Introduction to Data Analysis

(Second Edition)

By Matt Teachout
College of the Canyons
Santa Clarita, CA, USA

Special thanks to all of the people that made this book possible.

Thanks to all of the *Intermediate Algebra for Statistics* students and teachers at College of the Canyons for pioneering this material and giving great suggestions for improvement.

Thanks to the COC statistics team (Joe Gerda, Kathy Kubo, Ambika Silva, Dustin Silva) for your leadership and unending work to improve statistics education. I will always be grateful to be part of the best team ever.

Thank you to the original "honey badgers" Myra Snell and Katie Hern at the California Acceleration Project. You inspired so many teachers and programs. You made us believe we could change the system and taught us how to truly help students.

Thank you to Yousef Alasfoor, Kayla Teachout, Kathy Kubo and Udani Ranasinghe for helping me navigate through massive amounts of social justice data.

Thank you to Udani Ranasinghe for your help with Canvas and making answer keys.

Thank you to our "resident statistician" and fearless leader Joe Gerda. Your statistics expertise and leadership have been incredible. We could not have done this without you.

Special thanks to Kathy Kubo, Ralph (Randy) Ades, Udani Ranasinghe, Rupa Sinha, and Joe Gerda for your support, encouragement, help and suggestions.

Thank you to James Glapa-Grossklag, Brian Weston and the COC OER office staff for your support and help.

Introduction

We live in the age of computers and the internet. We are exposed to huge volumes of data every day. How do we make sense of this massive amount of information? How can we tell the difference between helpful and misleading information? How can businesses know what their customers want and need, or hospitals analyze various types of infections and which treatments are working and which are not? All of these questions revolve around the study of data and statistics. A good understanding of statistics is vital to anyone living in the modern world, however very few people understand how to analyze data. The shortage of trained statisticians, data analysts, and data scientists is a huge problem worldwide.

There are many fabulous books on statistics and analyzing data. Unfortunately, they are extremely expensive and most people cannot afford the cost. I wrote this book to help people learn to analyze data. It is free to use the material in this book, update it, add to it, print it or just read it. It is an open educational resource (OER) and so anyone can use it.

Many college students struggle to balance work and family with their education. One of the biggest roadblocks for many students is the cost of textbooks. Students today cannot afford the cost of textbooks and so chose to attend classes without purchasing books and materials needed for the class. It goes without saying, that this is a major impediment to passing their classes, but the students have no choice. They simply cannot afford $100-$200 for a textbook. For this reason, I believe strongly in open educational resources (OER). Open source materials like this book are available and are virtually free for students.

Notes about OER and Creative Commons Licensing

This textbook is licensed through Creative Commons as "Attribution CC-BY". Creative Commons describes this license as follows: "This license lets others distribute, remix, tweak, and build upon (the author's) work, even commercially, as long as they (give) credit (to the author) for the original creation." This is the most accommodating of licenses offered. Recommended for maximum dissemination and use of licensed materials." If you need to see the license deed or legal code, please go to https://creativecommons.org/licenses/ and look under the "CC-By" section.

Pre-Statistics or Intermediate Algebra for Statistics

I tell my beginning statistics students all the time that the study of statistics is a deep well of knowledge, and they are only playing in the puddle. Statisticians, data analysts and data scientists are life-long learners and spend years and years studying this subject.

This is an introduction to some very basic data analysis techniques. It is a book designed for anyone new to statistics. It can be used with a pre-statistics class or an intermediate algebra for statistics class.

Pre-statistics classes focus on helping students understand and analyze categorical and quantitative data sets.

Intermediate algebra for statistics has the same information as a pre-statistics class but often includes some intermediate algebra curve analysis and regression techniques. Many statisticians and statistics educators feel that curve analysis and regression is a topic better addressed in more advanced level statistics classes since this is a topic explored by many graduate level statistics students.

If your college requires intermediate algebra for statistics, I have included that material in chapter 6. If your college is using a pre-statistics class, then chapters 1-5 should suffice.

Important Note about Technology

We live in the age of computers, internet and a huge volume of data. No practicing statistician or data scientist uses a calculator or tables to analyze data. You cannot even begin to analyze a data set with a million values by hand with a calculator. You need high-powered computer software. There are many statistics software programs on the market, but very few of them are free.

If you read the history of statistics, you will find brilliant scientists, mathematicians and people in business who had to try to figure out data, but had no access to a computer. (Computers had not been invented yet.) Our pioneers of statistics dreamed of the day that they could compute statistics and graphs and analyze data with the touch of a button. They invented complicated techniques for analyzing data because they had no choice. Today, computers can calculate statistics and graphs directly.

Here is the problem. Most statistics classes taught in high schools, community colleges and even some universities are teaching statistics as if computers have not been invented yet. They are teaching the techniques developed by our pioneers of statistics before the computer age. This is a terrible approach to the subject, especially for the thousands of students that actually want to work in the field. A statistics class should be a study of how to practically collect and analyze data with a computer, not a class on what to do if computers have not been invented yet.

Are formulas important in statistics? Yes. We study formulas to understand what they tell us about the data and the world around us. The pioneers of statistics did an amazing job of addressing the major ideas of statistics with formulas and inventive calculations. However, we should not use a formula and a calculator to calculate a statistic for a data set with 10,000 values or use charts that list critical values and P-values. High-powered computers with statistics software can calculate the statistic and make graphs directly. Then students can focus on the analysis part, and explore and discover the meaning behind the data.

This book will show students how to use statistics software to calculate statistics and graphs. I want students to learn to analyze the data and not spend all their time just trying to calculate something. Remember, no one pays a data analyst to calculate something a computer can already do. They are paid to explore and explain what the data may be telling us.

Data Sets

The national (GAISE) guidelines for teaching statistics recommend that you use real data. Allowing students to learn statistics principles through analysis of real data is key. With that being said, there are many places where raw data can found and used. The key data sets throughout this book are located at my website www.matt-teachout.org. Just click on "Int Alg for Stats" and then "Data Sets".

The Computer Dilemma

Face to Face Classes

A statistics or pre-statistics class should be taught in a computer lab. It is important to allow the computers to do the difficult calculations. Students need to focus on interpretation and discovering the meaning behind the data. They cannot do that if they spend all their time trying to calculate with a formula or making graphs by hand.

If your school wants to teach statistics or pre-statistics, but you cannot teach in a computer lab, here are some suggestions for you.

1. Reserve unused computer labs. Some schools may have a couple computer labs that are not always in use. Schedule your statistics and pre-statistics classes in order to use the computer lab. Even if you can only reserve the lab once a week or once every two weeks, it will be a huge help to students.

2. Have groups of students share computers. If you do have a few computers in your classroom, you can dived the class up into groups and share computers. This has many advantages like encouraging explanations to one another and teamwork.

3. Teachers can use their own computer or laptop to project statistics software on a screen and have the class analyze the data with you. Teachers without any computer can make printed copies of the software printouts for your class and for exams. It is a poor substitute for a computer lab, but it is much better than teaching statistics as if computers have not been invented yet.

Free Statistics Software

Teaching statistics with computer software is very important, but many schools and students cannot afford to pay for software. If you are teaching pre-stats, intermediate algebra for stats, or intro stats online or face-to-face, it is vital that students have access to statistics software. I highly recommend a free statistics software that is easy to use. Most free software is not OER licensed, but they are still free for students. My favorite free statistics software programs are StatKey (www.lock5stat.com) and Statcato (www.statcato.org). I use both of these programs throughout the textbook.

Notes about the 2nd Edition

The second edition of Introduction to Data Analysis is similar to the first edition, but there are a few key differences.

1. **Change in software:** The 1st edition only used Statcato. Statcato is a great program and free for students, but is difficult to use with online classes. It works great in the classroom when every computer in the class has Statcato installed. However, some students have a hard time downloading it on their home computers. So for calculations, I moved the 2nd edition to using the free program StatKey (www.lock5stat.com). It is free, online hosted and so does not need to be downloaded. It works great on MAC and PC. It is ideal for online classes. There is still some Statcato in the book, but students only have to analyze Statcato printouts provided and not calculate. If they are calculating, they are using StatKey. The book shows students and faculty how to use StatKey.

2. **Added chapter 1 on Data:** I wanted to expand on the ideas of types of data, random, bias, collecting data and experimental design in the textbook, so I included a new chapter 1 on data. The 2nd edition of the book has 7 chapters while the 1st edition has 6 chapters. Chapters 2-7 in the 2nd edition have similar content as chapters 1-6 in the 1st edition.

3. **Added social justice questions:** The 2nd edition of the book has social justice questions dealing with racism and discrimination.

4. **Updated Projects:** There are optional projects available for chapters 2-6 in the 2nd edition of the textbook. The projects have updated directions for both online and face-to-face classes.

Table of Contents
Introduction to Data Analysis (2nd edition)
