1

Statistics and Data Science

Statistical methods first came into use before homes had electricity, and had several phases of rapid growth:

- The first big boost came from manufacturers and farmers who were able to decrease costs, produce better products, and improve crop yields via statistical experiments.
- Similar experiments helped drug companies graduate from snake oil purveyors to makers of scientifically proven remedies.
- In the late 20th century, computing power enabled a new class of computationally intensive methods, like the resampling methods that we will study.
- In the early decades of the current millennium, organizations discovered that the rapidly growing repositories of data they were collecting ("big data") could be mined for useful insights.

As with any powerful tool, the more you know about it the better you can apply it and the less likely you will go astray. The lurking dangers are illustrated when you type the phrase "How to lie with..." into a web search engine. The likely autocompletion is "statistics."

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

- How to assess the role that random chance can play 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

We will start our study in the next chapter with a look at how to design experiments, but, before we dive in, let's look at some statistical wins and losses from different arenas.

Statistics for Data Science and Analytics, First Edition. Peter C. Bruce, Peter Gedeck, and Janet Dobbins.

Companion website: www.wiley.com/go/Wiley_Statistics_for_Data

1.1 Big Data: Predicting Pregnancy

In 2010, a statistician from Target described how the company used customer transaction data to make educated guesses about whether customers are 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 are not pregnant, so data on their purchases is 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—the 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 non-obvious patterns in the data that distinguish pregnant from non-pregnant customers. One clue that emerged was shifts in the pattern of supplement purchases—e.g. a customer who was not buying supplements 60 days ago but is buying them now.

1.2 Phantom Protection from Vitamin E

In 1993, researchers examining a database on nurses' health found that nurses who took vitamin E supplements had 30% to 40% fewer heart attacks than those who didn't. 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, which were one driver of the nutritional supplements industry.

However, 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 268 mg 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. Since 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 happened to make the researchers in 1993 think they had found a link between vitamin E and disease inhibition? In reviewing a vast quantity of data, researchers thought 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.

1.3 Statistician, Heal Thyself

In 1993, Mathsoft Corp., the developer of Mathcad mathematical software, acquired StatSci, the developer of S-PLUS statistical software, the precursor to R. 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 looking to boost revenues, Mathsoft turned to an established marketing principle—cross-selling. In other words, try 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 1992 revenue.

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. Mathcad users turned out not to be likely 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 (1) determine whether the list is productive and (2) test different headlines, copy, pricing, etc., to see what works best.
2) Using predictive modeling techniques to identify which names on a list are most likely to turn into customers.

1.4 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 in 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 a screening process that is applied to everyone. The process is largely the same for each individual. Potential terrorists can see the process and its workings in detail and identify 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 used enhanced screening for a limited number of randomly selected travelers. While it adds some uncertainty to the screening process, which acts as a deterrent to attackers, random selection does nothing to focus attention on high-risk individuals.

Determining who is at 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, since it does tend to merge with stereotyping and is based on logic and reason, enables terrorist organizations to engineer attackers that 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, since 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.

1.5 Looking Ahead

We'll be studying many things in this book, but several important themes will be

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 a medical therapy works, which marketing message generates a better response, and which management technique or industrial process produces fewer errors.
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 ("What factors contribute to a drug's success, or the response to a marketing message?").
   - Make predictions ("Will a given subscriber leave this year?" "Is a given insurance claim fraudulent?")

1.6 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 researcher 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

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.

1.6.1 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.

The science of computer programming and details of database administration lie beyond the scope of this book, but these fields now lie within the scope of statistical work. The statistician must be conversant with the data, as well as how to get it and work with it.

- 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 their databases possess value that is strategic, not just administrative, and goes well beyond the original reason for collecting the data.

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 this chapter with a probability experiment:

Try It Yourself

1) Write down a series of 50 random coin flips without actually flipping the coins. That is, write down a series of 50 made-up H's and T's selected in such a way that they appear random.
2) Now actually flip a coin 50 times.

If you look at the series of made-up tosses and compare them to the real tosses, the longest streaks of either H or T generally occur in the ACTUAL tosses. When a person is asked to make up random tosses, they will rarely "allow" more than four H's or T's in a row. By the time they have written down four H's in a row, they think it is time to switch over to T, or else the series would not appear random. By contrast, instructors who teach this exercise in class often see a streak of 8 T's or H's in a row. Most people think that this is not random, and yet it clearly is.

In 1913, a roulette wheel at the Monte Carlo casino landed on black 26 times in a row. As the streak developed, gamblers, convinced that the wheel would most certainly have to end the streak, increasingly bet heavily on red—they lost millions.

The message here is that random variation reliably produces patterns that *appear* non-random.

Why is this significant? Just as with coin tosses, there is a significant component of random variation (engineers call it noise) in the data that routinely flow through life—whether business life, government affairs, the education world, or personal life. So much data…so much random variation…how do we know what is real and what is random?

We can't know for certain, though we do know that random behavior can appear to be real. One purpose of this book is to teach you about probability, and help you evaluate the potential random component in data, and provide ways of modeling it. This gives us a benchmark against which to measure patterns, and form educated guesses about whether observed events or patterns of interest might be really due to chance. When we understand randomness better, we can curb the tendency to chase after random patterns, and produce more reliable analyses of data.