Introduction to Data Analysis

(First Edition)

By Matt Teachout

College of the Canyons
Santa Clarita, CA, USA

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

Most especially all of the *Intermediate Algebra for Statistics* students and teachers at College of the Canyons. Your work to pioneer this material and improve statistics education continues to inspire education reform.

I would also like to thank Kathy Kubo, Joe Gerda, Dustin Silva, Ambika Silva, Ralph (Randy) Ades, Udani Ranasinghe, and Rupa Sinha,

Myra Snell and Katie Hern at the California Acceleration Project,

James Glapa-Grossklag and Brian Weston and the COC OER office staff,

The incredible COC math department, reprographics, and the rest of the COC faculty and staff.

Your help and support made this possible.

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 $150-$200 textbooks. 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

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 of 100,000 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 look at 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.

Statcato

Teaching statistics with computer software is very important, but many schools and students cannot afford to pay for software. For this reason, I used the free statistics software Statcato throughout the book. Statcato is a JAVA based software, which can be saved to any computer (MAC or PC) and is easy to use.

Statcato Website: www.statcato.org

Statcato Download for MAC: http://mac.softpedia.com/get/Utilities/Statcato.shtml

Statcato Download for PC: https://sourceforge.net/projects/statcato/

You can of course use the book with any statistics software. Most basic statistics software programs are very similar to Statcato. You will just need to supplement software directions.

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 under "Int Alg for Stats" and "Data Sets".

Matt Teachout's Website: www.matt-teachout.org

The Computer Dilemma

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.