1

INTRODUCTION

The world suddenly has become awash in data! A great many popular books have been written recently that extol "big data" and the information derived for decision makers. These data are considered "big" because a certain "catalog" of data may be so large that traditional ways of managing and analyzing such information cannot easily accommodate it. The data originate from you and me whenever we use certain social media, or make purchases online, or have information derived from us through radio frequency identification (RFID) readers attached to clothing and cars, even implanted in animals, and so on. The result is a massive avalanche of information that exists for businesses leaders, decision makers, and researchers to use for predicting related behaviors and attitudes.

BIG DATA ANALYSIS

Decision makers are trying to figure out how to manage and use the information available. Typical computer software used for statistical decision making is currently limited to a number of cases far below that which is available for consideration of big data. A traditional approach to address this issue is known as "data mining" in which a number of techniques, including statistics, are used to discover patterns in a large set of data.

Researchers may be overjoyed with the availability of such rich data, but it provides both opportunities and challenges. On the opportunity side, never before have

Using Statistics in the Social and Health Sciences with SPSS® and Excel®, First Edition.
Martin Lee Abbott.

such large amounts of information been available to assist researchers and policy makers understand widespread public thinking and behavior. On the challenge side however are several difficult questions:

- How are such data to be examined?
- Do current social science methods and processes provide guidance to examining data sets that surpass historical data-gathering capacity?
- Are big data representative?
- Do data sets so large obviate the need for probability-based research analyses?
- Do decision makers understand how to use social science methodology to assist in their analyses of emerging data?
- Will the decisions emerging from big data be used ethically, within the context to social science research guidelines?
- Will effect size considerations overshadow questions of significance testing?

Social scientists can rely on existing statistical methods to manage and analyze big data, but the *way in which the analyses are used for decision making will change*. One trend is that prediction may be hailed as a more prominent method for understanding the data than traditional hypothesis testing. We will have more to say about this distinction later in the book, but it is important at this point to see that researchers will need to adapt statistical approaches for analyzing big data.

VISUAL DATA ANALYSIS

Another emerging trend for understanding and managing the swell of data is the use of visuals. Of course, visual descriptions of data have been used for centuries. It is commonly acknowledged that the first "pie chart" was published by Playfair (1801). Playfair's example in Figure 1.1 compares the dynamics of nations over time.

Figure 1.1 compared nations using size, color, and orientation over time. Using this method for comparing information has been useful for viewing the patterns in data not readily observable from numerical analysis.

As with numerical methods, however, there are opportunities and challenges in the use of visual analyses:

- Can visual means be used to convey complex meaning?
- Are there "rules" that will help to insure a standard way of creating, analyzing, and interpreting such visual information?
- Will visual analyses become divorced from numerical analysis so that observers have no way of objectively confirming the meaning of the images?

Several visual data software analysis programs have appeared over the last several years. Simply running an online search will yield several possibilities including many that offer free (initial) programs for cataloging and presenting data from the user. I offer one very important caveat (see the final bullet point earlier), which is that it is

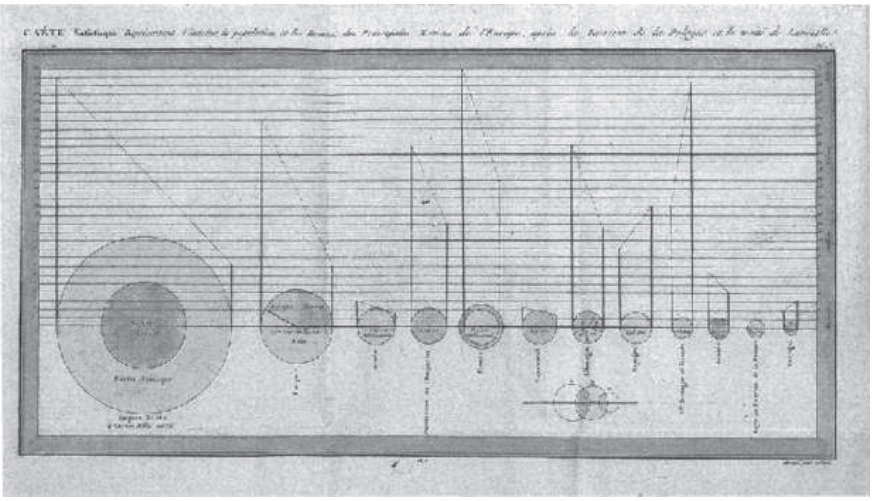

Figure 1.1 William Playfair's pie chart. *Source*: https://commons.wikimedia.org/wiki/File:Playfair_piecharts.jpg. Public domain.

important to perform visual data analysis in concert with numerical analysis. As we will see later in the book, it is easy to intentionally or unintentionally mislead readers using visual presentations when these are divorced from numerical statistical means that discuss the "significance" and "meaningfulness" of the visual data.

IMPORTANCE OF STATISTICS FOR THE SOCIAL AND HEALTH SCIENCES AND MEDICINE

The presence of so much rich information presents meaningful opportunities for understanding many of the processes that affect the social world. While much of the time big data analyses are used for understanding business dynamics and economic trends, it is also important to focus on those data patterns that can affect the social sphere beyond these indicators: social and psychological behavior and attitudes, changes in understanding health and medicine, and educational progress. These social indicators have been the subject of a great deal of analyses over the decades and now may make significant advances depending on how big data are analyzed and managed. On a related note, the social sciences (especially sociology and psychology) are now areas included in the new Medical College Admission Test (MCAT), which also includes greater emphasis upon "Scientific Inquiry & Reasoning Skills." The material we will learn from this book will help to support study in these areas for aspiring health and medical professionals.

In this book, I intend to focus on how to use and analyze data of all sizes and shapes. While we will be limited in our ability to dive into the world of big data fully, we can study the basics of how to recognize, generate, interpret, and critique analyses of data for decision making. One of the first lessons is that *data can be understood both numerically and visually*. When we describe information, we are attempting to

see and convey underlying meaning in the numbers and visual expressions. If I have a collection of data, I cannot recognize its meaning by simply looking at it. However, if I apply certain numerical and visual methods to organize the data, I can see what patterns lay below the surface.

HISTORICAL NOTES: EARLY USE OF STATISTICS

Statistics as a field has had a long and colorful history. Students will recognize some prominent names as the field developed its mathematical identity: Pearson, Fisher, Bayes, Laplace, and others. But it is important to note that some of the earliest statistical studies were based in solving social and political problems.

One of the earliest of such studies was developed by John Graunt who compiled information from Bills of Mortality to detect, among other things, the impact and origins of deaths by plague. Parish records documented christenings, weddings, and burials at the time, so Graunt's study tracked the number of deaths in the parishes as

Figure 1.2 John Snow's map showing deaths in the London cholera epidemic of 1854. *Source*: https://commons.wikimedia.org/wiki/File:Snow-cholera-map-1.jpg. Public domain.

a way to understand the dynamics of the plague. His broader goal was to predict the population of London using extant data from the parish records.

Another early use of statistics was Dr John Snow's map showing deaths in the houses of London's Soho District during the 1854 cholera epidemic, as popularized by Johnson's book, *The Ghost Map* (2006). In order to investigate the reasons for the spread of cholera other than odor ("miasma theory"), Snow created a map showing each death as a black line outside each household, along with features of the neighborhood including the water sources located throughout the district. The map created a visual picture of the concentration of deaths across the district and led to hypotheses about cholera spreading by waterborne contamination rather than smell. (If you were to walk across the same London district today, you will see that the great social theorist Karl Marx lived just a few streets away from the center of the cholera deaths.)

Figure 1.2 shows Snow's map. You can see that near the center of the map is the "Broad Street Pump" which Snow determined to be the source for the spread of cholera. (At the time, Karl Marx lived on Dean Street, just to the east of the Broad Street Pump.) Notice that the houses nearest this pump recorded the highest numbers of deaths.

Figure 1.2 example not only shows how descriptive statistics underscored the use of visual means of representing data, but it also helped to clarify possible reasons for an epidemic. Graunt's tables based on the Bills of Mortality were rudimentary visuals, but Snow's map was a more effective means of portraying complex data by visual means. A still later statistician made even greater advancements in using visual information to communicate trends in data.

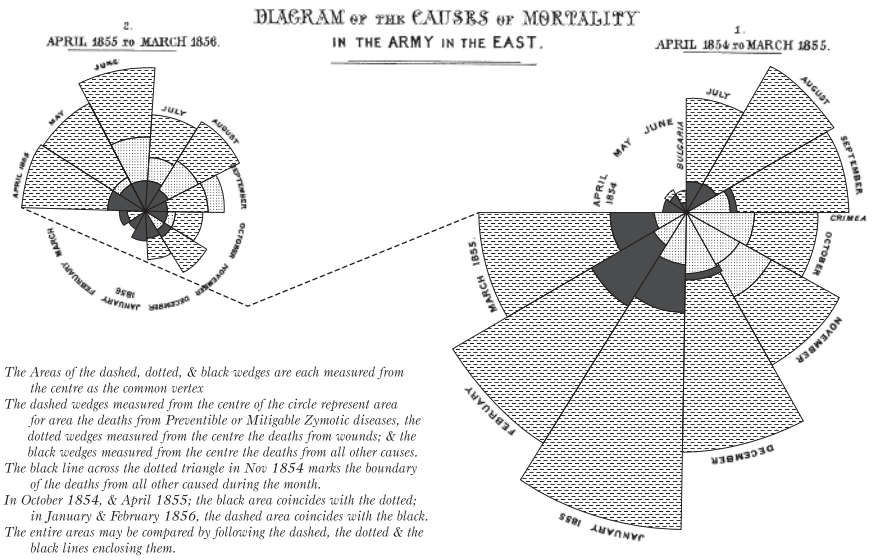

Figure 1.3 Florence Nightingale's polar chart comparing battlefield and nonbattlefield deaths. *Source*: https://en.wikipedia.org/wiki/Pie_chart#/media/File:Nightingale-mortality.jpg. Public domain.

Note: The original color version of this figure can be found in the online version of this book.

Nightingale (1858) is most often remembered as the founder of modern nursing. She is often represented in paintings as "the lady with the lamp," since she was known to walk among the bedsides checking on the sick and wounded of the war. But Nightingale was also an astute statistician who used statistics to capture the dramatic need in hospitals during the Crimean War. She is credited as being one of the first to use a "pie chart" (more accurately, a "polar chart"). Figure 1.3 shows comparisons in her original polar chart of differences between soldiers who died of battlefield wounds ("dotted" wedges near the center) and those who died from other causes ("dashed" wedges measured from the center of the graph) over time. The original color version of Figure 1.3 can be found in the online version of this book. The relationship between these groups fueled Nightingale's efforts to obtain further funding for sanitary hospital conditions since those who died of infections were greater in number than those dying of battlefield wounds.

APPROACH OF THE BOOK

Many students and researchers are intimidated by statistical procedures, which may be due to fear of math, problematic math teachers in earlier education, or the lack of exposure to a "discovery" method for understanding difficult procedures. This book is an introduction to understanding statistics in a way that allows students to discover patterns in data and developing skill at making interpretations from data analyses. I describe how to use statistical programs (SPSS and Excel) to make the study more understandable and to teach students how to approach problem solving. Ordinarily, a first course in statistics leads students through the worlds of descriptive and inferential statistics by highlighting the formulas and sequential procedures that lead to statistical decision making. We will do all this in this book, but I place a good deal more attention on conceptual understanding. Thus, rather than memorizing a specific formula and using it in a specific way to solve a problem, I want to make sure the student first understands the nature of the problem, why a specific formula is needed, and how it will result in the appropriate information for decision making.

By using statistical software, we can place more attention on understanding how to *interpret findings*. Statistics courses taught in mathematics departments, and in some social science departments, often place primary emphases on the formulas/processes themselves. In the extreme, this can limit the usefulness of the analyses to the practitioner. My approach encourages students to focus more on how to understand and make applications of the results of statistical analyses. SPSS and other statistical programs are much more efficient at performing the analyses; the key issue in my approach is how to interpret the results in the context of the research question.

Beginning with my first undergraduate course teaching statistics with conventional textbooks, I have spent countless hours demonstrating how to conduct statistical tests manually and teaching students to do likewise. This is not always a bad strategy; performing the analysis manually can lead the student to understand how formulas treat data and yield valuable information. However, it is often the case that the student gravitates to memorizing the formula or the steps in an analysis. Again, there is nothing wrong with this approach as long as the student does not stop there. *The*

outcome of the analysis is more important than memorizing the steps to the outcome. Examining the appropriate output derived from statistical software shifts the attention from the nuances of a formula to the wealth of information obtained by using it.

It is important to understand that I do indeed teach the student the nuances of formulas, understanding why, when, how, and under what conditions they are used. But in my experience, forcing the student to scrutinize statistical output files accomplishes this and teaches them the appropriate use and limitations of the information derived.

Students in my classes are always surprised (ecstatic) to realize they can use their textbooks and notes on my exams. But they quickly find that, unless they really understand the principles and how they are applied and interpreted, an open book is not going to help them. Over time, they come to realize that the analyses and the outcomes of statistical procedures are simply the ingredients for what comes next: building solutions to research problems. Therefore, their role is more detective and constructor than number juggler.

This approach mirrors the recent national and international debate about math pedagogy. In our recent book, *Winning the Math Wars* (2010), my colleagues and I addressed these issues in great detail, suggesting that, while traditional ways of teaching math are useful and important, the emphases of reform approaches are not to be dismissed. Understanding and memorizing detail are crucial, but problem solving requires a different approach to learning.

CASES FROM CURRENT RESEARCH

I focus on using real-world data in this book. There are several reasons for doing so, primarily because students need to be grounded in approaches for using data from the real world with all their problems and "grittiness." When people respond to surveys or interviews, they inevitably fill out information in ways not asked by interviewers (e.g., respondents may choose two possible answers when one is required, etc.). Moreover, transferring data to electronic form may result in miscoded responses or categorization problems. Researchers always confront these issues, and I believe it is important for students to leave the classroom aware of the range of possible problems with real-world data and prepared for dealing with them. Of course, much of the data we will examine will already have been put in standard forms, but other research issues will arise (e.g., how do I recategorize data, assign missing cases, compute new variables, etc.?).

Another reason I use real-world data is to familiarize students with contemporary research questions in the social and health science fields. Classroom data often are contrived to make a certain point or show a specific procedure, which are both helpful. But I believe it is important to draw the focus away from the procedure per se and understand how the procedure will help the researcher resolve a research question. The research questions are important. Policy reflects the available information on a research topic, to some extent, so it is important for students to be able to generate that information as well as to understand it. This is an "active" rather than "passive" learning approach to understanding statistics.

Data Labs are a very important part of this course since they allow students to take charge of their learning. This is the heart of discovery learning. Understanding a statistical procedure in the confines of a classroom is necessary and helpful. However, learning that lasts is best accomplished by students directly engaging the processes with actual data and observing what patterns emerge in the findings that can be applied to real research problems.

Some practice problems may use data created for classroom use, but real-world data from actual research databases will enable a deepening of understanding. In addition to national databases, I use results from my own research for classroom learning. In every case, researchers know that they will discover knotty problems and unusual, sometimes idiosyncratic, information in their data. If students are not exposed to this real-world aspect of research, it will be confusing when they engage in actual research beyond the confines of the classroom.

In this course, we will have several occasions to complete Data Labs that pose research problems with actual data. Students take what they learn from the book material and conduct a statistical investigation using SPSS and Excel. Then, they have the opportunity to examine the results, write research summaries, and compare findings with the solutions presented at the end of the book.

The project labs also introduce students to two software approaches for solving statistical problems. These are quite different in many regards, as we will see in the chapters that follow. SPSS provides additional advanced procedures educational researchers utilize for more complex and extensive research questions. Excel is widely accessible and provides a wealth of information to researchers about many statistical processes they encounter in actual research. The Data Labs provide solutions in both formats so the student can learn the capabilities and approaches of each.

This book makes use of publically available research data. The General Social Survey or GSS[1] is a nationally representative survey designed to be part of a program of social research to monitor changes in Americans' social characteristics and attitudes. Funded through the National Science Foundation and administered by the National Opinion Research Center (NORC), the GSS has been administered annually or biannually since 1972. As a general survey, the GSS asks a variety of questions on a series of topics designed to track the opinions of Americans over the last four decades.

Other databases we will use in the book include the following:

- The Centers for Disease Control and Prevention (CDC) conducts the Behavioral Risk Factor Surveillance System (BRFSS) as a health-related telephone survey to measure American residents' health conditions, health behaviors, and use of preventative services.[2]

[1]Tom W. Smith, Peter Marsden, Michael Hout, and Jibum Kim. General Social Surveys, 1972–2012 [machine-readable data file]/Principal Investigator, Tom W. Smith; Coprincipal Investigator, Peter V. Marsden; Coprincipal Investigator, Michael Hout; Sponsored by National Science Foundation. – NORC ed. – Chicago: National Opinion Research Center [producer]; Storrs, CT: The Roper Center for Public Opinion Research, University of Connecticut [distributor], 2013. 1 data file (57,061 logical records) + 1 codebook (3432 pp.). -- (National Data Program for the Social Sciences, No. 21).

[2]Centers for Disease Control and Prevention (CDC) (2013). *Behavioral Risk Factor Surveillance System Survey Data*. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.

- Association of Religion Data Archives (ARDA) presents a series of databases on a variety of religion topics from the sociological perspective. In addition to other databases, the ARDA presents GSS databases on special modules (sets of questions) relevant to religion. By visiting the ARDA (www.thearda.com), you can peruse the codebook for the latest GSS file (www.thearda.com/Archive/GSS.asp) to get a fuller sense of the types of questions a general survey asks. You can also visit the ARDA's "Learning Center" to take a survey that allows you to compare yourself to a larger national profile. The "Compare Yourself to the Nation" survey allows you to see how you compare to others based on the results from the 2005 Baylor Religion Survey (addressing religious identity, beliefs, experiences, paranormal views, etc.).

RESEARCH DESIGN

Researchers who write statistics books have a dilemma with respect to research design. Typically, statistics and research design are taught separately in order for students to understand each in greater depth. The difficulty with this approach is that the student is left on their own to synthesize the information; this is often not done successfully.

Colleges and universities attempt to manage this problem differently. Some require statistics as a prerequisite for a research design course or vice versa. Others attempt to synthesize the information into one course, which is difficult to do given the eventual complexity of both "sets" of information. Adding somewhat to the problem is the approach of multiple courses in both domains.

I do not offer a perfect solution to this dilemma. My approach focuses on an in-depth understanding of statistical procedures for actual research problems. What this means is that I cannot devote a great deal of attention in this book to research design apart from the statistical procedures which are an integral part of it. (You may wish to consult a separate book on research design I authored with my colleague Jennifer McKinney, *Understanding and Applying Research Design*, 2013.)

I try to address the problem in two ways. First, wherever possible, I connect statistics with specific research designs. This provides an additional context in which students can focus on using statistics to answer research questions. The research question drives the decision about which statistical procedures to use; it also calls for discussion of appropriate design in which to use the statistical procedures. We will cover essential information about research design in order to show how these might be used.

Second, I have an online course in research design that can be accessed to continue your exploration from this book. In addition to databases and other research resources, you can follow the web address in the preface to gain access to the online course as additional preparation in research design.

FOCUS ON INTERPRETATION

I call attention to problem solving and interpretation as the important elements of statistical analysis. It is tempting for students to focus so much on using statistical

procedures to create meaningful results (a critical matter!) that they do not focus on what the results mean for the research question. They stop after they use a formula and decide whether or not a finding is statistically significant. I strongly encourage students to think about the findings in the context and words of the research question. This is not an easy thing to do because the meaning of the results is not always cut and dried. It requires students to think beyond the formula.

Statisticians and practitioners have devised rules to help researchers with this dilemma by creating criteria for decision making. For example, as we will see in Chapter 11, squaring a correlation yields the "coefficient of determination," which represents the amount of variance in one variable that is accounted for by the other variable (this is known as "effect size," a topic which we will spend a great deal of time with in this book). But the next question is, how much of the "accounted for variance" is meaningful? This consideration is key to understanding how to use and make decisions on the basis of big data.

In many ways, interpretation of results is an art undergirded by the cannons of science. Much of the ability to develop expertise in interpretation comes by long hours of tutelage with researchers who have done it for many years. We cannot hope to emerge from our study with this expertise, but through constant focus on interpretation, we can become aware of the acceptable ways of understanding and using statistical results.

Statisticians have suggested different ways of helping with interpretation. For example, when dealing with the "accounting of variance" example presented earlier, statisticians have created criteria that determine 0.01 (1%) of the variance accounted for is considered "small" while 0.05 (5%) is "medium" and so forth. (And, much to the dismay of many students, there are more than one set of these criteria.) Therefore, if we determine that the correlation between two variables reach these criteria levels, we can feel secure in sticking to good interpretation guidelines. Problems exist however in how to view these statistical results within the context of the research problem.

For example, if a research question is, "Does class size affect math achievement?" and the results suggest that class size accounts for 1% of the variance in math achievement, many researchers might agree the results represent a small and perhaps even inconsequential impact. However, if a research question is, "Does drug X affect Ebola survival rates?," researchers might consider 1% of the variance to be much more consequential than "small!" This is not to say that math achievement is any less important than Ebola survival rates (although that is another of those debatable questions researchers face), but the researcher must consider a range of factors in determining meaningfulness: the intractability of the research problem, the discovery of new dimensions of the research focus, whether or not the findings represent life and death, and so on. The material point is that statistical criteria are important for establishing meaningfulness of results, but overall interpretation involves the larger context within which the research takes place.

I have found that students have the most difficult time with these matters. Using a formula to create numerical results is often much preferable to understanding what the results mean in the context of the research question. Students have been conditioned to stop after they get the right numerical answer. They typically do not get to the difficult work of what the right answer *means* because it isn't always apparent.

I emphasize "practical significance" (effect size) in this book as well as statistical significance. In many ways, this is a more comprehensive approach to uncertainty, since effect size is a measure of "impact" in the research evaluation. It is important to measure the likelihood of chance findings (statistical significance), but the extent of influence represented in the analyses affords the researcher another vantage point to determine the relationship among the research variables.

Coverage of Statistical Procedures

The statistical applications we will discuss in this book are "workhorses." This is an introductory treatment, so we need to spend time discussing the nature of statistics and basic procedures that allow you to use more sophisticated procedures. We will not be able to examine advanced procedures in much detail. I will provide some references for students who wish to continue their learning in these areas. Hopefully, as you learn the capability of SPSS and Excel, you can explore more advanced procedures on your own, beyond the end of our discussions.

Some readers may have taken statistics coursework previously. If so, my hope is that they are able to enrich what they previously learned and develop a more nuanced understanding of how to address problems in educational research through the use of SPSS and Excel. Whether readers are new to the study or experienced practitioners, my hope is that statistics becomes meaningful as a way of examining problems and debunking prevailing assumptions in the social and health sciences.

Often, well-intentioned people can, through ignorance of appropriate processes, promote ideas that may not be true. Further, policies might be offered that would have a negative impact even though the policy was not based on sound statistical analyses. Statistics are tools that can be misused and influenced by the value perspective of the wielder. However, policies are often generated in the absence of compelling research. Students need to become "research literate" in order to recognize when statistical processes should be used and when they are being used incorrectly.