LEARNING MADE EASY

2nd Edition

Statistics II

for dummies®
A Wiley Brand

Learn to analyze big data sets

Explore important intermediate statistical techniques

Use statistical software for real-world applications

Deborah J. Rumsey, PhD

Associated Professor of Statistics
The Ohio State University

Statistics II

for dummies®
A Wiley Brand

Statistics II

2nd Edition

by Deborah J. Rumsey, PhD

for dummies®
A Wiley Brand

Statistics II For Dummies®, 2nd Edition

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

Published simultaneously in Canada

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: 2021946535

ISBN 978-1-119-82739-9 (pbk); ISBN 978-1-119-82740-5 (ebk); ISBN 978-1-119-82741-2 (ebk)

Contents at a Glance

Table of Contents

Introduction

So you've gone through some of the basics of statistics. Means, medians, and standard deviations all ring a bell. You know about surveys and experiments and the basic ideas of correlation and simple regression. You've studied probability, margin of error, and a few hypothesis tests and confidence intervals. Are you ready to load your statistical toolbox with a new level of tools? *Statistics II For Dummies,* 2nd Edition, picks up right where *Statistics For Dummies,* 2nd Edition, (John Wiley & Sons) leaves off and keeps you moving along the road of statistical ideas and techniques in a positive, step-by-step way.

The focus of *Statistics II For Dummies,* 2nd Edition, is on finding more ways of analyzing data. I provide step-by-step instructions for using techniques such as multiple regression, nonlinear regression, one-way and two-way analysis of variance (ANOVA), and Chi-square tests, and I give you some practice with big data sets, which are all the rage right now. Using these new techniques, you estimate, investigate, correlate, and congregate even more variables based on the information at hand, and you see how to put the tools together to create a great story about your data (nonfiction, I hope!).

About This Book

This book is designed for those who have completed the basic concepts of statistics through confidence intervals and hypothesis testing (found in *Statistics For Dummies,* 2nd Edition) and are ready to plow ahead to get through the final part of Stats I, or to tackle Stats II. However, I do pepper in some brief overviews of Stats I as needed, just to remind you of what was covered and to make sure you're up to speed. For each new technique, you get an overview of when and why it's used, how to know when you need it, step-by-step directions on how to apply it, and tips and tricks from a seasoned data analyst (yours truly). Because it's very important to be able to know which method to use when, I emphasize what makes each technique distinct and what the results tell you. You will also see many applications of the techniques used in real life.

I also include interpretation of computer output for data analysis purposes. I show you how to use the software to get the results, but I focus more on how to interpret the results found in the output, because you're more likely to be interpreting this kind of information than doing the programming specifically. Because the equations and calculations can get too involved if you are solving them by hand, you often use a computer to get your results. I include instructions for using Minitab to conduct many of the calculations in this book. Most statistics teachers who cover these topics use this approach as well. (What a relief!)

This book is different from the other Stats II books in many ways. Notably, this book features the following:

- **Full explanations of Stats II concepts.** Many statistics textbooks squeeze all the Stats II topics at the very end of their Stats I coverage; as a result, these topics tend to get condensed and presented as if they're optional. But no worries; I take the time to clearly and fully explain all the information you need to survive and thrive.
- **Dissection of computer output.** Throughout the book, I present many examples that use statistical software to analyze the data. In each case, I present the computer output and explain how I got it and what it means.
- **An extensive number of examples.** I include plenty of examples to cover the many different types of problems you'll face. Some examples are short, and some are quite extensive and include multiple variables.
- **Lots of tips, strategies, and warnings.** I share with you some trade secrets, based on my experience teaching and supporting students and grading their papers.
- **Understandable language.** I try to keep things conversational to help you understand, remember, and put into practice statistical definitions, techniques, and processes.
- **Clear and concise, step-by-step procedures.** In most chapters, you can find steps that intuitively explain how to work through Stats II problems — and remember how to do it on your own later on.

Throughout this book, I've used several conventions that I want you to be aware of:

- I indicate multiplication by using a times sign, indicated by a lowered asterisk *.
- I indicate the null and alternative hypotheses as H_o (for the null hypothesis) and H_a (for the alternative hypothesis).
- The statistical software package I use and display throughout the book is Minitab 18, but I simply refer to it as Minitab.
- Whenever I introduce a new term, I *italicize* it.
- Keywords and numbered steps appear in **boldface.**

At times I get into some of the more technical details of formulas and procedures for those individuals who may need to know about them — or just really want to get the full story. These minutiae are marked with a Technical Stuff icon. I also include sidebars along with the essential text, usually in the form of a real-life statistics example or some bonus information you may find interesting. You can feel free to skip those icons and sidebars because you won't miss any of the main information you need (but by reading them, you may just be able to impress your stats professor with your above-and-beyond knowledge of Stats II!).

Foolish Assumptions

Because this book deals with Stats II, I assume you have one previous course in introductory statistics under your belt (or at least have read *Statistics For Dummies,* 2nd Edition), with topics taking you up through the Central Limit Theorem and perhaps an introduction to confidence intervals and hypothesis tests (although I review these concepts briefly in Chapter 4). Prior experience with simple linear regression isn't necessary. Only college algebra is needed for the math details. Some experience using statistical software is also a plus, but not required.

As a student, you may be covering these topics in one of two ways: either at the tail end of your Stats I course (perhaps in a hurried way, but in some way nonetheless); or through a two-course sequence in statistics in which the topics in this book are the focus of the second course. If so, this book provides you the information you need to do well in those courses.

You may simply be interested in Stats II from an everyday point of view, or perhaps you want to add to your understanding of studies and statistical results presented in the media. If this sounds like you, you can find plenty of real-world examples and applications of these statistical techniques in action, as well as cautions for interpreting them.

Icons Used in This Book

I use icons in this book to draw your attention to certain text features that occur on a regular basis. Think of the icons as road signs that you encounter on a trip. Some signs tell you about shortcuts, and others offer more information that you may need; some signs alert you to possible warnings, while others leave you with something to remember.

COMPUTER OUTPUT

When you see this icon, it means I'm explaining how to carry out that particular data analysis using Minitab. I also explain the information you get in the computer output so you can interpret your results.

REMEMBER

I use this icon to reinforce certain ideas that are critical for success in Stats II, such as things I think are important to review as you prepare for an exam.

TECHNICAL STUFF

When you see this icon, you can skip over the information if you don't want to get into the nitty-gritty details. They exist mainly for people who have a special interest or obligation to know more about the technical aspects of certain statistical issues.

TIP

This icon points to helpful hints, ideas, or shortcuts that you can use to save time; it also includes alternative ways to think about a particular concept.

WARNING

I use warning icons to help you stay away from common misconceptions and pitfalls you may face when dealing with ideas and techniques related to Stats II.

Beyond the Book

In addition to all the great content included in the book itself, you can find even more content online. Check out this book's online Cheat Sheet on dummies.com. It covers the major formulas needed for Statistics II. You can access it by going to `www.dummies.com` and then typing "Statistics II For Dummies Cheat Sheet" into the search bar.

I've also included two major data sets that are analyzed in Chapters 20 and 21, so you can follow along with me or do your own analysis (not required!). Go to `www.dummies.com/go/statisticsIIfd2e` to access these files.

Where to Go from Here

This book is written in a nonlinear way, so you can start anywhere and still understand what's happening. However, I can make some recommendations if you want some direction on where to start.

If you're thoroughly familiar with the ideas of hypothesis testing and simple linear regression, start with Chapter 5 (multiple regression). Use Chapter 1 if you need a reference for the jargon that statisticians use in Stats II.

If you've covered all topics up through the various types of regression (simple, multiple, nonlinear, and logistic) or a subset of those as your professor deemed important, proceed to Chapter 10, the basics of analysis of variance (ANOVA).

Chapter 15 is the place to begin if you want to tackle categorical (qualitative) variables before hitting the quantitative stuff. You can work with the Chi-square test there.

Nonparametric statistics are presented starting in Chapter 17. Start there if you want the full details on the most common nonparametric procedures, used when you do not necessarily have an assumed distribution (for example, a normal).

If you want to see a bunch of Stats II ideas put into practice right off the bat, head to Chapter 19 where I discuss a multi-stage approach to analyzing a big data set, or Chapter 21, where you look into a big data set on refrigerators and see how it's analyzed in a multi-stage approach.

1

Tackling Data Analysis and Model-Building Basics

IN THIS PART . . .

Understand why data analysis is both a science and an art.

Make sure you use the right type of analysis for the job.

Work with the normal and binomial distribtions.

Reaquaint yourself with confidence intervals and hypothesis tests.

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.