LEARNING MADE EASY

R

ALL-IN-ONE

for dummies®
A Wiley Brand

5 Books in one!

Joseph Schmuller, PhD
Author of *Statistical Analysis with R For Dummies*

R

ALL-IN-ONE

by Joseph Schmuller

for dummies®
A Wiley Brand

R All-in-One For Dummies®

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

Published simultaneously in Canada

Library of Congress Control Number: 2022950749

ISBN: 978-1-119-98369-9 (pbk); 978-1-119-98370-5 (ebk); 978-1-119-98371-2 (ebk)

Contents at a Glance

Table of Contents

Introduction

In this book, I've brought together all the information you need to hit the ground running with R. It's heavy on statistics, of course, because R's creators built this language to analyze data.

So it's necessary that you learn the foundations of statistics. Let me tell you at the outset: This *All-in-One* is not a cookbook. I've never taught statistics that way and I never will. Before I show you how to use R to work with a statistical concept, I give you a strong grounding in what that concept is all about.

In fact, Books 2 and 3 of this 5-book compendium are something like an introductory statistics text that happens to use R as a way of explaining statistical ideas.

Book 4 follows that path by teaching the ideas behind machine learning before you learn how to use R to implement them. Book 5 gives you a set of projects that give you a chance to exercise your newly minted R skill set.

Want some more details? Read on.

About This All-in-One

The volume you're holding (or the e-book you're viewing) consists of five books that cover a lot of the length and breadth of R.

Book 1: Introducing R

As I said earlier in this introduction, R is a language that deals with statistics. Accordingly, Book 1 introduces you to the fundamental concepts of statistics that you just *have* to know in order to progress with R.

You then learn about R and RStudio, a widely used development environment for working with R. I begin by describing the rudiments of R code, and I discuss R functions and structures.

R truly comes alive when you use its specialized packages, which you learn about early on.

Book 2: Describing Data

Part of working with statistics is to summarize data in meaningful ways. In Book 2, you find out how to do just that.

Most people know about averages and how to compute them. But that's not the whole story. In Book 2, I tell you about additional descriptive statistics that fill in the gaps, and I show you how to use R to calculate and work with those statistics. You also learn to create graphics that visualize the data descriptions and analyses you encounter in Books 2 and 3.

Book 3: Analyzing Data

Book 3 addresses the fundamental aim of statistical analysis: to go beyond the data and help you make decisions. Usually, the data are measurements of a sample taken from a large population. The goal is to use these data to figure out what's going on in the population.

This opens a wide range of questions: What does an average mean? What does the difference between two averages mean? Are two things associated? These are only a few of the questions I address in Book 3, and you learn to use the R tools that help you answer them.

Book 4: Learning from Data

Effective machine learning model creation comes with experience. Accordingly, in Book 4 you gain experience by completing machine learning projects. In addition to the projects you complete along with me, I suggest additional projects for you to try on your own.

I begin by telling you about the University of California-Irvine Machine Learning Repository, which provides the data sets for most of the projects you encounter in Book 4.

To give you a gentle on-ramp into the field, I show you the `Rattle` package for creating machine learning applications. It's a friendly interface to R's machine learning functionality. I like `Rattle` a lot, and I think you will, too. You use it to learn about and work with decision trees, random forests, support vector machines, k-means clustering, and neural networks.

You also work with fairly large data sets — not the terabytes and petabytes data scientists work with, but large enough to get you started. In one project, you analyze a data set of more than 500,000 airline flights. In another, you complete a customer segmentation analysis of over 300,000 customers of an online retailer.

Book 5: Harnessing R: Some Projects to Keep You Busy

As its title suggests, Book 5 is also organized around projects.

In these projects, you create applications that respond to users. I show you the `shiny` package for working with web browsers and the `shinydashboard` package for creating dashboards.

All this is a little far afield from R's original mission in life, but you get an idea of R's potential to expand in new directions.

After you've worked with R for a while, maybe you can discover some of those new directions!

What You Can Safely Skip

Any reference book throws a lot of information at you, and this one is no exception. I intended it all to be useful, but I didn't aim it all at the same level. So if you're not deeply into the subject matter, you can avoid paragraphs marked with the Technical Stuff icon, and you can also skip the sidebars.

Foolish Assumptions

I'm assuming that you

- Know how to work with Windows or the Mac. I don't go through the details of pointing, clicking, selecting, and so forth.
- Can install R and RStudio (I show you how in Book 1) and follow along with the examples. I use the Windows version of RStudio, but you should have no problem if you're working on a Mac.

Icons Used in This Book

As is the case in all *For Dummies* books, icons help guide you through your journey. Each one is a little picture in the margin that lets you know something special about the paragraph it's next to.

This icon points out a hint or a shortcut that helps you in your work and makes you an all-around better person.

This one points out timeless wisdom to take with you as you continue on the path to enlightenment.

Pay attention to this icon. It's a reminder to avoid something that might gum up the works for you.

As I mention in "What You Can Safely Skip," this icon indicates material you can blow past if it's just too technical. (I've kept this content to a minimum.)

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 *R All-in-One For Dummies Cheat Sheet* in the Search box.

If you've read any of my earlier books, welcome back!

Where to Go from Here

Time to hit the books! You can start from anywhere, but here are a couple of hints. Want to introduce yourself to R and packages? Book 1 is for you. Has it been a while (or maybe never?) since your last statistics course? Hit Book 2. For anything else, find it in the table of contents or in the index and go for it.

If you prefer to read from cover to cover, just turn the page. . . .

1 Introducing R

Contents at a Glance

IN THIS CHAPTER

- Introducing statistics
- Getting R and RStudio on your computer
- Starting a session with R
- Working with R functions
- Working with R structures

Chapter 1

R: What It Does and How It Does It

So you're ready to journey into the wonderful world of R! Designed by and for statisticians and data scientists, R has a short but illustrious history.

In the 1990s, Ross Ihaka and Robert Gentleman developed R at the University of Auckland, New Zealand. The R Core Team and the R Foundation for Statistical Computing support R, which has a huge worldwide user base.

Before I tell you about R, however, I have to introduce you to the world that R lives in — the world of data and statistics.

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

The analytical tools that R provides are based on statistical concepts I help you explore in this section. As you'll see, these concepts are based on common sense.

Samples and populations

If you watch TV on election night, you know that one of the main events is the prediction of the outcome immediately after the polls close (and before all the votes are counted). How is it that pundits almost always get it right?

The idea is to talk to a *sample* of voters right after they vote. If they're truthful about how they marked their ballots, and if the sample is representative of the *population* of voters, analysts can use the sample data to draw conclusions about the population.

That, in a nutshell, is what statistics is all about — using the data from samples to draw conclusions about populations.

Here's another example. Imagine that your job is to find the average height of 10-year-old children in the United States. Because you probably wouldn't have the time or the resources to measure every child, you'd measure the heights of a representative sample. Then you'd average those heights and use that average as the 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 later section "Inferential Statistics: Testing Hypotheses."


REMEMBER

Here's some important terminology: Properties of a population (like the population average) are called *parameters*, and properties of a sample (like the sample average) are called *statistics*. If your only concern is the sample properties (like the heights of the children in your sample), the statistics you calculate are *descriptive*. (I discuss descriptive statistics in Book 2.) If you're concerned about estimating the population properties, your statistics are *inferential*. (I discuss inferential statistics in Book 3.)


REMEMBER

Now for an important convention about notation: 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.

Variables: Dependent and independent

A *variable* is something that can take on more than one value — like your age, the value of the dollar against other currencies, or the number of games your favorite sports team wins. Something that can have only one value is a *constant*. Scientists tell us that the speed of light is a constant, and we use the constant π to calculate the area of a circle.

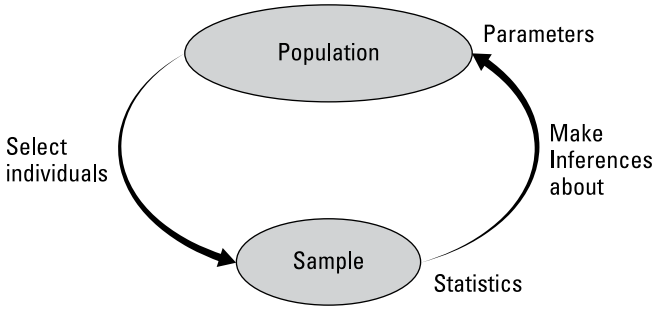

FIGURE 1-1: The relationship between populations, samples, parameters, and statistics.

Statisticians work with *independent* variables and *dependent* variables. In any study or experiment, you'll find both kinds. Statisticians assess the relationship between them.

For example, imagine a computerized training method designed to increase a person's IQ. How would a researcher find out whether this method does what it's supposed to do? First, the researcher would randomly assign a sample of people to one of two groups. One group would receive the training method, and the other would complete another kind of computer-based activity — like reading text on a website. Before and after each group completes its activities, the researcher measures each person's IQ. What happens next? I discuss that topic in the later section "Inferential Statistics: Testing Hypotheses."

For now, understand that the independent variable here is Type of Activity. The two possible values of this variable are IQ Training and Reading Text. The dependent variable is the change in IQ from Before to After.

REMEMBER

A dependent variable is what a researcher *measures*. In an experiment, an independent variable is what a researcher *manipulates*. In other contexts, a researcher can't manipulate an independent variable. Instead, they note naturally occurring values of the independent variable and how they affect a dependent variable.

REMEMBER

In general, the objective is to find out whether changes in an independent variable are associated with changes in a dependent variable.

In examples that appear throughout this book, I show you how to use R to calculate characteristics of groups of scores, or to compare groups of scores. Whenever I show you a group of scores, I'm talking about the values of a dependent variable.

Types of data

When you do statistical work, you can run into four kinds of data. And when you work with a variable, the way you work with it depends on what kind of data it is:

The first kind is *nominal* data. If a set of numbers happens to be nominal data, the numbers are labels — their values don't signify anything. On a sports team, the jersey numbers are nominal. They just identify the players.

The next kind is *ordinal* data. In this data type, the numbers are more than just labels. As the name *ordinal* might tell you, the order of the numbers is important. If I ask you to rank ten foods from the one you like best (1) to the one you like least (10), we'd have a set of ordinal data.

But the difference between your third-favorite food and your fourth-favorite food might not be the same as the difference between your ninth-favorite and your tenth-favorite. So this type of data lacks equal intervals and equal differences.

Interval data gives us equal differences. The Fahrenheit scale of temperature is a good example. The difference between 30° and 40° is the same as the difference between 90° and 100°. So each degree is an interval.

People are sometimes surprised to find out that on the Fahrenheit scale a temperature of 80° is not twice as hot as 40°. For ratio statements ("twice as much as," "half as much as") to make sense, *zero* has to mean the complete absence of the thing you're measuring. A temperature of 0° F doesn't mean the complete absence of heat — it's just an arbitrary point on the Fahrenheit scale. (The same holds true for Celsius.)

The fourth kind of data, *ratio*, provides a meaningful zero point. On the Kelvin scale of temperature, *zero* means absolute zero, where all molecular motion (the basis of heat) stops. So 200° Kelvin is twice as hot as 100° Kelvin. Another example is length. Eight inches is twice as long as 4 inches. *Zero inches* means a complete absence of length.


REMEMBER

An independent variable or a dependent variable can be either nominal, ordinal, interval, or ratio. The analytical tools you use depend on the type of data you work with.

A little probability

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

What do we mean by *probability?* Mathematicians and philosophers might give you complex definitions. In my experience, however, the best way to understand probability is in terms of examples.

Here's a simple example: If you toss a coin, what's the probability that it turns up heads? If the coin is fair, you might figure that you have a 50-50 chance of heads and a 50-50 chance of tails. And you'd be right. In terms of the kinds of numbers associated with probability, that's ½.

Think about rolling a fair die (one member of a pair of dice). What's the probability that you roll a 4? Well, a die has six faces and one of them is 4, so that's ⅙.

Still another example: Select one card at random from a standard deck of 52 cards. What's the probability that it's a diamond? A deck of cards has four suits, so that's ¼.

These examples tell you that if you want to know the probability that an event occurs, count how many ways that event can happen and divide by the total number of events that can happen. In the first two examples (heads, 4), the event you're interested in happens only one way. For the coin, we divide 1 by 2. For the die, we divide 1 by 6. In the third example (diamond), the event can happen 13 ways (Ace through King), so we divide 13 by 52 (to get ¼).

Now for a slightly more complicated example. Toss a coin and roll a die at the same time. What's the probability of tails and a 4? Think about all the possible events that can happen when you toss a coin and roll a die at the same time. You could have tails and 1 through 6, or heads and 1 through 6. That adds up to 12 possibilities. The tails-and-4 combination can happen only one way. So the probability $\frac{1}{12}$.

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

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

At the beginning of this section, I say that statisticians express their confidence about their conclusions in terms of probability, which is why I brought all this up in the first place. This line of thinking leads to *conditional* probability — the probability that an event occurs given that some other event occurs. Suppose that I roll a die, look at it (so that you don't see it), and tell you that I rolled an odd number. What's the probability that I've rolled a 5? Ordinarily, the probability of a 5 is ⅙, but "I rolled an odd number" narrows it down. That piece of information eliminates the three even numbers (2, 4, 6) as possibilities. Only the three odd numbers (1,3, 5) are possible, so the probability is ⅓.

What's the big deal about conditional probability? What role does it play in statistical analysis? Read on.

Inferential statistics: Testing hypotheses

Before a statistician does a study, they draw up a tentative explanation — a *hypothesis* that tells why the data might come out a certain way. After gathering all the data, the statistician has to decide whether to reject the hypothesis.

That decision is the answer to a conditional probability question — what's the probability of obtaining the data, given that this hypothesis is correct? Statisticians have tools that calculate the probability. If the probability turns out to be low, the statistician rejects the hypothesis.

Back to coin-tossing for an example: Imagine that you're interested in whether a particular coin is fair — whether it has an equal chance of heads or tails on any toss. Let's start with "The coin is fair" as the hypothesis.

To test the hypothesis, you'd toss the coin a number of times — let's say 100. These 100 tosses are the sample data. If the coin is fair (as per the hypothesis), you'd expect 50 heads and 50 tails.

If it's 99 heads and 1 tail, you'd surely reject the fair-coin hypothesis: The conditional probability of 99 heads and 1 tail given a fair coin is very low. Of course, the coin could still be fair and you could, quite by chance, get a 99-1 split, right? Sure. You never really know. You have to gather the sample data (the 100 toss results) and then decide. Your decision might be right, or it might not.

Juries make these types of decisions. In the United States, the starting hypothesis is that the defendant is not guilty ("innocent until proven guilty"). Think of the evidence as data. Jury members consider the evidence and answer a conditional probability question: What's the probability of the evidence, given that the defendant is not guilty? Their answer determines the verdict.

Null and alternative hypotheses

Think again about that coin-tossing study I just mentioned. The sample data are the results from the 100 tosses. I said that we can start with the hypothesis that the coin is fair. 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 sample data is out of the ordinary.

An alternative hypothesis is possible — that the coin isn't a fair one and it's biased to produce an unequal number of heads and tails. This hypothesis says that any heads-tails split is consistent with an unfair coin. This alternative hypothesis is