INTRODUCTORY STATISTICS AND ANALYTICS

A RESAMPLING PERSPECTIVE

PETER C. BRUCE

WILEY

INTRODUCTORY STATISTICS AND ANALYTICS

A Resampling Perspective

PETER C. BRUCE
Institute for Statistics Education
Statistics.com
Arlington, VA

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data is available.

ISBN: 978-1-118-88135-4

Printed in the United States of America

INTRODUCTORY STATISTICS AND ANALYTICS

PREFACE

This book was developed by Statistics.com to meet the needs of its introductory students, based on experience in teaching introductory statistics online since 2003. The field of statistics education has been in ferment for several decades. With this book, which continues to evolve, we attempt to capture three important strands of recent thinking:

1. Connection with the field of *data science*—an amalgam of traditional statistics, newer machine learning techniques, database methodology, and computer programming to serve the needs of large organizations seeking to extract value from "big data."
2. Guidelines for the introductory statistics course, developed in 2005 by a group of noted statistics educators with funding from the American Statistical Association. These Guidelines for Assessment and Instruction in Statistics Education (GAISE) call for the use of real data with active learning, stress statistical literacy and understanding over memorization of formulas, and require the use of software to develop concepts and analyze data.
3. The use of resampling/simulation methods to develop the underpinnings of statistical inference (the most difficult topic in an introductory course) in a transparent and understandable manner.

We start off with some examples of statistics in action (including two of statistics gone wrong) and then dive right in to look at the proper design of studies and account for the possible role of chance. All the standard topics of introductory statistics are here (probability, descriptive statistics, inference, sampling, correlation, etc.), but sometimes, they are introduced not as separate standalone topics but rather in the context of the situation in which they are needed.

Throughout the book, you will see "Try It Yourself" exercises. The answers to these exercises are found at the end of each chapter after the homework exercises.

BOOK WEBSITE

Data sets, Excel worksheets, software information, and instructor resources are available at the book website: www.introductorystatistics.com

PETER C. BRUCE

ACKNOWLEDGMENTS

STAN BLANK

The programmer for Resampling Stats, Stan has participated actively in many sessions of Statistics.com courses based on this work and has contributed well both to the presentation of regression and to the clarification and improvement of sections that deal with computational matters.

MICHELLE EVERSON

Michelle Everson, editor (2013) of the *Journal of Statistics Education*, has taught many sessions of the introductory sequence at Statistics.com and is responsible for the material on decomposition in the ANOVA chapter. Her active participation in the statistics education community has been an asset as we have strived to improve and perfect this text.

ROBERT HAYDEN

Robert Hayden has taught early sessions of this course and has written course materials that served as the seed from which this text grew. He was instrumental in getting this project launched.

In the beginning, Julian Simon, an early resampling pioneer, first kindled my interest in statistics with his permutation and bootstrap approach to statistics, his Resampling Stats software (first released in the late 1970s), and his statistics text on the same subject. Simon, described as an "iconoclastic polymath" by Peter Hall in his "Prehistory of the Bootstrap," (*Statistical Science*, 2003, vol. 18, #2), is the intellectual forefather of this work.

Our Advisory Board—Chris Malone, William Peterson, and Jeff Witmer (all active in GAISE and the statistics education community in general) reviewed the overall concept and outline of this text and offered valuable advice.

Thanks go also to George Cobb, who encouraged me to proceed with this project and reinforced my inclination to embed resampling and simulation more thoroughly than what is found in typical college textbooks.

Meena Badade also teaches using this text and has also been very helpful in bringing to my attention errors and points requiring clarification and has helped to add the sections dealing with standard statistical formulas.

Kuber Deokar, Instructional Operations Supervisor at Statistics.com, and Valerie Troiano, the Registrar at STatisticscom, diligently and carefully shepherded the use of earlier versions of this text in courses at Statistics.com.

The National Science Foundation provided support for the Urn Sampler project, which evolved into the Box Sampler software used both in this course and for its early web versions. Nitin Patel, at Cytel Software Corporation, provided invaluable support and design assistance for this work. Marvin Zelen, an early advocate of urn-sampling models for instruction, shared illustrations that sharpened and clarified my thinking.

Many students at The Institute for Statistics Education at Statistics.com have helped me clarify confusing points and refine this book over the years.

Finally, many thanks to Stephen Quigley and the team at Wiley, who encouraged me and moved quickly on this project to bring it to fruition.

INTRODUCTION

As of the writing of this book, the fields of statistics and data science are evolving rapidly to meet the changing needs of business, government, and research organizations. It is an oversimplification, but still useful, to think of two distinct communities as you proceed through the book:

1. The traditional academic and medical *research communities* that typically conduct extended research projects adhering to rigorous regulatory or publication standards, and
2. Business and large organizations that use statistical methods to extract value from their data, often on the fly. Reliability and value are more important than academic rigor to this *data science community*.

IF YOU CAN'T MEASURE IT, YOU CAN'T MANAGE IT

You may be familiar with this phrase or its cousin: if you can't measure it, you can't fix it. The two come up frequently in the context of Total Quality Management or Continuous Improvement programs in organizations. The flip side of these expressions is the fact that if you do measure something and make the measurements available to decision-makers, the something that you measure is likely to change.

Toyota found that placing a real-time gas-mileage gauge on the dashboard got people thinking about their driving habits and how they relate to gas consumption. As a result, their gas mileage—miles they drove per gallon of gas—improved.

In 2003, the Food and Drug Administration began requiring that food manufacturers include trans fat quantities on their food labels. In 2008, it was found from a study that

blood levels of trans fats in the population had dropped 58% since 2000 (reported in the *Washington Post*, February 9, 2012, A3).

Thus, the very act of measurement is, in itself, a change agent. Moreover, measurements of all sorts abound—so much so that the term Big Data came into vogue in 2011 to describe the huge quantities of data that organizations are now generating.

Big Data: If You Can Quantify and Harness It, You Can Use It

In 2010, a statistician from Target described how the company used customer transaction data to make educated guesses about whether customers were pregnant or not. On the strength of these guesses, Target sent out advertising flyers to likely prospects, centered around the needs of pregnant women.

How did Target use data to make those guesses? The key was data used to "train" a statistical model: data in which the outcome of interest—pregnant/not pregnant—was known in advance. Where did Target get such data? The "not pregnant" data was easy—the vast majority of customers were not pregnant so the data on their purchases was easy to come by. The "pregnant" data came from a baby shower registry. Both datasets were quite large, containing lists of items purchased by thousands of customers.

Some clues are obvious—purchase of a crib and baby clothes is a dead giveaway. But, from Target's perspective, by the time a customer purchases these obvious big ticket items, it was too late—they had already chosen their shopping venue. Target wanted to reach customers earlier, before they decided where to do their shopping for the big day. For that, Target used statistical modeling to make use of nonobvious patterns in the data that distinguish pregnant from nonpregnant customers. One such clue was shifts in the pattern of supplement purchases—for example, a customer who was not buying supplements 60 days ago but is buying them now. Crafting a marketing campaign on the basis of educated guesses about whether a customer is pregnant aroused controversy for Target, needless to say.

Much of the book that follows deals with important issues that can determine whether data yields meaningful information or not:

- The role that random chance plays in creating apparently interesting results or patterns in data.
- How to design experiments and surveys to get useful and reliable information.
- How to formulate simple statistical models to describe relationships between one variable and another.

PHANTOM PROTECTION FROM VITAMIN E

In 1993, researchers examining a database on nurses' health found that nurses who took vitamin E supplements had 30–40% fewer heart attacks than those who did not. These data fit with theories that antioxidants such as vitamins E and C could slow damaging processes within the body. Linus Pauling, winner of the Nobel Prize in Chemistry in 1954, was a major proponent of these theories. The Linus Pauling Institute at Oregon State University is still actively promoting the role of vitamin E and other nutritional supplements in inhibiting

disease. These results provided a major boost to the dietary supplements industry. The only problem? The heart health benefits of vitamin E turned out to be illusory. A study completed in 2007 divided 14,641 male physicians randomly into four groups:

1. Take 400 IU of vitamin E every other day
2. Take 500 mg of vitamin C every day
3. Take both vitamin E and C
4. Take placebo.

Those who took vitamin E fared no better than those who did not take vitamin E. As the only difference between the two groups was whether or not they took vitamin E, if there were a vitamin E effect, it would have shown up. Several meta-analyses, which are consolidated reviews of the results of multiple published studies, have reached the same conclusion. One found that vitamin E at the above dosage might even increase mortality.

What made the researchers in 1993 think that they had found a link between vitamin E and disease inhibition? After reviewing a vast quantity of data, researchers thought that they saw an interesting association. In retrospect, with the benefit of a well-designed experiment, it appears that this association was merely a chance coincidence. Unfortunately, coincidences happen all the time in life. In fact, they happen to a greater extent than we think possible.

STATISTICIAN, HEAL THYSELF

In 1993, Mathsoft Corp., the developer of Mathcad mathematical software, acquired StatSci, the developer of S-PLUS statistical software, predecessor to the open-source R software. Mathcad was an affordable tool popular with engineers—prices were in the hundreds of dollars, and the number of users was in the hundreds of thousands. S-PLUS was a high-end graphical and statistical tool used primarily by statisticians—prices were in the thousands of dollars, and the number of users was in the thousands.

In an attempt to boost revenues, Mathsoft turned to an established marketing principle—cross-selling. In other words, trying to convince the people who bought product A to buy product B. With the acquisition of a highly regarded niche product, S-PLUS, and an existing large customer base for Mathcad, Mathsoft decided that the logical thing to do would be to ramp up S-PLUS sales via direct mail to its installed Mathcad user base. It also decided to purchase lists of similar prospective customers for both Mathcad and S-PLUS.

This major mailing program boosted revenues, but it boosted expenses even more. The company lost over $13 million in 1993 and 1994 combined—significant numbers for a company that had only $11 million in revenue in 1992.

What Happened?

In retrospect, it was clear that the mailings were not well targeted. The costs of the unopened mail exceeded the revenue from the few recipients who did respond. In particular, Mathcad users turned out to be unlikely users of S-PLUS. The huge losses could have been avoided through the use of two common statistical techniques:

1. Doing a test mailing to the various lists being considered to (a) determine whether the list is productive and (b) test different headlines, copy, pricing, and so on, to see what works best.
2. Using predictive modeling techniques to identify which names on a list are most likely to turn into customers.

IDENTIFYING TERRORISTS IN AIRPORTS

Since the September 11, 2001 Al Qaeda attacks in the United States and subsequent attacks elsewhere, security screening programs at airports have become a major undertaking, costing billions of dollars per year in the United States alone. Most of these resources are consumed by an exhaustive screening process. All passengers and their tickets are reviewed, their baggage is screened, and individuals pass through detectors of varying sophistication. An individual and his or her bag can only receive a limited amount of attention in an exhaustive screening process. The process is largely the same for each individual. Potential terrorists can see the process and its workings in detail and identify its weaknesses.

To improve the effectiveness of the system, security officials have studied ways of focusing more concentrated attention on a small number of travelers. In the years after the attacks, one technique enhanced the screening for a limited number of randomly selected travelers. Although it adds some uncertainty to the process, which acts as a deterrent to attackers, random selection does nothing to focus attention on high-risk individuals.

Determining who is of high risk is, of course, the problem. How do you know who the high-risk passengers are?

One method is passenger profiling—specifying some guidelines about what passenger characteristics merit special attention. These characteristics were determined by a reasoned, logical approach. For example, purchasing a ticket for cash, as the 2001 hijackers did, raises a red flag. The Transportation Security Administration trains a cadre of Behavior Detection Officers. The Administration also maintains a specific no-fly list of individuals who trigger special screening.

There are several problems with the profiling and no-fly approaches.

- Profiling can generate backlash and controversy because it comes close to stereotyping. American National Public Radio commentator Juan Williams was fired when he made an offhand comment to the effect that he would be nervous about boarding an aircraft in the company of people in full Muslim garb.
- Profiling, as it does tend to merge with stereotype and is based on logic and reason, enables terrorist organizations to engineer attackers who do not meet profile criteria.
- No-fly lists are imprecise (a name may match thousands of individuals) and often erroneous. Senator Edward Kennedy was once pulled aside because he supposedly showed up on a no-fly list.

An alternative or supplemental approach is a statistical one—separate out passengers who are "different" for additional screening, where "different" is defined quantitatively across many variables that are not made known to the public. The statistical term is "outlier." Different does not necessarily prove that the person is a terrorist threat, but the theory is that outliers may have a higher threat probability. Turning the work over to a statistical

algorithm mitigates some of the controversy around profiling as security officers would lack the authority to make discretionary decisions.

Defining "different" requires a statistical measure of distance, which we will learn more about later.

LOOKING AHEAD IN THE BOOK

We will be studying many things, but several important themes will be the following:

1. Learning more about random processes and statistical tools that will help quantify the role of chance and distinguish real phenomena from chance coincidence.
2. Learning how to design experiments and studies that can provide more definitive answers to questions such as whether vitamin E affects heart attack rates and whether to undertake a major direct mail campaign.
3. Learning how to specify and interpret statistical models that describe the relationship between two variables or between a response variable and several "predictor" variables, in order to
   - explain/understand phenomena and answer research questions ("Does a new drug work?" "Which offer generates more revenue?")
   - make predictions ("Will a given subscriber leave this year?" "Is a given insurance claim fraudulent?")

RESAMPLING

An important tool will be resampling—the process of taking repeated samples from observed data (or shuffling that data) to assess what effect random variation might have on our statistical estimates, our models, and our conclusions. Resampling was present in the early days of statistical science, but, in the absence of computers, was quickly superceded by formula approaches. It has enjoyed a resurgence in the last 30 years.

Resampling in Data Mining: Target Shuffling

John Elder is the founder of the data mining and predictive analytics services firm Elder Research. He tests the accuracy of his data mining results through a process he calls "target shuffling". It's a method Elder says is particularly useful for identifying false positives, or when events are perceived to have a cause-and-effect relationship, as opposed to a coincidental one.

"The more variables you have, the easier it becomes to "over-search" and identify (false) patterns among them," Elder says—what he calls the 'vast search effect.'

As an example, he points to the Redskins Rule, where for over 70 years, if the Washington Redskins won their last home football game, the incumbent party would win the presidential election. "There's no real relationship between those two things," Elder says, "but for generations, they just happened to line up."

> As hypotheses generated by automated search grow in number, it becomes easy to make inferences that are not only incorrect, but dangerously misleading. To prevent this problem, Elder Research uses target shuffling with all of their clients. It reveals how likely it is that results as strong as you found could have occurred by chance.
>
> "Target shuffling is a computer simulation that does what statistical tests were designed to when they were first invented," Elder explains. "But this method is much easier to understand, explain, and use than those mathematical formulas."

Here's how the process works. On a set of training data:

1. Build a model to predict the target variable (output) and note its strength (e.g., R-squared, lift, correlation, explanatory power).
2. Randomly shuffle the target vector to "break the relationship" between each output and its vector of inputs.
3. Search for a new best model – or "most interesting result" - and save its strength. (It is not necessary to save the model; its details are meaningless by design.)
4. Repeat steps 2 and 3 many times and create a distribution of the strengths of all the bogus "most interesting" models or findings.
5. Evaluate where your actual results (from step 1) stand on (or beyond) this distribution. This is your "significance" measure or probability that a result as strong as your initial model can occur by chance.

Let's break this down: imagine you have a math class full of students who are going to take a quiz. Before the quiz, everyone fills out a card with specified personal information, such as name, age, how many siblings they have, and what other math classes they've taken. Everyone then takes the quiz and receives their score.

To discover why certain students scored higher than others, you could model the target variable (the grade each student received) as a function of the inputs (students' personal information) to identify patterns. Let's say you find that older sisters have the highest quiz scores, which you think is a solid predictor of which types of future students will perform the best.

But depending on the size of the class and the number of questions you asked everyone, there's always a chance that this relationship is not real, and therefore won't hold true for the next class of students. (Even if the model seems reasonable, and facts and theory can be brought to support it, the danger of being fooled remains: "Every model finding seems to cause our brains to latch onto corroborating explanations instead of generating the critical alternative hypotheses we really need.")

With target shuffling, you compare the same inputs and outputs against each other a second time to test the validity of the relationship. This time, however, you randomly shuffle the outputs so each student receives a different quiz score—Suzy gets Bob's, Bob gets Emily's, and so forth.

All of the inputs (personal information) remain the same for each person, but each now has a different output (test score) assigned to them. This effectively breaks the relationship between the inputs and the outputs without otherwise changing the data.

You then repeat this shuffling process over and over (perhaps 1000 times, though even 5 times can be very helpful), comparing the inputs with the randomly assigned outputs each

time. While there should be no real relationship between each student's personal information and these new, randomly assigned test scores, you'll inevitably find some new false positives or "bogus" relationships (e.g. older males receive the highest scores, women who also took Calculus receive the highest scores, etc.).

As you repeat the process, you record these "bogus" results over the course of the 1000 random shufflings. You then have a comparison distribution that you can use to assess whether the result that you observed in reality is truly interesting and impressive or to what degree it falls in the category of "might have happened by chance."

Elder first came up with target shuffling 20 years ago, when his firm was working with a client who wasn't sure if he wanted to invest more money into a new hedge fund. While the fund had done very well in its first year, it had been a volatile ride, and the client was unsure if the success was due to luck or skill. A standard statistical test showed that the probability of the fund being that successful in a chance model was very low, but the client wasn't convinced.

So Elder performed 1,000 simulations where he shuffled the results (as described above) where the target variable was the daily buy or hold signal for the next day. He then compared the random results to how the hedge fund had actually performed.

Out of 1,000 simulations, the random distribution returned better results in just 15 instances—in other words, there was only a 1.5% chance that the hedge fund's success could occur just as the result of luck. This new way of presenting the data made sense to the client, and as a result he invested 10 times as much in the fund.[1]

"I learned two lessons from that experience," Elder says. "One is that target shuffling is a very good way to test non-traditional statistical problems. But more importantly, it's a process that makes sense to a decision maker. Statistics is not persuasive to most people—it's just too complex.

"If you're a business person, you want to make decisions based upon things that are real and will hold up. So when you simulate a scenario like this, it quantifies how likely it is that the results you observed could have arisen by chance in a way that people can understand."

BIG DATA AND STATISTICIANS

Before the turn of the millennium, by and large, statisticians did not have to be too concerned with programming languages, SQL queries, and the management of data. Database administration and data storage in general was someone else's job, and statisticians would obtain or get handed data to work on and analyze. A statistician might, for example,

- Direct the design of a clinical trial to determine the efficacy of a new therapy
- Help a psychology student determine how many subjects to enroll in a study
- Analyze data to prepare for legal testimony
- Conduct sample surveys and analyze the results
- Help a scientist analyze data that comes out of a study
- Help an engineer improve an industrial process

[1] The fund went on to do well for a decade; the story is recounted in chapter 1 of Eric Siegel's *Predictive Analytics* (Wiley, 2013)

All of these tasks involve examining data, but the number of records is likely to be in the hundreds or thousands at most, and the challenge of obtaining the data and preparing it for analysis was not overwhelming. So the task of obtaining the data could safely be left to others.

Data Scientists

The advent of big data has changed things. The explosion of data means that more interesting things can be done with data, and they are often done in real time or on a rapid turnaround schedule. FICO, the credit-scoring company, uses statistical models to predict credit card fraud, collecting customer data, merchant data, and transaction data 24 hours a day. FICO has more than two billion customer accounts to protect, so it is easy to see that this statistical modeling is a massive undertaking.

Computer programming and database administration lie beyond the scope of this course but not beyond the scope of statistical studies. See the book website for links to over 100 online courses, to get an idea of what statistics covers now. The statistician must be conversant with the data, and the data keeper now wants to learn the analytics:

- Statisticians are increasingly asked to plug their statistical models into big data environments, where the challenge of wrangling and preparing analyzable data is paramount, and requires both programming and database skills.
- Programmers and database administrators are increasingly interested in adding statistical methods to their toolkits, as companies realize that they have strategic, not just clerical value hidden in their databases.

Around 2010, the term *data scientist* came into use to describe analysts who combined these two sets of skills. Job announcements now carry the term *data scientist* with greater frequency than the term *statistician*, reflecting the importance that organizations attach to managing, manipulating, and obtaining value out of their vast and rapidly growing quantities of data.

We close with a probability experiment:

> ***Try It Yourself 1.1***
>
> Let us look first at the idea of randomness via a classroom exercise.
>
> 1. Write down a series of 50 random coin flips without actually flipping the coins. That is, write down a series of 50 Hs and Ts selected in such a way that they appear random.
> 2. Now, actually flip a coin 50 times.
>
> If you are reading this book in a course, please report your results to the class for compilation—specifically, report two lists of Hs and Ts like this: My results—Made up flips: HTHHHTT, and so on. Actual flips: TTHHTHTHTH, and so on.