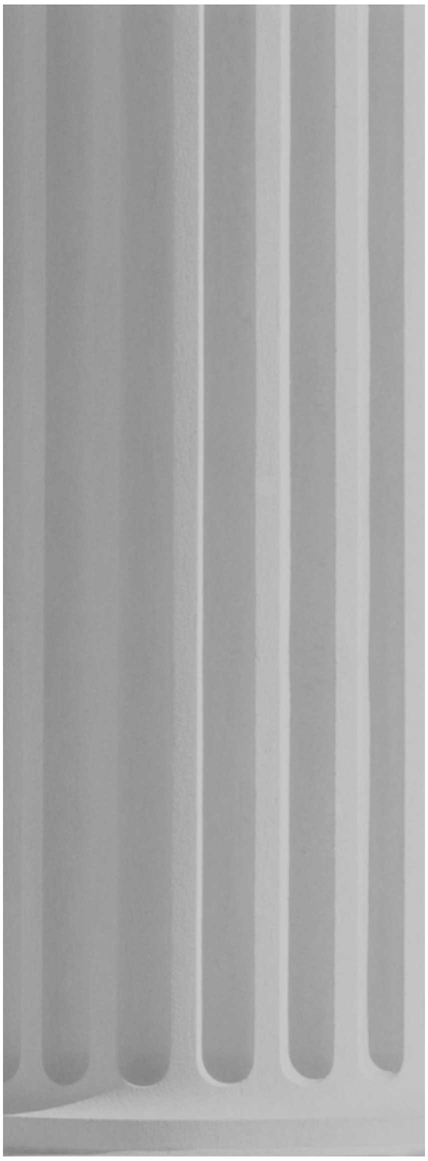
CHAPTER 12

Analysis

KEY TERMS

.05 level of significance
alpha level
boxplot
causal
central tendency
codebook
conclusion validity
construct validity
correlation
correlation matrix
cross-tabulation
degrees of freedom (*df*)
descriptive statistics
dispersion
distribution
double entry
exploratory data analysis
external validity
fishing and the error rate problem
frequency distribution
hypothesis
imputation
internal validity
mean
median
mode
Pearson product moment correlation
qualitative
quantitative
range
reliability
standard deviation
statistical power
stem and leaf diagram
symmetric matrix
threat to conclusion validity
variance



OUTLINE

By the time you get to the analysis of your data, most of the really difficult work has been done. It's much more difficult to define the research problem; develop and implement a sampling plan; conceptualize, operationalize, and test your measures; and develop a design structure. If you have done this work well, the analysis of the data is usually straightforward.

In most social research, data analysis involves three major steps, performed in roughly this order:

- Data preparation involves checking or logging the data in, checking the data for accuracy, entering the data into the computer, transforming the data, and developing and documenting a database structure that integrates the various measures.
- Descriptive statistics describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every **quantitative** analysis of data. With descriptive statistics, you are simply describing what is—what the data shows.
- Statistical analysis of the research design tests your research hypotheses. In experimental and quasi-experimental designs, you use statistics to determine whether the program or treatment has a statistically detectable effect.

quantitative
The numerical representation of some object. A quantitative variable is any variable that is measured using numbers.

You should note that the term *statistics* encompasses both descriptive analyses of your data and inferential analyses designed to test formal hypotheses. The descriptive statistics that you actually look at can be voluminous. In most write-ups, you carefully select and organize these statistics into summary tables and graphs that show only the most relevant or important information. After you describe the data, you construct specific analyses for each of the research questions or hypotheses raised in your research design. In most analysis write-ups, it's especially critical that you not miss the forest for the trees. If you present too much detail, the reader may not be able to follow the central line of the results. Often extensive analysis details are appropriately relegated to appendices, reserving only the most critical analysis summaries for the body of the report itself.

This chapter discusses the basics of data analysis. I save the topic of data analysis for your research design for the next chapter. However, I'll warn you right now that this is not a statistics text. I'll cover lots of statistics, some elementary and some advanced, but I'm not trying to teach you statistics here. Instead, I'm trying to get you to think about data analysis and how it fits into the broader context of your research.

I'll begin this chapter by discussing **conclusion validity**, the validity of inferences you draw from your data analyses. This will give you an understanding of some of the key principles involved in any research analysis. Then I'll cover the

conclusion validity
The degree to which conclusions you reach about relationships in your data are reasonable.

often-overlooked issue of data preparation. This includes all of the steps involved in cleaning and organizing the data for analysis. I then introduce the basic descriptive statistics and consider some general analysis issues that set the stage for consideration of the analysis of the major research designs in the Chapter 14.

12-1 Conclusion Validity

Of the four types of validity (see also **internal validity**, **construct validity**, and **external validity**), conclusion validity is undoubtedly the least considered and most misunderstood—probably due to the fact that it was originally labeled statistical conclusion validity and you know how even the mere mention of the word *statistics* will scare off most of the human race!

internal validity
The approximate truth of inferences regarding cause-effect or causal relationships.

construct validity
The degree to which inferences can legitimately be made from the operationalizations in your study to the theoretical constructs on which those operationalizations are based.

external validity
The degree to which the conclusions in your study would hold for other persons in other places and at other times.

In many ways, conclusion validity is the most important of the four validity types because it is relevant whenever you are trying to decide whether there is a relationship in your observations (and that's one of the most basic aspects of any analysis). Perhaps I should start with an attempt at a definition:

> Conclusion validity is the degree to which conclusions you reach about relationships in your data are reasonable.

For instance, if you're doing a study that looks at the relationship between socioeconomic status (SES) and attitudes about capital punishment, you eventually want to reach some conclusion. Based on your data, you might conclude that there is a positive relationship—that persons with higher SES tend to have a more positive view of capital punishment, whereas those with lower SES tend to be more opposed. Conclusion validity in this case is the degree to which that conclusion or inference is credible or believable.

Although conclusion validity was originally thought to be a statistical-inference issue, it has become more apparent that it is also relevant in **qualitative** research. For example, in an observational field study of homeless adolescents, a researcher might, on the basis of field notes, see a pattern that suggests that teenagers on the street who use drugs are more likely to be involved in more complex social networks and to interact with a more varied group of people than the nondrug users. Although this conclusion or inference may be based entirely on qualitative observational data, you can ask whether it has conclusion validity, that is, whether it is a reasonable conclusion about the relationship inferred from the observations.

qualitative
The descriptive nonnumerical characteristic of some object. A qualitative variable is a descriptive nonnumerical observation.

Whenever you investigate a relationship, you essentially have two possible conclusions: either there is a relationship in your data, or there isn't. In either case, however, you could be wrong in your conclusion. You might conclude that there is a relationship when in fact, there is not; or you might infer that no relationship exists when in fact one does (but you didn't detect it).

It's important to realize that conclusion validity is an issue whenever you are talking about a relationship, even when the relationship is between some program (or treatment) and some outcome. In other words, conclusion validity also pertains to **causal** relationships. How do you distinguish it from internal validity, which is also involved with causal relationships? Conclusion validity is concerned only with whether there is a relationship; internal validity assumes you have demonstrated a relationship and is concerned with whether that relationship is causal. For instance, in a program evaluation, you might conclude that there is a positive relationship between your educational program and achievement test scores; students in the program get higher scores and students not in the program get lower ones. Conclusion validity is essentially concerned with whether that relationship is a reasonable one or not, given the data. However, it is possible to conclude that, while a relationship exists between the program and outcome, the program didn't cause the outcome. Perhaps some other factor, and not your program, was responsible for the outcome in this study. For instance, the observed differences in the outcome could be due to the fact that the program group was smarter than the comparison group

causal
Pertaining to a cause-effect relationship.

to begin with. Your observed posttest differences between these groups could be due to this initial difference and not be the result of your program. This issue—the possibility that some factor other than your program caused the outcome—is what internal validity is all about. So, it is possible that in a study you can conclude that your program and outcome are related (conclusion validity) and also conclude that the outcome was caused by some factor other than the program (you don't have internal validity).

reliability
The repeatability or consistency of a measure. More technically, reliability is the ratio of the variability in true scores to the variability in the observed scores. In more approximate terms, reliability is the proportion of truth in what you measure as opposed to the proportion of error in measurement.

statistical power
The probability of correctly concluding that there is a treatment or program effect in your data.

I'll begin this discussion by considering the major threats to conclusion validity—the different reasons you might be wrong in concluding that there is or isn't a relationship. You'll see that there are several key reasons why reaching conclusions about relationships is so difficult. One major problem is that it is often hard to see a relationship because your measures or observations have low **reliability**; they are too weak relative to all of the noise in the environment. Another issue is that the relationship you are looking for may be a weak one and seeing it is a bit like looking for a needle in the haystack. Sometimes the problem is that you just didn't collect enough information to see the relationship even if it is there. All of these problems are related to the idea of **statistical power**, so I'll spend some time trying to explain what power is in this context. Finally, you need to recognize that you have some control over your ability to detect relationships, and I'll conclude with some suggestions for improving conclusion validity.

12-1a Threats to Conclusion Validity

threat to conclusion validity
Any factor that can lead you to reach an incorrect conclusion about a relationship in your observations.

A **threat to conclusion validity** is any factor that can lead you to reach an incorrect conclusion about a relationship in your observations. You can essentially make two kinds of errors about relationships:

- You can conclude that there is no relationship when in fact there is. (You missed the relationship or didn't see it.)
- You can conclude that there is a relationship when in fact there is not. (You're seeing things that aren't there!)

Most threats to conclusion validity have to do with the first problem. Why? Maybe it's because it's so hard in most research to find relationships in data in the first place that it's not as big or frequent a problem; researchers tend to have more problems finding the needle in the haystack than seeing things that aren't there! So, I'll divide the threats by the type of error with which they are associated.

Type I Error: Finding a Relationship When There Is Not One (or Seeing Things That Aren't There)

In anything but the most trivial research study, the researcher spends a considerable amount of time analyzing the data for relationships. Of course, it's important to conduct a thorough analysis, but most people are well aware of the fact that if you play with the data long enough, you can often turn up results that support or corroborate your hypotheses. In more everyday terms, you fish for a specific result by analyzing the data repeatedly under slightly differing conditions or assumptions.

.05 level of significance
The significance level. Specifically, alpha is the Type I error, or the probability of concluding that there is a treatment effect when, in reality, there is not.

In statistical analysis, you attempt to determine the probability that your finding is a real one or a chance finding. In fact, you often use this probability to decide whether to accept the statistical result as evidence that there is a relationship. In the social sciences, researchers often use the rather arbitrary value, known as the **.05 level of significance**, to decide whether their result is credible or could be considered a fluke. Essentially, the value .05 means that the result you got could be expected to occur by chance at least 5 times out of every 100 times you ran the statistical analysis.

The probability assumption that underlies most statistical analyses assumes that each analysis is independent of the other. However, that may not be true when you conduct multiple analyses of the same data. For instance, let's say you conduct

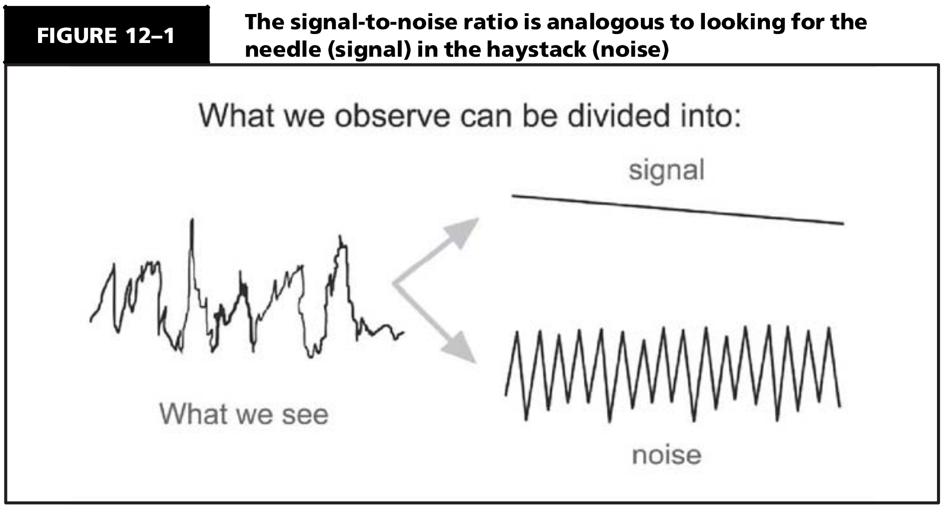
20 statistical tests and for each one you use the .05 level criterion for deciding whether you are observing a relationship. For each test, the odds are 5 out of 100 that you will see a relationship even if there is not one there. (That's what it means to say that the result could be due to chance.) Odds of 5 out of 100 are equal to the fraction 5/100, which is also equal to 1 out of 20. Now, in this example, you conduct 20 separate analyses. Let's say that you find that of the 20 results, only 1 is statistically significant at the .05 level. Does that mean you have found a statistically significant relationship? If you had done only the one analysis, you might conclude that you found a relationship in that result. However, if you did 20 analyses, you would expect to find one of them significant by chance alone, even if no real relationship exists in the data. This threat to conclusion validity is called the **fishing and the error rate problem**. The basic problem is that you were fishing by conducting multiple analyses and treating each one as though it was independent. Instead, when you conduct multiple analyses, you should adjust the error rate (the significance level) to reflect the number of analyses you are doing. The bottom line here is that you are more likely to see a relationship when there isn't one when you keep reanalyzing your data and don't take your fishing into account when drawing your conclusions.

fishing and the error rate problem
A problem that occurs as a result of conducting multiple analyses and treating each one as independent.

Type II Error: Finding No Relationship When There Is One (or Missing the Needle in the Haystack) When you're looking for the needle in the haystack, you essentially have two basic problems: the tiny needle and too much hay. You can view this as a signal-to-noise ratio problem (Figure 12–1). The signal is the needle—the relationship you are trying to see. The noise consists of all of the factors that make it hard to see the relationship.

There are several important sources of noise, each of which is a threat to conclusion validity. One important threat is *low reliability of measures* (see Section 3-2, Reliability). This can be due to many factors, including poor wording of questions, bad instrument design or layout, illegibility of field notes, and so on. In studies where you are evaluating a program, you can introduce noise through *poor reliability of treatment implementation.* If the program doesn't follow the prescribed procedures or is inconsistently carried out, it will be harder to see relationships between the program and other factors like the outcomes. Noise caused by *random irrelevancies in the setting* can also obscure your ability to see a relationship. In a classroom context, the traffic outside the room, disturbances in the hallway, and countless other irrelevant events can distract the researcher or the participants. The types of people you have in your study can also make it harder to see relationships. The threat here is due to the *random heterogeneity of respondents.* If you have a diverse group of respondents, group members are likely to vary more widely on your measures or

FIGURE 12–1 **The signal-to-noise ratio is analogous to looking for the needle (signal) in the haystack (noise)**



observations. Some of their variability may be related to the phenomenon you are looking at, but at least part of it is likely to constitute individual differences that are irrelevant to the relationship you observe.

All of these threats add variability into the research context and contribute to the noise relative to the signal of the relationship you are looking for. But noise is only one part of the problem. You also have to consider the issue of the signal—the true strength of the relationship. One broad threat to conclusion validity tends to subsume or encompass all of the noise-producing factors mentioned and also takes into account the strength of the signal, the amount of information you collect, and the amount of risk you're willing to take in making a decision about whether a relationship exists. This threat is called *low statistical power.* Because this idea is so important in understanding how to make decisions about relationships, I have included a separate discussion of statistical power later in this chapter.

Problems That Can Lead to Either Conclusion Error Every analysis is based on a variety of assumptions about the nature of the data, the procedures you use to conduct the analysis, and the match between these two. If you are not sensitive to the assumptions behind your analysis, you are likely to draw erroneous conclusions about relationships. In quantitative research, this threat is referred to as the *violated assumptions of statistical tests.* For instance, many statistical analyses are based on the assumption that the data is distributed normally—that the population from which it is drawn would be distributed according to a normal or bell-shaped curve. If that assumption is not true for your data and you use that statistical test, you are likely to get an incorrect estimate of the true relationship. It's not always possible to predict what type of error you might make—seeing a relationship that isn't there or missing one that is.

I believe that the same problem can occur in qualitative research as well. There are assumptions, some of which you may not even realize, behind all qualitative methods. For instance, in interview situations you might assume that the respondents are free to say anything they wish. If that is not true—if the respondent is under covert pressure from supervisors to respond in a certain way—you may erroneously see relationships in the responses that aren't real and/or miss ones that are.

The threats discussed in this section illustrate some of the major difficulties and traps that are involved in one of the most basic areas of research—deciding whether there is a relationship in your data or observations. So, how do you attempt to deal with these threats? Section 12-1c details a number of strategies for improving conclusion validity through minimizing or eliminating these threats.

12-1b Statistical Power

Warning! I am about to launch into some technical, statistical gibberish. I think I can explain statistical power in a way that is understandable, but you will need to have a little patience. This is probably a good section for you read while intermittently applying some classic relaxation techniques—deep breathing, meditation, and so on. I highly recommend reading this in small doses with frequent meditative breaks. So here goes . . .

Four interrelated components influence the conclusions you might reach from a statistical test in a research project:

- *Sample size,* or the number of units (people) accessible to the study
- *Effect size,* or the salience of the treatment relative to the noise in measurement
- **Alpha level** (α, or significance level), or the odds that the observed result is due to chance
- *Power,* or the odds that you will observe a treatment effect when it occurs

alpha level
The significance level. Specifically, alpha is the Type I error, or the probability of concluding that there is a treatment effect when, in reality, there is not.

If you know the values for any three of these components, it is possible to compute the value of the fourth. For instance, you might want to determine what a

FIGURE 12–2 **The statistical inference decision matrix**

In reality → What we conclude ↓	Null true Alternative false In reality... • There is no real program effect • There is no difference, gain • Our theory is wrong	Null false Alternative true In reality... • There is a real program effect • There is a difference, gain • Our theory is correct
Accept null Reject alternative We say... • There is no real program effect • There is no difference, gain • Our theory is wrong	$1 - \alpha$ (e.g., .95) THE CONFIDENCE LEVEL The odds of saying there is no effect or gain when in fact there is none # of times out of 100 when there is no effect, we'll say there is none	β (e.g., .20) TYPE II ERROR The odds of saying there is no effect or gain when in fact there is one # of times out of 100 when there is an effect, we'll say there is none
Reject null Accept alternative We say... • There is a real program effect • There is a difference, gain • Our theory is correct	α (e.g., .05) TYPE I ERROR The odds of saying there is an effect or gain when in fact there is none # of times out of 100 when there is no effect, we'll say there is one	$1 - \beta$ (e.g., .80) POWER The odds of saying there is an effect or gain when in fact there is one # of times out of 100 when there is an effect, we'll say there is one

reasonable sample size would be for a study. If you could make reasonable estimates of the effect size, alpha level, and power, it would be simple to compute (or, more likely, look up in a table) the sample size. Fortunately, there are conventional guidelines to give you a starting point for your alpha level and power. These will be presented later in this chapter. However, since the advent of a research synthesis strategy called *meta-analysis* in the late 1970s, the importance of carefully considering effect sizes in both the planning and interpretive stages of research has been elevated considerably. In addition, there is much more appreciation of the distinction between practical or clinical significance and statistical significance. You will learn more about these matters when you get to Chapter 16, but for now I advise you to look carefully at prior studies in your topic area to see if there might be any consensus about what constitutes a meaningful change on a measure you are considering using, or how strong a relationship between constructs would be predicted by a reasonable theory or prior studies.

Some of these components are easier to manipulate than others, depending on the project's circumstances. For example, if the project is an evaluation of an educational program or counseling program with a specific number of available consumers, the sample size is set or predetermined, or, if the drug dosage in a program has to be small due to its potential negative side effects, the effect size may consequently be small. The goal is to achieve a balance of the four components that allows the maximum level of power to detect an effect if one exists, given programmatic, logistical, or financial constraints on the other components.

Figure 12–2 shows the basic decision matrix involved in any statistical conclusion. What do I mean by a *decision matrix*? It is a table that shows what decisions or conclusions you can reach from any statistical analysis and how these are related to reality. All statistical conclusions involve constructing two mutually exclusive hypotheses, termed the null (labeled H_0) and alternative (labeled H_1) hypothesis (see Chapter 1). Together, the **hypotheses** describe all possible outcomes with respect to the inference. The central decision involves determining which hypothesis to accept and which to reject. (Because the two are mutually exclusive and

hypothesis
A specific statement of prediction.

exhaustive, you will always have to accept one and reject the other.) For instance, in the typical case, the null hypothesis might be:

H_0: Program Effect = 0

whereas the alternative might be:

H_1: Program Effect <> 0

When you conduct a statistical analysis to test this hypothesis, you have to accept one of these: either your program works (H_1) or it doesn't (H_0). When you accept one, you automatically reject the other. This is what you conclude, but things are a little more complicated than this. Just because you conclude something doesn't make it true. (Remember your parents telling you this at some point?) Reality often has a way of being different from what we think it is. So, the other aspect of your decision has to do with the reality of the conclusion. Like your statistical conclusion, this can be expressed in only two ways—the null hypothesis is true or the alternative one is true. That's it. Those are the only options.

You should now be getting an inkling of where I am going with this somewhat convoluted presentation. The statistical decision matrix shown in Figure 12–2 shows the four possible options made by combining each possible conclusion with each possible reality.

Figure 12–2 is a complex figure that you should take some time to study. In fact, I think you should prop it up on a table, sit down in front of it cross-legged, and just stare at it for a few hours.

First, look at the header row (the shaded area). This row depicts reality—whether there really is a program effect, difference, or gain. Of course, the problem is that you never know for sure what is really happening (unless you're God). Nevertheless, because you have set up mutually exclusive hypotheses, one must be right and one must be wrong. Therefore, consider this the view from God's position, knowing which hypothesis is correct—isn't it great to get a chance to play God? The first column of the 2 × 2 table shows the case where the program does not have an effect; the second column shows where it does have an effect or make a difference.

The left header column describes the world mortals live in. Regardless of what's true, you have to make decisions about which of your hypotheses is correct. This header column describes the two decisions you can reach—that your program had no effect (the first row of the 2 × 2 table) or that it did have an effect (the second row).

Now, let's examine the cells of the 2 × 2 table. The first thing to recognize is that two of the cells represent a correct conclusion and two of them represent an error. If you say there is no relationship or effect (accept the null) and there is in reality no relationship or effect, you're in the upper-left cell and you are correct. If you say a program effect exists (accept the alternative) and there is in reality a program effect, you're in the lower-right cell and you are correct. Those are the two possible correct conclusions. Now consider the two errors. If you say there is a relationship or effect and there is not, you're in the cell on the lower left and you're wrong. We call this type of error a Type I error. (Pretty original, huh?) It is like seeing things that aren't there (as described earlier in this chapter). You're seeing an effect but you're wrong. If you say there is no effect and in fact there is an effect, you're in the cell on the upper right and you're wrong. We call this type of error—guess what—a Type II error. This type of error is like not seeing the needle in the haystack as described earlier in this chapter. There is an effect in reality, but you couldn't see it.

Each cell shows the Greek symbol used to name that cell. (You knew there had to be Greek symbols here. Statisticians can't even write their own names without using Greek letters.) Notice that the columns sum to 1 ($\alpha + (1 - \alpha) = 1$ and $\beta + (1 - \beta) = 1$). (Having trouble adding in Greek? Just keep in mind that $\alpha - \alpha = 0$, no matter what language you use for the symbol α.) Why can you sum down the columns, but not across the rows? Because if one column is true, the other is irrelevant; if the program has a real effect (the right column), it can't, at the same time, not have one. Reality can be in only one column or the other (even though, given the reality, you could be in either row). Therefore, the odds or probabilities have to sum to 1 for each column because the two rows in each column describe the only possible decisions (accept or reject the null/alternative) for each possible reality.

Below the Greek symbol is a typical value for that cell. You should especially note the values in the bottom two cells. The value of α is typically set at .05 in the social sciences. A newer, but growing, tradition is to try to achieve a standard for statistical power of at least .80. Below the typical values is the name typically given for that cell (in caps). If you weren't paying attention a few paragraphs ago, I'll give you one more chance to note that two of the cells describe errors—you reach the wrong conclusion—and in the other two cells, you reach the correct conclusion. Sometimes it's hard to remember which error is Type I and which is Type II. If you keep in mind that Type I is the same as the α, or significance level, it might help you remember that both involve seeing things that aren't there. People are more likely to be susceptible to a Type I error because they almost always want to conclude that their program works. If they find a statistical effect, they tend to advertise it loudly. On the other hand, people probably check more thoroughly for Type II errors because when they find that the program was not demonstrably effective, they immediately want to find out why. (In this case, you might hope to show that you had low power and high β—that the odds of saying there was no treatment effect even when there was were too high.) Following the capitalized common name are two ways of describing the value of each cell: one in terms of outcomes and one in terms of theory testing. In italics, I give an example of how to express the numerical value in words.

To better understand the strange relationships between the two columns, think about what happens if you want to increase your power in a study. As you increase power, you increase the chances that you are going to find an effect if it's there (wind up in the bottom row). However, if you increase the chances of winding up in the bottom row, you must, at the same time, increase the chances of making a Type I error! Although you can't sum to 1 across rows, there is clearly a relationship. Since you usually want high power *and* low Type I error, you should be able to appreciate that you have a built-in tension here. (Now might be a good moment for a meditation break. Reread the last paragraph over and over until it begins to make sense!)

We often talk about alpha (α) and beta (β) using the language of higher and lower. For instance, you might talk about the advantages of a higher or lower α level in a study. You have to be careful about interpreting the meaning of these terms. When you talk about *higher* α levels, you mean that you are *increasing* the chance of a Type I error. Therefore, a *lower* α level actually means that you are conducting a *more rigorous* test.

With all of this in mind, let's consider a few common associations evident in the table. You should convince yourself of the following (each of these is it's own little meditation exercise):

- The lower the α, the lower the power. The higher the α, the higher the power.
- The lower the α, the less likely it is that you will make a Type I error (reject the null when it's true).
- The lower the α, the more rigorous the test.
- An α of .01 (compared with .05 or .10) means the researcher is being relatively careful and is willing to risk being wrong only 1 in a 100 times in rejecting the null when it's true (saying there's an effect when there really isn't).
- An α of .01 (compared with .05 or .10) limits the chances of ending up in the bottom row, of concluding that the program has an effect. This means that statistical power and the chances of making a Type I error are lower.
- An α of .01 means there is a 99 percent chance of saying there is no difference when there in fact is no difference (being in the upper-left box).

- Increasing α (for example from .01 to .05 or .10) increases the chances of making a Type I error (saying there is a difference when there is not), decreases the chances of making a Type II error (saying there is no difference when there is), and decreases the rigor of the test.
- Increasing α (for example from .01 to .05 or .10) increases power because you will be rejecting the null more often (accepting the alternative) and consequently, when the alternative is true, there is a greater chance of accepting it (power).

12-1c Improving Conclusion Validity

So let's say you have a potential problem ensuring that you reach credible conclusions about relationships in your data. What can you do about it? Here are some general guidelines you can follow in designing your study that will help improve conclusion validity.

- *Good statistical power.* The rule of thumb in social research is that you want statistical power to be greater than 0.8 in value (see the previous discussion on statistical power). That is, you want to have at least 80 chances out of 100 of finding a relationship when there is one. As pointed out in the discussion of statistical power, several factors interact to affect power. One thing you can usually do is collect more information—use a larger sample size. Of course, you have to weigh the gain in power against the time and expense of having more participants or gathering more data. The second thing you can do is increase your risk of making a Type I error—increase the chance that you will find a relationship when it's not there. In practical terms, you can do that statistically by raising the alpha level. For instance, instead of using a 0.05 significance level, you might use 0.10 as your cutoff point. Finally, you can increase the effect size. Since the effect size is a ratio of the signal of the relationship to the noise in the context, there are two broad strategies here. To raise the signal, you can increase the salience of the relationship itself. This is especially true in experimental contexts where you are looking at the effects of a program or treatment. If you increase the dosage of the program (for example, increase the hours spent in training or the number of training sessions), it will be easier to see the effect when the treatment is stronger. The other option is to decrease the noise (or, put another way, increase reliability).
- *Good reliability.* Reliability (see Section 3-30) is related to the idea of noise or error that obscures your ability to see a relationship. In general, you can improve reliability by doing a better job of constructing measurement instruments, by increasing the number of questions on a scale, or by reducing situational distractions in the measurement context. When you improve reliability, you reduce noise, which increases your statistical power and improves conclusion validity.
- *Good implementation.* When you are studying the effects of interventions, treatments, or programs, you can improve conclusion validity by ensuring good implementation. You accomplish this by training program deliverers and standardizing the protocols for administering the program.

12-2 Data Preparation

Data preparation involves checking or logging the data in, checking the data for accuracy, entering the data into the computer, transforming the data, and developing and documenting a database structure that integrates the various measures.

12-2a Logging the Data

In any research project, you might have data coming from several different sources at different times as in the following examples:

- Mail survey returns
- Coded-interview data
- Pretest or posttest data
- Observational data

In all but the simplest of studies, you need to set up a procedure for logging the information and keeping track of it until you are ready to do a comprehensive data analysis. Different researchers differ in how they keep track of incoming data. In most cases, you will want to set up a database that enables you to assess, at any time, which data is already entered and which still needs to be entered. You could do this with any standard computerized database program (such as Microsoft Access or Claris Filemaker), although this requires familiarity with such programs, or you can accomplish this using standard statistical programs (for example, SPSS, SAS, Minitab, or Datadesk) and running simple descriptive analyses to get reports on data status. It is also critical that the data analyst retain the original data records—returned surveys, field notes, test protocols, and so on—for a reasonable time. Most professional researchers retain such records for at least 5 to 7 years. For important or expensive studies, the original data might be stored in a data archive. The data analyst should always be able to trace a result from a data analysis back to the original forms on which the data was collected. A database for logging incoming data is a critical component in good research record keeping.

12-2b Checking the Data for Accuracy

As soon as you receive the data, you should screen it for accuracy. In some circumstances, doing this right away allows you to go back to the sample to clarify any problems or errors. You should ask the following questions as part of this initial data screening:

- Are the responses legible/readable?
- Are all important questions answered?
- Are the responses complete?
- Is all relevant contextual information included (for example, date, time, place, and researcher)?

In most social research, quality of measurement is a major issue. Ensuring that the data collection process does not contribute inaccuracies helps ensure the overall quality of subsequent analyses.

12-2c Developing a Database Structure

The database structure is the manner in which you intend to store the data for the study so that it can be accessed in subsequent data analyses. You might use the same structure you used for logging in the data; or in large complex studies, you might have one structure for logging data and another for storing it. As mentioned previously, there are generally two options for storing data on computer: database programs and statistical programs. Usually database programs are the more complex of the two to learn and operate, but they allow you greater flexibility in manipulating the data.

In every research project, you should generate a printed **codebook** that describes the data and indicates where and how it can be accessed. At a minimum, the codebook should include the following items for each variable:

codebook
A written description of the data that describes each variable and indicates where and how it can be accessed.

- Variable name
- Variable description

- Variable format (number, data, text)
- Instrument/method of collection
- Date collected
- Respondent or group
- Variable location (in database)
- Notes

The codebook is an indispensable tool for the analysis team. Together with the database, it should provide comprehensive documentation that enables other researchers who might subsequently want to analyze the data to do so without any additional information.

12-2d Entering the Data into the Computer

You can enter data into a computer in a variety of ways. Probably the easiest is to just type the data in directly. To ensure a high level of data accuracy, you should use a procedure called **double entry**. In this procedure, you enter the data once. Then, you use a special program that allows you to enter the data a second time and checks the second entries against the first. If there is a discrepancy, the program notifies you and enables you to determine which is the correct entry. This double-entry procedure significantly reduces entry errors. However, these double-entry programs are not widely available and require some training. An alternative is to enter the data once and set up a procedure for checking the data for accuracy. For instance, you might spot-check records on a random basis.

double entry
An automated method for checking data-entry accuracy in which you enter data once and then enter it a second time, with the software automatically stopping each time a discrepancy is detected until the data enterer resolves the discrepancy. This procedure assures extremely high rates of data entry accuracy, although it requires twice as long for data entry.

After you enter the data, you will use various programs to summarize the data that enable you to check that all the data falls within acceptable limits and boundaries. For instance, such summaries enable you to spot whether there are persons whose age is 601 or whether anyone entered a 7 where you expect a 1 to 5 response.

12-2e Data Transformations

After the data is entered, it is almost always necessary to transform the raw data into variables that are usable in the analyses. This is often accomplished by using a transformation, that is, by transforming the original data into a form that is more useful or usable. There are a variety of transformations that you might perform. The following are some of the more common ones:

- *Missing values.* Many analysis programs automatically treat blank values as missing. In others, you need to designate specific values to represent missing values. For instance, you might use a value of −99 to indicate that the item is missing. You need to check the specific program you are using to determine how to handle missing values.
- *Item reversals.* On scales and surveys, the use of reversal items (see Chapters 4 and 5) can help reduce the possibility of a response set. When you analyze the data, you want all scores for scale items to be in the same direction, where high scores mean the same thing and low scores mean the same thing. In these cases, you have to reverse the ratings for some of the scale items. For instance, let's say you had a 5-point response scale for a self-esteem measure where 1 meant strongly disagree and 5 meant strongly agree. One item is "I generally feel good about myself." If respondents strongly agree with this item, they will put a 5, and this value would be indicative of higher self-esteem. Alternatively, consider an item like "Sometimes I feel like I'm not worth much as a person." Here, if a respondent strongly agrees by rating this a 5, it would indicate low self-esteem. To compare these two items, you would reverse the scores. (Probably you'd reverse the latter item so that high values always indicate higher self-esteem.) You want a transformation where if the original value was 1, it's changed to 5; 2 is changed to 4; 3 remains the same; 4 is changed to 2; and 5 is

changed to 1. Although you could program these changes as separate statements in most programs, it's easier to do this with a simple formula like the following:

$$\text{New Value} = (\text{High Value} + 1) - \text{Original Value}$$

In our example, the *high value* for the scale is 5; so to get the new (transformed) scale value, you simply subtract the *original value* on each reversal item from 6 (that is, 5 + 1).

- *Scale totals.* After you transform any individual scale items, you will often want to add or average across individual items to get a total score for the scale.
- *Categories.* You will want to collapse many variables into categories. For instance, you may want to collapse income estimates (in dollar amounts) into income ranges.

12-2f Dealing with Missing Data

Missing data can be a major threat to conclusion validity, depending on how much data is missing and why it is missing. As always, prevention is better than cure, and when it comes to research, good planning can help prevent missing data from becoming a major issue. If you have carefully planned your study so that it is feasible (for example, you have good access to your sample and your study is not too long, boring, or difficult, and might even be rewarding in some way), then you may have an inconsequential problem and can proceed with your analysis as planned (for example, 5 percent or fewer missing cases from a well-powered large study). If your study is at the other extreme and you end up with a relatively sparse representation of your sampling frame, then it will be difficult to convince anyone that what you find in your data is a valid representation of the sample much less the population, no matter how clever your analysis is. In that case you might want to go back to the design drawing board and consider this a valuable pilot. More often than not, you will be somewhere in the middle, in a situation where you have invested available resources and think you have enough data to analyze. The questions that need to be addressed in such cases are: "Who or what is missing, why, and what effect might it have on the results of the analysis?" So in effect you have to address an empirical question about the quality of your data before you address the substantive questions of your study.

There are many reasons that data may be missing, and you may find missing data issues at the level of the item (for example, some items might not be relevant, may be badly worded, or may be so sensitive that people do not want to answer) or at the level of the case (for example, data forms may have been missing or lost; some respondents may have limited reading ability or be fatigued, ill, or disabled; or some may have low motivation to help). For example, imagine the challenges of getting complete data if you happen to want to study a 10-session method of therapy for families with troubled adolescents. In such a study, you may have to deal with missing data for certain questions, forms, therapists, dates, families, or individual family members. It is wise to examine early returns on a study, especially a longitudinal study. In addition to alerting you to any potential abnormalities in data **distributions**, you may detect an obvious problem (for example, the last page of a questionnaire). If you have found such an issue early in the data collection process, you can possibly take steps to correct it.

distribution
The manner in which a variable takes different values in your data.

If there is nothing that can be done procedurally to get more complete reports, then you must work with what you have. Fortunately, methodologists have provided us with some good ideas about how to assess the impact of "missingness" as well as procedures for replacing missing values with estimates. The essential issue is whether the missing data can be considered to be randomly missing. That is, if there is a systematic pattern in the missing data related to important characteristics of your sample or relationships among variables, then your conclusions will be biased (less valid). This type of situation is often referred to as one of *nonignorable missing data*. The default mechanism for handling missing data in most statistical programs is to delete all cases that have missing data on any variable. This procedure, known as *listwise deletion*, leaves you with just the complete cases to analyze. If the resulting complete cases are essentially a random subsample of your entire dataset, then you can proceed with the analysis you planned to do and not worry about introducing bias. This may be the rare case, however, and most researchers will want to consider alternatives to use as many cases as possible. Let's first consider the case of the missing scale item, and then the situation when one or more variables have missing data.

If you are using well-developed multi-item scales and have the scoring manuals or original articles describing the measure, then you may already have a standardized way to estimate the total score from the items that were completed. However, if you do not have the benefit of standardized procedures for handling missing data for a particular scale, or if your missing data is related to individual items that are not part of larger summary scales, you will need to examine the pattern of missingness in your data and consider your options. SPSS includes a missing values analysis module that allows you to examine patterns of data completion in a descriptive way, and it also includes a test that will allow you to see if your data significantly depart from the missing completely at random assumption (Little's MCAR test). If the chi-square test statistic is not significant, then you can assume that the data are missing completely at random and proceed with confidence using the listwise deletion procedure.

imputation
Substituting an estimated value for a missing one so that an analysis can include the variable.

If it appears that the missing data are not random, then a method of substituting an estimate can be considered. Substituting an estimated value is called **imputation**. Many people have heard the rule of thumb advising that you "plug in the **mean**," but research has shown that this method is biased, as are other methods, with some important exceptions. Two methods that can provide relatively unbiased estimates for missing values are maximum likelihood procedures and multiple imputation methods (Allison, 2002). Popular software, such as SPSS and SAS, include procedures for generating substitute values based on one or both of these methods. Finally, structural equation modeling programs such as AMOS can be used to generate unbiased estimates.

12-3 Descriptive Statistics

descriptive statistics
Statistics used to describe the basic features of the data in a study.

Descriptive statistics describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data.

Descriptive statistics present quantitative descriptions in a manageable form. In a research study, you may have many measures, or you might measure a large number of people on any given measure. Descriptive statistics help you summarize large amounts of data in a sensible way. Each descriptive statistic reduces data into a simpler summary. For instance, consider a simple number used to summarize how well a batter is performing in baseball—the batting average. This single number is the number of hits divided by the number of times at bat (reported to three significant digits). A batter who is hitting .333 is getting a hit one time in every three at bats. One batting .250 is hitting one time in four. The single number describes a large number of discrete events. Or consider the scourge of many students—the grade point average (GPA). This single number describes the general performance of a student across a potentially wide range of course experiences.

Every time you try to describe a large set of observations with a single indicator, you run the risk of distorting the original data or losing important detail. The batting average doesn't tell you whether batters hit home runs or singles. It doesn't tell whether they've been in a slump or on a streak. The GPAs don't tell you whether the students were in difficult courses or easy ones, or whether the courses were in their major field or in other disciplines. Even given these limitations, descriptive statistics provide a powerful summary that enables comparisons across people or other units.

A single variable has three major characteristics that are typically described as follows:

- Distribution
- **Central tendency**
- **Dispersion**

In most situations, you would describe all three of these characteristics for each of the variables in your study.

central tendency
An estimate of the center of a distribution of values. The most usual measures of central tendency are the mean, median, and mode.

dispersion
The spread of the values around the central tendency. The two common measures of dispersion are the range and the standard deviation.

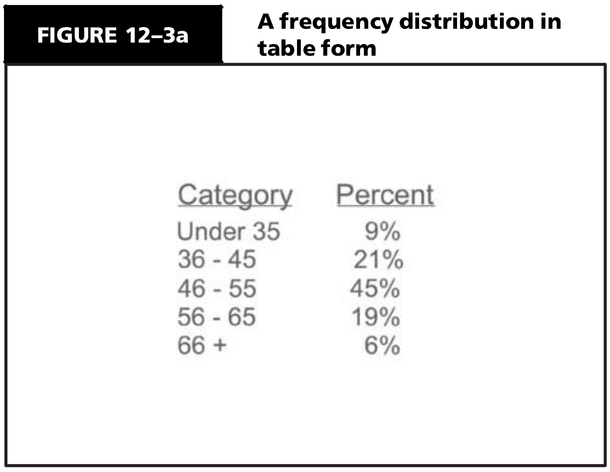
12-3a The Distribution

The distribution is a summary of the frequency of individual values or ranges of values for a variable. The simplest distribution lists every value of a variable and the number of persons who had each value. For instance, a typical way to describe the distribution of college students is by year in college, listing the number or percent of students at each of the four years. Or, you describe gender by listing the number or percent of males and females. In these cases, the variable has few enough values that you can list each one and summarize how many sample cases had the value. But what do you do for a variable like income or GPA? These variables have a large number of possible values, with relatively few people having each one. In this case, you group the raw scores into categories according to ranges of values. For instance, you might look at GPA according to the letter grade ranges, or you might group income into four or five ranges of income values.

One of the most common ways to describe a single variable is with a **frequency distribution**. Depending on the particular variable, all of the data values might be represented, or you might group the values into categories first. (For example, with age, price, or temperature variables, it is usually not sensible to determine the frequencies for each value. Rather, the values are grouped into ranges and the frequencies determined.) Frequency distributions can be depicted in two ways, as a table or as a graph. Figure 12–3a shows an age frequency distribution with five categories of age ranges defined. The same frequency distribution can be depicted in a graph, as shown in Figure 12–3b. This type of graph is often referred to as a histogram or bar chart.

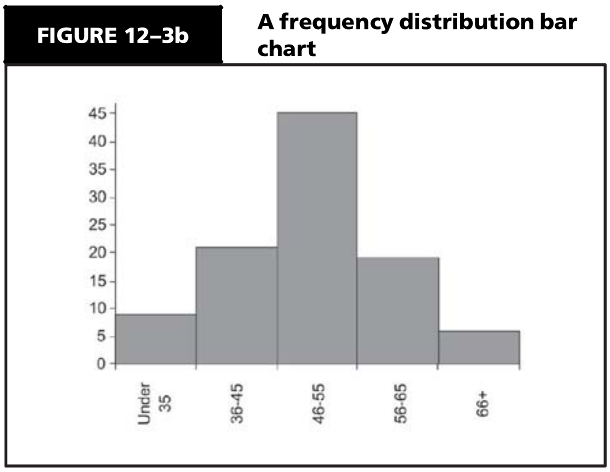
frequency distribution
A summary of the frequency of individual values or ranges of values for a variable.

FIGURE 12–3a A frequency distribution in table form



Category	Percent
Under 35	9%
36 - 45	21%
46 - 55	45%
56 - 65	19%
66 +	6%

FIGURE 12–3b A frequency distribution bar chart



Distributions can also be displayed using percentages. For example, you could use percentages to describe the following:

- Percentage of people in different income levels
- Percentage of people in different age ranges
- Percentage of people in different ranges of standardized test scores

12-3b Central Tendency

The central tendency of a distribution is an estimate of the center of a distribution of values. There are three major types of estimates of central tendency:

- Mean
- Median
- Mode

mean
A description of the central tendency in which you add all the values and divide by the number of values.

The **mean**, or average is probably the most commonly used method of describing central tendency. To compute the mean, all you do is add up all the values and divide by the number of values. For example, the mean, or average, quiz score is determined by summing all the scores and dividing by the number of students taking the exam. Consider the test score values:

15, 20, 21, 20, 36, 15, 25, 15

The sum of these eight values is 167, so the mean is 167/8 = 20.875.

median
The middle number in a series of numbers or the score found at the exact middle or fiftieth percentile of the set of values. One way to compute the median is

The **median** is the score found at the exact middle of the set of values. One way to compute the median is to list all scores in numerical order and then locate the score in the center of the sample. For example, if there are 500 scores in the list, score number 250 would be the median. If you order the eight scores shown previously, you would get

15, 15, 15, 20, 20, 21, 25, 36

There are eight scores and score number 4 and number 5 represent the halfway point. Since both of these scores are 20, the median is 20. If the two middle scores had different values, you would have to interpolate to determine the median.

mode
The most frequently occurring value in the set of scores.

The **mode** is the most frequently occurring value in the set of scores. To determine the mode, you might again order the scores as shown previously and then count each one. The most frequently occurring value is the mode. In our example, the value 15 occurs three times and is the mode. In some distributions, there is more than one modal value. For instance, in a bimodal distribution, two values occur most frequently.

Notice that for the same set of eight scores, we got three different values—20.875, 20, and 15—for the mean, median, and mode, respectively. If the distribution is truly normal (bell-shaped), the mean, median, and mode are all equal to each other.

range
The highest value minus the lowest value.

12-3c Dispersion or Variability

standard deviation
The spread or variability of the scores around their average in a *single sample.* The standard deviation, often abbreviated sd, is mathematically the square root of the variance. The standard deviation and variance both measure dispersion, but because the standard deviation is measured in the same units as the original measure and the variance is measured in squared units, the standard deviation is usually more directly interpretable and meaningful.

Dispersion refers to the spread of the values around the central tendency. The two common measures of dispersion are the **range** and the **standard deviation**. The range is simply the highest value minus the lowest value. In the previous example distribution, the high value is 36 and the low is 15, so the range is 36 − 15 = 21.

The standard deviation is a more accurate and detailed estimate of dispersion because an outlier can greatly exaggerate the range (as was true in this example where the single outlier value of 36 stands apart from the rest of the values). The standard deviation shows the relation that set of scores has to the mean of the sample. Again, let's take the set of scores:

15, 20, 21, 20, 36, 15, 25, 15

To compute the standard deviation, you first find the distance between each value and the mean. You know from before that the mean is 20.875. So, the differences from the mean are:

$15 - 20.875 = -5.875$

$20 - 20.875 = -0.875$

$21 - 20.875 = +0.125$

$20 - 20.875 = -0.875$

$36 - 20.875 = +15.125$

$15 - 20.875 = -5.875$

$25 - 20.875 = +4.125$

$15 - 20.875 = -5.875$

Notice that values that are below the mean have negative discrepancies and values above it have positive ones. Next, you square each discrepancy:

$-5.875 \times -5.875 = 34.515625$

$-0.875 \times -0.875 = 0.765625$

$+0.125 \times +0.125 = 0.015625$

$-0.875 \times -0.875 = 0.765625$

$+15.125 \times 15.125 = 228.765625$

$-5.875 \times -5.875 = 34.515625$

$+4.125 \times +4.125 = 17.015625$

$-5.875 \times -5.875 = 34.515625$

Now, you take these squares and sum them to get the sum of squares (SS) value. Here, the sum is 350.875. Next, you divide this sum by the number of scores minus 1. Here, the result is $350.875/7 = 50.125$. This value is known as the **variance**. To get the standard deviation, you take the square root of the variance (remember that you squared the deviations earlier). This would be = 7.079901129253.

variance
A statistic that describes the variability in the data for a variable. The variance is the spread of the scores around the mean of a distribution. Specifically, the variance is the sum of the squared deviations from the mean divided by the number of observations minus 1.

Although this computation may seem convoluted, it's actually quite simple. To see this, consider the formula for the standard deviation shown in Figure 12–4.

In the top part of the ratio, the numerator, notice that each score has the mean subtracted from it, the difference is squared, and the squares are summed. In the bottom part, the denominator, you take the number of scores minus 1. The ratio is

FIGURE 12–4 **Formula for the standard deviation**

$$\sqrt{\frac{\Sigma(X - \bar{X})^2}{(n-1)}}$$

where:
X = each score
$\bar{X}$ = the mean or average
n = the number of values
Σ means we sum across the values

TABLE 12–1 **Table of Descriptive Statistics**

N	8
Mean	20.8750
Median	20.0000
Mode	15.00
Standard deviation	7.0799
Variance	50.1250
Range	21.00

the variance and the square root is the standard deviation. In English, the standard deviation is described as follows:

> The square root of the sum of the squared deviations from the mean divided by the number of scores minus one.

Although you can calculate these univariate statistics by hand, it becomes quite tedious when you have more than a few values and variables. Every statistics program is capable of calculating them easily for you. For instance, I put the eight scores into SPSS and got the results shown in Table 12–1.

This table confirms the calculations I did by hand previously.

The standard deviation allows you to reach some conclusions about specific scores in your distribution. Assuming that the distribution of scores is normal or bell-shaped (or close to it), you can reach the following conclusions:

- Approximately 68 percent of the scores in the sample fall within one standard deviation of the mean.
- Approximately 95 percent of the scores in the sample fall within two standard deviations of the mean.
- Approximately 99 percent of the scores in the sample fall within three standard deviations of the mean.

For instance, since the mean in our example is 20.875 and the standard deviation is 7.0799, you can use the statement listed previously to estimate that approximately 95 percent of the scores will fall in the range of 20.875 (2 × 7.0799) to 20.875 + (2 × 7.0799) or between 6.7152 and 35.0348. This kind of information is critical in enabling you to compare the performance of individuals on one variable with their performance on another, even when the variables are measured on entirely different scales.

12-3d Correlation

correlation
A single number that describes the degree of relationship between two variables.

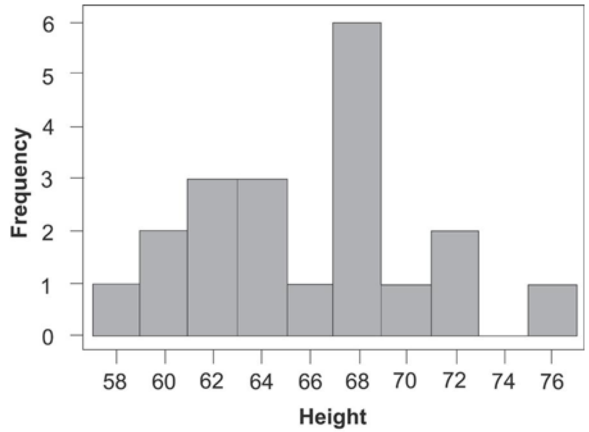
The **correlation** is one of the most common and most useful statistics. A **correlation** is a single number that describes the degree of relationship between two variables. Let's work through an example to show you how this statistic is computed.

Correlation Example Let's assume that you want to look at the relationship between two variables: height (in inches) and self-esteem. Perhaps you have a hypothesis that how tall you are affects your self-esteem. (Incidentally, I don't think you have to worry about the direction of causality here; it's not likely that self-esteem causes your height.) Let's say you collect some information on twenty individuals—all male. (The average height differs for males and females, so to keep this example simple, I'll just use males.) Height is measured in inches. Self-esteem is measured based on the average of 10, 1 to 5 rating items (where higher scores mean

TABLE 12–2 **Hypothetical Data to Demonstrate the Correlation between Height and Self-Esteem**

Person	Height	Self-Esteem
1	68	4.1
2	71	4.6
3	62	3.8
4	75	4.4
5	58	3.2
6	60	3.1
7	67	3.8
8	68	4.1
9	71	4.3
10	69	3.7
11	68	3.5
12	67	3.2
13	63	3.7
14	62	3.3
15	60	3.4
16	63	4.0
17	65	4.1
18	67	3.8
19	63	3.4
20	61	3.6

FIGURE 12–5 **Histogram for the height variable in the example correlation calculation**



higher self-esteem). See Table 12–2 for the data for the 20 cases. (Don't take this too seriously; I made this data up to illustrate what a correlation is.)

Now, let's take a quick look at the histogram for each variable (Figure 12–5 and Figure 12–6).

Table 12–3 shows the descriptive statistics.

Finally, look at the simple bivariate (two-variable) plot (Figure 12–7).

You should immediately see in the bivariate plot that the relationship between the variables is a positive one because if you were to fit a single straight line through the dots it would have a positive slope or move up from left to right. (If you can't see the positive relationship, review Section 1-1c, Types of Relationships.) Since the

TABLE 12–3 **Descriptive Statistics for Correlation Calculation Example**

	Variable						
	Mean	Std. Dev.	Variance	Sum	Minimum	Maximum	Range
Height	65.4	4.40574	19.4105	1308	58	75	17
Self-esteem	3.755	0.426090	0.181553	75.1	3.1	4.6	1.5

FIGURE 12–6 **Histogram for the self-esteem variable in the example correlation calculation**

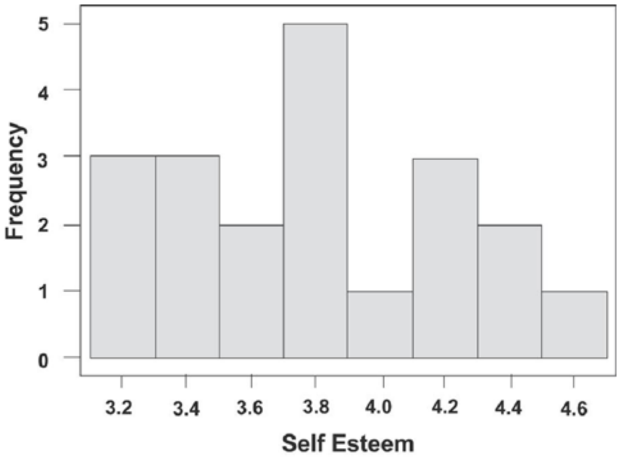
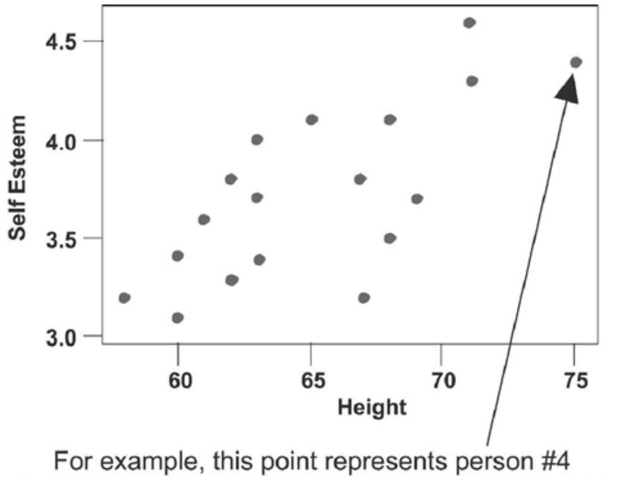


FIGURE 12–7 **Bivariate plot for the example correlation calculation**

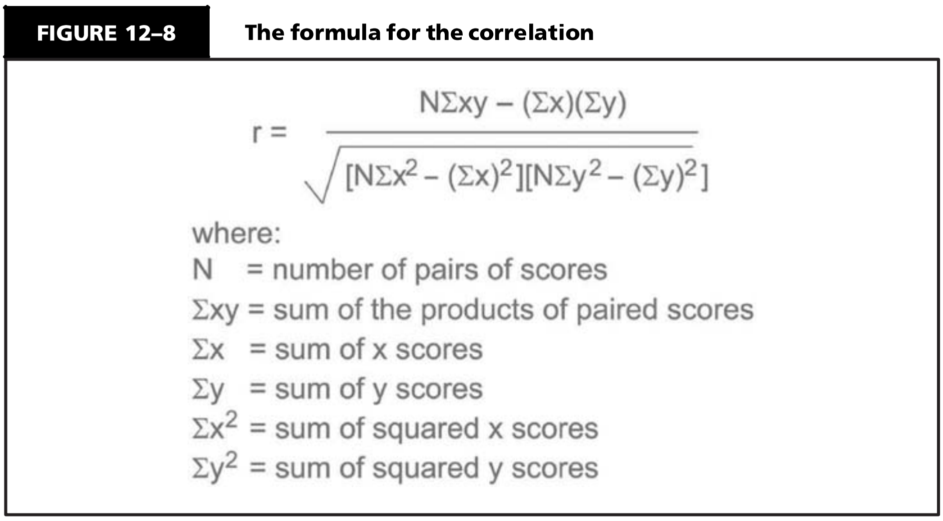


For example, this point represents person #4
who had a height of 75 inches and self esteem of 4.4.

correlation is nothing more than a quantitative estimate of the relationship, you would expect a positive correlation.

What does a positive relationship mean in this context? It means that, in general, higher scores on one variable tend to be paired with higher scores on the other and that lower scores on one variable tend to be paired with lower scores on the other. You should confirm visually that this is generally true in the plot in Figure 12–7.

FIGURE 12–8 **The formula for the correlation**



$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

where:

N = number of pairs of scores

Σxy = sum of the products of paired scores

Σx = sum of x scores

Σy = sum of y scores

Σx^2 = sum of squared x scores

Σy^2 = sum of squared y scores

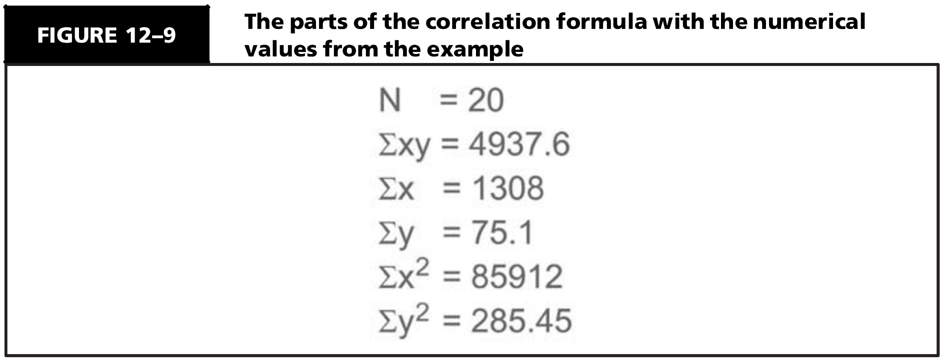
TABLE 12–4 **Computations for the Example Correlation Calculation**

Person	Height (*x*)	Self-Esteem (*y*)	*xy*	x^2	y^2
1	68	4.1	278.8	4624	16.81
2	71	4.6	326.6	5041	21.16
3	62	3.8	235.6	3844	14.44
4	75	4.4	330	5625	19.36
5	58	3.2	185.6	3364	10.24
6	60	3.1	186	3600	9.61
7	67	3.8	254.6	4489	14.44
8	68	4.1	278.8	4624	16.81
9	71	4.3	305.3	5041	18.49
10	69	3.7	255.3	4761	13.69
11	68	3.5	238	4624	12.25
12	67	3.2	214.4	4489	10.24
13	63	3.7	233.1	3969	13.69
14	62	3.3	204.6	3844	10.89
15	60	3.4	204	3600	11.56
16	63	4.0	252	3969	16
17	65	4.1	266.5	4225	16.81
18	67	3.8	254.6	4489	14.44
19	63	3.4	214.2	3969	11.56
20	61	3.6	219.6	3721	12.96
Sum =	**1308**	**75.1**	**4937.6**	**85912**	**285.45**

Calculating the Correlation Now you're ready to compute the correlation value. The formula for the correlation is shown in Figure 12–8.

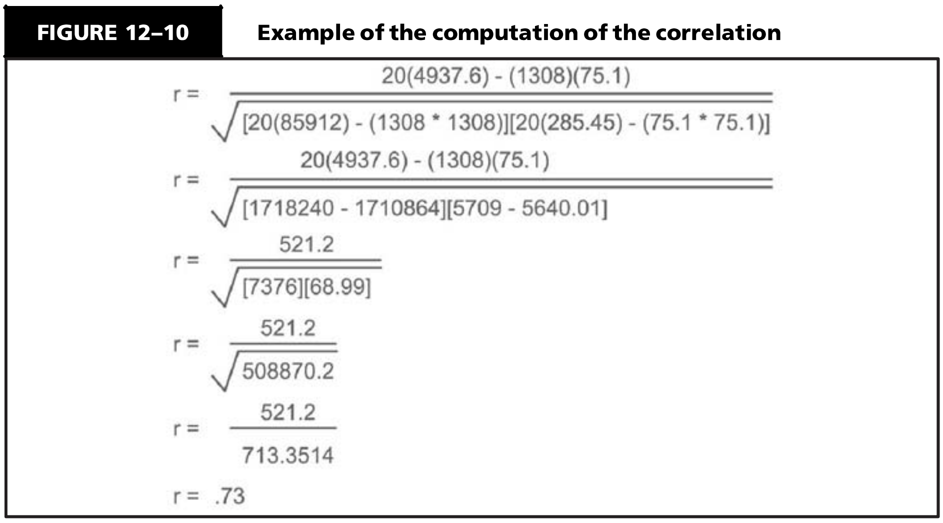
The symbol *r* stands for the correlation. Through the magic of mathematics, it turns out that *r* will always be between −1.0 and +1.0. If the correlation is negative, you have a negative relationship; if it's positive, the relationship is positive. (Pretty clever, huh?) You don't need to know how I came up with this formula unless you want to be a statistician. But you probably will need to know how the formula relates to real data—how you can use the formula to compute the correlation. Let's look at the data you need for the formula. Table 12–4 shows the original data with the other necessary columns.

FIGURE 12–9 The parts of the correlation formula with the numerical values from the example



$$N = 20$$
$$\Sigma xy = 4937.6$$
$$\Sigma x = 1308$$
$$\Sigma y = 75.1$$
$$\Sigma x^2 = 85912$$
$$\Sigma y^2 = 285.45$$

FIGURE 12–10 Example of the computation of the correlation



$$r = \frac{20(4937.6) - (1308)(75.1)}{\sqrt{[20(85912) - (1308 * 1308)][20(285.45) - (75.1 * 75.1)]}}$$

$$r = \frac{20(4937.6) - (1308)(75.1)}{\sqrt{[1718240 - 1710864][5709 - 5640.01]}}$$

$$r = \frac{521.2}{\sqrt{[7376][68.99]}}$$

$$r = \frac{521.2}{\sqrt{508870.2}}$$

$$r = \frac{521.2}{713.3514}$$

$$r = .73$$

The first three columns are the same as those in Table 12–2. The next three columns are simple computations based on the height and self-esteem data in the first three columns. The bottom row consists of the sum of each column. This is all the information you need to compute the correlation. Figure 12–9 shows the values from the bottom row of the table (where $N = 20$ people) as they are related to the symbols in the formula:

Now, when you plug these values into the formula in Figure 12–8, you get the following. (I show it here tediously, one step at a time in Figure 12–10.

So, the correlation for the 20 cases is .73, which is a fairly strong positive relationship. I guess there is a relationship between height and self-esteem, at least in this made-up data!

Testing the Significance of a Correlation After you've computed a correlation, you can determine the probability that the observed correlation occurred by chance. That is, you can conduct a significance test. Most often, you are interested in determining the probability that the correlation is a real one and not a chance occurrence. When you are interested in that, you are testing the mutually exclusive hypotheses:

$$H_0 : r = 0$$
$$H_1 : r \neq 0$$

The easiest way to test this hypothesis is to find a statistics book that has a table of critical values of *r*. (Most introductory statistics texts would have a table like this.) As in all hypothesis testing, you need to first determine the significance level you will use for the test. Here, I'll use the common significance level of $\alpha = .05$. This

TABLE 12–5 **Hypothetical Correlation Matrix for 10 Variables**

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
C1	1.000									
C2	0.274	1.000								
C3	−0.134	−0.269	1.000							
C4	0.201	−0.153	0.075	1.000						
C5	−0.129	−0.166	0.278	−0.011	1.000					
C6	−0.095	0.280	−0.348	−0.378	−0.009	1.000				
C7	0.171	−0.122	0.288	0.086	0.193	0.002	1.000			
C8	0.219	0.242	−0.380	−0.227	−0.551	0.324	−0.082	1.000		
C9	0.518	0.238	0.002	0.082	−0.015	0.304	0.347	−0.013	1.000	
C10	0.299	0.568	0.165	−0.122	−0.106	−0.169	0.243	0.014	0.352	1.000

means that I am conducting a test where the odds that the correlation occurred by chance are no more than 5 out of 100. Before I look up the critical value in a table, I also have to compute the **degrees of freedom (*df*)**. The *df* for a correlation is simply equal to $N - 2$ or, in this example, $20 - 2 = 18$. Finally, I have to decide whether I am doing a one-tailed or two-tailed test (see the discussion in Chapter 1). In this example, because I have no strong prior theory to suggest whether the relationship between height and self-esteem would be positive or negative, I'll opt for the two-tailed test. With these three pieces of information—the significance level ($\alpha = .05$), degrees of freedom ($df = 18$), and type of test (two-tailed)—I can now test the significance of the correlation I found. When I look up this value in the handy little table at the back of my statistics book, I find that the critical value is .4438. This means that if my correlation is greater than .4438 or less than −.4438 (remember, this is a two-tailed test), I can conclude that the odds are less than 5 out of 100 that this is a chance occurrence. Since my correlation of .73 is actually quite a bit higher, I conclude that it is not a chance finding and that the correlation is statistically significant (given the parameters of the test). I can reject the null hypothesis and accept the alternative—I have a statistically significant correlation.

degrees of freedom (*df*)
A statistical term that is a function of the sample size. In the *t*-test formula, for instance, the df is the number of persons in both groups minus 2.

The Correlation Matrix All I've shown you so far is how to compute a correlation between two variables. In most studies, you usually have more than two variables. Let's say you have a study with 10 interval-level variables and you want to estimate the relationships among all of them (between all possible pairs of variables). In this instance, you have 45 unique correlations to estimate (more later on how I knew that). You could do the computations just completed 45 times to obtain the correlations, or you could use just about any statistics program to automatically compute all 45 with a simple click of the mouse.

I used a simple statistics program to generate random data for 10 variables with 20 cases (persons) for each variable. Then, I told the program to compute the correlations among these variables. The results are shown in Table 12–5.

This type of table is called a **correlation matrix**. It lists the variable names (in this case, C1 through C10) down the first column and across the first row. The diagonal of a correlation matrix (the numbers that go from the upper-left corner to the lower right) always consists of ones because these are the correlations between each variable and itself (and a variable is always perfectly correlated with itself). The statistical program I used shows only the lower triangle of the correlation matrix. In every correlation matrix, there are two triangles: the values below and to the left of the diagonal (lower triangle) and above and to the right of the diagonal (upper triangle). There is no reason to print both triangles because the two triangles of a correlation matrix are always mirror images of each other. (The correlation of variable

correlation matrix
A table of correlations showing all possible relationships among a set of variables. The diagonal of a correlation matrix (the numbers that go from the upper-left corner to the lower right) always consists of 1s because these are the correlations between each variable and itself (and a variable is always perfectly correlated with itself). Off-diagonal elements are the correlations of variables represented by the relevant row and column in the matrix.

FIGURE 12–11 **Formula for determining the number of unique correlations, given the number of variables**

$$\frac{N * (N - 1)}{2}$$

x with variable *y* is always equal to the correlation of variable *y* with variable *x*.) When a matrix has this mirror-image quality above and below the diagonal, it is referred to as a **symmetric matrix**. A correlation matrix is always a symmetric matrix.

symmetric matrix
A square (as many rows as columns) table of numbers that describes the relationships among a set of variables, where each variable represents a row or column. Each value in the table represents the relationship between the row and column variable for that cell of the table. The table is "symmetric" when the relationship between a specific row and column variable is identical to the relationship between the same column and row. A correlation matrix is a symmetric matrix.

To locate the correlation for any pair of variables, find the value in the table for the row and column intersection for those two variables. For instance, to find the correlation between variables C5 and C2, look for where row C2 and column C5 is (in this case, it's blank because it falls in the upper triangle area) and where row C5 and column C2 is and, in the second case, the correlation is –.166.

Okay, so how did I know that there are 45 unique correlations when there are 10 variables? There's a simple little formula that tells how many pairs (correlations) there are for any number of variables (Figure 12–11).

N is the number of variables. In the example, I had 10 variables, so I know I have (10 × 9)/2 = 90/2 = 45 pairs.

Other Correlations The specific type of correlation I've illustrated here is known as the **Pearson product moment correlation**. It is appropriate when both variables are measured at an interval level (see Section 3-3a, Why Is Level of Measurement Important?). However, there are other types of correlations for other circumstances. For instance, if you have two ordinal variables, you could use the Spearman rank order correlation (rho) or the Kendall rank order correlation (tau). When one measure is a continuous, interval level one and the other is dichotomous (two-category), you can use the point-biserial correlation. The formulas for these various correlations differ because of the type of data you're feeding into the formulas, but the idea is the same; they estimate the relationship between two variables as a number between –1 and +1.

Pearson product moment correlation
A particular type of correlation used when both variables can be assumed to be measured at an interval level of measurement.

12-3e Cross-Tabulations

You've just learned about the correlation, a very handy statistic that summarizes the relationship between two continuous variables (and sometimes one categorical and one continuous variable). What if you need to examine the relationship between two categorical (nominal) variables? This situation is very common and can be addressed in a straightforward way using something known as **cross-tabulation** (often abbreviated as cross-tabs, and more formally known as *contingency table analysis*). In this analysis, we calculate the number and/or percentage of each possible combination of the two variables to get a sense of both specific combinations of values and the overall association of the variables. No doubt you have seen many, many such tables already, but here we will add a statistical test to help us draw a conclusion. Once again, an example will help to make this clear.

cross-tabulation
A table that describes the frequency and/or percentage for all combinations of two or more nominal or categorical variables.

In Table 12–6, I reproduced the height and self-esteem data, but added two more columns showing each person's gender and also their overall grade-point average, converted from the usual 1.00 to 4.00 scale to just *A* or *B*. This gives us two new variables with just two values possible for each variable. When we cross-tabulate the data, we can see the joint distribution of gender and grades and examine the four resulting cells via the number of cases in each or the percentage of the row, column, or total. In Table 12–7, the simplest of cross-tabulation tables is presented, with just the counts for each cell. In the table we see that there are the same

TABLE 12–6 Hypothetical Height and Self-Esteem Data with Gender and Grades Added

	Person	Height	Self-Esteem	Gender	Grades
1	1.00	68.00	4.10	F	*A*
2	2.00	71.00	4.60	M	*B*
3	3.00	62.00	3.80	F	*A*
4	4.00	75.00	4.40	M	*B*
5	5.00	58.00	3.20	F	*B*
6	6.00	60.00	3.10	M	*B*
7	7.00	67.00	3.80	F	*A*
8	8.00	68.00	4.10	M	*B*
9	9.00	71.00	4.30	F	*B*
10	10.00	69.00	3.70	M	*B*
11	11.00	68.00	3.50	F	*A*
12	12.00	67.00	3.20	M	*B*
13	13.00	63.00	3.70	F	*B*
14	14.00	62.00	3.30	M	*A*
15	15.00	60.00	3.40	F	*B*
16	16.00	63.00	4.00	M	*B*
17	17.00	65.00	4.10	F	*A*
18	18.00	67.00	3.80	M	*A*
19	19.00	63.00	3.40	F	*B*
20	20.00	61.00	3.60	M	*B*

TABLE 12–7 Hypothetical Cross-Tabulation of Two Variables

		Grades		
		A	B	Total
Gender	Female	5	5	10
	Male	2	8	10
Total		7	13	20

The last column and the last row are called the "marginal totals" because they give us totals across the categories and are shown at the margins of the table.

numbers of females and males, 10 of each. However, we might wonder if the *A*s and *B*s are equally distributed because the females obtained more As and fewer Bs than the males. We might next want to consider these numbers as percentages to understand the pattern better, and then we might want a test to help us conclude whether the relationship we think we see is statistically significant (that is, not attributable to chance).

In Table 12–8, you see the same data with row and column percentages added for every cell and for the marginal totals. What do you see as you examine this table? You might notice that the females obtained more than 70 percent of the *A*s earned by this group of students, which is the column percentage of females within grades. Or you might notice that 80 percent of the males obtained *B*s, which is the row percentage of grades within males. Either way, you might be detecting a pattern that makes you wonder about whether there is a "real" (statistically significant) difference between males and females when it comes to grades. A slightly more formal

TABLE 12–8 **Hypothetical Cross-Tabulation of Two Variables with Row and Column Percentages Added**

			Grades		
			A	B	Total
Gender	Female	Count	5	5	10
		% within gender	50.0%	50.0%	100.0%
		% within grades	71.4%	38.5%	50.0%
		% of total	25.0%	25.0%	50.0%
	Male	Count	2	8	10
		% within gender	20.0%	80.0%	100.0%
		% within grades	28.6%	61.5%	50.0%
		% of total	10.0%	40.0%	50.0%
Total		Count	7	13	20
		% within gender	35.0%	65.0%	100.0%
		% within grades	100.0%	100.0%	100.0%
		% of total	35.0%	65.0%	100.0%

TABLE 12–9 **Observed and Expected Frequencies**

			Grades		
			A	B	Total
Gender	Female	Count	5	5	10
		Expected count	3.5	6.5	10.0
	Male	Count	2	8	10
		Expected count	3.5	6.5	10.0
Total		Count	7	13	20
		Expected count	7.0	13.0	20.0

The expected cell frequency is calculated by multiplying the observed row total by the observed column total and dividing by the total *N*. For example, in this table the expected count of males obtaining *A*s is 3.5, which is (7 × 10)/20.

statistical way to state this question would be: Do the observed percentages (or frequencies or proportions) differ from those that would be expected to occur by chance?

Several statistics could be used to summarize the strength of this relationship and to test whether it is statistically significant, but by far the most frequently used statistic in this situation is the chi-square test. The chi-square statistic directly tests whether the observed frequencies differ from those that would be expected to occur by chance. The idea of the test is very straightforward. Calculate the average difference between the observed and expected frequencies across all of the cells and then compare this number to the critical value associated with a particular probability level, given the degrees of freedom for your data. The degrees of freedom for a chi-square test is the product of the number of rows times minus 1 times the number of columns minus 1 [in our example: $(2 - 1)(2 - 1) = 1$]. In Table 12–9, the observed and expected frequencies are shown and the calculation of the expected frequency for one of the cells is highlighted. I imagine that you now have the idea but are very curious about the final result. It turns out that the critical value for this analysis (1 degree of freedom, at the .05 level of significance) is 3.84, but the computed value of chi-square for this data is only 1.98 with an associated probably of .16, above our .05 cutoff. So we have to conclude that for this data, the pattern observed does not differ from that expected by chance. This example demonstrates a basic way to examine a simple two-variable situation. Of course, there are far more complex possibilities and many more ways to examine relationships between nominal variables, including multivariate model-building techniques.

The expected cell frequency is calculated by multiplying the observed row total by the observed column total and dividing by the total *N*. For example, in this table the expected count of males obtaining *A*s is 3.5, which is (7 × 10)/20.

12-4 Exploratory Data Analysis and Graphics

John Tukey provided us with important ways of obtaining mathematical, verbal, and graphic insight about our data. He coined the term **exploratory data analysis** (EDA) to describe methods that would reestablish the exploration of data as a necessary companion to the confirmatory statistical techniques that became dominant in the second half of the last century. (Tukey also coined the term *software* and is regarded as one of the great minds of the 20th century, especially in statistics.) I present the topic of EDA and graphics toward the end of this chapter, but that does not imply it should be done last or least. In fact, if I have not done some EDA and graphic analysis of a dataset I'm working on, then I cannot convince *myself* of the conclusion validity of any other analysis I might have done let alone anyone else. EDA is as much an attitude of wanting to understand your data, to see below the surface, as it is mastering the tools. The goal is to effectively describe patterns in data so that you will understand what meaning there may be. Simplicity and clarity are valued over complexity and abstraction. We have already used some of the most common graphic methods when we examined the histogram (Figures 12–5 and 12–6) and scatterplot (Figure 12–7). In this section, I will review only some of the classic techniques, but I encourage you to explore every variable carefully before conducting your formal hypothesis tests. You may have even practiced some of these techniques early in your education because people who design curricula for elementary and secondary math courses realize how important they are in facilitating understanding of data and relationships.

exploratory data analysis
The use of graphic and other methods to examine relationships in a data set. EDA is especially helpful when trying to develop hypotheses about relationships and when examining distributions of variables by themselves or in relation to other variables.

12-4a The Stem and Leaf Plot

Earlier in the chapter you read about ways of summarizing central tendency and variability with statistical indicators such as the mean and standard deviation. Tukey created a graphic approach that allows us to see both dimensions of a variable, called the **stem and leaf diagram**. In Figure 12–12, you will see a stem and leaf diagram for the previously studied height data. The left-hand column represents the "stem," which is the unique part of each value after removing the last digit. The last digit for every value is then shown with the other "leaves" on the right side of the figure. Note that you can see the "big picture" on this variable while retaining every single case. This particular figure was done using SPSS and gives you the frequency count for each row as well as a legend to help interpret the figure at the bottom.

stem and leaf diagram
In a stem-and-leaf plot, each observed value is divided into two components—leading digits (the stem) and trailing digits (the leaves). Like the histogram, it shows the entire distribution of a variable, but in addition preserves all of the individual values in the display. The stem-and-leaf plot is one of the many graphic techniques developed by John Tukey.

12-4b The Boxplot (or Box and Whisker Plot)

The **boxplot** is another way to effectively display several characteristics of a variable in a simple graphic format. In Figure 12–13, you see a boxplot of the height data. This plot shows you all of the following about height in this sample: (1) the median

boxplot
A boxplot (or box and whisker plot) is a graphic display invented by John Tukey that summarizes the distribution of a numeric variable by showing the median and quartiles as a box, and the extreme values as "whiskers" extending from the box.

FIGURE 12–12 Stem and Leaf Plot of Height Data

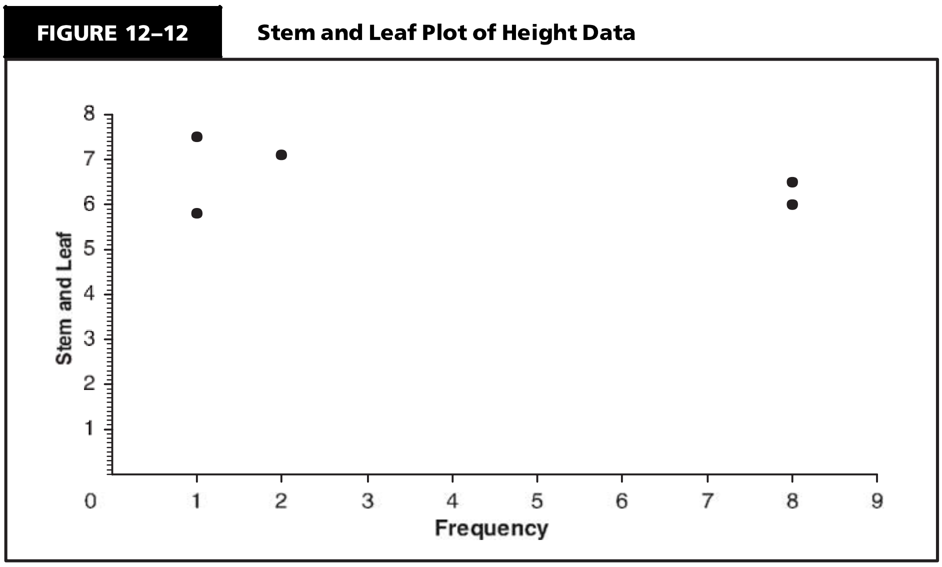
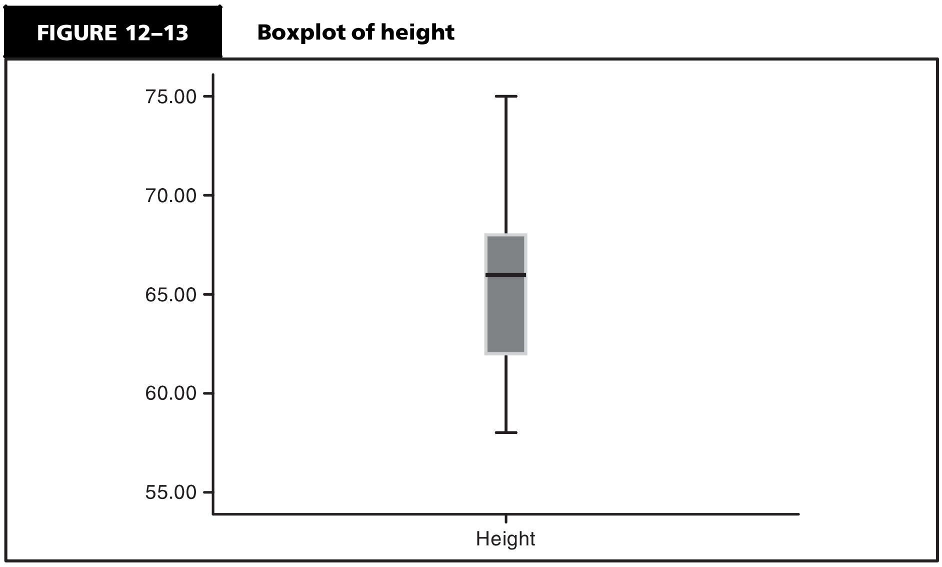


FIGURE 12–13 Boxplot of height



(the black line in middle of the box); (2) the 25th and 75th percentile values (the bottom and top boundaries of the box); (3) the interquartile range (the middle 50% of the values represented by the box itself); and (4) the highest and lowest values (the "whiskers"; the ends of the lines extending from the box). If the variable had extreme values (outliers), then you could also add these to the plot as dots (or other symbols) beyond the end of the whiskers. The boxplot can also be used to explore subgroups on variables. For example, in Figure 12–14, you can see the self-esteem variable broken down by gender and grades. What do you notice about this figure? I hope you notice that quite a bit of information about the relationship of three variables can be gleaned from this one picture, including the central tendencies, variability, and very likely some ideas about what may be going on below the surface of the data, as Professor Tukey would say.

12-4c Anscombe's Quartet

Anscombe's quartet is a dataset that clearly illustrates the importance of looking at your data graphically, and not just assuming that summary statistics tell the whole

FIGURE 12–14 **Boxplot of self-esteem by gender and grades**

Caption: Case 9, a female with a B average, has unusually high self-esteem. If you go back and look at the data table, you'll see she is also unusually tall for this group.

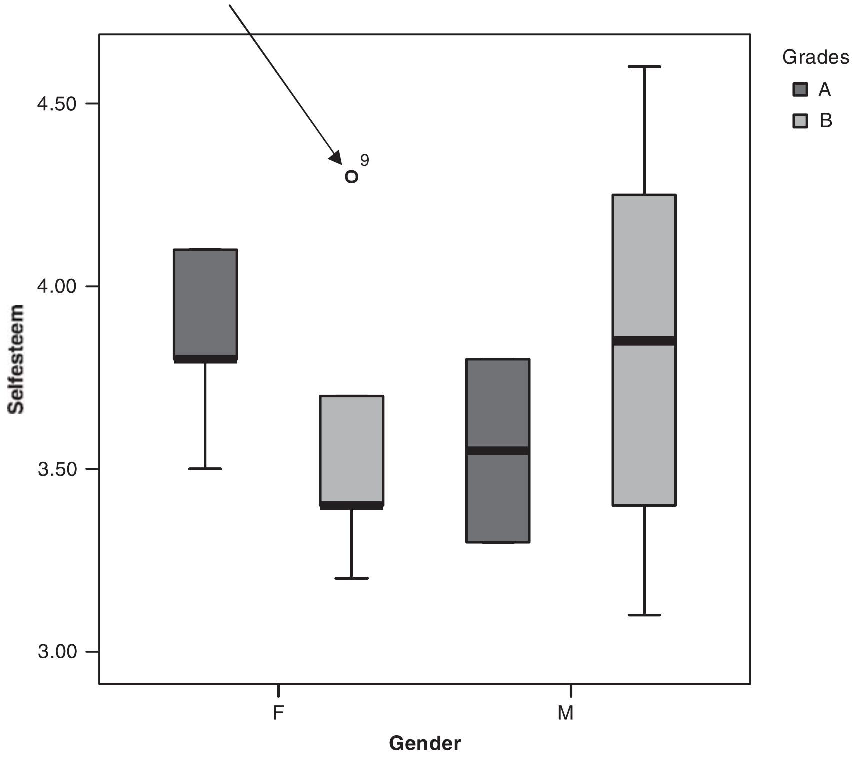


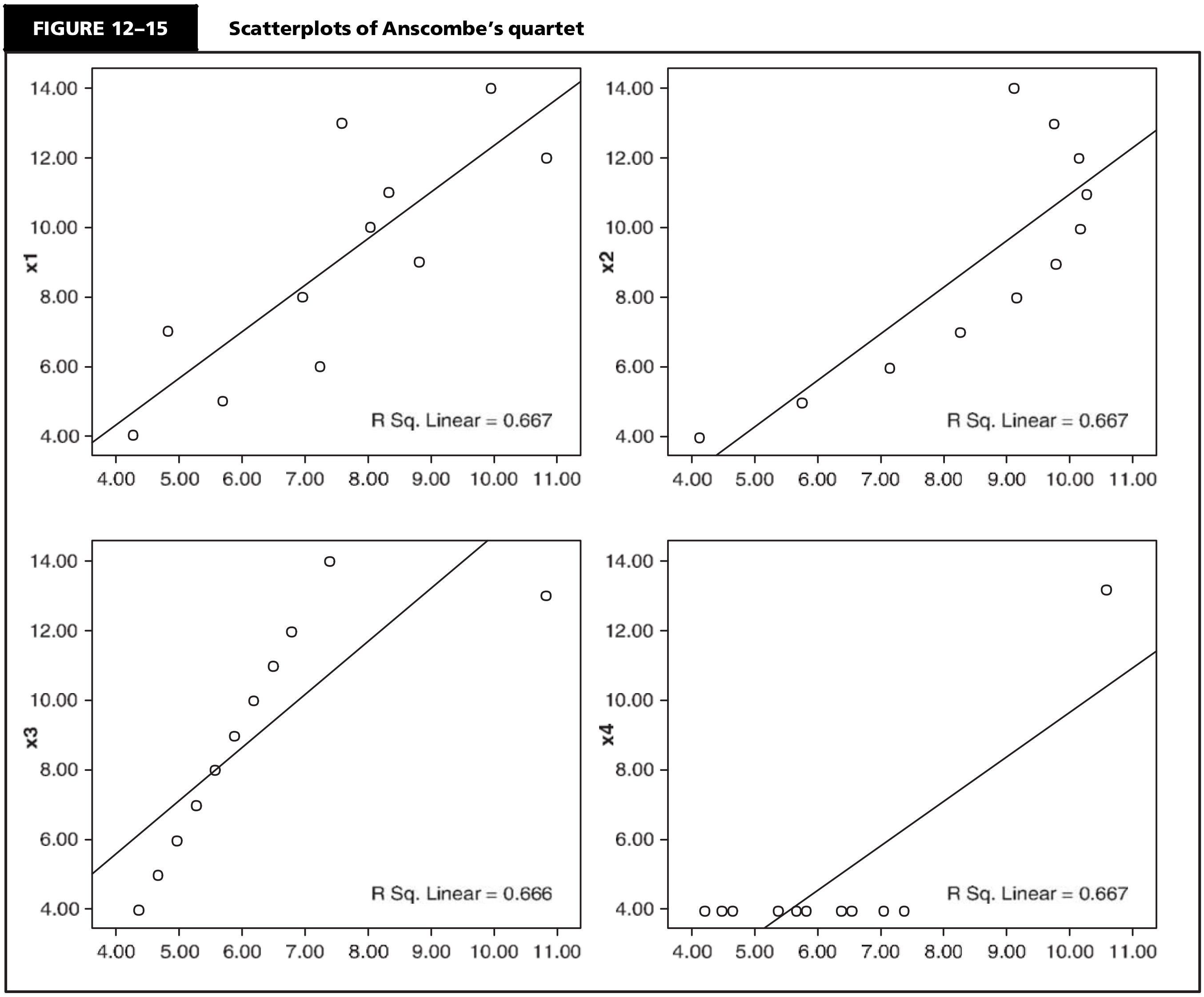
TABLE 12–10 **Anscombe's Quartet with Summary Statistics**

*x*1	*y*1	*x*2	*y*2	*x*3	*y*3	*x*4	*y*4
10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58
8.00	6.95	8.00	8.14	8.00	6.77	8.00	5.76
13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71
9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
4.00	4.26	4.00	3.10	4.00	5.39	19.00	12.50
12.00	10.84	12.00	9.13	12.00	8.15	8.00	5.56
7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91
5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89

Summary statistics for all *X*-*Y* pairs:

- Mean of the *x* values = 9.0
- Mean of the *y* values = 7.5
- Equation of the least-squared regression line is: $y = 3 + 0.5x$
- Sums of squared errors (about the mean) = 110.0
- Regression sums of squared errors (variance accounted for by *x*) = 27.5
- Residual sums of squared errors (about the regression line) = 13.75
- Correlation coefficient = 0.82
- Coefficient of determination = 0.67

FIGURE 12–15 Scatterplots of Anscombe's quartet



story. Anscombe published the data in 1973, and it was later utilized very effectively by Edward Tufte in his book, The Visual Display of Quantitative Information (2001), one of the most insightful and frankly beautiful books I've ever seen (his Web site is listed at the end of this chapter, too). In Table 12–10 (see p. 279), you'll see four *X-Y* pairs. Remarkably, the summary statistics for all for *X-Y* pairs are equivalent. The lesson from Anscombe's quartet is seen strikingly when we plot the data. In Figure 12–15, scatterplots of all of the data pairs are shown.

Summary

This chapter introduced the basics involved in data analysis. Conclusion validity is the degree to which inferences about relationships in data are reasonable. Conclusions from data involve accepting one hypothesis and thereby rejecting its mutually exclusive and exhaustive alternative, and in reaching a conclusion, you can either be correct or incorrect. You can make two types of errors. A Type I error occurs when you conclude there is a relationship when in fact there is not (seeing something that's not there). A Type II error occurs when you conclude there is no effect when in fact there is (missing the needle in the haystack). Data preparation involves checking or logging the data in, checking the data for accuracy, entering the data into the computer, transforming the data, and developing and documenting a database structure that integrates the various measures. Missing data should be assessed before any analysis, especially whether the missing data is random. Descriptive statistics describe the basic features of the data in a study. The basic descriptive statistics include descriptions of the data distributions, measures of central tendency and dispersion or variability, and the different forms of correlation. EDA and statistical graphics can greatly enhance your understanding of your data as well as your ability to communicate that understanding to others.

Login to the Online Edition of your text at www.atomicdog.com to find additional resources located in the Study Guide at the end of each chapter.