1

Statistics Suck; So Why Do I Need to Learn About It?

1.1 Introduction

The question is a fair one, considering how many people feel about the subject. In this chapter, we go over some interesting aspects of the topic to convince the reader of the need to learn about statistics. Then we provide a formal definition.

Statistics Are All Around Us

Nowadays, we see and hear about statistics everywhere. Whether it is in statements about the median student loan amount of college students, the probability of rain tomorrow, association between two financial variables, or how an environmental factor may increase the risk of being diagnosed with a medical condition. These are all brief examples that rely on statistics. Information extracted from data is so ubiquitous these days that not understanding statistics limits our comprehension of the information coming to us. Therefore, understanding the main concepts from statistics has become imperative. In the world of business, making decisions based on evidence leads to a competitive advantage. For example, retailers can use data from customers to determine appropriate marketing campaigns, and manufacturing experiments can help establish the best settings to reduce costs. Investors can combine assets according to an adequate expected return on investment with an appropriate risk. We go over some specific applications of statistical methods in business in Section 1.2.

The Importance of Understanding the Concepts

Suppose you are invited to invest in a fund with an expected yearly return of 10%. Therefore, if you invest $100 at 10% yearly return, you expect to have $110 by the end of the year. So you decide to invest money; but at the end of the year, your portfolio is worth $90 (ignoring all fees). Specifically, the return on your money has been negative. That raises the question, based on the information you were given about the return you would receive, was your investment a mistake? Think about this question before reading ahead. The trick is in understanding the meaning of expected value. The information provided to you was that the fund had an “expected” yearly return.
of 10%. In some scenarios, such as a deposit account in a bank, the return given to you is guaranteed, and your principal (the money you start with) is protected. But in other scenarios, such as the stock market, the return you receive is not guaranteed, and, in fact, your principal may not be protected. Consequently, in any given year, you may have an outstanding return, much higher than the expected 10%, or you may get a return lower than the expected return of 10%. To counter this risk, possible returns are higher in these investments than safer alternatives, such as a bank savings account. The expected return is what the company has determined you should receive over the long term. Because of this aspect, an investment such as a stock has to be looked at differently than a bank savings account, and if after a year your portfolio is worth $90, your decision to invest in the portfolio was not necessarily a bad decision. What matters is the return you obtain over the long term, say, 10 or 20 years, and how much risk you are willing to take as an investor.

Another relevant component in understanding statistics is being able to distinguish between good information and bad information. An example is a spurious relationship. A spurious relationship between two variables occurs when an association is not due to a direct relationship between them, but from their association with other variables or coincidence. Strange associations are not the only way we can wrongly use statistics. Many media outlets (and politicians) often misquote scientific studies, which often rely on statistical methods, to reach their conclusions.

Example 1.1  In the summer of 2014, several media outlets reported on a new scientific study; one of the titles was “Study: Smelling farts may be good for your health.” In reality the study was about a developed compound that delivered small amounts of hydrogen sulfide that helped protect cells. Hydrogen sulfide is known to be a foul-smelling gas; hence the misunderstanding of the study results occurred. It is not enough to be able to correctly perform statistical analysis, but understanding the results is also key.

The issues brought up are rather common and tend to occur for two reasons:

• Inappropriate use of statistics.
• Wrong interpretation of results (poor statistical literacy skills).

Either of these two issues will potentially lead to reaching a wrong conclusion, hence making the wrong managerial decision. In contrast, avoiding these issues offers valuable knowledge for a given context.

1 Actually, in the case of investments such as stocks, it is often possible to obtain a measure of the risk of that financial asset. Not all assets offering an expected 10% return are the same. We will go into this in more detail later.
1.1 Introduction

Did you know that judges used to give more lenient decisions after meals compared to before meals? The finding comes from a study that looked at 1112 judicial rulings on parole. It was found that judge rulings were more likely to be favorable to the prisoner earlier in the day and that there was a jump in probability of parole ruling after a break. The study even considered the length of sentence, incarceration history of the prisoner, etc. Of course, the time of day of a parole hearing should not be a factor in determining a decision. The findings of this study helped establish a protocol to help eliminate the impact of time of day on parole ruling. Thus, an unintentional benefit of reading this book is that if you are facing the possibility of jail time, you should ask your lawyer to delay the sentencing until after lunch, just in case!

Case Study 1  A professor wants to answer the following question: Do quizzes help business undergraduate students get a better grade in their introductory statistics course? In academia, it is often simply assumed that quizzes will force students to study, leading to better course grades. The goal here is to answer the question based on evidence. There are two possible answers:

- They do not work: No difference between grades of students who took quizzes and students who did not take quizzes, or students who took quizzes perform worse.
- They do work: Grades of students who took quizzes are better than students who did not take quizzes.

Next is to determine how the comparison will be made. More specifically, what aspects associated with students will be taken into account? Hours of study per week, undergrad concentration, high school they went to, and the education level of parents are just some examples of “variables” that could be considered. Yet, the simplest way to make the comparison is based on two groups of students:

- One group that took quizzes during the course.
- The other group that did not take quizzes during the course.

Then, it is assumed that both groups are homogeneous in all their attributes, except that one group of students took the course with quizzes, while the other did not. This way, the average exam scores of both groups can be compared.

To compare average exam scores, two semesters were selected: one with quizzes and the other without. It was verified that the number of students that

---

4 For the statistical evidence see www.pnas.org (accessed July 26, 2019).
dropped the course or stopped attending class for both groups was similar (why do you think it is necessary to check this?).

- The group that took quizzes had an average test score of 79.95.
- The average of the group that had no quizzes was 75.54.

Now, these averages are just estimates, and different semesters will have different students in each group leading to new average values. This leads to uncertainty. Informally, what is known as statistical inference accounts for this uncertainty. When the inferential procedure was implemented, it was found that in reality there was no difference in average exam score among the two groups. Thus, there was no evidence that quizzes helped students improve their test scores.

Now, let’s not get too excited about this. The interpretation is that quizzes do not help improve the performance of students who take the class with that professor. The conclusion does not necessarily extend to quizzes in other subjects, and it doesn’t even necessarily extend to all introductory business statistics courses.

## Practice Problems

1.1 In Case Study 1, it was stated that “It was verified that the number of students that dropped the course or stopped attending class for both groups were similar.” Why do you think it is necessary to check this condition?

1.2 In September 20, 2017, Puerto Rico was struck by Hurricane Maria, a powerful storm with 155 mph sustained winds. In May 2018, multiple media outlets reported the findings of a study where researchers used September to December 2017 data to estimate that 4645 people had died in Puerto Rico, directly or indirectly, due to Hurricane Maria. The news shocked many, since at the time the local government had stated a total of 64 people had died due to the storm. Indeed, the study was one of several indicating the official estimate of 64 was too low. Before the news broke, the Puerto Rican government had not publicly shared death certificate data. Due to growing pressure from the study results, the government released data on the number of death certificates (Table 1.1). How does the released data suggest that the 4645 number may have been wrong?

*Hint:* Compare death certificates after Hurricane Maria to the ones before.

---


6 A government must create a death certificate for every single death.
### Table 1.1  
Death certificates in Puerto Rico by month and year. Causes of death not provided. Numbers in bold include death certificates after Hurricane Maria.

<table>
<thead>
<tr>
<th>Month</th>
<th>2015</th>
<th>2016</th>
<th>2017</th>
<th>2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>2744</td>
<td>2742</td>
<td>2894</td>
<td>2821</td>
</tr>
<tr>
<td>February</td>
<td>2403</td>
<td>2592</td>
<td>2315</td>
<td>2448</td>
</tr>
<tr>
<td>March</td>
<td>2427</td>
<td>2458</td>
<td>2494</td>
<td>2643</td>
</tr>
<tr>
<td>April</td>
<td>2259</td>
<td>2241</td>
<td>2392</td>
<td>2218</td>
</tr>
<tr>
<td>May</td>
<td>2340</td>
<td>2312</td>
<td>2390</td>
<td>–</td>
</tr>
<tr>
<td>June</td>
<td>2145</td>
<td>2355</td>
<td>2369</td>
<td>–</td>
</tr>
<tr>
<td>July</td>
<td>2382</td>
<td>2456</td>
<td>2367</td>
<td>–</td>
</tr>
<tr>
<td>August</td>
<td>2272</td>
<td>2427</td>
<td>2321</td>
<td>–</td>
</tr>
<tr>
<td>September</td>
<td>2258</td>
<td>2367</td>
<td>2928</td>
<td>–</td>
</tr>
<tr>
<td>October</td>
<td>2393</td>
<td>2357</td>
<td>3040</td>
<td>–</td>
</tr>
<tr>
<td>November</td>
<td>2268</td>
<td>2484</td>
<td>2671</td>
<td>–</td>
</tr>
<tr>
<td>December</td>
<td>2516</td>
<td>2854</td>
<td>2820</td>
<td>–</td>
</tr>
</tbody>
</table>


### 1.2 Data-Based Decision Making: Some Applications

Data-based decision making is a relatively new principle in management. Some people also use the term evidence-based management to mean data-based decision making, while others state that the latter is part of the former. We will use the term data-based decision making to avoid confusion. Researchers have found that the more data driven a firm is, the more productive it is, and the benefit can be substantial.

#### Making Open Data Profitable

In 2006, The Climate Corporation was founded. They obtained government data from weather stations, Doppler radar, US Geological Survey, and other freely accessible sources. Originally, the company provided weather insurance to clients. The company analyzed weather data, forecast future weather, and created an insurance policy from their analysis. A business would go to The Climate Corporation’s website and buy insurance against bad weather nearby. Since 2010, The Climate Corporation focuses exclusively on agriculture. Monsanto acquired the company in 2013 for approximately $1 billion.

In marketing, statistical methods are applied to better understand the needs of customers, to create a profile of consumers, or to market products and services efficiently to potential consumers. Auditors sample company
transactions to infer about its internal control system. In finance, statistics play a role in the actuarial field, where actuaries help control risk for banks, insurance companies, and other entities. Financial engineering, a field where investment strategies are developed, draws heavily from statistics.

CityScore – Boston

CityScore\(^7\) is an initiative to inform city managers about the overall health of Boston. After its launch, the Boston Transportation Department noticed that the on-time percentage for sign installation from December 1, 2015 through January 14, 2016 was 73%, 7% below target. Looking into it, administrators discovered a backlog of 90 sign installation requests. Within seven days, the request backlog was reduced from 90 to 7 requests.

Making decisions based on data does come with challenges. Besides the obvious need to sometimes deal with massive amounts of data in a computationally effective and theoretically sound way, privacy issues have been raised, and failures have been pointed out.

Moreover, although nowadays there is access to data through open data portals, universities, government agencies, companies, and other entities, it is quite common to make data available without describing the exact source of the data and how it was gathered. These intricacies of the data set must not be taken for granted.

Intricacies of Data: Google Trends

Studies have exemplified the use of search query data extracted from Google Trends (www.google.com/trends) to forecast processes of interest. Search query data has been used to model tourism demand, auto sales, home sales, initial unemployment claims, influenza activity in the United States, consumer behavior, dengue fever, and more. However, the search query volume data given by Google Trends for a fixed past time period, in fact, varies periodically. That is, if one uses the tool today to gather search volume data for a given query in January 2006, one may find a volume of, say, 58. But when using the same tool under the same settings next week to gather search volume data in January 2006 for the same given query, one may find a volume of, say, 65, and the following week some other result. In their website Google only mentions that the search query volume data comes from a sample of the total search queries. Yet no information is given on sample size, sampling technique, and other important details of the search query volume data. Some researchers have found that even when

\(^7\) www.boston.gov (accessed July 26, 2019).
accounting for the search query volume data uncertainty, from a practical point of view, incorporating search query volume data is not universally better than simpler forecasting alternatives.⁸

Search query volume is an example of proprietary data, data shared by a private company. Companies have two strong reasons not to provide too much information about their proprietary data. First, it can threaten their competitiveness. Second, it would give the opportunity to others to “game the model.”

**Leadership and Statistics** We have presented a wealth of examples to convince the reader of the importance of statistics in making managerial decisions. In addition to the importance of evidence-based decisions being emphasized, we must acknowledge that to ensure the best gain from data, a company must have the proper environment. Designing a study, gathering data, and analyzing the results must be done efficiently. To do all these tasks adequately, good leadership is needed. Bad leadership leads to poor organization, and poor organization makes it harder to base decisions on evidence since it would be harder to filter out unwanted “noise.”

**Example 1.2** Let’s suppose that a company wants to determine the association between square feet of land and its value. There are two separate sources, source 1 and source 2, of land value and land size data from 75 purchase transactions. Each source collects land value differently so the entries vary. Both sources should imply the same type of association between these land metrics. But source 1 has been careful in their data collection procedure, while source 2 had poor record keeping, and the data was not preprocessed for errors. Figure 1.1 presents the association between land size (x-axis) and land value (y-axis) using data from source 1 and source 2, respectively. According to the data from source 1, there is approximately a linear association between land size and land value: as land size increases, on average, land value increases. On the other hand, due to the inefficiency-driven errors in the land value, source 2 has a lot of “noise,” which has led to a wider “scatter” of the data. As a result, it is hard to tell if there is an association between land size and land value at all when using source 2 data.

We cannot always control the error in measurements, but this example shows us that if we could, this would have a dramatic impact on the usefulness of our data to make decisions.

**Setting up the Proper Environment** Once the leadership has recognized the benefits of evidence-based decisions, they must ensure the firm has the right structure to exploit data. Some points to consider are as follows:

Figure 1.1 Land size versus land value from source 1, with little error in measuring land value (left), and from source 2, with large error in measuring land value (right).

- Does the entity have the right personnel? (To design a study, analyze data, etc.)
- Are hardware and software capabilities adequate?
- Is there good communication between offices that will need to work together?
- Has personnel been properly trained or oriented about any changes in procedures that may occur?
- Has a data management protocol been developed? (See Section 2.5.)

**Ethics and Data-Based Decision Making** One cannot discuss making decisions based on data without talking about ethics. Several aspects must be considered to make decisions ethically, based on data. Some of the aspects apply in general, while others only apply in some situations. Specifically, the team must ensure the data is not misrepresented (intentionally or not), and they should avoid misinterpreting the results. For some analysis, researchers require the approval of boards that carefully evaluate the design of the process or the implementation of the work to make sure human subjects and animals are safe.
1.3 Statistics Defined

Statistics can be defined as the science and art of collecting, organizing, presenting, summarizing, and interpreting data. The inclusion of the word “art” in the definition is not an error. As we will see, there is strong theoretical evidence backing much of the statistical procedures covered in this book. However, in practice, the implementation of statistical methods requires decisions on building plots, organizing data, and relying on rules of thumb that make statistics also an art, not just a science.

The statistical tools at our disposal fall into two branches: descriptive statistics and inferential statistics. Descriptive statistics organize, present, and summarize data. For example, consider domestic violence data\(^9\) from Edmonton, Canada. It provides number of reported criminal and noncriminal occurrences among intimate partners. The information is provided in a quarterly basis, from 2009 until the third quarter of 2018. Figure 1.2 displays a snapshot of the original data online.

Summarizing the data helps us make sense of it. Which type of data description to apply depends on the type of data, a topic that we discuss in Chapter 2. One useful description of this data is the mean criminal domestic violence cases reported per each quarter. Visual descriptive summaries are also extremely useful. One alternative here is to chart the number of criminal domestic violence cases through time, allowing us to explore whether the cases are changing in time.

Informally, inferential statistics are procedures that allows us to use a subset from a group of interest to reach conclusions about the group overall. This

\(^9\) dashboard.edmonton.ca (accessed December 12, 2018).
subset is called a **sample**. Statistical inference is incredibly valuable to make decisions based on evidence, and we have seen many examples of how this is performed in the real world. In fact, even babies have been found to perform statistical inference.

### Babies Doing Statistics
Laura Schulz and colleagues at MIT performed an experiment where 15-month-old babies were randomly assigned to two groups. Randomization guaranteed that the characteristics of the babies were generally similar in both groups. Each group of toddlers saw how a research assistant:

- Removed three *blue* balls one at a time from a see-through box.
- Squeezed each *blue* ball for their squeak effect.
- And then handed a *yellow* ball from the box to the child.

The babies did not know that *yellow* balls did not squeak.

- In group 1, the baby could see that the see-through box had mostly *blue* balls.
- While in group 2, the baby could see that the see-through box had mostly *yellow* balls.

The study uses a sample of babies to infer about 15-month-old babies overall. It was concluded that, on average, babies from group 1 were more likely to try to squeeze the *yellow* ball handed to them than those from group 2. The researchers argue that this happens because, since the box from group 1 had mostly *blue* balls, babies saw the assistant’s three *blue* ball sample as representative of the population. In contrast, since the box from group 2 had mostly *yellow* balls, babies did not see the assistant’s three *blue* ball sample as representative of the population. In a way, babies in group 2 were able to determine that the assistant was “cherry-picking” the *blue* balls that could squeak.11

Descriptive statistics and inferential statistics are complementary. The former is always performed on data, while the latter tends to be performed when the data does not represent the entire group of interest.

**Statistics and Critical Thinking** Critical thinking is clear, rational thinking that is informed by evidence. Evidence can be generated through

- observations
- experiments
- memory
- reflection (opinion)
- reason (theory)

---

10 See [www.ted.com](http://www.ted.com) (accessed July 26, 2019).
Critical thinking pertains to the evaluation of the evidence to build – perhaps adjust – a belief or take action. Our emphasis is evidence gathered through data, and within this context, statistics is an essential part of critical thinking, allowing us to test an idea against empirical evidence. Also, as implied by some of the examples encountered earlier in this chapter, statistical literacy is required for proper interpretation of results, and critical thinking plays an important role in evaluating the assumptions made to conduct statistical inference. Given that critical thinking evaluates information, statistics allows us to do better critical thinking. At the same time, statistical inference requires making assumptions about the population of interest, and hence, critical thinking will allow us to do better statistics.

1.4 Use of Technology and the New Buzzwords: Data Science, Data Analytics, and Big Data

There is a wide selection of software available to perform statistical analysis. Computers have also become faster, allowing for new methodologies to be used and applications on bigger data sets. These technologies make the use of statistical procedures much easier, sometimes too easy. Specifically, just because a computer software creates a figure does not mean the figure is correct. It is up to the user to ensure that they choose the right type of chart for the data. Similarly, a computer software may calculate a statistic for you, but it will not necessarily determine if the statistic is adequate for your type of data or objective. Furthermore, computer software does not check all the assumptions that a statistical procedure entails. And the more complex the statistical procedure is, the greater the need for an expert to implement the procedure correctly. The misconception that with just the right computer software (or adequate programming skills), one can perform any type of statistical procedure without feedback from an expert in statistics is common in business, and it has become a topic of debate in the big data/data analytics world. Overall, advanced procedures should always be performed by a qualified data analyst.

1.4.1 A Quick Look at Data Science: Some Definitions

**Virus X Cure Found!**

Remarkably, recently a team of scientists was able to cure Virus X, a deadly antibiotic resistant disease in chickens. They reported that:

- Thirty-three percent of the chickens in the study showed improvement after treatment against the disease.
• One-third of the chickens showed no change.
• Unfortunately, the third chicken got away.

If you are confused with these results, do not fret. You should be confused. The results of the study, which did not really happen, help point out that evidence from very small data sets should not be trusted. Also, the joke is a way to introduce the concepts of data science and big data. A decade ago, business intelligence – methods to process data and influence decision making – was all the talk. But in the last few years, data science, big data, and data analytics have overtaken business intelligence in popularity.

**Big data** is defined as data that cannot be processed using traditional methods. Big data can hold petabytes\(^\text{12}\) of information or may be sizable enough to not be processed through usual procedures. The field finds ways to query the data, store it, and analyze it.

**Example 1.3**  *Data from the computer-aided emergency management system (EMS) for New York City is available online.*\(^\text{13}\) The raw data includes almost five million records for incidents from January 1, 2013 until June 30, 2016.

Although not massive in size (the original data file is a little over 1 gigabyte in size), this data is large enough to cause trouble. In fact, the raw data cannot be opened in most computers with software such as Excel or Minitab. It requires special software, and even then it is best to filter out unnecessary information from this data set at the beginning stages of the analysis.

**Data analytics** involve statistical methods to retrieve useful information from data. The main distinction it has from statistics is that in the latter, data is used to answer a very specific question, while in data analytics, data may be “mined” to gain insight. A data analytics branch called **predictive analytics** has become popular to forecast future observations.

**Example 1.4**  *In a 2018 working paper,*\(^\text{14}\) David Andrew Finer, a graduate student at University of Chicago’s Booth School of Business, used open data on over one billion New York City taxi trips to assess if there was systematic information leakage from the Federal Reserve Bank. Specifically, he extracted cab trips starting at commercial banks and at the New York Federal Bank that converged on the same destination around lunchtime and those directly from banks to the New York Fed late in the evening.\(^\text{15}\) He found evidence of those journeys rising sharply

\(^{12}\) A petabyte consists of 1024 terabytes, which in turn consists of 1204 gigabytes.

\(^{13}\) data.cityofnewyork.us (accessed July 26, 2019).

\(^{14}\) What insights do taxi rides offer into federal reserve leakage?

around the dates of meetings when interest rates were determined by the Federal Reserve’s monetary-policy committee in Washington. This is one way data mining works, using data collected by the New York City Taxi and Limousine Commission for official purposes, to gain insight on something else.

**Data science** is an umbrella term that refers to the science that enables data management to convert data into knowledge while providing computational capabilities to carry out statistical methods. Data science merges aspects from statistics and computer science into one field. The fields of big data and data analytics fall within data science.

Figure 1.3 outlines the data science process. It starts with a problem statement, which may be rather ambiguous (Example 1.4). This is followed by getting the raw data and if necessary data wrangling, steps that sometimes demand big data methods. **Data wrangling** are methods to process the data for analysis. This includes filtering unwanted information out, sorting, aggregating information, or transforming variables into new ones. Preliminary data analysis serves as a first attempt to summarize the data, numerically and visually. The data must be screened for errors or any issues that may arise. For example, the data must be cleaned up when categories are not clearly defined, or numerical information is nonsensical, or missing values are present. If the data issue limits the usefulness of some information, more data wrangling is implemented. The exploratory data analysis will summarize the processed data, serving as a first look of its usefulness. Further statistical procedures are applied on the basis of the problem statement, followed by interpreting the outcome. Next, the results

![Figure 1.3](image.png)
are communicated to pundits. The importance of this step should not be underestimated, for it affects how pundits will make decisions from the results. This book highlights many parts of the exploratory data analysis and statistical procedure stages of data science, with parts of acquiring raw data, data wrangling, preliminary analysis, and data cleanup examined from time to time, especially in “Quick Look at Data Science” sections throughout the book.

Chapter Problems

1.3 University administrators record data from their students. For example, using admission records, they determine that 1075 was the mean SAT score of the latest business program freshmen. Is the number provided part of descriptive statistics or inferential statistics?

1.4 A political candidate wants to know what are her chances of being elected in the coming elections. Her team takes a random sample of 2500 registered voters and ask them who they will vote for. Will the candidate’s team need to perform statistical inference?

1.5 Provide examples of applying descriptive and inferential statistics.

1.6 In September 2018, Google launched a beta version of a search tool for data. Go to https://toolbox.google.com/datasetsearch, and enter domestic violence statistics. Download in CSV form one of the results.

1.7 Search online to see if cities near to your school have an open data portal. Download in CSV form data you consider interesting.

Further Reading