

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

Copyright © 2022 by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without the prior written permission of the Publisher. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

Trademarks: Wiley, For Dummies, the Dummies Man logo, Dummies.com, Making Everything Easier, and related trade dress are trademarks or registered trademarks of John Wiley & Sons, Inc., and may not be used without written permission. All other trademarks are the property of their respective owners. Wiley Publishing, Inc., is not associated with any product or vendor mentioned in this book.

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

Introduction	1
Part 1: Tackling Data Analysis and Model-Building Basics	7
CHAPTER 1: Beyond Number Crunching: The Art and Science of Data Analysis	9
CHAPTER 2: Finding the Right Analysis for the Job	21
CHAPTER 3: Having the Normal and Sampling Distributions in Your Back Pocket. . .	37
CHAPTER 4: Reviewing Confidence Intervals and Hypothesis Tests	49
Part 2: Using Different Types of Regression to Make Predictions	65
CHAPTER 5: Getting in Line with Simple Linear Regression	67
CHAPTER 6: Multiple Regression with Two X Variables	95
CHAPTER 7: How Can I Miss You If You Won't Leave? Regression Model Selection. . .	117
CHAPTER 8: Getting Ahead of the Learning Curve with Nonlinear Regression . . .	129
CHAPTER 9: Yes, No, Maybe So: Making Predictions by Using Logistic Regression	153
Part 3: Analyzing Variance with ANOVA	167
CHAPTER 10: Testing Lots of Means? Come On Over to ANOVA!	169
CHAPTER 11: Sorting Out the Means with Multiple Comparisons	189
CHAPTER 12: Finding Your Way through Two-Way ANOVA. . .	205
CHAPTER 13: Regression and ANOVA: Surprise Relatives!	221
Part 4: Building Strong Connections with Chi-Square Tests and Nonparametrics	233
CHAPTER 14: Forming Associations with Two-Way Tables	235
CHAPTER 15: Being Independent Enough for the Chi-Square Test	257
CHAPTER 16: Using Chi-Square Tests for Goodness-of-Fit (Your Data, Not Your Jeans)	279
CHAPTER 17: Rebels Without a Distribution —Nonparametric Procedures. . .	291
CHAPTER 18: All Signs Point to the Sign Test	299

Part 5: Putting it All Together: Multi-Stage Analysis of a Large Data Set	309
CHAPTER 19: Conducting a Multi-Stage Analysis of a Large Data Set	311
CHAPTER 20: A Statistician Watches the Movies	325
CHAPTER 21: Looking Inside the Refrigerator	347
Part 6: The Part of Tens	361
CHAPTER 22: Ten Common Errors in Statistical Conclusions	363
CHAPTER 23: Ten Ways to Get Ahead by Knowing Statistics	375
CHAPTER 24: Ten Cool Jobs That Use Statistics	385
Appendix A: Reference Tables	395
Index	409

Table of Contents

INTRODUCTION	1
About This Book	1
Foolish Assumptions	3
Icons Used in This Book	3
Beyond the Book	4
Where to Go from Here	4
PART 1: TACKLING DATA ANALYSIS AND MODEL-BUILDING BASICS	7
CHAPTER 1: Beyond Number Crunching: The Art and Science of Data Analysis	9
Data Analysis: Looking before You Crunch	9
Nothing (not even a straight line) lasts forever	10
Data snooping isn't cool	11
No (data) fishing allowed	12
Getting the Big Picture: An Overview of Stats II	13
Population parameter	13
Sample statistic	13
Confidence interval	14
Hypothesis test	14
Analysis of variance (ANOVA)	15
Multiple comparisons	15
Interaction effects	16
Correlation	16
Linear regression	17
Chi-square tests	18
CHAPTER 2: Finding the Right Analysis for the Job	21
Categorical versus Quantitative Variables	22
Statistics for Categorical Variables	23
Estimating a proportion	23
Comparing proportions	24
Looking for relationships between categorical variables	25
Building models to make predictions	26
Statistics for Quantitative Variables	27
Making estimates	27
Making comparisons	28
Exploring relationships	28
Predicting y using x	30

	Avoiding Bias	31
	Measuring Precision with Margin of Error	33
	Knowing Your Limitations	35
CHAPTER 3:	Having the Normal and Sampling Distributions in Your Back Pocket	37
	Recognizing the VIP Distribution — the Normal	38
	Characterizing the normal	38
	Standardizing to the standard normal (Z-) distribution	38
	Using the normal table	40
	Finding probabilities for the normal distribution	41
	Finally Getting Comfortable with Sampling Distributions	42
	The mean and standard error of a sampling distribution	42
	Sampling distribution of \bar{X}	43
	Sampling distribution of \hat{p}	44
	Heads Up! Building Confidence Intervals and Hypothesis Tests	45
	Confidence interval for the population mean	45
	Confidence interval for the population proportion	46
	Hypothesis test for population mean	46
	Hypothesis test for the population proportion	47
CHAPTER 4:	Reviewing Confidence Intervals and Hypothesis Tests	49
	Estimating Parameters by Using Confidence Intervals	50
	Getting the basics: The general form of a confidence interval	50
	Finding the confidence interval for a population mean	51
	What changes the margin of error?	52
	Interpreting a confidence interval	55
	What's the Hype about Hypothesis Tests?	56
	What H_0 and H_a really represent	56
	Gathering your evidence into a test statistic	57
	Determining strength of evidence with a p-value	57
	False alarms and missed opportunities: Type I and II errors	58
	The power of a hypothesis test	60
	PART 2: USING DIFFERENT TYPES OF REGRESSION TO MAKE PREDICTIONS	65
CHAPTER 5:	Getting in Line with Simple Linear Regression	67
	Exploring Relationships with Scatterplots and Correlations	68
	Using scatterplots to explore relationships	69
	Collating the information by using the correlation coefficient	70
	Building a Simple Linear Regression Model	71
	Finding the best-fitting line to model your data	72
	The y-intercept of the regression line	73

The slope of the regression line	74
Making point estimates by using the regression line.	75
No Conclusion Left Behind: Tests and Confidence Intervals for Regression	75
Scrutinizing the slope	76
Inspecting the y-intercept.	78
Building confidence intervals for the average response	80
Making the band with prediction intervals	81
Checking the Model's Fit (The Data, Not the Clothes!)	83
Defining the conditions	84
Finding and exploring the residuals	85
Using r^2 to measure model fit	89
Scoping for outliers	90
Knowing the Limitations of Your Regression Analysis	92
Avoiding slipping into cause-and-effect mode	92
Extrapolation: The ultimate no-no	93
Sometimes you need more than one variable	94
CHAPTER 6: Multiple Regression with Two X Variables	95
Getting to Know the Multiple Regression Model	96
Discovering the uses of multiple regression	96
Looking at the general form of the multiple regression model	96
Stepping through the analysis.	97
Looking at x's and y's.	97
Collecting the Data	98
Pinpointing Possible Relationships.	100
Making scatterplots.	100
Correlations: Examining the bond	101
Checking for Multicollinearity.	104
Finding the Best-Fitting Model for Two x Variables	105
Getting the multiple regression coefficients	106
Interpreting the coefficients	107
Testing the coefficients.	108
Predicting y by Using the x Variables	110
Checking the Fit of the Multiple Regression Model	111
Noting the conditions	112
Plotting a plan to check the conditions	112
Checking the three conditions.	114
CHAPTER 7: How Can I Miss You If You Won't Leave? Regression Model Selection.	117
Getting a Kick out of Estimating Punt Distance.	118
Brainstorming variables and collecting data.	118
Examining scatterplots and correlations.	120

	Just Like Buying Shoes: The Model Looks Nice, But Does It Fit?	123
	Assessing the fit of multiple regression models	124
	Model selection procedures	125
CHAPTER 8:	Getting Ahead of the Learning Curve with Nonlinear Regression	129
	Anticipating Nonlinear Regression	130
	Starting Out with Scatterplots	131
	Handling Curves in the Road with Polynomials	133
	Bringing back polynomials	134
	Searching for the best polynomial model	136
	Using a second-degree polynomial to pass the quiz	138
	Assessing the fit of a polynomial model	141
	Making predictions	143
	Going Up? Going Down? Go Exponential!	145
	Recollecting exponential models	145
	Searching for the best exponential model	146
	Spreading secrets at an exponential rate	148
CHAPTER 9:	Yes, No, Maybe So: Making Predictions by Using Logistic Regression	153
	Understanding a Logistic Regression Model	154
	How is logistic regression different from other regressions?	154
	Using an S-curve to estimate probabilities	155
	Interpreting the coefficients of the logistic regression model . . .	156
	The logistic regression model in action	157
	Carrying Out a Logistic Regression Analysis	158
	Running the analysis in Minitab	158
	Finding the coefficients and making the model	160
	Estimating p	161
	Checking the fit of the model	162
	Fitting the movie model	162
	PART 3: ANALYZING VARIANCE WITH ANOVA	167
CHAPTER 10:	Testing Lots of Means? Come On Over to ANOVA!	169
	Comparing Two Means with a t-Test	170
	Evaluating More Means with ANOVA	171
	Spitting seeds: A situation just waiting for ANOVA	172
	Walking through the steps of ANOVA	173

Checking the Conditions	174
Verifying independence	174
Looking for what's normal	174
Taking note of spread	176
Setting Up the Hypotheses	178
Doing the F-Test	179
Running ANOVA in Minitab	180
Breaking down the variance into sums of squares	180
Locating those mean sums of squares	182
Figuring the F-statistic	183
Making conclusions from ANOVA	184
What's next?	186
Checking the Fit of the ANOVA Model	186

CHAPTER 11: Sorting Out the Means with Multiple Comparisons

Following Up after ANOVA	190
Comparing cellphone minutes: An example	190
Setting the stage for multiple comparison procedures	192
Pinpointing Differing Means with Fisher and Tukey	193
Fishing for differences with Fisher's LSD	194
Separating the turkeys with Tukey's test	197
Examining the Output to Determine the Analysis	198
So Many Other Procedures, So Little Time!	199
Controlling for baloney with the Bonferroni adjustment	200
Comparing combinations by using Scheffé's method	201
Finding out whodunit with Dunnett's test	202
Staying cool with Student Newman-Keuls	202
Duncan's multiple range test	202

CHAPTER 12: Finding Your Way through Two-Way ANOVA

Setting Up the Two-Way ANOVA Model	206
Determining the treatments	206
Stepping through the sums of squares	207
Understanding Interaction Effects	209
What is interaction, anyway?	209
Interacting with interaction plots	210
Testing the Terms in Two-Way ANOVA	213
Running the Two-Way ANOVA Table	214
Interpreting the results: Numbers and graphs	214
Are Whites Whiter in Hot Water? Two-Way ANOVA Investigates	217

CHAPTER 13: Regression and ANOVA: Surprise Relatives!	221
Seeing Regression through the Eyes of Variation.	222
Spotting variability and finding an “x-planation”.	222
Getting results with regression	223
Assessing the fit of the regression model	225
Regression and ANOVA: A Meeting of the Models.	226
Comparing sums of squares	226
Dividing up the degrees of freedom.	228
Bringing regression to the ANOVA table	229
Relating the F- and t-statistics: The final frontier	230

PART 4: BUILDING STRONG CONNECTIONS WITH CHI-SQUARE TESTS AND NONPARAMETRICS 233

CHAPTER 14: Forming Associations with Two-Way Tables	235
Breaking Down a Two-Way Table	236
Organizing data into a two-way table.	236
Filling in the cell counts.	237
Making marginal totals.	238
Breaking Down the Probabilities.	239
Marginal probabilities.	239
Joint probabilities.	241
Conditional probabilities	242
Trying To Be Independent	247
Checking for independence between two categories	247
Checking for independence between two variables	249
Demystifying Simpson’s Paradox	250
Experiencing Simpson’s Paradox	250
Figuring out why Simpson’s Paradox occurs	253
Keeping one eye open for Simpson’s Paradox	254

CHAPTER 15: Being Independent Enough for the Chi-Square Test	257
The Chi-Square Test for Independence	258
Collecting and organizing the data	259
Determining the hypotheses.	261
Figuring expected cell counts	261
Checking the conditions for the test.	262
Calculating the Chi-square test statistic.	263
Finding your results on the Chi-square table	266
Drawing your conclusions	269
Putting the Chi-square to the test.	271
Comparing Two Tests for Comparing Two Proportions	272

	Getting reacquainted with the Z-test for two population proportions	273
	Equating Chi-square tests and Z-tests for a two-by-two table . . .	274
CHAPTER 16:	Using Chi-Square Tests for Goodness-of-Fit (Your Data, Not Your Jeans)	279
	Finding the Goodness-of-Fit Statistic	280
	What's observed versus what's expected	280
	Calculating the goodness-of-fit statistic	282
	Interpreting the Goodness-of-Fit Statistic Using a Chi-Square	284
	Checking the conditions before you start	285
	The steps of the Chi-square goodness-of-fit test	286
CHAPTER 17:	Rebels Without a Distribution — Nonparametric Procedures	291
	Arguing for Nonparametric Statistics	292
	No need to fret if conditions aren't met	292
	The median's in the spotlight for a change	293
	So, what's the catch?	295
	Mastering the Basics of Nonparametric Statistics	296
	Sign	296
CHAPTER 18:	All Signs Point to the Sign Test	299
	Reading the Signs: The Sign Test	300
	Testing the median in real estate	302
	Estimating the median	304
	Testing matched pairs	306
PART 5: PUTTING IT ALL TOGETHER: MULTI-STAGE ANALYSIS OF A LARGE DATA SET		309
CHAPTER 19:	Conducting a Multi-Stage Analysis of a Large Data Set	311
	Steps Involved in Working with a Large Data Set	311
	Wrangling Data	313
	Discovery	313
	Structuring	314
	Cleaning	315
	Enriching	315
	Validating	316
	Publishing	317
	Visualizing Data	317
	Exploring the Data	319
	Looking for Relationships	319

Building Models and Making Inferences	320
Sharing the Story	321
Who is the audience?	322
Make an outline	322
Include an executive summary	323
Check your writing	323
CHAPTER 20: A Statistician Watches the Movies	325
Examining the Movie Variables and Asking Questions	326
Visualizing the Movie Data	327
Categorical movie variables	328
Quantitative movie variables	329
Doing Descriptive Dirty Work	332
Looking for Relationships	333
Relationships between quantitative movie variables	333
Relationships between two categorical variables	337
Relationships between quantitative and categorical variables	338
Building a Model for Predicting U.S. Revenue	340
Writing It Up	343
CHAPTER 21: Looking Inside the Refrigerator	347
Refrigerator Data — The Variables	348
Exploring the Data	348
Analyzing the Data	350
Writing It Up	358
PART 6: THE PART OF TENS	361
CHAPTER 22: Ten Common Errors in Statistical Conclusions	363
Claiming These Statistics Prove	363
It's Not Technically Statistically Significant, But	364
Concluding That x Causes y	365
Assuming the Data Was Normal	366
Only Reporting "Important" Results	366
Assuming a Bigger Sample Is Always Better	367
It's Not Technically Random, But	369
Assuming That 1,000 Responses Is 1,000 Responses	369
Of Course the Results Apply to the General Population	371
Deciding Just to Leave It Out	372
CHAPTER 23: Ten Ways to Get Ahead by Knowing Statistics	375
Asking the Right Questions	375
Being Skeptical	376
Collecting and Analyzing Data Correctly	377

Calling for Help	378
Retracing Someone Else’s Steps	379
Putting the Pieces Together	379
Checking Your Answers	380
Explaining the Output	381
Making Convincing Recommendations	382
Establishing Yourself as the Statistics Go-To Person	383
CHAPTER 24: Ten Cool Jobs That Use Statistics	385
Pollster	386
Data Scientist	387
Ornithologist (Bird Watcher)	387
Sportscaster or Sportswriter	388
Journalist	390
Crime Fighter	390
Medical Professional	391
Marketing Executive	392
Lawyer	393
APPENDIX A: REFERENCE TABLES	395
INDEX	409

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_0 (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 distributions.

Reacquaint yourself with confidence intervals and hypothesis tests.

- » 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.