology, you would hope that all the sexual attraction items loaded on one factor, and all the sacrifice items loaded on a different factor. Specifically, because factor loadings are like a correlation between the test question and the factor, you would want all your sexual attraction items to have high loadings (above .5) on the factor you want to label sexual attraction and near zero loadings on the factor that you want to label sacrifice. Conversely, you would want all your sacrifice items to have very low factor loadings on the factor that you want to label sexual attraction and loadings on the factor you want to label sacrifice.