LEARNING MADE EASY

5th Edition

Statistical Analysis with Excel®

for dummies®

A Wiley Brand

Perform statistical analyses on Windows®, Mac®, and iPad®

Download the book's spreadsheets to accelerate your learning

Understand common (and uncommon) statistical terms

Joseph Schmuller, PhD

Teaching statistics through the use of Excel since 2005

Statistical Analysis with Excel®

5th Edition

by Joseph Schmuller, PhD

for dummies®
A Wiley Brand

Statistical Analysis with Excel® For Dummies®, 5th Edition

Published by: **John Wiley & Sons, Inc.,** 111 River Street, Hoboken, NJ 07030-5774, www.wiley.com

LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY: WHILE THE PUBLISHER AND AUTHORS HAVE USED THEIR BEST EFFORTS IN PREPARING THIS WORK, THEY MAKE NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS WORK AND SPECIFICALLY DISCLAIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES REPRESENTATIVES, WRITTEN SALES MATERIALS OR PROMOTIONAL STATEMENTS FOR THIS WORK. THE FACT THAT AN ORGANIZATION, WEBSITE, OR PRODUCT IS REFERRED TO IN THIS WORK AS A CITATION AND/OR POTENTIAL SOURCE OF FURTHER INFORMATION DOES NOT MEAN THAT THE PUBLISHER AND AUTHORS ENDORSE THE INFORMATION OR SERVICES THE ORGANIZATION, WEBSITE, OR PRODUCT MAY PROVIDE OR RECOMMENDATIONS IT MAY MAKE. THIS WORK IS SOLD WITH THE UNDERSTANDING THAT THE PUBLISHER IS NOT ENGAGED IN RENDERING PROFESSIONAL SERVICES. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR YOUR SITUATION. YOU SHOULD CONSULT WITH A SPECIALIST WHERE APPROPRIATE. FURTHER, READERS SHOULD BE AWARE THAT WEBSITES LISTED IN THIS WORK MAY HAVE CHANGED OR DISAPPEARED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ. NEITHER THE PUBLISHER NOR AUTHORS SHALL BE LIABLE FOR ANY LOSS OF PROFIT OR ANY OTHER COMMERCIAL DAMAGES, INCLUDING BUT NOT LIMITED TO SPECIAL, INCIDENTAL, CONSEQUENTIAL, OR OTHER DAMAGES.

For general information on our other products and services, please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993, or fax 317-572-4002. For technical support, please visit https://hub.wiley.com/community/support/dummies.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at http://booksupport.wiley.com. For more information about Wiley products, visit www.wiley.com.

Library of Congress Control Number: 2021950426

ISBN: 978-1-119-84454-9; 978-1-119-84455-6 (ebk); 978-1-119-84456-3 (ebk)

Table of Contents

Introduction

What? Yet another statistics book? Well, this is a statistics book, all right — but in my humble (and thoroughly biased) opinion, it's *still*, after four editions, not *just* another statistics book.

What? Yet another Excel book? Same thoroughly biased opinion (still, after four editions) — it's not just another Excel book. What? Yet another edition of a book that's not just another statistics book and not just another Excel book? Well . . . yes. For the fifth time, you got me there.

Here's the story — for the previous four editions and for this one. Many statistics books teach you the concepts but don't give you a way to apply them — which often leads to a lack of understanding. With Excel, you have a ready-made package for applying statistics concepts.

Looking at it from the opposite direction, many Excel books show you Excel's capabilities but don't tell you about the concepts behind them. Before I tell you about an Excel statistical tool, I give you the statistical foundation it's based on. That way, you understand the tool when you use it — and you use it more effectively. I didn't want to write a book that's just "select this menu" and "click this button." Some of that is necessary, of course, in any book that shows you how to use a software package. My goal was to go way beyond that.

Neither did I want to write a statistics "cookbook" — when-faced-with-problem-#862-use-statistical-procedure-#412. My goal was to go way beyond that, too.

This book isn't just about statistics or just about Excel — it sits firmly at the intersection of the two. In the course of telling you about statistics, I cover every Excel statistical feature. (Well, *almost* every one. I did leave one out and, truth be told, I left it out of the first four editions, too. It's called "Fourier Analysis." All the necessary math to understand it would take an entire book to present, and you may never use this tool anyway. Perhaps I'll cover it in the infinitieth edition. . . .)

About This Book

Although statistics involves a logical progression of concepts, I've organized this book so that you can open it up in any chapter and start reading. The idea is for you to find what you're looking for in a hurry and use it immediately — whether it's a statistical concept or an Excel tool.

On the other hand, reading from cover to cover is okay if you're so inclined. If you're a statistics newbie and you have to use Excel for statistical analysis, I recommend you begin at the beginning — even if you know Excel pretty well.

What's New in This Edition

I wanted to add a dimension or two to this fifth edition, and I think I've done just that.

In addition to the usual material on Windows and on the Mac, I also cover — wait for it — the iPad! The iPad doesn't support Excel's major statistical package (the Analysis Toolpak), but other packages fill the void, as you'll see. I think you'll find that Excel on the iPad is a powerful tool for statistics. (I'm working with a fourth generation, 12.9-inch iPad Pro. If you're working with a different model, your mileage may vary.)

Throughout this book, then, you see material about Excel on the Mac and on the iPad — particularly when MacOS and iPadOS differ substantially from Windows. Otherwise, it's been my experience that Mac users and iPad users are a hearty lot and know how to adapt. (So? Enough . . . Apple polishing?)

Making its debut in this edition is the increasingly popular topic of logistic regression (see Chapter 21).

In Chapter 20, I've added a section on simulating a business. It's my first foray into Excel's What If analysis tools.

In this edition, I've moved the discussion about some of the lesser used charts from Chapter 3 to an online appendix you can download at `www.dummies.com/go/statisticalanalysiswexcelfd5e`.

Speaking of online appendixes, I've moved an online appendix from the fourth edition, "When Your Data Live Elsewhere," into this edition. I think it's useful information you should have at your fingertips. Also moving from online to printed page is an appendix called "Tips for Teachers (And Learners)."

And finally, due to popular demand, you can download this edition's spreadsheets! (Again, available at `www.wiley.com/en-us/Statistical+Analysis+with+Excel+For+Dummies%2C+5th+Edition-p-9781119844563`.) The spreadsheets contain just the data. You still have to follow the steps I provide to complete the analyses.

What's New in Excel (Microsoft 365)

I work with the cloud-based subscription version of Excel, which is part of Microsoft 365. As part of the subscription, I receive updates you may not have if you work with Excel 2021 — a stand-alone product. Were you to examine the two incarnations of Excel up close and personal, you probably wouldn't find much difference in functionality. (But see the section on array functions in Chapter 2.)

Although Excel hasn't added any new statistical functions, the Windows version (of 365) has an exciting new feature called *linked data,* which enables you to look up information about a variety of topics (movies, universities, stocks, and more) without leaving Excel. (Exciting as it is, we won't be working with this one.)

The Mac and iPad have added Data from Pictures. As its name suggests, this feature looks at a picture of a data table and puts the data into a spreadsheet. I cover this topic in Chapter 2.

A new feature in Windows and Mac called Analyze Data offers insights about your data. I cover this feature in Chapter 2 as well.

An add-in called XLMiner Analysis ToolPak mimics Excel's Analysis Toolpak (an extensive set of analytical tools) and adds logistic regression, which, as I mentioned, I cover in Chapter 21.

Foolish Assumptions

This isn't an introductory book about Excel or about Windows, Mac, or iPad, so I'm assuming that you

- **Know how to use your computer:** I don't spell out the details of pointing, clicking, selecting, and other basic actions.

- **Have Excel (Microsoft 365 subscription) installed on your machine and can work along with the examples:** I don't walk you through the steps of Excel installation. Incidentally, I work with the 32-bit version — it seems to get the updates more quickly than the 64-bit version does. Excel 2021 should work for most of the examples, but the subscription version receives the latest updates.
- **Have worked with Excel:** I don't go into the essentials of worksheets and formulas. I do fill you in on a few Excel fundamentals in Chapter 1, however.

If you don't know much about Excel, consider looking into Greg Harvey's excellent Excel books in the *For Dummies* series.

Icons Used in This Book

As is the case with all *For Dummies* books, icons appear all over the place. Each one is a little picture in the margin that lets you know something special about the paragraph it sits next to.

TIP

This icon points out a hint or a shortcut that can help you in your work and make you an all-around better human being.

REMEMBER

This one points out timeless wisdom to take with you long after you finish this book, young Jedi.

WARNING

Pay attention to this icon — it's a reminder to avoid an action that may gum up the works for you.

TECHNICAL STUFF

This icon indicates material you can blow right past if statistics and Excel aren't your passion.

Where to Go from Here

You can start the book anywhere, but here are a few hints. Want to learn the foundations of statistics? Turn the page. Introduce yourself to Excel's statistical features? That's Chapter 2. Want to start with graphics? Hit Chapter 3. For anything else, find it in the table of contents or in the index and go for it.

Beyond This Book

In addition to what you're reading right now, this book comes with a free, access-anywhere Cheat Sheet that will help you quickly use the tools I discuss. To find this Cheat Sheet, visit `www.dummies.com` and search for *Statistical Analysis with Excel For Dummies Cheat Sheet* in the Search box. And don't forget to check out the bonus content on this book's companion website, at `http://www.dummies.com/go/statisticalanalysiswexcelfd5e`.

If you've read any of the previous editions, welcome back!

If not, it's nice to meet you.

1

Getting Started with Statistical Analysis with Excel: A Marriage Made in Heaven

IN THIS PART . . .

Find out about Excel's statistical capabilities

Explore how to work with populations and samples

Test your hypotheses

Understand errors in decision-making

Determine independent and dependent variables

IN THIS CHAPTER

- » Introducing statistical concepts
- » Generalizing from samples to populations
- » Getting into probability
- » Making decisions
- » Understanding important Excel fundamentals

Chapter 1

Evaluating Data in the Real World

The field of statistics is all about decision-making — decision-making based on groups of numbers. Statisticians constantly ask questions: What do the numbers tell us? What are the trends? What predictions can we make? What conclusions can we draw?

To answer these questions, statisticians have developed an impressive array of analytical tools. These tools help us make sense of the mountains of data that are out there waiting for us to delve into, and to understand the numbers we generate in the course of our own work.

The Statistical (and Related) Notions You Just Have to Know

Because intensive calculation is often part and parcel of the statistician's tool set, many people have the misconception that statistics is about number crunching. Number crunching is just one small step on the path to sound decisions, however.

By shouldering the number crunching load, software increases your speed of travel down that path. Some software packages are specialized for statistical analysis and contain many of the tools that statisticians use. Although not marketed specifically as a statistical package, Excel provides a number of these tools, which is why I wrote this book.

I just said that number crunching is a small step on the path to sound decisions. The most important part are the concepts statisticians work with, and that's what I talk about for most of the rest of this chapter.

Samples and populations

On election night, TV commentators routinely predict the outcome of elections before the polls close. Most of the time they're right. How do they do that?

The trick is to interview a sample of voters right after they cast their ballots. Assuming the voters tell the truth about whom they voted for, and assuming the sample truly represents the population, network analysts use the sample data to generalize to the population of voters.

This is the job of a statistician — to use the findings from a sample to make a decision about the population from which the sample comes. But sometimes those decisions don't turn out the way the numbers predict. History buffs are probably familiar with the memorable photo of President Harry Truman holding up a copy of the *Chicago Daily Tribune* with the famous, but incorrect, headline "Dewey Defeats Truman" after the 1948 election. Part of the statistician's job is to express how much confidence they have in the decision.

Another election-related example speaks to the idea of the confidence in the decision. Pre-election polls (again, assuming a representative sample of voters) tell you the percentage of sampled voters who prefer each candidate. The polling organization adds how accurate it believes the polls are. When you hear a newscaster say something like "accurate to within 3 percent," you're hearing a judgment about confidence.

Here's another example. Suppose you've been assigned to find the average reading speed of all fifth grade children in the United States but you haven't got the time or the money to test them all. What would you do?

Your best bet is to take a sample of fifth-graders, measure their reading speeds (in words per minute), and calculate the average of the reading speeds in the sample. You can then use the sample average as an estimate of the population average.

Estimating the population average is one kind of *inference* that statisticians make from sample data. I discuss inference in more detail in the upcoming section "Inferential Statistics: Testing Hypotheses."

REMEMBER

Here's some terminology you have to know: Characteristics of a population (like the population average) are called *parameters,* and characteristics of a sample (like the sample average) are called *statistics.* When you confine your field of view to samples, your statistics are *descriptive.* When you broaden your horizons and concern yourself with populations, your statistics are *inferential.*

REMEMBER

And here's a notation convention you have to know: Statisticians use Greek letters (μ, σ, ρ) to stand for parameters, and English letters ($\bar{X}$, s, r) to stand for statistics. Figure 1-1 summarizes the relationship between populations and samples, and between parameters and statistics.

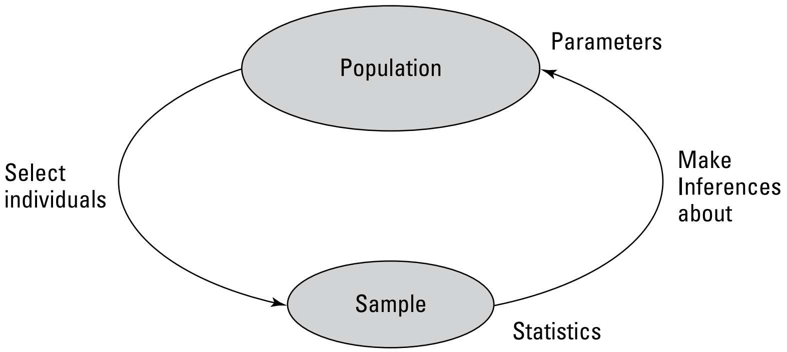

FIGURE 1-1: The relationship between populations and samples, and between parameters and statistics.

Variables: Dependent and independent

Simply put, a *variable* is something that can take on more than one value. (Something that can have only one value is called a *constant.*) Some variables you might be familiar with are today's temperature, the Dow Jones Industrial Average, your age, and the value of the dollar against the euro.

Statisticians care about two kinds of variables: *independent* and *dependent.* Each kind of variable crops up in any study or experiment, and statisticians assess the relationship between them.

Imagine a new way of teaching reading that's intended to increase the reading speed of fifth-graders. Before putting this new method into schools, it's a good idea to test it. To do that, a researcher randomly assigns a sample of fifth-grade students to one of two groups: One group receives instruction via the new method, and the other receives instruction via traditional methods. Before and after both groups receive instruction, the researcher measures the reading speeds of all the children in this study. What happens next? I get to that in the upcoming section "Inferential Statistics: Testing Hypotheses."

For now, understand that the independent variable here is the method of instruction. The two possible values of this variable are new and traditional. The dependent variable is the improvement in reading speed (a child's speed after instruction minus that child's speed before instruction) — which you would measure in words per minute.

REMEMBER

In general, the idea is to find out if changes in the independent variable are associated with changes in the dependent variable.

REMEMBER

In the examples that appear throughout the book, I show you how to use Excel to calculate characteristics of groups of scores. Keep in mind that each time I show you a group of scores, I'm really talking about the values of a dependent variable.

Types of data

Data come in four kinds. When you work with a variable, the way you work with it depends on what kind of data it is.

The first variety is called *nominal* data. If a number is a piece of nominal data, it's just a name. Its value doesn't signify anything. A good example is the number on an athlete's jersey. It's just a way of identifying the athlete. The number has nothing to do with the athlete's level of skill.

Next come ordinal data. *Ordinal* data are all about order, and numbers begin to take on meaning over and above just being identifiers. A higher number indicates the presence of more of a particular attribute than a lower number. One example is the *Mohs scale:* Used since 1822, it's a scale whose values are 1 through 10; mineralogists use this scale to rate the hardness of substances. Diamond, rated at 10, is the hardest. Talc, rated at 1, is the softest. A substance that has a given rating can scratch any substance that has a lower rating.

What's missing from the Mohs scale (and from all ordinal data) is the idea of equal intervals and equal differences. The difference between a hardness of 10 and a hardness of 8 is not the same as the difference between a hardness of 6 and a hardness of 4.

Interval data provide equal differences. Fahrenheit temperatures provide an example of interval data. The difference between 60 degrees and 70 degrees is the same as the difference between 80 degrees and 90 degrees.

Here's something that might surprise you about Fahrenheit temperatures: A temperature of 100 degrees isn't twice as hot as a temperature of 50 degrees. For ratio statements (twice as much as, half as much as) to be valid, zero has to mean the complete absence of the attribute you're measuring. A temperature of 0 degrees F doesn't mean the absence of heat — it's just an arbitrary point on the Fahrenheit scale.

The last data type, *ratio* data, includes a meaningful zero point. For temperatures, the Kelvin scale gives ratio data. One hundred degrees Kelvin is twice as hot as 50 degrees Kelvin. This is because the Kelvin zero point is *absolute zero,* where all molecular motion (the basis of heat) stops. Another example is a ruler. Eight inches is twice as long as four inches. A length of zero means a complete absence of length.

REMEMBER

Any of these data types can form the basis of an independent variable or a dependent variable. The analytical tools you use depend on the type of data you're dealing with.

A little probability

When statisticians make decisions, they express their confidence about those decisions in terms of probability. They can never be certain about what they decide. They can only tell you how probable their conclusions are.

So, what is probability? The best way to attack this is with a few examples. If you toss a coin, what's the probability that it comes up heads? Intuitively, you know that if the coin is fair, you have a 50-50 chance of heads and a 50-50 chance of tails. In terms of the kinds of numbers associated with probability, that's ½.

How about rolling a die? (That's one member of a pair of dice.) What's the probability that you roll a 3? Hmm. . . . A die has six faces and one of them is 3, so that ought to be ⅙, right? Right.

Here's one more. You have a standard deck of playing cards. You select one card at random. What's the probability that it's a club? Well, a deck of cards has four suits, so that answer is ¼.

I think you're getting the picture. If you want to know the probability that an event occurs, figure out how many ways that event can happen and divide by the total number of events that can happen. In each of the three examples, the event we're interested in (heads, 3, or club) happens only one way.

Things can get a bit more complicated. When you toss a die, what's the probability you roll a 3 or a 4? Now you're talking about two ways the event you're interested in can occur, so that's $(1+1)/6 = {}^2\!/_6 = {}^1\!/_3$. What about the probability of rolling an even number? That has to be 2, 4, or 6, and the probability is $(1+1+1)/6 = {}^3\!/_6 = {}^1\!/_2$.

On to another kind of probability question. Suppose you roll a die and toss a coin at the same time. What's the probability you roll a 3 and the coin comes up heads? Consider all the possible events that can occur when you roll a die and toss a coin at the same time. The outcome can be a head and 1-6 or a tail and 1-6. That's a total of 12 possibilities. The head-and-3 combination can happen only one way, so the answer is ${}^1\!/_{12}$.

In general, the formula for the probability that a particular event occurs is

$$\Pr(\mathit{Event}) = \frac{\mathit{Number\ of\ ways\ the\ event\ can\ occur}}{\mathit{Total\ number\ of\ possible\ events}}$$

I begin this section by saying that statisticians express their confidence about their decisions in terms of probability, which is really why I brought up this topic in the first place. This line of thinking leads me to *conditional* probability — the probability that an event occurs given that some other event occurs. For example, suppose I roll a die, take a look at it (so that you can't see it), and tell you I've rolled an even number. What's the probability that I've rolled a 2? Ordinarily, the probability of a 2 is ⅙, but I've narrowed the field. I've eliminated the three odd numbers (1, 3, and 5) as possibilities. In this case, only the three even numbers (2, 4, and 6) are possible, so now the probability of rolling a 2 is ⅓.

Exactly how does conditional probability play into statistical analysis? Read on.

Inferential Statistics: Testing Hypotheses

In advance of doing a study, a statistician draws up a tentative explanation — a *hypothesis* — of why the data might come out a certain way. After the study is complete and the sample data are all tabulated, the statistician faces the essential decision every statistician has to make: whether or not to reject the hypothesis.

That decision is wrapped in a conditional probability question — what's the probability of obtaining the sample data, given that this hypothesis is correct? Statistical analysis provides tools to calculate the probability. If the probability turns out to be low, the statistician rejects the hypothesis.

Suppose you're interested in whether or not a particular coin is fair — whether it has an equal chance of coming up heads or tails. To study this issue, you'd take the coin and toss it a number of times — say, 100. These 100 tosses make up your sample data. Starting from the hypothesis that the coin is fair, you'd expect that the data in your sample of 100 tosses would show around 50 heads and 50 tails.

If it turns out to be 99 heads and 1 tail, you'd undoubtedly reject the fair coin hypothesis. Why? The conditional probability of getting 99 heads and 1 tail given a fair coin is very low. Wait a second. The coin could still be fair and you just happened to get a 99-1 split, right? Absolutely. In fact, you never really know. You have to gather the sample data (the results from 100 tosses) and make a decision. Your decision might be right, or it might not.

Juries face this dilemma all the time. They have to decide among competing hypotheses that explain the evidence in a trial. (Think of the evidence as data.) One hypothesis is that the defendant is guilty. The other is that the defendant is not guilty. Jury members have to consider the evidence and, in effect, answer a conditional probability question: What's the probability of the evidence given that the defendant is not guilty? The answer to this question determines the verdict.

Null and alternative hypotheses

Consider once again the coin tossing study I mention in the preceding section. The sample data are the results from the 100 tosses. Before tossing the coin, you might start with the hypothesis that the coin is a fair one so that you expect an equal number of heads and tails. This starting point is called the *null hypothesis.* The statistical notation for the null hypothesis is H_o. According to this hypothesis, any heads-tails split in the data is consistent with a fair coin. Think of it as the idea that nothing in the results of the study is out of the ordinary.

An alternative hypothesis is possible: The coin isn't a fair one, and it's loaded to produce an unequal number of heads and tails. This hypothesis says that any heads-tails split is consistent with an unfair coin. The alternative hypothesis is called, believe it or not, the *alternative hypothesis.* The statistical notation for the alternative hypothesis is H_1.

With the hypotheses in place, toss the coin 100 times and note the number of heads and tails. If the results are something like 90 heads and 10 tails, it's a good idea to reject H_o. If the results are around 50 heads and 50 tails, don't reject H_o. Similar ideas apply to the reading speed example I give earlier, in the section "Samples and populations." One sample of children receives reading instruction under a new method designed to increase reading speed, and the other learns via a traditional method. Measure the children's reading speeds before and after instruction and tabulate the improvement for each child. The null hypothesis, H_o, is that one method isn't different from the other. If the improvements are greater with the new method than with the traditional method — so much greater that it's unlikely that the methods aren't different from one another — reject H_o. If they're not greater, don't reject H_o.

REMEMBER

Notice that I did *not* say "accept H_o." The way the logic works, you *never* accept a hypothesis. You either reject H_o or don't reject H_o.

Here's a real-world example to help you understand this idea. Whenever a defendant goes on trial, that person is presumed innocent until proven guilty. Think of *innocent* as H_o. The prosecutor's job is to convince the jury to reject H_o. If the jurors reject, the verdict is *guilty.* If they don't reject, the verdict is *not guilty.* The verdict is never *innocent.* That would be like accepting H_o.

Back to the coin tossing example. Remember I said "around 50 heads and 50 tails" is what you could expect from 100 tosses of a fair coin. What does *around* mean? Also, I said if it's 90-10, reject H_o. What about 85-15? 80-20? 70-30? Exactly how much different from 50-50 does the split have to be for you to reject H_o? In the reading speed example, how much greater does the improvement have to be to reject H_o?

I don't answer these questions now. Statisticians have formulated decision rules for situations like this, and you explore those rules throughout the book.

Two types of error

Whenever you evaluate the data from a study and decide to reject H_o or to not reject H_o, you can never be absolutely sure. You never really know what the true state of the world is. In the context of the coin tossing example, that means you never know for certain if the coin is fair or not. All you can do is make a decision based on the sample data you gather. If you want to be certain about the coin, you'd have to have the data for the entire population of tosses — which means you'd have to keep tossing the coin until the end of time.

Because you're never certain about your decisions, it's possible to make an error regardless of what you decide. As I mention earlier in this chapter, the coin could be fair and you just happen to get 99 heads in 100 tosses. That's not likely, and that's why you reject H_o. It's also possible that the coin is biased, yet you just happen to toss 50 heads in 100 tosses. Again, that's not likely and you don't reject H_o in that case.