IN THIS CHAPTER

- Realizing your role as a data analyst
- Avoiding statistical faux pas
- Delving into the jargon of Stats II

Chapter 1

Beyond Number Crunching: The Art and Science of Data Analysis

Because you're reading this book, you're likely familiar with the basics of statistics and you're ready to take it up a notch. That next level involves using what you know, picking up a few more tools and techniques, and finally putting it all to use to help you answer more realistic questions by using real data. In statistical terms, you're ready to enter the world of the *data analyst.*

In this chapter, you review the terms involved in statistics as they pertain to data analysis at the Stats II level. You get a glimpse of the impact that your results can have by seeing what these analysis techniques can do. You also gain insight into some of the common misuses of data analysis and their effects.

Data Analysis: Looking before You Crunch

It used to be that statisticians were the only ones who really analyzed data because the only computer programs available were very complicated to use, requiring a great deal of knowledge about statistics to set up and carry out analyses.

The calculations were tedious and at times unpredictable, and they required a thorough understanding of the theories and methods behind the calculations to get correct and reliable answers.

Today, anyone who wants to analyze data can do it easily. Many user-friendly statistical software packages are made expressly for that purpose — Microsoft Excel, Minitab, and SAS are just a few. Free online programs are available, too, such as R, which helps you do just what it says — crunch your numbers and get an answer.

Each software package has its own pros and cons (and its own users and protesters). My software of choice and the one I reference throughout this book is Minitab, because it's very easy to use, the results are precise, and the software's loaded with all the data-analysis techniques used in Stats II. Although a site license for Minitab isn't cheap, the student version is available for rent for only a few bucks a semester.

REMEMBER

The most important idea when applying statistical techniques to analyze data is to know what's going on behind the number crunching so you (not the computer) are in control of the analysis. That's why knowledge of Stats II is so critical.

WARNING

Many people don't realize that statistical software can't tell you when and when not to use a certain statistical technique. You have to determine that on your own. As a result, people think they're doing their analyses correctly, but they can end up making all kinds of mistakes. In the following sections, I give examples of some situations in which innocent data analyses can go wrong and why it's important to spot and avoid these mistakes before you start crunching numbers.

Bottom line: Today's software packages really are too good to be true if you don't have a clear and thorough understanding of the Stats II that's beneath the surface.

Nothing (not even a straight line) lasts forever

Bill Prediction is a statistics student who is studying the effect of study time on a student's exam score. Bill collects data on statistics students and uses his trusty software package to predict exam scores based on study time. His computer comes up with the equation $y = 10x + 30$, where y represents the test score you get if you study for a certain number of hours (x). Notice that this model is the equation of a straight line with a y-intercept of 30 and a slope of 10.

So using this model, Bill predicts that if you don't study at all, you'll get a 30 on the exam (plugging $x = 0$ into the equation and solving for y; this point represents

the *y*-intercept of the line). He also predicts, using this model, that if you study for 5 hours, you'll get an exam score of $y = (10 * 5) + 30 = 80$ So, the point (5,80) is also on this line.

But then Bill goes a little crazy and wonders what would happen if you studied for 40 hours (because it always seems that long when he's studying). The computer tells him that if he studies for 40 hours, his test score is predicted to be $(10 * 40) + 30 = 430$ points. Wow, that's a lot of points! Problem is, the exam only goes up to a total of 100 points. Bill wonders where his computer went wrong.

But Bill puts the blame in the wrong place. He needs to remember that there are limits on the values of *x* that make sense in this equation. For example, because *x* is the amount of study time, *x* can never be a number less than zero. If you plug a negative number in for *x*, say $x = -10$, you get $y = (10 * -10) + 30 = -70$, which makes no sense. However, the equation itself doesn't know that, nor does the computer that found it. The computer simply graphs the line you give it, assuming it'll go on forever in both the positive and negative directions.

WARNING

After you get a statistical equation or model, you need to specify for what values the equation applies. Equations don't know when they work and when they don't; it's up to the data analyst to determine that. This idea is the same for applying the results of any data analysis that you do.

Data snooping isn't cool

WARNING

Statisticians have come up with a saying that you may have heard: "Figures don't lie. Liars figure." Make sure that you find out about all the analyses that were performed on a data set, not just the ones reported as being statistically significant.

Suppose Bill Prediction (from the previous section) decides to try to predict scores on a biology exam based on study time, but this time his model doesn't fit. Not one to give in, Bill insists there must be some other factors that predict biology exam scores besides study time, and he sets out to find them.

Bill measures everything from soup to nuts. His set of 20 possible variables includes study time, GPA, previous experience in statistics, math grades in high school, and whether you chew gum during the exam. After his multitude of various correlation analyses, the variables that Bill finds to be related to exam score are study time, math grades in high school, GPA, and gum chewing during the exam. It turns out that this particular model fits pretty well (by criteria I discuss in Chapter 6 on multiple linear regression models).

But here's the problem: By looking at all possible correlations between his 20 variables and the exam score, Bill is actually doing 20 separate statistical analyses. Under typical conditions that I describe in Chapter 4, each statistical analysis has a 5 percent chance of being wrong just by chance. I bet you can guess which one of Bill's correlations likely came out wrong in this case. And hopefully, he didn't rely on a stick of gum to boost his grade in biology.

WARNING

Looking at data until you find something in it is called *data snooping.* Data snooping results in giving the researcher his five minutes of fame but then leads him to lose all credibility because no one can repeat his results.

No (data) fishing allowed

Some folks just don't take no for an answer, and when it comes to analyzing data, that can lead to trouble.

Sue Gonnafindit is a determined researcher. She believes that her horse can count by stomping his foot. (For example, she says "2" and her horse stomps twice.) Sue collects data on her horse for four weeks, recording the percentage of time the horse gets the counting right. She runs the appropriate statistical analysis on her data and is shocked to find no significant difference between her horse's results and those you would get simply by guessing.

Determined to prove her results are real, Sue looks for other types of analyses that exist and plugs her data into anything and everything she can find (never mind that those analyses are inappropriate to use in her situation). Using the famous hunt-and-peck method, at some point she eventually stumbles upon a significant result. However, the result is bogus because she tried so many analyses that weren't appropriate and ignored the results of the appropriate analysis because it didn't tell her what she wanted to hear.

Funny thing, too. When Sue went on a late-night TV program to show the world her incredible horse, someone in the audience noticed that whenever the horse got to the correct number of stomps, Sue would interrupt him and say "Good job!" and the horse quit stomping. He didn't know how to count; all he knew to do was to quit stomping when she said, "Good job!"

WARNING

Redoing analyses in different ways in order to try to get the results you want is called *data fishing,* and folks in the stats biz consider it to be a major no-no. (However, people unfortunately do it all too often to verify their strongly held beliefs.) By using the wrong data analysis for the sake of getting the results you desire, you mislead your audience into thinking that your hypothesis is actually correct when it may not be.

Getting the Big Picture: An Overview of Stats II

Stats II is an extension of Stats I (introductory statistics), so the jargon follows suit and the techniques build on what you already know. In this section, you get an introduction to the terminology you use in Stats II along with a broad overview of the techniques that statisticians use to analyze data and find the story behind it. (If you're still unsure about some of the terms from Stats I, you can consult your Stats I textbook or see my other book, *Statistics For Dummies, 2nd Edition* [Wiley], for a complete rundown.)

Population parameter


REMEMBER

A *parameter* is a number that summarizes the *population*, which is the entire group you're interested in investigating. Examples of parameters include the mean of a population, the median of a population, or the proportion of the population that falls into a certain category.

Suppose you want to determine the average length of a cellphone call among teenagers (ages 13–18). You're not interested in making any comparisons; you just want to make a good guesstimate of the average time. So you want to estimate a population parameter (such as the mean or average). The population is all cellphone users between the ages of 13 and 18 years old. The parameter is the average length of a phone call this population makes.

Sample statistic

Typically you can't determine population parameters exactly; you can only estimate them. But all is not lost; by taking a representative *sample* (a well-chosen subset of individuals) from the population and studying it, you can come up with a good estimate of the population parameter. A *sample statistic* is a single number that summarizes that subset.

For example, in the cellphone scenario from the previous section, you select a sample of teenagers and measure the duration of their cellphone calls over a period of time (or look at their cellphone records if you can gain access legally). You take the average of the cellphone call duration. For example, the average duration of 100 cellphone calls may be 12.2 minutes — this average is a statistic. This particular statistic is called the *sample mean* because it's the average value from your sample data.

Many different statistics are available to study different characteristics of a sample, such as the proportion, the median, and standard deviation.

Confidence interval

A *confidence interval* is a range of likely values for a population parameter. A confidence interval is based on a sample and the statistics that come from that sample. The main reason you want to provide a range of likely values rather than a single number is that sample results vary.

For example, suppose you want to estimate the percentage of people who eat chocolate. According to the Simmons Market Research Bureau, 78 percent of adults reported eating chocolate, and of those, 18 percent admitted eating sweets frequently. What's missing in these results? These numbers are only from a single sample of people, and those sample results are guaranteed to vary from sample to sample. You need some measure of how much you can expect those results to move if you were to repeat the study.

This expected variation in your statistic from sample to sample is measured by the *margin of error*, which reflects a certain number of standard deviations of your statistic that you add and subtract to have a certain confidence in your results (see Chapter 4 for more on margin of error). If the chocolate-eater results were based on 1,000 people, the margin of error would be approximately 3 percent. This means the actual percentage of people who eat chocolate in the entire population is expected to be 78 percent, ± 3 percent (that is, between 75 percent and 81 percent).

Hypothesis test

A *hypothesis test* is a statistical procedure that you use to test an existing claim about the population, using your data. The claim is noted by H_0 (the null hypothesis). If your data support the claim, you fail to reject H_0. If your data don't support the claim, you reject H_0 and conclude an alternative hypothesis, H_a. The reason most people conduct a hypothesis test is not to merely show that their data support an existing claim, but rather to show that the existing claim is false, in favor of the alternative hypothesis.

The Pew Research Center studied the percentage of people who turn to ESPN for their sports news. Its statistics, based on a survey of about 1,000 people, found that in 2000, 23 percent of people said they went to ESPN; in 2020, only 20.9 percent reported going to ESPN. The question is this: Does this 2.1 percent reduction in viewers represent a significant trend that ESPN should worry about?

To test these differences formally, you can set up a hypothesis test. You set up your null hypothesis as the result you have to believe without your study, H_0 = No difference exists between 2000 and 2020 data for ESPN viewership. Your alternative hypothesis (H_a) is that a difference is there. To run a hypothesis test, you look at the difference between your statistic from your data and the claim that has been already made about the population (in H_0), and you measure how far apart they are in units of standard deviations.

With respect to the example, using the techniques from Chapter 4, the hypothesis test shows that 23 percent and 20.9 percent aren't far enough apart in terms of standard deviations to dispute the claim (H_0). You can't say the percentage of viewers of ESPN in the entire population changed from 2000 to 2020.

REMEMBER

As with any statistical analysis, your conclusions can be wrong just by chance, because your results are based on sample data, and sample results vary. In Chapter 4, I discuss the types of errors that can be made in conclusions from a hypothesis test.

Analysis of variance (ANOVA)

ANOVA is the acronym for *analysis of variance.* You use ANOVA in situations where you want to compare the means of more than two populations. For example, say you want to compare the lifetimes of four brands of tires in number of miles. You take a random sample of 50 tires from each group, for a total of 200 tires, and set up an experiment to compare the lifetime of each tire, and record it. You now have four means and four standard deviations, one for each data set.

Then, to test for differences in average lifetime for the four brands of tires, you basically compare the variability between the four data sets to the variability within the entire data set, using a ratio. This ratio is called the *F-statistic.* If this ratio is large, the variability between the brands is more than the variability within the brands, giving evidence that not all the means are the same for the different tire brands. If the *F*-statistic is small, not enough difference exists between the treatment means compared to the general variability within the treatments (here the brands) themselves. In this case, you can't say that the means are different for the groups. (I give you the full scoop on ANOVA plus all the jargon, formulas, and computer output in Chapters 10 and 11.)

Multiple comparisons

Suppose you conduct ANOVA, and you find a difference in the average lifetimes of the four brands of tire (see the preceding section). Your next questions would probably be, "Which brands are different?" and "How different are they?" To answer these questions, you use multiple-comparison procedures.

A *multiple-comparison procedure* is a statistical technique that compares means to each other and finds out which ones are different and which ones aren't. With this information, you're able to put the groups in order from those with the largest mean to those with the smallest mean, realizing that sometimes two or more groups were too close to tell and are placed together in a group.

Many different multiple-comparison procedures exist to compare individual means and come up with an ordering in the event that your *F*-statistic does find that some difference exists. Some of the multiple-comparison procedures include Tukey's test, LSD (least significant difference), and pairwise *t*-tests. Some procedures are better than others, depending on the conditions and your goal as a data analyst. I discuss multiple-comparison procedures in detail in Chapter 11.

WARNING

Never take that second step to compare the means of the groups if the ANOVA procedure doesn't find any significant results during the first step. Computer software will never stop you from doing a follow-up analysis, even if it's wrong to do so.

Interaction effects

An *interaction effect* in statistics operates the same way that it does in the world of medicine. Sometimes if you take two different medicines at the same time, the combined effect is much different than if you were to take the two individual medications separately.

REMEMBER

Interaction effects can come up in statistical models that use two or more variables to explain or compare outcomes. In this case you can't automatically study the effect of each variable separately; you have to first examine whether or not an interaction effect is present.

For example, suppose medical researchers are studying a new drug for depression and want to know how this drug affects the change in blood pressure for a low dose versus a high dose. They also compare the effects for children versus adults. It could also be that dosage level affects the blood pressure of adults differently than the blood pressure of children. This type of model is called a *two-way ANOVA model*, with a possible interaction effect between the two factors (age group and dosage level). Chapter 12 covers this subject in depth.

Correlation

The term *correlation* is often misused. Statistically speaking, the correlation measures the strength and direction of the linear relationship between two *quantitative variables* (variables that represent counts or measurements only).

REMEMBER

You aren't supposed to use correlation to talk about relationships unless the variables are quantitative. For example, it's wrong to say that a correlation exists between eye color and hair color. (In Chapter 14, you explore associations between two categorical variables.)

Correlation is a number between −1.0 and +1.0. A correlation of +1.0 indicates a perfect positive relationship; as you increase one variable, the other one increases in perfect sync. A correlation of −1.0 indicates a perfect negative relationship between the variables; as one variable increases, the other one decreases in perfect sync. A correlation of zero means you found no linear relationship at all between the variables. Most correlations in the real world fall somewhere in between −1.0 and +1.0; the closer to −1.0 or +1.0, the stronger the relationship is; the closer to 0, the weaker the relationship is.

Figure 1-1 shows a plot of the number of coffees sold at football games in Buffalo, New York, as well as the air temperature (in degrees Fahrenheit) at each game. This data set seems to follow a downhill straight line fairly well, indicating a negative correlation. The correlation turns out to be −0.741; the number of coffees sold has a fairly strong negative relationship with the temperature of the football game. This makes sense because on days when the temperature is low, people get cold and want more coffee. I discuss correlation further, as it applies to model building, in Chapter 5.

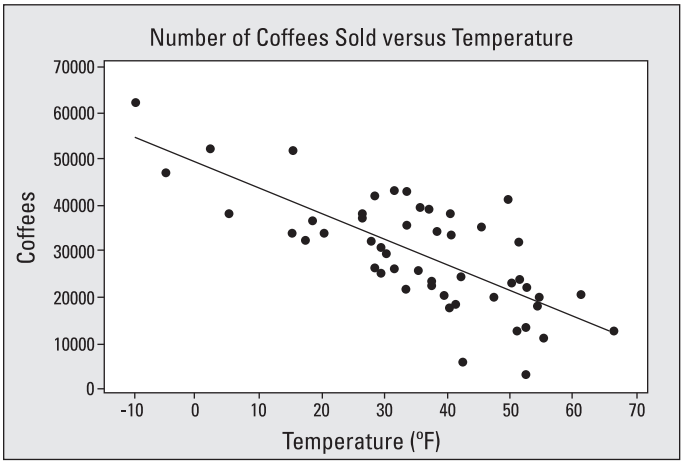

FIGURE 1-1: Coffees sold at various air temperatures on football game day.

Linear regression

After you've found a correlation and determined that two variables have a fairly strong linear relationship, you may want to try to make predictions for one variable based on the value of the other variable. For example, if you know that a fairly

strong negative linear relationship exists between coffees sold and the air temperature at a football game (see the previous section), you may want to use this information to predict how much coffee is needed for a game, based on the temperature. This method of finding the best-fitting line is called *linear regression.*

Many different types of regression analyses exist, depending on your situation. When you use only one variable to predict the response, the method of regression is called *simple linear regression* (see Chapter 5). Simple linear regression is the best known of all the regression analyses and is a staple in the Stats I course sequence.

However, you use other flavors of regression for other situations.

- If you want to use more than one variable to predict a response, you use *multiple linear regression* (see Chapter 6).
- If you want to make predictions about a variable that has only two outcomes, yes or no, you use *logistic regression* (see Chapter 9).
- For relationships that don't follow a straight line, you have a technique called (no surprise) *nonlinear regression* (see Chapter 8).

Chi-square tests

Correlation and regression techniques all assume that the variable being studied in most detail (the response variable) is quantitative — that is, the variable measures or counts something. You can also run into situations where the data being studied isn't quantitative, but rather categorical — that is, the data represents categories, not measurements or counts. To study relationships in categorical data, you use a Chi-square test for independence. If the variables are found to be unrelated, they're declared independent. If they're found to be related, they're declared dependent.

Suppose you want to explore the relationship between age group and eating breakfast. Because each of these variables is categorical, or qualitative, you use a Chi-square test for independence. You survey 70 adults and 70 children and find that 25 adults eat breakfast and 45 do not; for the children, 35 do eat breakfast and 35 do not. Table 1-1 organizes this data and sets you up for the Chi-square test for this scenario.

TABLE 1-1 **Table Setup for the Breakfast and Age Group Question**

	Do Eat Breakfast	Don't Eat Breakfast	Total
Adult	25	45	70
Child	35	35	70

REMEMBER

A Chi-square test first calculates what you expect to see in each cell of the table if the variables are independent (these values are brilliantly called the *expected cell counts*). The Chi-square test then compares these expected cell counts to what you observed in the data (called the *observed cell counts*) and compares them using a Chi-square statistic.

In the breakfast age-group comparison, fewer adults than children eat breakfast ($25/70 = 35.7$ percent compared to $35/70 = 50$ percent). Even though you know results will vary from sample to sample, this difference turns out to be enough to declare a relationship between age group and eating breakfast, according to the Chi-square test of independence. Chapter 15 reveals all the details of doing a Chi-square test.

You can also use the Chi-square test to see whether your theory about what percent of each group falls into a certain category is true or not. For example, can you guess what percentage of M&M'S fall into each color category? You can find more on these Chi-square variations, as well as the M&M'S question, in Chapter 16.